Understanding the acoustic implications of digital transmission on fricatives

Forståelse af akustiske implikationer af digital transmission på frikativer

KRESTINA V. CHRISTENSEN

A THESIS
SUBMITTED TO
AARHUS UNIVERSITY AND UNIVERSITY OF YORK
FOR THE JOINT DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF ARTS
AARHUS UNIVERSITY

AND

LANGUAGE AND LINGUISTIC SCIENCE
UNIVERSITY OF YORK

July 2023

Table of contents

Abstract (English)

The aim of this thesis is to provide a better understanding of the acoustic implications of digital transmission on fricatives relevant across research fields. This is motivated by the increasing amount of digital transmitted speech across the world, and the limited knowledge on the effects of digital transmission on consonants. The thesis investigates the fricatives /f/, /θ/, /s/, /ʃ/, /z/, /ð/ and [fʲ]. Fricatives were expected to be particularly affected by codec compression because of their noise-like and aperiodic structure, which might be mistaken for noise by the codecs.

The thesis investigates the effects of the AMR-WB-, Opus-, and MP3 codec using three different bitrates and in live transmission. The acoustic implications were measured as the first four spectral moments, peak frequency, and via spectrographic analysis. These measures were compared between baseline uncompressed WAV files and each of the codec compressed versions. This resulted in three studies. The first two are in *controlled* conditions i.e. the WAV files are codec compressed via a computer, whereas the third study is *live* with the speech transmitted between two mobile phones with and without background noise.

The findings indicate significant effects of the codec compressions on the spectral measures with segment, codec and bitrate dependent tendencies. The live transmission and background noise generally produced larger effects than the controlled conditions. Intensity played a key role in the magnitude of the effects of the codec compressions and live transmission.

This has implications when using codec compressed speech as data, but especially in socio- and forensic phonetics with possible diffusion of sound changes and speaker comparisons. In addition, the results have implications beyond linguistics e.g. in psychology, where clarity of speech plays a role in perceived charisma, and in hearing aid and cochlear implant technology, which both approach speech digitally and incorporate noise reduction.

Abstract (Danish)

Formålet med denne afhandling er at give en bedre forståelse af de akustiske implikationer af digital transmission på frikativer, der er relevant på tværs af forskningsfelter. Dette er motiveret af den stigende mængde ad digital transmitteret tale over hele verden, og den begrænsede viden om effekten af digital transmission på konsonanter. Afhandlingen undersøger frikativerne /f/, /θ/, /s/, /ʃ/, /z/, /ð/ og [fʲ]. Frikativer forventedes at blive særligt påvirket af codec-komprimeringen på grund af deres støjlignende og aperiodiske struktur, som under komprimeringen kan forveksles med støj.

Afhandlingen undersøger virkningerne af AMR-WB-, Opus- og MP3-codec'et ved hjælp af tre forskellige bithastigheder og i live transmission. De akustiske implikationer blev målt som de første fire spektrale momenter, topfrekvens og via spektrografisk analyse. Disse målinger blev sammenlignet mellem baseline ukomprimerede WAV-filer og hver af de codec-komprimerede versioner. Dette resulterede i tre studier. De to første er under kontrollerede forhold, det vil sige, at WAV-filerne er codec-komprimerede via en computer, hvorimod det tredje studie er live med talen transmitteret mellem to mobiltelefoner med og uden baggrundsstøj.

Resultaterne indikerer signifikante effekter af codec-komprimeringerne på de spektrale mål med segment-, codec- og bithastigheds-afhængige tendenser. Live transmission og baggrundsstøj producerede generelt større effekter end de kontrollerede forhold. Intensitet spillede en nøglerolle i omfanget af virkningerne af codec-komprimeringer og live-transmission.

Dette har implikationer ved brug af codec-komprimeret tale som data, men især i socio- og retsmedicinsk fonetik med mulig spredning af lydændringer og sammenligning af talere. Derudover har resultaterne betydning ud over lingvistik, f.eks. i psykologi, hvor taleklarhed spiller en rolle i vurderingen af karisma, samt i høreapparat- og cochlear implantatteknologi, som begge tilgår tale digitalt og inkorporerer støjreduktion.

Acknowledgements

Once upon a time, when I was 15 and passed my English exam in the Danish *Folkeskole*, my teacher smiled at me, while he pointed out there, somewhere in the future, and said ‘university is that way’. Martin you were right, and what a road it has been to travel.
Fast forward 10 years, and a bit, and here we are. One bachelor’s degree, one Master’s degree and now a PhD project later.

This project has taken me on a journey. A journey academically, as a person, and to a new country and a new home. It has at times been travelled alone, but it has also been travelled with so many good people. It is an adventure, which I will be forever grateful for, and which require a thank you or two.

I cannot mention everyone who has been part of this project over the last three years, and to whom I owe a thank you, but some stand out. So thank you to Mette, who introduced me to forensic phonetics and mobile phone transmitted speech way back when; to Olli, who inspired me to go beyond linguistics and sparked a new passion for my research; and to Jonas and Chris who showed me and helped me through the world of statistics. I also wish to express my gratitude for the inspiration, talks, and collaborations to everyone in the Sounds of Language and Speech research group at Aarhus University and the Forensic Speech Science research group at the University of York.

I would also like to thank my supervisors for all the support. Phil, for the many conversations about the technical aspects of the project, and to both him and Virginie for being there for me, when I first got to York and for helping out, when I was stuck in a dorm room with Covid 19. Paul, for the valuable input and the wise words, which picked me up in moments, where I was about to give up.
And Míša , my supervisor in Aarhus, who has been with me on this journey since my MA, for the encouragements, conversations, knowledge, critical questions, and for being my rock through this project.

The people who have travelled with me, have not only travelled, a lot of them have danced. So to all of you, who danced with me through this project, thank you. You have kept me sane and reminded me that there is more to life than whether the R-script messed up, or Word deleted an entire day’s work.

Finally, I am forever grateful to my friends in Denmark and England, who have been there for me with conversations, walks in the forest, *fyraftens-øl* (you know, who you are SRAU), and hugs when I needed them the most. And to my family back in Denmark, the biggest thank you for all the love and support, regardless of where in the world I have been.

With all of that said, this journey, like so many others, has been about courage. Courage to believe in yourself. Courage to take a change. Courage to share your ideas with the world. I am proud that I had that courage, and that I am now here with a finished PhD.

'Have no fear of perfection – you'll never reach it'

Salvador Dalí (1904 – 1989)

List of Tables

List of Figures

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Chapter 1 : Introduction

The motivation for the research presented in this thesis is a combination of two main aspects of digital transmission and their relation to speech. These are: a) the increasing amount of digitally transmitted speech in the everyday life of people all over the world, and b) the very limited knowledge of how this transmission affects one of the key elements of speech, namely, consonants and for the purpose of this thesis specifically the fricatives /f/, [fʲ], /θ/, /s/, /ʃ/, /z/, and /ð/. The distinction between /f/and [fʲ] is allophonic, and is made for reasons related to the forced alignment (see Chapter 3, section 3.3.4 for details). These fricatives are all English fricatives, as the data analysed will be in English. However, this does not mean that the results are only relevant for English, as the potential effects of the digital transmission of these fricatives are broadly applicable across languages. The thesis is in that way relevant for anyone who wishes to work with fricatives, particularly in relation to digital transmission. The rest of this section will be structured around three subsections, followed by three main research questions.

1.1. Digital background

In 2022, almost 73 percent of the world's population owned a mobile phone (International Telecommunication Union (ITU) 2022b). In addition, digitally transmitted speech has become a bigger part of everyday life even more so due to the Covid 19 pandemic. The International Telecommunication Union (ITU) further informs that in 2020 6,023 million people had an active mobile phone subscription and 7,505 million were covered by cellular networks (International Telecommunication Union (ITU) 2022a). In 2020, the number of daily active Skype users reached 40 million and Zoom reported 300 million meeting participants the same year (Galov 2023; Warren 2020).

Newer technologies as well as online meeting services are internet based, and the number of people with access to the internet is only increasing (see figure 1.1). ITU informs that '[…] approximately 4.9 billion people – or 63 per cent of the world's population – are using the Internet in 2021. This represents an increase of 17 per cent since 2019, with 782 million people estimated to have come online during that period (International Telecommunication Union (ITU) 2022b). It is thus clear that the number of users of speech transmission technologies, together with the spread of the internet, underlines the relevance of this research.

From a linguistic perspective, this is of interest as it heightens the individual's exposure to the acoustic variants of speech sounds affected by this type of transmission.

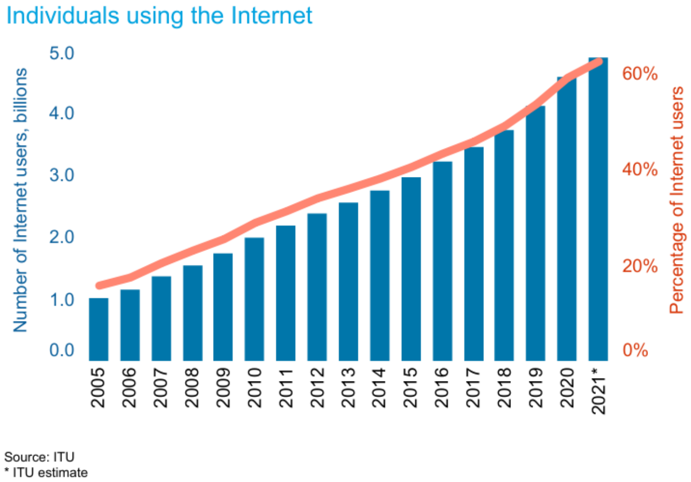


Figure 1.1. Individuals using the internet in billions

(International Telecommunication Union (ITU) 2022a)

In order to understand the possible effects of digital transmission on fricatives and the methodological approach taken in this thesis, it is necessary to understand some of the basic workings of digital transmission and essentially the codec compression. Codec is an amalgamation of *encoder* and *decoder*.

In short, the technical constrictions (e.g. limited data carrying capacity) placed on the transmission prevent all acoustic information from the original speech input to be maintained in the digitally transmitted output, and renders the signals *lossy* (Chakraborty, Misra, and Prasad 2019, 3; 3GPP 2020b). It is possible to have digitally converted speech as non-compressed (*lossless*) audio. However, this kind of audio is not fit for the technology and requirements of digital transmission of speech, because of the amount of data required to represent the complete signal. The choice of which information to include in the *lossy* transmission is made by the codecs (*encoder* and *decoder*). The design of these is based on intelligibility and noise removal (Chakraborty, Misra, and Prasad 2019; Guillemin and Watson 2006; Jamieson et al. 2002). This is illustrated by the way the codecs are built: '… each being designed with the overall goal of achieving the best perceptual speech quality, rather than maintaining the integrity of the individual acoustic parameters that make up the speech signal'

(Guillemin and Watson 2008, 484). This underlines the core idea behind the lossy codecs and the focus on intelligibility and acceptability.

The limited data carrying capacity has one other essential consequence, namely a limited sampling rate, and thus a limitation in bandwidth. The exact technical details of the codecs and the acoustic characteristics of fricatives will be described in depth in Chapter 2 in section 2.5 on digital conversion.

At the time of writing, three lossy codecs are prominent for speech and voice related purposes and will be the basis for this thesis. This is the Adaptive Multi-Rate Wideband (AMR-WB) codec used for mobile phones over the 3G (as well as in modified versions over newer generations e.g. 4G) Universal Mobile Telecommunication System (UMTS) network (3GPP 2020b; The 3rd Generation Partnership Project 2021); Opus used for Voice-Over-Internet Protocol (VoIP) including Skype, Zoom, and similar services (Valin, Vos, and Terriberry 2012; Chakraborty, Misra, and Prasad 2019; Goode 2002), and MPEG audio Layer-3 (MP3) used for e.g. digital audio broadcasting, electronic distribution of audio on online services (i.e. music) (Gayer, Lohwasser, and Lutzky 2003). The quality of these lossy encodings for all three codecs is dependent on the *bitrate*. The *bitrate* can be seen as the number of building blocks available to the network to build and represent the signal; the more blocks the more accurate the representation. Again, this will be elaborated on in section 2.5 to 2.8 in Chapter 2.

The codecs can work under two main conditions for the purpose of this thesis termed *controlled* and *live*. The *controlled* condition entails a high quality recording simply compressed with the codec via a computer, which means that speech signal is never actually transmitted. This provides a fairly controlled environment without influence from background noise or network capacity. However, while this is advantageous for research purposes, this is far from the reality under which digital transmission is experienced on an everyday basis. In everyday scenarios, the speech is *live* transmitted and the codecs have to deal with e.g. background noise and network access. This introduces a number of variables from equipment hardware to level of potential background noise. Nevertheless, from a linguistic perspective this is an essential condition to investigate, as this is the type of digital transmitted speech actually experienced by speakers/listeners and in forensic phonetic casework (Jessen 2018).

1.2 Interdisciplinary perspectives on digital transmitted speech

Consonants have been said to be less socially meaningful in comparison to vowels (e.g. Trousdale 2010, 116). However, consonants have in fact been found to carry both idiosyncratic and dialectal information in a variety of languages (e.g. Stuart-Smith et al. 2019 (English); Ayyad, Bernhardt, and Stemberger 2016 (Arabic); Torreira 2012 (Spanish); Pharao and Maegaard 2017 (Danish); Kong and Kang 2021 (Korean)), as well as essential phonemic contrasts e.g. in Mandarin Chinese (Luo 2020). This research on different languages illustrate the broader relevance of investigating consonantal sounds including fricatives. As previously mentioned, this thesis focuses on fricatives in particular because of their acoustic properties. These properties include aperiodic noise-like structure as well as typically high frequency energy components. The aperiodic structure means that they are more likely to be mistaken for noise, while the high frequency energy makes them more acoustically vulnerable to potential bandwidth limitations.

In linguistics, few fields have worked in depth with digitally transmitted speech, apart from forensic phonetics where the research, however, has mainly pertained to vowels, voice comparisons and earwitness accuracy (e.g. Hughes et al. 2020; Öhman, Eriksson, and Granhag 2010; Byrne and Foulkes 2004). The majority of speech analysed in forensic phonetic casework was mobile phone transmitted (Jessen 2018). Thus, forensic phonetics is one of the key motivations behind this project and will be described in further detail in section 2.9 in Chapter 2.

Another field where digital transmission has potential to be of interest is sociolinguistics. For sociolinguistics, the consequences of digital transmission on consonants is particularly relevant in terms of the potential phonetic variation caused and potentially diffused by digitally transmitted speech. This is together with the increasing level of exposure to this type of speech in people's everyday life. This will be expanded on in section 2.10 on sociolinguistics.

More broadly, digitally transmitted speech could aid and ease data gathering for any type of linguistic research, as it would eliminate many geographical and practical complications often encountered in such research (e.g. Labov 2000; Busso, Lee, and Narayanan 2007; Leemann et al. 2020; Sanker et al. 2021). However, it is essential to understand the exact implications of the digital transmission before engaging with such procedures.

These fields and a number of other fields, including studies of e.g. charisma and vocal attraction in psychology as well as hearing aid and cochlear implant development, could benefit from this research in different ways. For forensic phonetics, the benefit is probably the most straightforward because of the amount of work done on digitally transmitted files, where the knowledge of the potential implications of the acoustic implications of the transmission is directly applicable. For applied sociolinguistics the perspective might be similar, however, from a more theoretical perspective, sociolinguistics might benefit from this research in a number of different ways. This is for example in relation to the potential of linguistic variants being formed by the acoustic implications of the transmission or variants more generally being diffused via digital devices e.g. mobile phones.

For charisma and vocal studies, research has already established how digital transmission affects the level of perceived charisma and emotional prosody (e.g. Siegert and Niebuhr 2021; Niebuhr and Siegert 2023). Research into the acoustic implications at a segmental level could help understand e.g. the effects of digital transmission on clarity of speech, which is one of the features known to affect charisma evaluations (Siegert and Niebuhr 2021).

From the perspective of hearing aids and cochlear implant technology, the benefits are related just as much to the methodological approach as the results in themselves. Both hearing aids and cochlear implants in different ways apply digital technology to amplify and communicate e.g. a speech signal to the user, while also incorporating noise reduction technology (see Dhawan and Mahalakshmi 2016; Chung, Zeng, and Waltzman 2004; Wesarg et al. 2020). In that way, research as what is done in this thesis, which focuses on acoustics at a segmental level, might together with perceptual information improve the user experience by pinpointing the acoustic aspects, which are essential to the best possible user experience. In addition, the investigation of the effects of background noise and noise reduction algorithms in Chapter 5 is more directly applicable from the perspective of applying noise reduction algorithms in this type of technology.

All of the above mentioned areas that might benefit from the research in this thesis, all bridge between different fields of research. They all have some level of sound engineering in common because of the general interest in the technical components of the digital transmission, but then branch out to different areas and levels of linguistic research.

In that way, the project seeks to provide a better understanding of digitally transmitted speech within linguistics as well as an initial take on guidelines for linguists who encounter this type of speech in their research. In addition, the interdisciplinary structure inspires communication between different branches of research including linguistics, but not limited to this field. This type of interdisciplinary work has the potential to improve research in a number of fields, while also making it accessible to a broader audience. This provides a strong underlying motivation for this project.

1.3 Research questions

The thesis is structured around three main research questions (RQ) that will form the basis for the individual Chapters. Each Chapter will then include further specific and detailed research questions. These main research questions are:

mainRQ1: How are fricatives affected acoustically by digital transmission by different codecs at different qualities (i.e. different bitrates) in comparison to direct high quality microphone recordings?

mainRQ2: In what way does live transmission affect the acoustic implications of digital transmission on fricatives in comparison to high quality recordings and codec compression under controlled conditions?

mainRQ3: How can the understanding of the acoustic implications of digital transmission on consonants be applied in linguistic research (e.g. in forensic phonetics and sociolinguistics) and beyond?

These research questions will be answered in a final discussion and conclusion based on a collective view of the results from each Chapter.

Chapter 2 : Theoretical Background and Literature Review

This section consists of a number of subsections and for reasons of space primarily, but not solely, focus on English fricatives. The first subsections include an introduction to the history of speech transmission, the fundamental aspects of the technical composition of digital transmission, and the different technologies in use today (i.e. mobile phones and the internet). The second part has a linguistic focus. It provides more detail on fricatives and expands on the forensic phonetic and sociolinguistic perspectives. Finally, previous findings on digital transmission's consequences for speech are presented along the way.

The terminology in this section will frequently include *quality*. In many instances in both technical and linguistic literature, the term is used without an exact definition. However, from an engineering perspective and in evaluation of communication systems *quality* is generally seen as a correlate of acceptability and intelligibility (Raake 2006, 13–22). On the other hand, in linguistics *quality* is used both to describe perceptual and acoustic accuracy depending on the field of study (Guillemin and Watson 2008; Alzqhoul, Nair, and Guillemin 2012). In this thesis, because of the acoustic focus, *quality* will, unless otherwise stated or mentioned in relation to perception specifically, refer to acoustic quality.

Another important distinction, which will become relevant as the thesis progresses, is the distinction between phonetic and phonological voicing. As the main focus of this thesis is the codecs, the fricatives will be divided based on phonetic voicing in the analysis, and if not otherwise explicitly stated, *voicing* including *voiced* and *voiceless* refers to the phonetic definition. However, as it will become evident, the forced aligner used for the following studies makes the voicing distinction based on phonological criteria. This will be addressed in the results in each Chapter as well as the main discussion and conclusion.

The source filter model is essential for both the fricatives and the digital transmission technology. It forms the foundation for the underlying understanding of the acoustic composition of fricatives, as well as the basic build of the digital speech transmission technologies. Therefore, the basics of the source filter model will be laid out.

2.1 Source-Filter model

The source-filter model is based on the articulatory stages of sound production and was first coined by Fant (1960). As the name suggests, it is based on the idea of a source, which produces the energy before then being filtered and finally being audible as a sound. The source is the glottis, while voicing is produced by vibrations of the vocal folds. The energy in the form of airwaves is passed through and filtered by the vocal tract and finally the lips. Ladefoged divides this process into three stages following the energy input and excluding the actual output. The first stage is the vocal fold shaping, followed by filtering by the vocal tract and thirdly the radiation factor caused by the lips (Ladefoged 1996). The source filter model is illustrated below in figure 2.1.

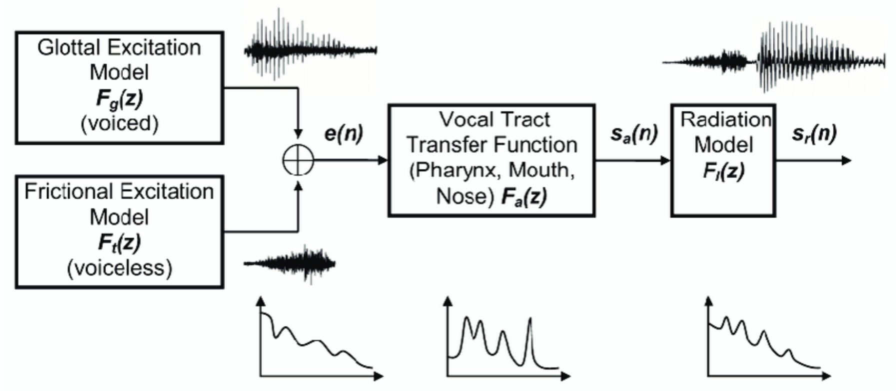

Glottal Excitation Model $F_g(z)$ (voiced)
Frictional Excitation Model $F_t(z)$ (voiceless)
$e(n)$
Vocal Tract Transfer Function (Pharynx, Mouth, Nose) $F_a(z)$
$s_a(n)$
Radiation Model $F_l(z)$
$s_r(n)$

Figure 2.1. Individuals using the internet in billions

(Muñoz-Mulas et al. 2013, 27).

On the surface, this process is simplified when adapted to speech synthesis and subsequently digital transmission of speech. This can be illustrated via linear predictive coding also known as LPC, which is one of the basic methods underlying analysis in e.g. Praat as well as speech synthesis and a number of digital speech transmission technologies (Weenik and Boersma 2004). For LPC the three steps above are combined into one step of filtering, which aims to obtain the filter characteristics from the speech signal (Ladefoged 1996).

The LPC and digital approach to speech thus work by separating the source (the energy of the sound waves travelling through the air within the vocal tract) from the filter. For the filters in these methods "[…] the spectral shaping characteristics of the glottal source and the lip radiation are incorporated into the same filter representing the characteristics of the vocal tract" (Ladefoged 1996, 182). More precisely, this filter is a mathematical approximation of the vocal tract determined by a number of parameters set by a combination of algorithms. Again, this will be elaborated on in further detail in section 2.5 below when the specific codecs are considered.

The important point to take away here is the fact that this approach naturally works better for vowels, with more regular resonances and harmonics than e.g. turbulent irregular fricatives. This is because the mathematical model of the filter will always only be an approximation and rely on the assumed redundancy of speech. Thus, the LPC analysis and any other mathematical approximation of the filter can be compared to shining a few torchlights into a cave, and predicting the outline of the cave from what you see. Increasing the number of torches will improve the quality of the prediction, but also require more resources, which are not necessarily available. Similarly, the amount of capacity available to the different speech handling technologies are finite and the quality will thus, be limited by the resources available to that specific service. This will be elaborated on in sections 2.6 to 2.7 in this chapter.

2.2 Fricatives

Fricatives are acoustically complex because of their constriction based articulation, which create an inherently aperiodic turbulent airflow (Shadle 1985). This means that no one single metric has been found to distinguish and define fricatives (Jongman, Wayland, and Wong 2000). Moreover, this complexity and focus on vowels in the literature mean that research on this topic is still sparse, and the information still relies heavily on a selected set of studies done before and around the millennium, as evident from the previous references. With this in mind, summarised below are some of the known acoustic characteristics of English fricatives.

Overall fricatives can be divided into a number of sub-groupings, but for this project three of these are particularly interesting. These three are based on: a) whether these are voiced vs. voiceless, b) place of articulation, and c) sibilant vs. non-sibilant.

Firstly, voicing is determined by the vibrations, or lack thereof, of the vocal folds. If the vocal folds are vibrating, the resulting sound is voiced, whereas if the airstream passes without any such vibrations, the sound will be voiceless. For the voiced fricatives a tendency has been found for them to be produced as approximants and specifically in the case of /ð/ as a plosive (Johnson 2011; Zhao 2010).

Secondly, the acoustic characteristics of the individual fricatives are naturally strongly related to their articulation. '[T]he longer the anterior cavity, the more defined the resulting spectrum' (Jongman, Wayland, and Wong 2000, 1253). Thus, the palato-alveolar fricatives (i.e. /s/, /z/, /ʃ/) have what can be considered well-defined spectra, whereas the labio-dental and inter- and dental fricatives have a flat spectra (i.e. /f/, /θ/, /ð/). Knowledge of this is essential when doing spectral measures of these sounds, as well as understanding why these measures might distinguish between certain groups and segments, and not others. This will be addressed in further detail below.

The distinction between the well-defined and flat spectra fricatives follow another categorisation of fricatives, namely the distinction between sibilants and non-sibilants. The sibilants include /s, z, ʃ, ʒ, tʃ, dʒ/which are characterised articulatorily by being produced with the tongue directing the airstream towards the teeth. On the other hand, the non-sibilants (e.g. /f/, /θ/, /ð/) are produced with an actual contact between lips, tongue and/or teeth. Acoustically this is of interest because the sibilants for this reason are characterised by relatively higher pitch and amplitude (Shadle 1985).

In addition, the fricatives can also be described in terms of their intensity level. This has been done by Strevens (1960). Of the fricatives analysed in this thesis, /θ/'s had the lowest intensity per unit air-pressure, while /ʃ/ had the highest. The fricatives in this ranking are illustrated in figure 2.2. below.

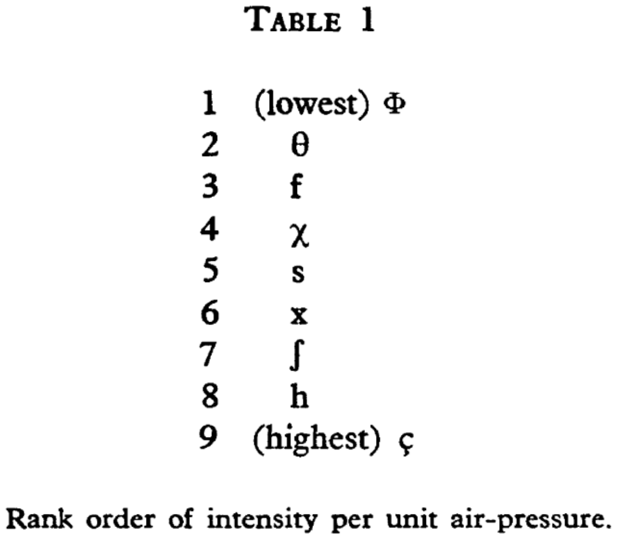


TABLE 1

1	(lowest) Φ
2	θ
3	f
4	χ
5	s
6	x
7	ʃ
8	h
9	(highest) ç

Rank order of intensity per unit air-pressure.

Table 2.1. Intensity levels per unit air-pressure
from Strevens 1960 (37)

2.3 Spectral measures

Despite methodological differences, a number of spectral measures have been found to characterise and distinguish fricatives, particularly sibilants from non-sibilants, as well as within-group differences for sibilants (e.g. Blacklock 2004; Shadle and Mair 1996; Jongman, Wayland, and Wong 2000).

Amongst these are the first four spectral moments. The first two moments describe the *mean* and *variance* in the energy distribution. The third and fourth moments, skewness and kurtosis, describe the spectral tilt and the peakedness of the distribution (Jongman, Wayland, and Wong 2000). Positive skewness indicates a concentration of energy above the mean and vice versa for negative skewness. A positive kurtosis value indicates relatively high peakedness, and a negative value a relatively flat distribution (Jongman, Wayland, and Wong 2000) (see figure 2.3 below). For skewness and kurtosis, Shadle and Mair (1996) note that these two moments again are found to distinguish sibilants from non-sibilants. However, research on skewness and kurtosis is still very limited.

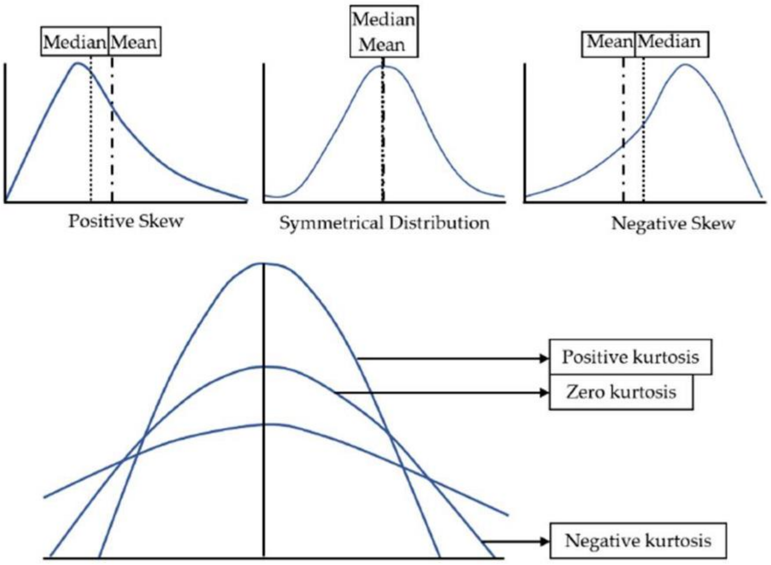


Figure 2.2. Illustration of possible distributions at different skewness (top) and kurtosis (bottom) values (Garrido et al. 2020, 3399)

The reports on spectral characteristics of fricatives are both limited and varied. This is as mentioned due to a number of factors, including the prevalence of vowels in research, as well as the aperiodic and highly varying acoustic structure of these sounds, and the use of different methodological approaches. The two first points entails that a lot of the studies of these sounds are not only related to the acoustic characteristics of these sounds, but also whether the measures are essentially useful to even characterise them.

The latter includes e.g. use of different window types for the analysis e.g. Hanning of Hamming windows. In short, in analysis of a section of a signal e.g. the central part of a segment, windowing is required to eliminate erroneous high frequency components appearing as a result of incomplete waves at the beginning and end of the section in question. To avoid this, a window is applied, which smooths the slopes and avoids this effect. These slopes can vary in their distance to zero, which is e.g. what distinguishes the Hanning and Hamming windows (see Ladefoged 1996).

Another methodological consideration is important when evaluating spectral measures, including peak values in downsampled audio. Any spectral measure in audio down-sampled to below the spectral peak, whether done as part of an analysis framework or by digital transmission, will result in the measures no longer representing their true value. This is in comparison to the value, which would have been obtained measuring from a full bandwidth recording.

With this in mind, a number of studies have been conducted and the results of some of these will be reported here to give an indication of the acoustic composition of the fricatives in question.

Shadle and Mair report on the spectral characteristics of /f/, /θ/, /s/ and /ʃ/ in three different positions in the segment (i.e. initial, mid, and final) in different vowel contexts, using FFT analysis measured by the first four spectral moment i.e. Centre of Gravity (CoG), Standard Deviation (SD), skewness (skew) and kurtosis (kurt) (1996). For CoG and SD, their findings are illustrated in figure 2.4.

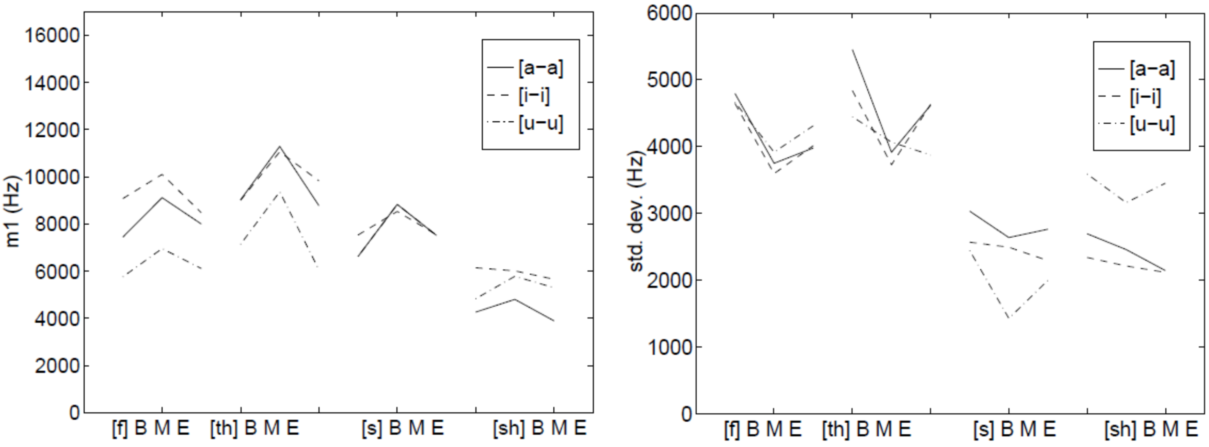


Figure 2.3: M1 i.e. CoG (left) and SD (right) for three vowel contexts and three positions beginning (B), mid (M), and end (E), [sh] = /ʃ/ from Shadle and Mair (1996, 394).

Their results illustrate how the first two spectral moments vary across the segment as well as dependent on vowel context. They measured skewness and kurtosis as dimensionless by normalising by the power of variance. They only report the results for skewness for sustained fricatives at different effort levels, and found little variation based on position in the segment and values from just below 0 to just below 2. For kurtosis, they found similar patterns for all the measured fricatives with values between 0 and 5 in all contexts, apart from /s/ in the [u-u] context, which had values at the midpoint just below 15. They concluded: ‘Spectral moments do not distinguish reliably between fricatives, but the variations were consistent with acoustic analysis. The moments proved to be […] relatively insensitive to position within the fricative (in VCV's) […]’ (Shadle and Mair 1996, 195). As this thesis is interested in digital transmission, which has a limited bandwidth, it is worth noting here that all the measured segments either had CoG values around or above 8 kHz at the midpoint.

Secondly, Maniwa et al. investigated differences in acoustic characteristics of fricatives between *clear* and conversational speech in VCV contexts (2009). Their *clear* speech is closest to the read speech dataset used in this thesis. Their measurements were done on 44.1 kHz files, using a Discrete Fourier Transform with a Hamming window. Their results for CoG and SD are illustrated below in figure 2.5 and 2.6.

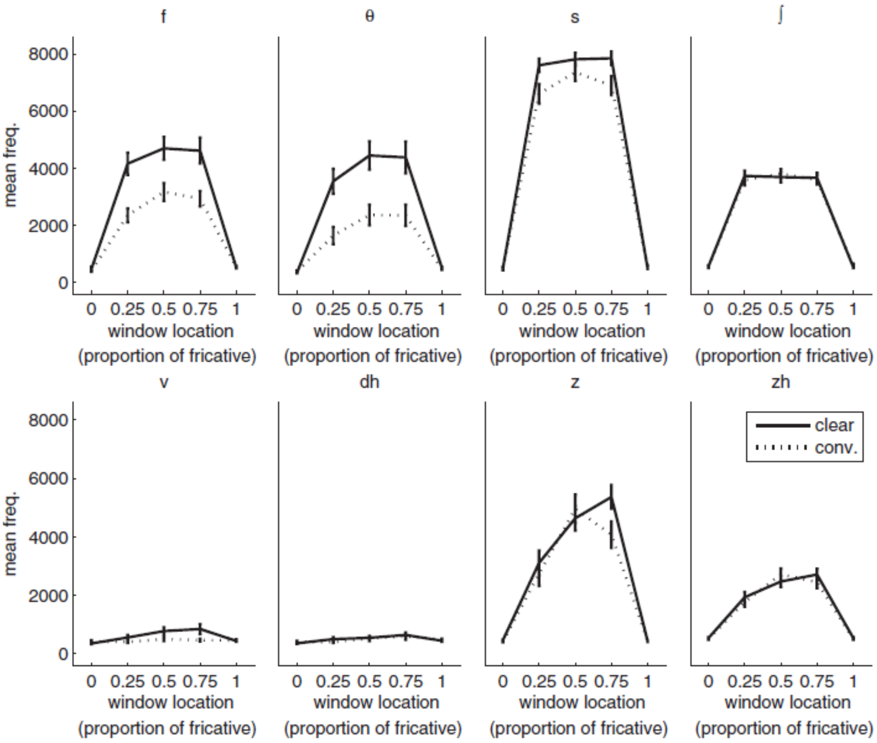


Figure 2.4. mean freq. (CoG) for each of the fricatives across window locations from Maniwa et. al dh = /ð/and zh = /ʒ/(2009, 3968).

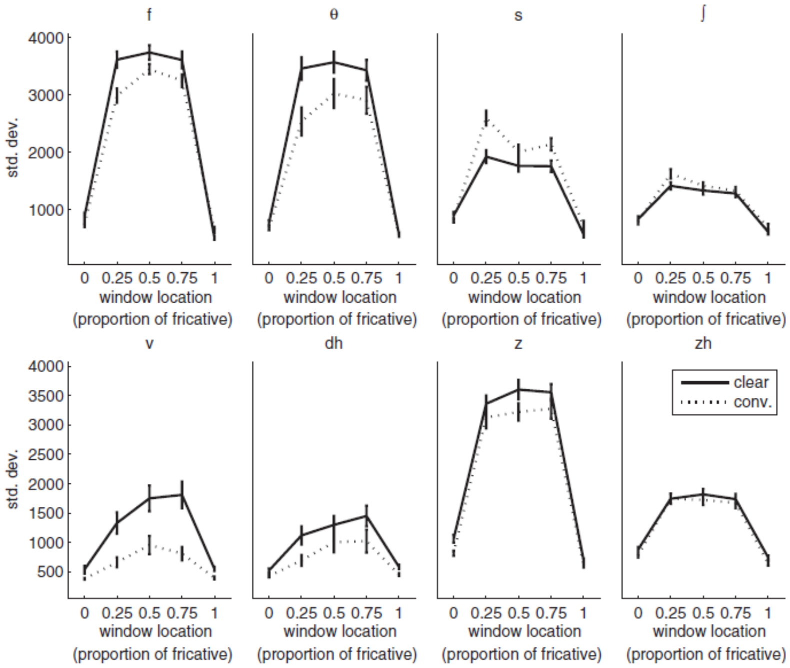


Figure 2.5. std. dev. (SD) for each of the fricatives across window locations from Maniwa et. al. *dh* = /ð/and *zh* = /ʒ/(2009, 3969).

It is worth noting here that only the CoG values for /s/is close to 8 kHz in their study, while the other fricatives had their values centred relatively lower in the spectrum e.g. between 4-5 kHz for both /f/, /θ/, and /z/. These three sounds also behaved similarly for SD. Of the fricatives, which will be investigated in this thesis, /ð/had the lowest CoG and SD, but for SD /ʃ/ presented similar values.

The mean values for skewness were found to be just around 0 for most segments, but went up to around 6 for /ð/. Finally, the mid-point mean values for kurtosis, like skewness, generally centred around 0 apart from /ð/, which had a mean value just over 50. The values for kurtosis at the initial and final point of a number of segments reached above 100 and for /ð/all the way up to 200 (Maniwa, Jongman, and Wade 2009). These values as previously mentioned indicates peakedness, and the higher the value above 0, the more peaked the spectrum. They concluded: 'In sum, this study demonstrates that there are systematic acoustic-phonetic modifications in the production of clear fricatives' (Maniwa, Jongman, and Wade 2009, 3972).

Thirdly, Jongman et al. used FFTs and a Hamming window (2000). They used files with a sample rate of 22 kHz sampled files and presented the fricatives in CVC context for both male and female speakers. The segments are, as in the previously presented studies, divided into onset, mid-point and offset (Jongman, Wayland, and Wong 2000, 1255). The results for the spectral moments were averaged across the voiceless and voiced segment pairs, as well as window location (e.g. one value for /f/and /v/). The results for CoG and SD are illustrated below in figure 2.7.

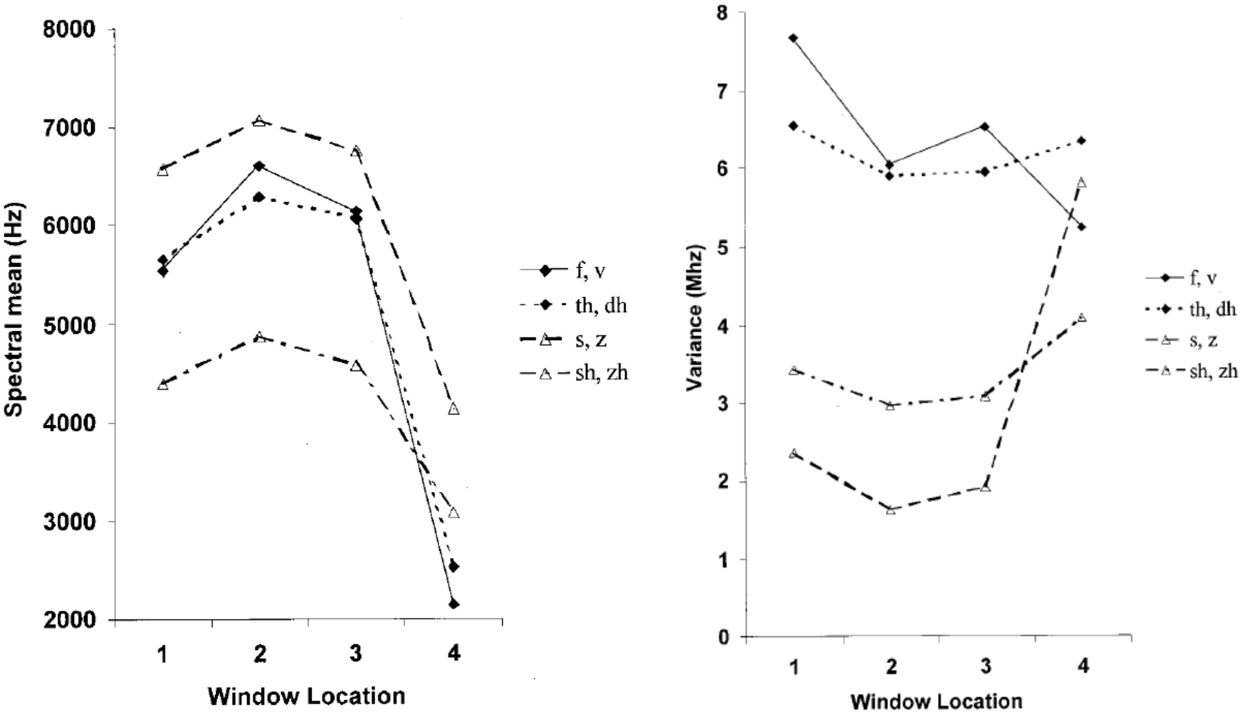


Figure 2.6. Spectral mean (CoG) (left) and Variance (SD) (right) across vowels, voicing, gender (male and female) for each window locations (1-3 first, middle, and final. 4 = final 20 ms of fricative and first 20 ms. of following vowel) as a function of place of articulation. t*h* = /θ/, *dh* = /ð /, *sh* = /ʃ/, and *zh* = /ʒ/(Jongman, Wayland, and Wong 2000, 1257)

The results are slightly less clear to interpret because the average is made across the segments, however /f/and /θ/behaved similarly, while /ʃ/ and /ʒ/had clearly lower CoG values.

The skewness values spanned from -1 to just above 0.5. /s, z/had the highest value and /f/the lowest. The kurtosis values are unclear due to an inconsistency between the results reported in the tables and in-text. However, it can be seen that the values span between just below 0 to just below 8.

In addition to the spectral moments, they also measured spectral peak and found both /f/and /θ/had a spectral peak just below 8 kHz, while /s/and /z/had close to identical peaks just below 7 kHz. /ʃ/had the lowest peak just below 4 kHz (Jongman, Wayland, and Wong 2000). They concluded that spectral peak distinguished sibilants from non-sibilants as well as /s,z/from /ʃ, ʒ /and /f,v/from /θ, ð/. SD and skewness were found to be the moments most effectively distinguishing the places of articulation, while the fricatives were most distinct in the initial and final part. They stated 'in general, the present data clearly show that four places of articulation were distinguished by most moments at most window locations' (Jongman, Wayland, and Wong 2000, 1261).

Based on these studies, it is clear that, moments can be used to distinguish sibilants from non-sibilants, as well as certain within-group differences for sibilants. However, even though spectral moments are good for sibilants, Blacklock states that '[the] very flat spectral shape is common amongst the non-sibilants, and this is clearly one of the root causes of the inability of spectral moments to differentiate them' (2004, 86). In Blacklock's research, despite the fact that the differences between the fricatives did not always reach significance, each fricative presented distinct patterns for almost every spectral moment.

Beyond English, this is supported by a cross-linguistic study by Gordon and Sands on voiceless fricatives in seven less widely spoken languages, including Gaelic, Apache and Toda, where they concluded that CoG can be used to distinguish the fricatives in each (2002).

These studies show how spectral moments and spectral peak cannot solely account for the variation in acoustic properties between individual fricatives, but do provide useful information especially in the distinction between sibilants and non-sibilants. In addition, the studies illustrate how different methodological approaches will produce notably different results e.g. the differences in CoG mean values from above 8 kHz in one setup, to around 5 kHz in a different setup. More specifically, it is clear that /f/and /θ/often behave similarly based on spectral moments, while /s/generally got the highest CoG values regardless of method and /ð/the lowest. Some of the results e.g. /z/behaving like /θ/ and /f/and /ð/having high kurtosis values, suggest some influence from the previously mentioned tendency for voiced fricatives to be produced e.g. as approximants and plosives.

In sum, despite the inconsistencies, spectral moments and spectral peak are well-founded measures of both local and global spectral characteristics of these types of sounds, and thereby useful for the current study to define any potential spectral changes caused by the codec compression.

2.4 History of speech transmission

The first communication systems were invented with one key objective: to connect people by transferring speech over long distances at sufficient perceptual quality (Chakraborty, Misra, and Prasad 2019, 1). The first to succeed in such communication was Graham Bell with the invention of the telephone in 1876 (Rutter 1987). The telephone originally worked by connecting two communication devices with a copper wire, but technology has since developed and become far more advanced (Rutter 1987) as a desire to simplify and improve the services provided by the telephone quickly arose. In the 1930's at the Graham Bell institute, Homer Dudley invented the speech vocoder (Atal 2006). This was the first attempt to synthesise speech by extracting individual acoustic elements of the signal. The vocoder would turn out to be the first step on the road to digital speech transmission (Atal 2006). These machines succeeded in compressing the speech signal in a way that left it intelligible at low amounts of data, which in turn provided an increased bandwidth and reduced delay in comparison to previous technology.

However, these original vocoders were not sufficient in perceptual quality for the long distance communication provided by telephones. They were however, used in some instances during the Second World War to provide secure lines of communication (Atal 2006). From vocoders, the technology progressed and in 1973 the first cell phone was introduced (Loeffler 2021).

The development of mobile phone technology accelerated rapidly (Dan 2014). The popularity of mobile phone technology can be clearly illustrated via the number of subscriptions across the world as also mentioned in the Introduction.

Mobile phone technology is divided into generations and today extends from 1st generation (1G) to 5th generation (5G). The first generation network was analogue, whereas the second generation was the first to be completely digital (Dan 2014). The digitisation was possible because of the implementation of effective speech compression-decompression algorithms (codecs) into the transmission process (Dan 2014) (see section 2.5 to 2.7 for further detail on this). The codecs have developed alongside the generations of mobile phone technology (The 3rd Generation Partnership Project 2021).

The 3rd Generation (3G) works without internet access and with technology based on the original Global System for Mobile Communication network (GSM network). One of the developments, which is based on this technology, is the Universal Mobile Telecommunications System (UMTS), which is the European standard (The 3rd Generation Partnership Project 2021).

If the data connection is switched on, and the phone is compatible with the technology, the transmission will access the 4G or 5G technology, which both use Voice over Long Term Evolution (VoLTE) (The 3rd Generation Partnership Project 2021; Sauter 2010, 206). This technology includes the same codecs as 3G but at better overall quality, because it makes use of the internet (Nguyen, Nguyen, and Renault 2016, 2). This thesis will be working with 3G as both 4G and 5G cannot be guaranteed to be consistent throughout a call depending on location and network access. In 2017, ITU published a Quality of Service Manual, which included the graph illustrated below in figure 2.8

indicating a clear trend. It should be noted here that the fact that the technology is available to people does not entail that it is also used by the same number of people.

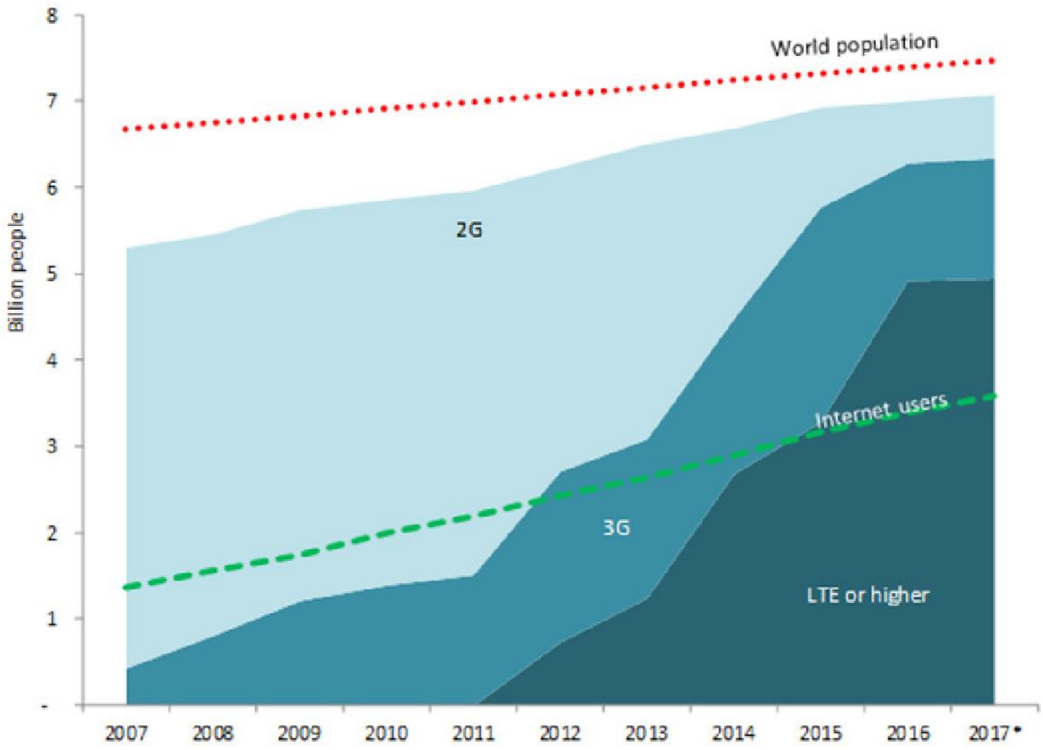


Figure 2.7. Figure of 'Mobile broadband network deployment trends' adapted from ITU Quality of Service Regulation Manual (Janevski, Markus, and Jankovic 2017, 3)

As the internet and the World Wide Web have expanded and just over 66 percent of the world's population had access in 2022 (International Telecommunication Union (ITU) 2022b), it ensured connectivity between regions across the globe, the implementation of communication technology using the internet has followed e.g. VoLTE (Chakraborty, Misra, and Prasad 2019, 1). Outside of mobile phone technology, this has happened in the form of a system called Voice over Internet Protocol (VoIP). VoIP combines internet and already existing communication technologies, such as those used in mobile phones again, 'in order to reduce the cost of communication and also merge the data services with voice' (Chakraborty, Misra, and Prasad 2019, 1). Chakraborty et al. explain this as follows, 'the goals of VoIP implementation are to achieve (a) significant savings in network maintenance and operations costs and (b) rapid rollout of new services' (Chakraborty, Misra, and Prasad 2019, 3).

All in all, the overall quality of communication systems have developed and improved with the speech handling mechanisms becoming increasingly sophisticated over the years. However, as it will become evident from the remainder of this Chapter and from this thesis as a whole, these developments cannot unconditionally be rendered an improvement from all linguistic perspectives.

2.5 Digital Conversion

Apart from physical appearance, in simple terms the main difference between landline telephones, mobile phones, and internet technology, is in how analogue and digital signal transmission is used. In this project, when talking about digital signals, it will refer to signals using pulse-code modulation (PCM) (see Garg 2007 for further information). In addition, it is important to make a distinction between *digitisation* of a signal and actual *digital transmission* of a signal. The former refers to the process in which a continuous wave is converted into a digital signal, whereas the latter is the transmission of this signal between two communicative devices.

To understand digital signals, it is useful to begin with its predecessor used for e.g. landline transmission: analogue signals. The term *analogue* refers to the fact that speech in itself is constituted of continuous waves propagating through air, and it is the transmission in this continuous form, which renders a signal analogue. Specifically, the analogue signal transmission works as a representation of the original sound wave. It does so by converting the time varying pressure changes of the sound wave into electromagnetic waves with analogue time varying voltage changes (Ladefoged 1996). Essentially, the signal stays in the form of waves and is never converted to a different medium e.g. digits.

In comparison, instead of the continuous form in analogue transmission, digital transmission converts the soundwave into digits and in consequence, linearly spaced levels. In other words, the original speech input is translated from a continuous wave into discrete numbers. These numbers represent a model of the speech input, which is, what is transferred between the communicating devices. The model is made by the encoder at the sending end, and then at the receiving end the model is translated from the discrete numbers back into an intelligible speech signal (Johnson 2011). An illustration of the difference between an analogue and a digital signal can be seen in figure 2.9 below.

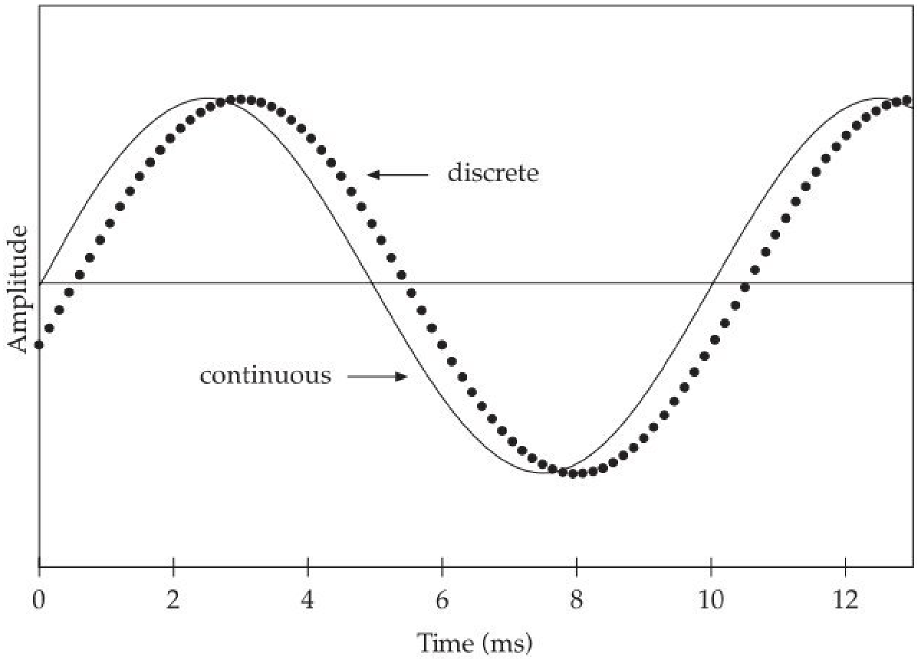


Figure 2.8. A continuous (analogue) sine wave plotted together with a discrete (digital) sine wave with time on the x-axes and amplitude on the y-axes from Johnson (2011, 50)

The model and the choice of information to include is, as previously mentioned, done by the codecs. The codecs are a part of the actual transmission and consist of a complex set of algorithms, which set the parameters for which parts of the speech signal are transferred between devices. The mathematical details of these algorithms are beyond the scope of this thesis.

Three codecs are investigated in this thesis, they are: AMR-WB, Opus, and MP3. These three codecs can be divided into a set of sub categories including: narrowband (NB) and wideband (WB), and speech and psychoacoustic codecs. The NB codecs are primarily used in 2G, and implement the telephone bandwidth of 300 Hz to 3,400 Hz, and is thus not of interest here. The WB codecs have the improved bandwidth from 50 Hz to 7-8 kHz and will be the type investigated in this thesis (3GPP 2018).

The human ear can perceive sounds between 20 Hz and 20 kHz. Speech typically consists of frequencies up to 10 kHz with the most relevant information below 8 kHz (Johnson 2011; Ladefoged 1996). Human speech is reported to have significant energy between 200 Hz and 4 kHz. In consequence, because of their fixed amount of data carrying and storage capacity, telephones, mobile phones and VoIP take advantage of this fact, and work within a limited bandwidth (see section 2.6 and 2.7 for specific information on the individual technologies).

For WB, this means that speech sounds between 7-8 kHz and 10 kHz are essentially excluded from the transmission. Maher states the following about this: ‘increasing the audio bandwidth generally results in listeners happy with the improved *quality* of the speech, but the *intelligibility* does not necessarily improve even if listeners perceive that the quality is to be better’ (2018, 25).
However, since speech related acoustic information is present at these excluded frequencies, the acoustic quality of the signal is potentially degraded, especially from a linguistic perspective.

From a technical perspective these frequencies are represented by and dependent on the sampling rate (number of sample points per second). The sampling rate is directly related to the frequency range of the signal, as the upper limit of the bandwidth (in Hz/kHz) in a digital signal will always be half the sampling rate. This maximum frequency is known as the Nyquist frequency (Johnson 2011). This means that to represent all speech sounds a sampling rate of 20 kHz is needed. So, as it is evident from the information on bandwidths above, newer technology typically samples the speech with up to 16,000 samples per second (sampling rate of 16 kHz), which means not all speech related frequency information can be included.

The speech and psychoacoustic distinction is slightly more complex. AMR-WB and Opus can both be classified as *speech* codecs, as they are designed to process speech and are based on source-filter modelling (see Sauter 2010). It should be noted that Opus is in fact a hybrid codec, but due to the way it is used in this thesis, will work as a speech codec.

For these codecs, the source is, instead of the vocal fold vibrations, the energy from the waves travelling through the air. The filter is then the mathematical approximation (i.e. model) constructed by the codecs to imitate the vocal tract. This creates an estimate of the original speech signal via a mathematical model of the original speech spectrum.

More specifically, these codecs incorporate a number of features aimed to identify speech, which is primarily based on voicing criteria (3GPP 2020b; Valin, Vos, and Terriberry 2012). Essentially, the speech codecs use a version of tone-detection, which is based on pitch period-related thresholds that need to be surpassed for a *frame* (typically 20 ms of audio) to be recognised as containing ‘a signalling tone, voiced speech, or other strongly periodic signal’ (3GPP 2020b, 8).
On the other hand, MP3 is a *psychoacoustic* codec (i.e. only what is audible to the human ear is to be encoded) designed for music rather than speech Thus, MP3 does not incorporate such speech specific features, but bases the encoding on principles of psychoacoustic modelling (Tan and Jiang 2019; Yost

2015; Herre and Dick 2019). It identifies and encodes any psychoacoustically relevant audible audio based on loudness criteria (Yost 2015). The primary feature in MP3 is frequency and temporal masking of non-tonal content based on FFT analysis, which is inferred to exclude any inessential information from the signal i.e. noise (see Herre and Dick 2019 for further details). More details on the technical specifications of these speech identification features for the individual codecs can be found in section 2.6 and 2.7.

With this in mind, the level of detail in a digital signal is dependent on a number of predefined parameters related to frequency and amplitude, one of them being the sampling rate mentioned above. Apart from the sampling rate, the representation of amplitude is, essentially, what determines the accuracy of the digitised waveform illustrated in figure 4 above.

The process, which generates the information about amplitude in the digital signal, is called *quantisation* and is expressed in bit depth. Bit depth is the number of bits (the smallest units in a digital signal expressed in 0 and 1s) per sample.

Quantisation is the way the digital signal constrains the infinite number of possible amplitudes in an analogue signal and approximates the continuous input via a relatively small set of fixed values (see Rumsey 2009). These values are the available bit depths and consist of a set of equidistant levels. In that way, this is what gives the digital signal its characteristic linearly spaced form.

The bit depth (i.e. the number of levels) is what determines the *dynamic range* (the range of amplitudes represented in the signal). The dynamic range is measured in decibel (dB) and increases with the bit depth used in the digitisation. The higher the dynamic range, the more information about the amplitude of the signal will be preserved in the digitised version. In other words, the bit depth used in the digitisation correlates with the dynamic range, which in turn describes the accuracy and resolution of the digital estimation of the original speech input. Again, it is possible to calculate the dynamic range from the bit depth:

$$number\ of\ bits\ \times 6dB\ = dynamic\ range$$

All now mentioned components of digitisation are separate values and relate a specific aspect of the digital signal. However, in digital speech transmission, the sampling rate and the bit depth are often combined to express the overall amount of data, which can be transferred or stored per second. This is expressed as the bitrate and calculated in uncompressed signals as:

$$bit\ depth\ \times sampling\ rate\ \times number\ of\ channels = Bitrate$$

The bits per second is then expressed in kilobytes per second (kbps). In that way, the bitrate is an expression of the overall quality of the signal as it determines the amount of data, which is lost during the transmission. In speech communication technology, it is the codecs that allocate the bitrates as part of the compression process. What bitrates are available for each type of transmission technology and its consequences for speech will be elaborated on in the two following sections (2.6 and 2.7) on mobile phone- and VoIP technology.

Lastly, it is important to know that the transmission can vary dynamically between bitrates over time in live transmission. It is of course also possible in controlled conditions if it is so decided. Dynamicity is one of the cornerstones of live digital speech transmission as it allows the signal to vary in overall quality, e.g. depending on location and amount of network traffic (Alzqhoul, Nair, and Guillemin 2012, 29; Guillemin and Watson 2008, 197). In practice, this means that the signal is divided into smaller segments called *packages,* where different bitrates may be allocated to different packages. It is these packages that are compressed (by the codecs), sent over the network, and decompressed at the receiving end (Chakraborty, Misra, and Prasad 2019, 7). This process is referred to as *packetisation* and the packages typically comprise of 20 ms.

The packages do not necessarily arrive to the receiver in the order they were produced, and hence it is not unusual to experience delays and lost packages in package-based networks such as 3-5G and VoIP (Chakraborty, Misra, and Prasad 2019, 9).

In summary, the digital transmitted speech signal will always only be an imitation of the original speech input, and some will be more successful replications than others.

2.6 Mobile Phone Technology

This section outlines three main types of mobile phone technology and the AMR-WB, which form the basis for both 3G, 4G, and 5G.

Initially, the UMTS network handles the actual transfer of the speech between the mobile phones via cell phone towers. The different subscription companies own these cell phone towers. In consequence, the overall quality of a mobile phone call is dependent on the proximity of a cell phone tower together with which towers are available to the company providing the network access (Sauter 2010).

The newer VoLTE used in 4G and 5G works differently from the UMTS network. It still handles the transfer of the speech between devices, but does so via the internet. VoLTE supports greater data carrying capacity and transfer speed. Nevertheless, it still uses the UMTS network as it would otherwise not be possible to switch back to 3G in cases where VoLTE is not available (Sauter 2010).

Non-optimal conditions and lowered overall quality can be experienced if either the call is placed in a rural area with insufficient cell tower coverage or internet connection, or if the network is used by a significant number of people at the same time e.g. in crowded areas. Therefore, in these cases, the network cannot provide the optimal overall quality while still maintaining service and the bitrate is limited to fit the demand (Whitrow 2019).

From a linguistic perspective there are three particularly interesting features of the AMR-WB. These are: Voice Activity Detection (VAD), handling and insertion of background noise, and handling of lost or corrupted frames.

The VAD works by filtering any non-speech sounds from the input signal before transmission (3GPP 2020b). It does so via a pre-set energy level threshold, which the input signal then needs to surpass to be recognised as speech and transferred. When the threshold is exceeded the VAD will indicate the presence of speech and the acoustic information is encoded (3GPP 2020b; ETSI 1992). The threshold is defined based on a number of parameters such as tone and periodicity (e.g. if the sound is voiced). The higher the level of background noise defined by signal to noise ratio, the lower the threshold will be set for the VAD (3GPP 2020b).

In cases where the input signal does not pass the threshold, the frame will be left silent. Here, for reasons related to background noise-handling, *silent* does not entail silence, but simply that the frames are without speech. The VAD also forms part of the background noise-handling scheme in the AMR codec.

Background noise is not only removed from the signal, it is also inserted. In discontinuous transmission, in frames where no speech is present or detected i.e. silent frames, a low frequency noise is inserted to prevent the frame from actually being silent (Besette and Salami 2002). This type of noise is called *comfort noise*. The reason for inserting the comfort noise is the fact that listeners report that a complete silent period within a speech signal interrupts and disturbs the experience of continuity (Warren 1970). This is a finding which is replicated in psychoacoustic studies e.g. of phonemic restoration (e.g. Kashino 2006).

It should also be noted here that the filtering of background noise to some extent also happens in the hardware of mobile phones. The different hardware have different types of active noise cancellation incorporated (ANC). ANC works by recording the background noise via a microphone placed on the outside of the phone. This recorded signal is then inverted and played back to the listener together with the original signal (Kottayi et al. 2016). In that way, the two waves cancel each other out because the soundwaves' linear form makes them subject to the physical principle of superposition. In consequence, the input will be silent to the listener (Kottayi et al. 2016).

Lastly, frames can be lost or corrupted during transmission. The codec can either insert comfort noise in the given frame or replace it with a previous frame. Such replacement can happen in up to sixteen frames or a total of 320 ms (Alzqhoul, Nair, and Guillemin 2012).

Taken together, it is evident that even though the bandwidth has improved, the compression codecs implement a number of mechanisms e.g. VAD, which have the potential to influence the speech output in mobile phone transmitted speech particularly under live conditions.

2.7 Voice over internet protocol (VoIP)

As previously mentioned, there are two different internet based transmission technologies around i.e. VoLTE and VoIP. The most noticeable difference between VoLTE and VoIP is the type of devices, which can use each of these and what codecs are used. VoLTE is as mentioned designed for mobile phone devices, whereas VoIP is designed for apps and PCs primarily, as it is used for programmes such as Zoom and Skype (Whitrow 2019). VoIP is completely internet based, but works in the same way as the digital 3-5G mobile phone technology in terms of packetisation (Chakraborty, Misra, and Prasad 2019, 2). Moreover, VoIP uses both VAD and insertion of background noise (silent suppression) in the same way as the mobile phone technology described above.

A number of other factors also affect the overall quality of the signal in VoIP transmission. These factors include: packet loss, delay, delay variation (jitter), echo cancellation, and general network design (Goode 2002). The first four of these will be considered here as the general network design is above the level of detail needed for the present project, as the primary focus is at a segmental level (For more detailed description of the network design see Goode 2002; Chakraborty, Misra, and Prasad 2019).

Firstly, Packet loss occurs regularly in VoIP transmission, and similarly to mobile phone transmission the gaps in such cases can be filled with a comfort noise (Chakraborty, Misra, and Prasad 2019, 8). In VoIP transmission, this is referred to as packet loss concealment. It is also possible with VoIP transmission to draw information from previous packages, and restore the lost or corrupted frames on the basis of these (Chakraborty, Misra, and Prasad 2019, 9).

Furthermore, it is not uncommon to experience delay caused by packet loss. This delay may then vary during a conversation, which can in turn lead to jitter, '[which] can result in choppy voice or temporary glitches and must be minimized' (Chakraborty, Misra, and Prasad 2019, 9). For this reason a jitter buffer is implemented in the VoIP codecs. The packetisation process including the components mentioned above is illustrated in figure 2.10.

.

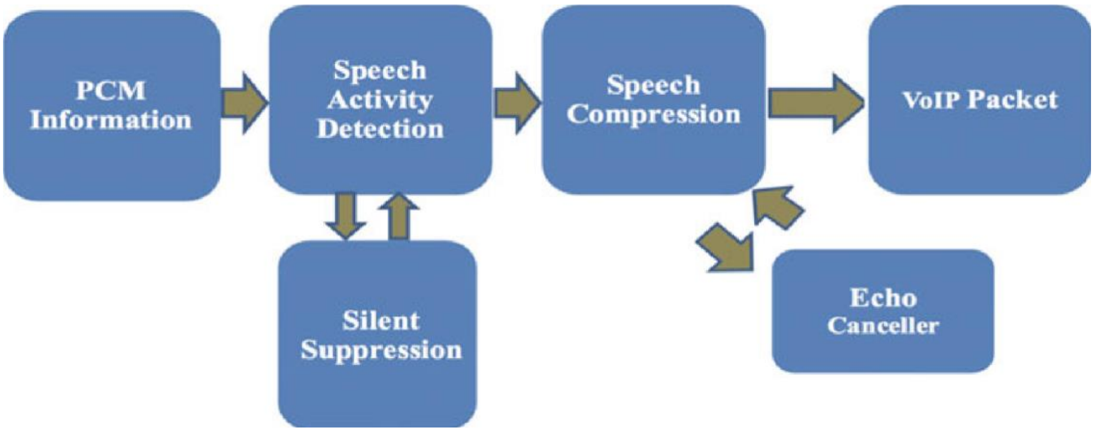


Figure 2.9. 'Development of VoIP packet'
(adapted from Chakraborty, Misra, and Prasad 2019, 10)

2.8 Speech sounds and digital transmission

The number of studies on the effects of digital transmission on consonants is limited, and even further so for the study of fricatives and all spectral moments. However, a number of studies have been done, and the results from some of these will be reported in this section to illustrate what is currently known about the effect of digital transmission on speech sounds.

Firstly, the fact that digital transmission and any kind of transmission involving a limited bandwidth (e.g. landline) do affect speech sounds has been established via studies of vowels, where formant frequencies are generally lowered (e.g. Zhang et al. 2013; Künzel 2001; Byrne and Foulkes 2004; Alzqhoul, Nair, and Guillemin 2012; Hughes et al. 2020). De Decker and Freeman investigated the effects on formant frequencies for vowels following transmission by Zoom and Skype (2021). They only investigated the speech of two speakers, one male and one female. They found that, for Skype the female speaker was the most affected, whereas Zoom had similar effects for both speakers. It is particularly interesting that they found the largest effects, when the recordings were done from smartphones and iPads, which suggests an influence of hardware. Specifically, they found a deviance in formant frequencies for these between 750 Hz and 1500 Hz.

A number of other studies have been conducted on remote data collection via digital media, e.g. Zoom or Skype, smartphones, and other remote recording technologies, especially following the pandemic. These mainly focus on the equipment, different recording opportunities via these, and experimental setup rather than the actual acoustic implications of the transmission on the speech (e.g. Leemann et al. 2020; Decker and Nycz 2011; van Son 2005). Regardless, they raise an important point, as these studies overall conclude an effect of type of hardware, e.g. more prominent effects on the vowel space when recording from smartphones. Despite the fact that these studies did not work with fricatives or consonants specifically, they highlight an essential aspect of the potential influence of hardware in live transmission.

One study which in fact aimed to investigate the acoustic and phonetic effects of digital transmission was done by Sanker et al. (2021). They measured CoG in different recording conditions and live transmission for three speakers. There are important caveats to their study, the limited amount of data and the fact it is unclear which effects are due to the codec compression and transmission, and which are due to the variation in equipment and experiment setup. With this in mind, they report that the higher the sampling rate, the higher the CoG. They further noted that fricatives are generally sensitive to the sampling rate. They used a Hanning window to generate the spectra, and investigate Facebook Messenger, Zoom and Skype. They find that Messenger failed to capture the difference between /s/ and /ʃ/, while Skype and Zoom, despite both using Opus, produce divergent CoG effects because of their unique noise handling mechanisms. They note that the contrasts between the segments largely stayed intact due to a consistent effect across each, but with a number of exceptions, where the contrasts were either exaggerated or underestimated. This included an overestimation of CoG for /f/, suggested to be caused by amplification of lower frequencies or filtering of higher frequencies. Intensity was not found to vary greatly between the conditions, however, a slight lowering was observed for two out of three speakers. In summary, they advise against comparison across recording types.

Another study was done by Siegert and Niebuhr (2021), and again the setup and investigated speech data is different from the current study as the study focuses on sentences, charisma and was done based on a German corpus. Nevertheless, the results are still relevant for this thesis as they measured CoG in speech encoded with the Opus, MP3, SPEEX and AMR-NB (narrow-band) and AMR-WB. They used a 16 kHz sampling rate, and found that AMR-NB, Opus and SPEEX increased CoG across the sentences, while AMR-WB decreased this measure. They also found that intensity was generally lowered also for AMR-WB. They further note a clear gender difference between male and female speakers, with female speakers being more affected than male speakers. They state how "…quality of a codec must not be determined in terms of word intelligibility alone" while they call for more research on specific sounds e.g. consonants (Siegert and Niebuhr 2021, 7-8).

Other studies which have investigated CoG across sentences included Van Son (2005), who amongst other codecs looked at MP3. He found that changes in CoG of up to 5 semitones following the codec compression, and that the effects are more prominent in a low bitrate (40 kbs). He concluded that compressed speech is useful for pitch and formant measurements, but that CoG measurements should be used cautiously (Van Son 2005). With this said a final note is made that the codec-compressed speech will never be of the same quality as a high quality studio recording.

In sum, it is thus clear as initially stated, that the number of studies on fricatives in contexts identical to the current project is very limited. Regardless, these studies give an indication of the relevance of consistent methodology as well as general effects of codec compression on speech sounds.

2.9 Forensic Phonetics

This linguistic field of forensic phonetics is highlighted here because it is one of the fields most directly impacted by digital transmission of speech. Exactly how will be clarified below.

Firstly, there are four key types of analysis in forensic phonetics. These are speaker comparison, speaker profiling, questioned utterances, and authentication (for an overview of the field see Jessen 2008). All of these are usually conducted as a combination of both acoustic and auditory analysis, which makes both acoustic and perceptual consequences of digital transmission relevant for the field (Morrison 2016).

It is never possible to identify a speaker with a hundred percent certainty because of the plasticity of the voice, 'That is, there are no unchanging, biologically-determined properties of voice, speech, or language' (Foulkes and French 2012, 561). Nevertheless, it is possible to limit the number of candidates significantly and express this based on likelihood ratio (Foulkes and French 2012; Nolan 2001). This is a key feature in understanding forensic phonetic work and research.

With this in mind, speaker comparisons are done, when both an incriminating recording and recording of a potential suspect are available. These two or more recordings are then compared to figure out the likelihood ratio i.e. the likelihoods that the suspect is the speaker or vice versa that the speaker is not the suspect.

This is the most common type of analysis in forensic phonetics. Foulkes and French estimate that around 70 percent of the forensic casework undertaken in the UK is speaker comparisons. Speaker comparison can be used e.g. in cases with hoax bomb calls or 999 calls.

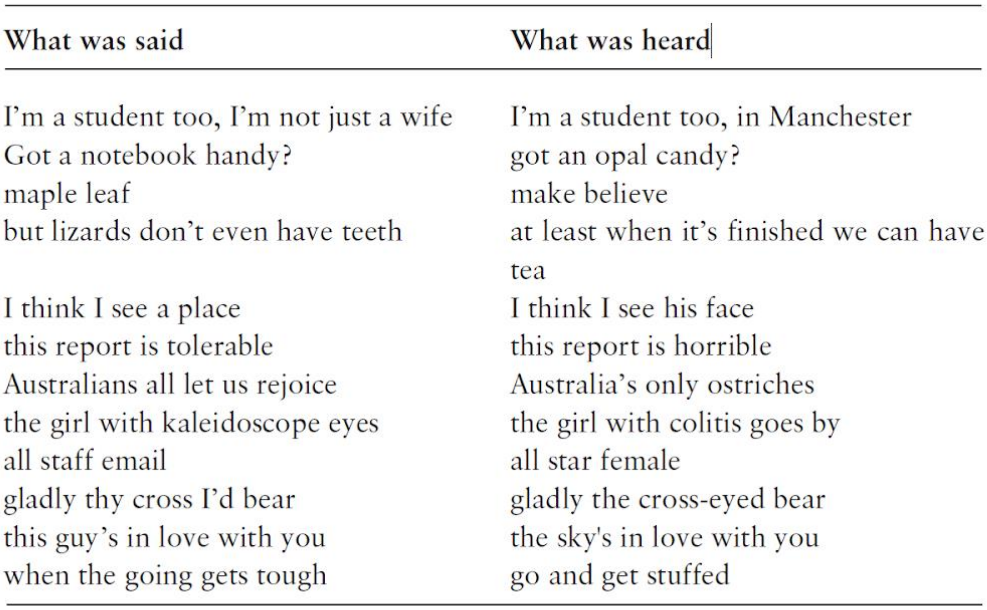
Speaker profiling on the other hand is done, when only an incriminating recording is available and the forensic phonetic analysis is then intended to limit the number of potential suspects. One of the most famous cases of speaker profiling in a criminal case was done by Ellis in 1979, when he successfully identified the geographical origin of, what was at the time believed to be, the Yorkshire Ripper from a tape recording (Ellis 1994; French, Harrison, and Lewis 2006).

A number of linguistic parameters can be used to make such an identification or comparison based on the individual's voice. This includes amongst others dialectal or group specific features, as well as other idiosyncratic features such as stutter (French and Harrison 2006)

For the current project, this is relevant because not only prosodic features and vowels have been shown to constitute unique dialectal features, so have consonants in a range of English dialects (Bauer 2002).

Furthermore, forensic phonetics is also concerned with questioned utterances or disputed utterances, which involves determining the exact content of a speech sample. This is often in cases where there is a significant amount of background noise or other types of interference, which allow multiple interpretations of the speech utterance due to the degraded quality. The forensic phonetician is then asked, again via both auditory and acoustic analysis, to present the most likely interpretation (French and Harrison 2006).

In a more perceptual perspective, consonants constitute minimal pairs in English and thus, it is possible to change the meaning of a word completely by simply changing a consonant e.g. *hat* vs. *cat*. The same is true for both word medial and word final consonants. Nevertheless, hearing errors related to consonants also occur outside of minimal pairs. Fraser provides the following real-life examples (2003):



What was said	What was heard
I'm a student too, I'm not just a wife	I'm a student too, in Manchester
Got a notebook handy?	got an opal candy?
maple leaf	make believe
but lizards don't even have teeth	at least when it's finished we can have tea
I think I see a place	I think I see his face
this report is tolerable	this report is horrible
Australians all let us rejoice	Australia's only ostriches
the girl with kaleidoscope eyes	the girl with colitis goes by
all staff email	all star female
gladly thy cross I'd bear	gladly the cross-eyed bear
this guy's in love with you	the sky's in love with you
when the going gets tough	go and get stuffed

Table 2.2. Hearing errors in real-life situations
adapted from Fraser (2003, 209).

The most robust features in perception are rhythm and stress, whereas consonants and particularly fricatives are confusable (Fraser 2003). Again, if the digital transmission alters the acoustic composition of a given consonant, it could lead to an increased number of such misinterpretations. In cases where a questioned utterance has been transferred digitally, knowledge of how the transmission alters individual consonants could aid in determining the consonant type via acoustic analysis and minimise misperceptions.

Lastly, authentication is the task to determine a recording's authenticity (Koenig 2009). For example, has a recording been edited or has it in fact been recorded the way claimed by the relevant parties? In these cases, more detailed acoustic knowledge on the acoustic consequences of digital transmission on individual sounds will enable the forensic phonetician to more accurately determine the origin of a digital recording. Moreover, authentication cases as well as other forensic phonetic cases often involve a replication of the original recording, which again will be made more accurate in correspondence to the level of acoustic knowledge available.

Accordingly, with the increased amount of digital transmitted speech shaping language today, it is essential to establish in exactly what way the transmission alters consonants in order to do correct and improved forensic phonetic analysis

In summary, the digital transmission of forensic phonetic casefiles entails that any such analysis will be undertaken based on speech, which cannot be expected to present the same level of quality as a direct recording. In that way, it is difficult, yet crucial, to identify whether a speech pattern is an artefact of the digital transmission or speaker variation. Hence, it is important to define what specific acoustic alterations that might be expected for consonants in such cases to achieve the best possible results and avoid misidentifications.

2.10 Sociolinguistics

The sociolinguistic perspective is included here with two main objectives, namely possible diffusion of acoustic variants occurring as a consequence of digital transmission, and the use of digital transmitted speech in data gathering. Firstly, social network theory, and how digital transmission could potentially work as a facilitator for the diffusion of sound changes are presented.

The theory of social networks will be introduced here based on Milroy, Kauhanen and Bermúdez-Otereo (L. Milroy and Milroy 1985; 1992; L. Milroy and Llamas 2013; Kauhanen 2016; Bermúdez-Otero 2017).

The introduction of social networks is attested as the beginning of the second wave of sociolinguistic research (Eckert 2012, 91). The social network models are primarily structured with the individual in the centre, connected to the other speakers in a language community through links called *ties*. The model of interest to the present study includes not only the individual, but the entire society. This model is illustrated below in figure 2.12.

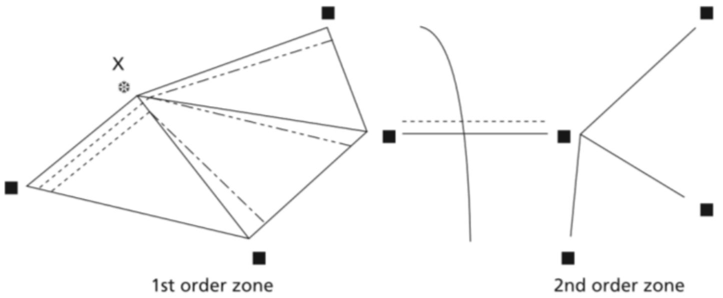


Figure 2.10 Complex social network model with the individual represented as X.

This model is a community model which do not only link the individual to other speakers, but also links these speakers to each other (adapted from L. Milroy and Llamas 2013, 411)

In this perspective, Milroy and Llamas state about social networks, ‘A social network may be seen as a boundless web of ties which reaches out through a whole society, linking people to one another, however remotely’ (L. Milroy and Llamas 2013, 411). This is essential because it clarifies how social networks are not bound by geographical proximity of the speakers. This is key to digital transmission of speech, which similarly is not bound by constraints of the geographical distance of the speakers, as long as a network access is ensured. In addition, the ties between speakers are divided into two sub-categories, namely strong and weak ties. The strong ties can be compared to the relationship between family and friends. The weak ties can in comparison be seen as the ties between acquaintances and colleagues (L. Milroy and Milroy 1985). Variants spread between speech communities via *innovators* who for various reasons e.g. their job form a number of weak ties with other communities. Through these ties, variants can diffuse between the communities, and when introduced by the innovators to what Milroy and Milroy defines as *the early adapters,* be implemented as new features in the primary speech community of these innovators (L. Milroy and Milroy 1985).

Originally, the ties in the 1st order zone (i.e. individuals who were directly linked to a given individual) were defined as physical encounters as research showed that innovations appeared to spread between areas connected by road and railway (L. Milroy and Llamas 2013). However, these assumptions can no longer be assumed to hold true in a globalised world (e.g. Sayers 2014; Tagliamonte 2014). Sayers points out the influence of media in sociolinguistics has mainly pertained to TV and film, but that research beyond this is relevant and suggests what he terms the *mediate innovation model* (2014). His motivation for this model is found in examples of sound changes spreading rapidly and in discontinuous geographical areas e.g. TH-fronting. In short, the *mediate innovation model* suggests media as an intermediate link between a source community and potentially multiple adopting communities, where diffusion happens via social networks. The relevance of such a model is debated (e.g. Trudgill 2014). Nevertheless, the fact that such a model is suggested underlines the potential role of media including digitally transmitted speech in diffusion of sound changes beyond the conventional approaches to diffusion based on face-to-face interactions.

As mentioned, the main research on media has been done on TV, where effects have been found, but in combination with traditional ties between innovators and adapters. One example of this is found in a study by Stuart-Smith et al. on the diffusion of dialectal variants via TV (2013). They find that speakers do indeed incorporate sociolinguistic variants into their language, which they are only exposed to via TV, but to a limited extent. Moreover, Del Tredici and Fernández found that social networks on social media, also conformed to Milroy and Milroy's concept of innovators and early adapters, despite the lack of any physical ties (Del Tredici and Fernández 2018).

Sayers states: 'my own sense is that using social media is much closer to interpersonal interaction and conventional diffusion than the communicative imbalance and parasocial interaction characterised by TV viewing' (2014, 207). For digital transmission of speech, this means that phones and computers are also likely to be a way of diffusion of sociolinguistic variants, e.g. across discontinuous geographical areas, and that the ties do not have to be physical for a variant to spread between speakers. In the same way, it is no longer a requirement for speakers to have physically met.

Lastly, this section will briefly consider some of the driving forces behind which variants are adopted by a speech community e.g. via the social networks. The traditional models suggest a number of social factors and individual preferences to be primary, in which innovations are diffused leading to sound changes. This could be e.g. the sex of the speaker, ethnicity, prestige and other individual traits (e.g. Trudgill 1972; Eckert 2012; L. Milroy and Milroy 1985).

In contrast, newer mechanical approaches have suggested that sound changes, to different extents, diffuse without influence from the social factors. This is known as neutral change (Kauhanen 2016; Bermúdez-Otero 2017). Kauhanen excluded all non-neutral factors i.e. biases, so that only frequency remained as a factor. He states 'language change is neutral if the probability of a language learner adopting any given linguistic variant only depends on the frequency of that variant in the learner's environment' (2016, 327). Bermúdez-Otero took a similar approach, but only excluded influence from individual differences and personal agency (Bermúdez-Otero 2017). Kauhanen finds that sound changes do occur based on only frequency of exposure, and that researchers should thus turn to this reasoning first, especially where no clear reason can be found for a biased explanation. Bermúdez-Otero concludes that based on his model, broader demographic factors together with network structure are much better predictors of sound change than individual differences and preferences (Bermúdez-Otero 2017).

This is of interest to the present study, because it emphasises how variants occurring as a consequence of the acoustic alterations of digital transmission, potentially diffuse between communities simply based on the frequency of occurrence. A specific example could, as also mentioned above, be TH-fronting, but in a different perspective to what is already mentioned. The primary interest of this study is not only the role of e.g. mobile phones as links between communities, but rather if the digital transmission via mobile phones could be the reason for the occurrence of linguistic variables and their following diffusion. For example, it might be that digital transmission results in reductions of /f/ and /θ/ to a point where the two become acoustically indistinguishable. If this is the case and adopting Sayers' model and the mechanical approaches, digital transmission potentially plays a role in both causing variants as well as their diffusion.

The second perspective of this section related to the possibilities of using digitally transmitted speech in data gathering. It is important here to make a distinction between data, e.g. speech recorded on a digital device, and speech transmitted between to digital devices and then recorded. It was previously pointed out how the main body of research on this has been concerned with the former (e.g. Leemann et al. 2020). In this type of research, the main factor of concern is the microphone of the recording mobile phone, or computer and the app used for the recording. In contrast, the scenario including transmission is the focus of this thesis, and is concerned with both the hardware as well as the transmission (e.g. Freeman and De Decker 2021). Very little research has been done on this, but should it turn out that digital transmission has minor acoustic implications, it would allow researchers to collect data e.g. via simple phone calls.

This will be investigated in Chapter 5 on live transmission. In the perspective of remote data collection via apps. The experimental setup in Chapter 5 will not allow the effects of the hardware to be distinguished from the effects of the transmission. Chapters 3 and 4 will more broadly give an indication of the relevance of awareness of the type of compressions, which might be used by apps both to store and potentially send audio files.

Chapter 3 : Baseline spectral implications of codec compression

3.1 Introduction

The first study of the thesis will be presented in this Chapter. The study aims to clarify and create a baseline for investigating the multifaceted acoustic implications of digital transmission of fricatives expressed through spectral measurements (e.g. spectral moments) in controlled conditions. The study works with the three codecs AMR-WB, Opus and MP3, but only using one average to good quality bitrate for each codec. In that way, the number of variables is minimised and the focus will be on the acoustic measurements and the effects on these in the comparison between high quality microphone recording and codec compressed signals. Overall, this thesis aims to illustrate real life scenarios including codec-compressed speech, thus, the chosen bitrates for this study were all within the span of what is expected in everyday use. The exact qualities were based on general quality estimates and their corresponding bitrates (3GPP 2022; Valin, Vos, and Terriberry 2012; Triton 2022).

Specifically, the fricatives investigated are /s, z, f, θ, ð, ʃ/and [fʲ]. The allophonic distinction between /f/ and [fʲ] was made based on the forced alignment of the files. More specifically, this was done with the Montreal Forced Aligner (MFA), which has this specific annotation as part of its analysis frame e.g. in initial position of the word *furiously* (McAuliffe et al. 2017) (see section 3.3.4 for details).

[v] and [h] were not included due to their high variation in articulation in English and correspondingly in their acoustic characteristics (Johnson 2012, 160), while [ʒ] was omitted as it only occurs as part of an affricate with /d/in the dataset. These three sounds are for these reasons both more and less likely to be affected by the codec-compression, however, accounting for the level of variation for each of these fricatives is beyond the scope of the present study.

The acoustic characteristics are measured by the first four spectral moments and frequency peak as these are some of the most widely used measures of fricatives (e.g. Blacklock 2004; Shadle 1985; Shadle and Mair 1996; Siegert and Niebuhr 2021. See section 2.2 for details). These are all static measures, as dynamic measures require a separate analysis e.g. using Generalised Adaptive Models (GAMs) (see S. N. Wood 2017), which is beyond the scope of this study. However, for technical reasons related to the design of the codecs, duration is expected to influence the effects of the codec compression, and will therefore still be measured, and considered as a variable in the statistical

modelling to provide the most accurate results. The study also includes spectrographic analysis to illustrate the observed tendencies.
Overall, the study will work to answer the following research questions:

BaselineRQ1 What do the included measures indicate about the effect of digital transmission on fricatives?

BaselineRQ2: Are the fricatives affected differently by the different codec compression? If so, how?

BaselineRQ3: In what way do these findings help create a baseline and understanding of the acoustic consequences of codec compression on fricatives in view of further research?

3.2 Predictions

The fricatives investigated here have almost all been found to have CoG between 5 and 8 kHz, which means that they all have frequency information centred around the upper cut-offs of the various codec compressions. Moreover, /s/, /z/ and /f/ are also reported to have their frequency peak above 8 kHz (Shadle & Mair 1996) (See section 2.2 and 2.3 for further details).

In addition to these differences, the differences in voicing between these fricatives are key to the current study. This is because the voiceless/voiced distinction presents a difference in wave structure (aperiodic vs. periodic) as well as a relative difference in intensity level. The voiceless sounds are less intense than the voiced counterparts (e.g. Strevens 1960).

Thus, for the spectral moments, a lowering of CoG was expected due to the limited bandwidth and potential decrease in intensity at the higher frequency bands. This effect was especially predicted to be prevalent in AMR-WB, where the 6-7 kHz frequency band is a reconstruction rather than a direct representation of the original speech input. The SD was expected to lower as well. This is because the energy recognised as speech is likely to be enhanced in comparison to the non-speech sounds following codec compression, which in turn limits the overall diffuseness of the energy. In combination with the limited bandwidth, this forms the basis for predicting a lower SD. For skewness and kurtosis, the predictions were less clear-cut as the effect of digital transmission on these measures is still to be investigated. Nevertheless, since broader acoustic changes were expected to be caused by the digital transmission, changes to skewness and kurtosis values were expected as well.

Moreover, as mentioned earlier a number of the fricatives in question often have their frequency peak and CoG in frequencies higher than the cut-off frequency employed by most digital transmissions (i.e. above 7 kHz). This is especially relevant for /s/, which from an acoustic and perceptual viewpoint was predicted to become more like an /ʃ/, if the spectral peak and CoG were lowered. As CoG is also one of the distinctive features that separates sibilants and non-sibilants, it was predicted that this distinction will be affected in codec compression. It was further predicted that the codec compression will affect the level of spectral distinctiveness between the fricatives due to the limited frequency range and concentration of energy.

In addition, due to their aperiodic structure, the fricatives are more likely to be mistaken for noise by the codec compression because of the codecs' focus on voicing and the similarity between the fricatives and e.g. potential background noise. Hence, it was predicted that especially the voiceless fricatives /f/ and /θ/, which are also lower in intensity (Strevens 1960, 37), were mistaken for noise and potentially not encoded. This was expected to be expressed as a CoG below 1 kHz in the codec compressions. Initially, frequency peak was considered based on previous research suggesting this as a distinctive measure for fricatives (e.g. Shadle and Mair 1996). However, as it will be evident from the results, this measure is not an illustration of the actual acoustic content of these fricatives. This was primarily because of the limited bandwidth, which prevents an accurate representation of the frequency peak.

Lastly, [fʲ] is as mentioned included due to the analysis frame of the forced aligner, and was expected to behave similarly to /f/ because of their allophonic relationship, while the effect on /ð/ was expected to be limited across spectral measures because of its predominantly low frequency content.

Overall, the speech codecs were predicted to perform better than the MP3 codec because of their speech specific design (See details in section 2.5 on digital conversion). Regardless of these predictions, the results have implications for a number of fields across linguistics in both data collection and speech analysis. This will be discussed more in depth in the main discussion and conclusion in Chapter 6.

3.3 Methodology

This section will present the methodology including procedure and data analysis for the baseline study. Philip Harrison assisted with the spectral analysis and generation of spectrograms as described in sections 3.3.3 and 3.3.5.

3.3.1 Corpus and Participants

The *You Came to Die?! corpus* (Best et al. 2012-2015) was used for this study. In total, the corpus consists of thirty male and thirty female speakers, all native speakers of English and aged between 18 and 41. For sociolinguistic purposes, specific information on the exact age each speakers would have been of interest. However, this information was not available.

The participants spoke five different accents of English with six speakers of each accent. These were Australian (AUS), New Zealand (NZL), London (LON), Newcastle (NCL), and York (YRK) English. Every speaker was recorded reading a set of nonsense words in /zVbə/context; reading real keywords, and reading a phonologically balanced version of the Chicken Little story (approximately 10 minutes) (see transcript in appendix A). The current study incorporates data from the 10 min Chicken Little story as produced by all 30 male speakers. Female speakers are reported to be more affected by codec compression than male speakers (Siegert and Niebuhr 2021).

Since, the aim was to establish a baseline (here understood as the minimum effect of codec compression) male speakers were chosen as they are reported to be less affected by the codec compression in comparison to female speakers (e.g. Siegert and Niebuhr 2021). It should be noted that the corpus was originally designed for research on the fundamental processes of speech perception and adaptation between accents in adults. Nevertheless, the fact that each participant produces the same material means that the data is comparable across participants. This, together with the fact that the material elicits a reasonable number of tokens of the segments investigated in this study, made it useful for the current study.

Overall, the fricatives investigated are not reported to vary between the investigated dialects. The only main exception to this is TH-fronting. TH-fronting is the linguistic phenomenon where /θ/is replaced by /f/ e.g. in a word like *think* (e.g. Wood 2003; Stuart-Smith et al. 2013). This feature is common across dialects of English and especially so in London and Glasgow English (Stuart-Smith, Timmins, and Tweedie 2007). It is usually used at a production term, but in the perspective of digital transmission just as relevant in terms of acoustics and later perception. Thus, TH-fronting and the acoustic similarity between /f/ and /θ/ are considered as the results are presented and discussed. Another dialectal feature, which is still only studied and attested in a few English dialects e.g. Estuary English, Glaswegian English, and Edinburgh English is /s/-retraction (Bailey et al. 2022; Stuart-Smith et al. 2019). Thus, this feature is not reported to be prevalent in any of the dialects investigated here. However, it results in a lowered CoG for /s/, which in turn makes /s/ more similar to /ʃ/. It occurs in /s/CC clusters, and particularly /stɹ/clusters (Bailey et al. 2022; Stuart-Smith et al. 2019). The present dataset contains only three cases of /str/-clusters (i.e. <stress>, <backstroke>, and <struggle>), thus the number of tokens in this combination specifically is far too limited to influence the results. Nevertheless, the fact that the codec compression is expected to lower CoG across the fricatives, makes it relevant to consider as it might be mimicked by the codec compression outside this specific context.

The fact that the corpus is read speech presents both advantages and limitations in terms of generalisability. Read speech allows control of the contents of the dataset and the number of tokens available for each speaker and dialect. This is an advantage in terms of generalisability looking from the perspective of the acoustic measurements across the codec compressions. On the other hand, read speech is known to present certain acoustic and perceptual differences in style and pronunciation compared to spontaneous speech (see Nakamura, Iwano, and Furui 2008; Mehta and Cutler 1988). However, since the current study is interested in the acoustic consequences of codec compression on fricatives, the production differences caused by the read speech are less essential, and the advantages presented by the content control more important. The read speech will in fact give a clearer picture of the exact effect of the compression on the individual sounds, as less elision and other spontaneous speech features diminishing the acoustic structure of the sounds, are likely to occur. In addition, from a practical viewpoint, the read speech makes the forced alignment less complicated and time consuming. This of course opens the door to future studies, when the more fundamental effects of the compression is better known, so that more variable data of spontaneous speech can be investigated.

3.3.2 Materials

The dataset elicited the following seven fricatives /s, z, f, θ, ð, ʃ/and [fʲ] in read speech from an approximately 10 minute reading of the Chicken Little story in varying phonetic context and in word-initial, word-medial, and word-final position (e.g. <fear>, <painful>, <safe>; see full transcript in appendix A).

The 16 kHz down-sampled files were compressed with the three codecs of interest (see section 3.3.3 on sound files for details) at an average quality achieved by three different bitrates specified in kilobits per second (kbps) (see section 3.3.3). This resulted in a dataset with a total of 85,280 fricatives with 21,320 in each codec compression and WAV baseline. The number of fricatives was not equal (e.g. more tokens of /s/ than /z/) as this was determined by the content of the *Chicken Little story*. Hence, including all codecs and baseline, the study consisted in total of 5,104 tokens of /ʃ/, 5,156 tokens of /θ/, 15,572 tokens of /ð/, 13,668 tokens of /f/, 960 tokens of [fʲ], 29,336 tokens of /s/, and 15,484 tokens of /z/. Due to mispronunciations, technical flaws etc. the total number of tokens across accents varied slightly, but not to an extent that was judged to affect the results significantly (i.e. average difference across segments of 206 tokens) (See section 3.3.4 on segmentation for details).

Furthermore, a separate dataset (below 1 kHz tokens) was made with all tokens with a CoG below 1 kHz and their counterparts in any other codec compression or WAV baseline. This dataset is analysed separately and the results presented in section 3.4.4 on 1 kHz tokens. These tokens were analysed separately to avoid skewing the main dataset, while also investigating the tokens with the potentially most apparent effect of the codec compression. See details in section 3.3.5 on data extraction and measurements (See section 3.3.5 on data extraction and measurements for details).

In sum, this meant that the final main dataset consisted of 79,860 tokens with 19,965 per codec compression and WAV baseline. See table 3.1 for additional details on number of tokens.

Segment	**Total** (without 1 kHz tokens)	**Initial position**	**Medial position**	**Final position**
/f/	13,604	8,960	2,524	2,120
[fʲ]	960	600	360	0
/s/	29,316	15,624	6,576	7,116
/z/	15,212	240	2,304	12,668
/ʃ/	5,104	3,336	1,168	600
/θ/	5,104	1,080	1,860	2,164
/ð/	10,560	9,524	448	588

Table 3.1. Number of tokens per segment in full dataset with different criteria

3.3.3 Sound files

The sound files were originally sampled at 44.1 kHz, and down-sampled for the present study to 16 kHz in order for AMR-WB and Opus to be applied as well as to match the sampling rate of the codec-compressed speech. The spectra for each fricative token were obtained via a MATLAB script (Harrison 2022; MathWorks Inc. 2010) (see details in section 3.3.5 on data extraction). An upper frequency limit of 8 kHz together with a lower frequency limit of 500 Hz was specified in the MATLAB functions that was used to extract the spectral measures. The lower cut-off was set at 500 Hz as this is the typical cut-off applied to avoid the influence of mains electricity hum on the measurements (Smorenburg and Heeren 2019).

The codec compressions were done using three different software systems. For MP3 this was the FFmpeg 64-bit static Windows 4.4.1-essentials_build (Bellard and FFmpeg Team 2000). For AMR-WB the 3GPP AMR-WB Floating-point Speech Coder, v16.0.0 (3GPP 2007) was used and finally, the opusenc/opusdec opus-tools 0.2 using libopus 1.3.1 was used for the Opus compression (Xiph.Org Foundation 2022).

The OPUS codec is slightly different from the other codecs as it works in three different modes (i.e. speech, hybrid and CELT), but as the study works with one bitrate and an upper cut-off at 8 kHz mainly the speech mode is activated. Each codec compression was done with a set bitrate as summarised in table 3.2.

Codec	Bitrate
MP3	32 kbps
AMR-WB	12.65 kbps
OPUS	24 kbps

Table 3.2. Overview of codecs and bitrates in kbps.

3.3.4 Segmentation

The corpus files were force aligned using the MFA (McAuliffe et al. 2017). This was done with a customised version of the English (UK) MFA dictionary (McAuliffe and Sonderegger 2022). The dictionary was customised as a number of words in the reading were not included in the original dictionary (see appendix B for list of added words). The forced alignment dictionary included [fʲ] as a separate annotation and in consequence had this as part of the segmentation e.g. in initial position of words like *furiously* and *feeling*. [fʲ] is thus an allophone of /f/, but no other such distinction was made by the forced aligner for any of the remaining fricatives. Instead of discarding these tokens or merging them with /f/, it was decided to keep this distinction and the forced alignment unaltered. This was done as it gives a potential indication of whether the codecs are sensitive to the allophonic distinction as well as the adjacent segments. This allophonic distinction gives an indication of the latter because the production of /f/ as [fʲ] is dependent on the following segment.

From the data and as it will be evident in the results, it was clear that the MFA works primarily with voicing as a phonological rather than phonetic distinction. Therefore, as mentioned earlier, a distinction will be made in the following Chapter between phonetic voicing (i.e. voicing) when considering the codecs and segments, and phonological voicing when considering the segmentation done by the MFA. This is key, as the codecs in contrast are sensitive to phonetic voicing. It was decided not to further manually correct the data and e.g. make a narrower transcription of voiced and voiceless realisations or plosive and approximant productions of the voiced fricatives. This was decision was made to maintain the best possible opportunities for replication of this study without influence of subject evaluations made by the author.

In that way, it influences the analysis of the results. This will be addressed in the results and discussion of this Chapter as well as in the main discussion and conclusion in Chapter 6.

Following the forced alignment, all files were inspected and the text grids manually corrected by the author in Praat (Boersma and Weenik 2021). The correction process of the segmentation largely followed Behrens and Blumstein (1988, 296). Corrections were only made in cases where the correction had an influence on fricative tokens. These corrections included: a) boundaries between two fricatives (especially between voiced and voiceless segments e.g. /z/ and /f/ to be placed at the change in voicing, intensity and/or frication, b) initial and final boundaries corrected if background noise, silence or other non-speech elements included in the fricative segment. Finally, c) places where speakers mispronounce resulting in wrong segmentation by the forced aligner were marked with NA in the segment tier. The choice was made to mark the wrongly aligned words with NA rather than manually correcting the segmentation as it concerned a limited number of tokens. In addition, this ensured complete consistency in the segmentation and transcription. No corrections were made to TextGrids following the codec compression to allow the direct comparison between WAV and codec compression. As an example, Figure 3.1 below illustrates a correct segmentation of the word *fluff.*

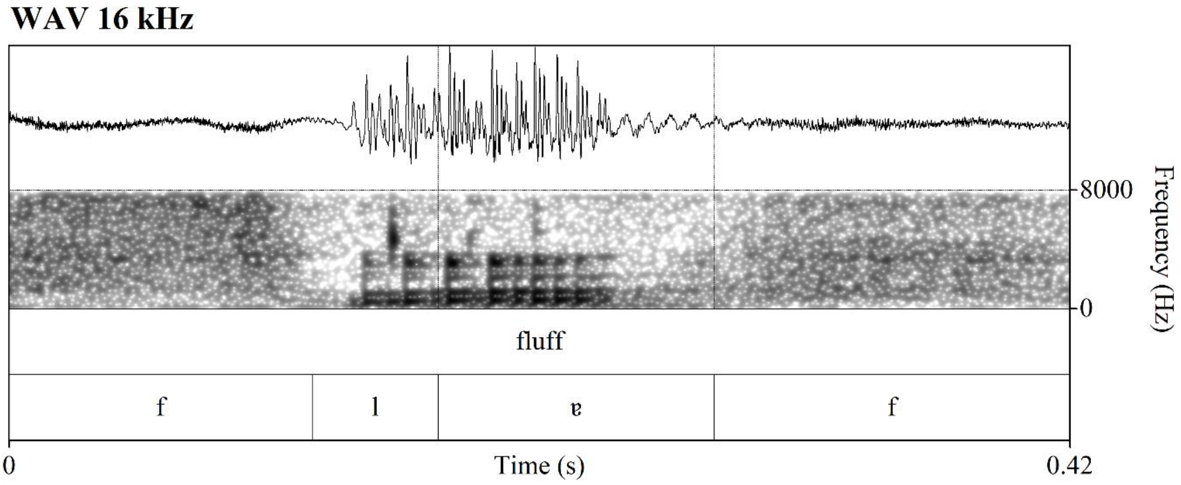


Figure 3.1. Illustration of segmentation from the word *fluff* in the WAV baseline

3.3.5 Data extraction and measurements

Firstly, a white noise signal was generated and subjected to each codec compression to indicate the actual upper-frequency limit of the codec compression, which was expected to be below the 8 kHz and vary for each codec. The results of this are reported in Chapter 3 section 3.4.

In order to illustrate how the codec compression would likely be accounted for in linguistic research without adjusting for the technical limitations introduced by the codecs, the actual limitation in bandwidth illustrated by the white noise signal was not accounted for. Instead, the subsequent measurements were done with the 8 kHz upper cut-off imposed by the 16 kHz sampling rate, which is also the maximum cut-off given for these codec compressions in their specifications (3GPP 2020b; Valin, Vos, and Terriberry 2012; Gayer, Lohwasser, and Lutzky 2003)

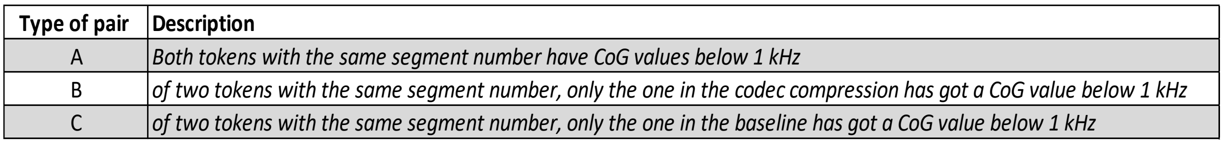
Based on the segmentation boundaries, the spectral measures (i.e. CoG, SD, skewness, kurtosis, and frequency peak) were extracted from the central frame via a MATLAB script written by Philip Harrison (Harrison 2022; MathWorks Inc. 2010). Rather the measurements being done across the entire segment, the 20 ms central frame was chosen to ensure consistency in measurements while minimising the effect of durational and contextual differences. This was also done as certain speaker specific effects have been found to be more prevalent in onset and offset of the sounds (e.g. Carrelo-Fernandez 2002; 109).

The measurements were obtained from multitaper spectral analysis (Prerau et al. 2017) rather than the more traditional LPC, periodogram and Fast Fourier Transform (FFT). Especially for noisy and aperiodic sounds, the multitaper method has proved useful as it smooth the signal via multiple tapers. Thus, it provides more stable and clear spectral moments for fricatives in comparison to the FFT (see Blacklock 2004; Shadle 1985). Multitaper analysis requires no additional windowing and no pre-emphasis was applied.

Spectrographic representations as well as spectra were generated for a random selection of segments for each speaker with the average bitrate (3 to 5 examples per speaker). These were visually inspected and used as examples in section 3.4 on results. The spectrograms were generated using regular FFT analysis, while the spectra were generated using multitaper.

Any token with a CoG below 1 kHz and above in any of the formats and vice versa were identified and excluded from the main dataset. These tokens are referred to as *1 kHz tokens*. This was done as these tokens were evaluated to be either a result of these fricatives in fact being fully voiced and having the structure of glides (Johnson 2011, 156) (see note on this in section 3.4 on results), not being encoded by the codecs, or measurement errors e.g. caused by wrong alignment. In order not to skew the dataset and ensure complete comparability between the baseline and codec compressions in the main dataset, the tokens with the same segment number as the 1 kHz tokens in any other condition (i.e. baseline or other codec compression) were likewise excluded. For example: the initial /f/ in <fluff> in the AMR-WB compression had a CoG value below 1 kHz. This specific /f/ is the 10^{th} segment in the reading and was thus, segment number 10 across all conditions. This /f/ was then removed from all conditions i.e. every codec compression and WAV baseline in the main dataset.

These 1 kHz tokens were subset into separate datasets consisting of only the baseline and one codec compression, and only including the tokens, which had one or both CoG values below 1 kHz in either of the two conditions. In the following, these tokens will be referred to as *pairs*. Taken together, this gives three types of pairs as illustrated in table 3.3 below.



Type of pair	Description
A	*Both tokens with the same segment number have CoG values below 1 kHz*
B	*of two tokens with the same segment number, only the one in the codec compression has got a CoG value below 1 kHz*
C	*of two tokens with the same segment number, only the one in the baseline has got a CoG value below 1 kHz*

Table 3.3. Description of types of pairs in the dataset including only the 1 kHz tokens.

Spectrograms were generated for each of the pairs in group B and C where only one token was above 1 kHz in either of the conditions. This was done to visually inspect the potential causes of this change in CoG. On the other hand, all three groups were included in the statistical analysis.

3.3.6 Statistical analysis

All statistical analysis was carried out in R (RStudio Team 2019; RCore Team 2020). To provide an overall summary of the spectral measures for each segment, the descriptive statistics (i.e. minimum, mean, median and maximum) for all the spectral measures across all speakers were calculated for all codecs as well as the baseline. To show the magnitude and direction of the change in values between the baseline and the codecs, the difference between the mean values was calculated for each spectral measure and fricative. In the mixed effects modelling, no comparison was made between codecs.

In order to examine the influence and interaction of the various variables and factors within this study, a statistical analysis was done in R using mixed effects modelling and ANOVAs (J. M. Chambers and Hastie 1992). The mixed effects models were maximal lmer based models (Kuznetsova, Brockhoff, and Christensen 2017) done individually for each codec compression type and WAV baseline with the spectral measures as dependent variables (i.e. CoG, SD, skewness, kurtosis and frequency peak), which gives a total of five final models per codec compression. For this part of the analysis, the datasets were further divided into two subsets including voiceless and voiced segments respectively. This was done as it was clear from the initial descriptive analysis that the two groups behaved substantially differently, and it was thus meaningful to make this distinction. In addition, this also follows the initial prediction of the codecs using voicing as a key parameter.

A range of models were compared for each codec and spectral measure including the following independent variables: format (with 4 levels (2 per model): baseline 16 kHz, and one of the following AMR-WB, MP3, or Opus), speaker (with 30 levels: individual speakers), word position (with three levels: initial, medial, and final), segment duration, and preceding segment (with 21 levels) and following segment (with 60 levels).

The evaluations of fit of the models, and the decisions of what to include as fixed and random effects, were based on inspections and calculations of residuals, standard deviation of residuals, and Akaike Information Criterion (AIC) values (lowest values = best fitted model) for each model and each dependent variable. Following this procedure, it was determined for each of the dependent variables (i.e. for every model), which independent variables were fixed and which were random effects. This resulted in the same model for all spectral measures with format, segment, preceding segment, following segment, and duration as fixed effects, and speaker and word as random effects.

For all models, only one interaction was used. This was between format and segment as this interaction was predicted based on the expected segment-dependent behaviour of the codec compression. This is together with the fact that more interactions would likely have resulted in over-fitting the models and complicated interpretation of the outputs. Different intercepts were also tested for the random effects, however with no changes to the fit of the models and thus, all random effects were fitted without any additional intercepts.

Finally, post-hoc tests were conducted using Emmeans and the pairs function to extract the significance values for the interactions in the mixed effects models and to plot the linear predictions of those same models (Lenth 2020).

It will become evident from the results section below that a number of spectral measures turned out to be problematic when modelling this way. These issues related to a number of statistical zeros (i.e. extreme values) which skewed the residuals and the randomness of the models and in the end made the models unable to reasonably predict kurtosis and frequency peak. Similar tendencies were found for the voiced tokens. All illustrations of results generated in R was made using the ggplot package (Wickham 2016).

3.4 Results

Firstly, as one of the key aspects of the codec compression is the limitation in bandwidth a white noise signal was codec compressed and based on this the actual cut-off values for each codec was determined by visual inspection of the resulting spectra. The result of this is illustrated in figure 3.2 below. These values are not obtained with the usual -3 dB definition for the cut-off and therefore should be seen as indicative. With this in mind, the cut-off for each codec is estimated to be as follows; 5,600 Hz for AMR-WB, 7,100 Hz for MP3, and 7,000 Hz for Opus. These values are determined as the point where the values trail off and no longer follow the trajectory of the white noise.

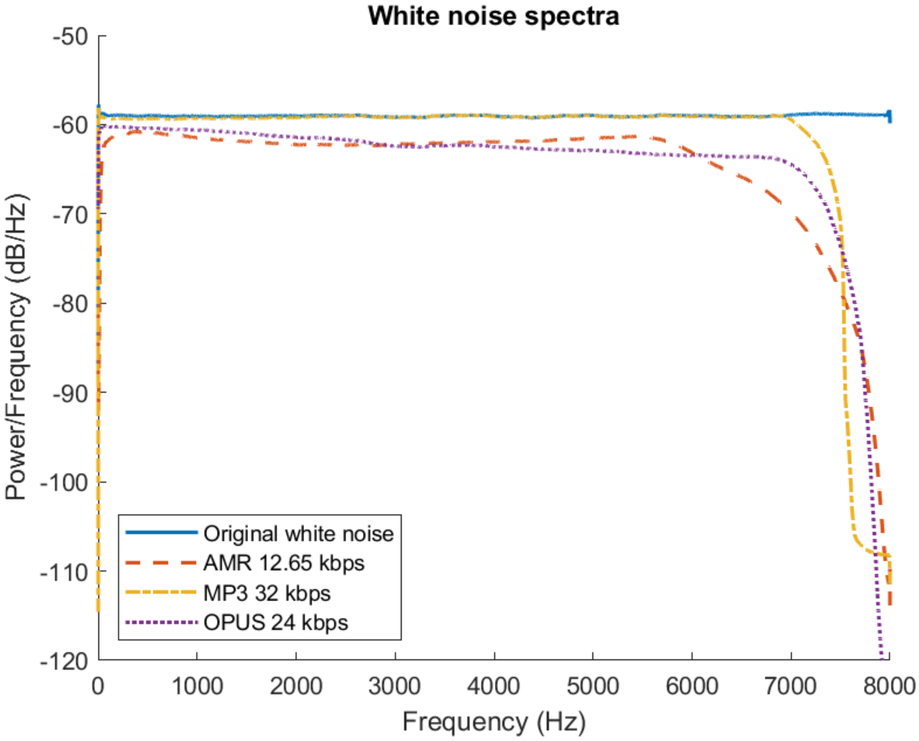


Figure 3.2. Frequency spectra of original white noise signal and codec compressed versions

All results reported in the following sections are from the main dataset excluding the 1 kHz tokens. These 1 kHz tokens will as mentioned be presented separately in section 3.4.4, where details on the exact numbers of tokens can also be found.

Overall, the mean values in Table 3.4 show that, as expected, for all segments, CoG, SD, and frequency peak are all affected in all codec compressions in comparison to the uncompressed 16 kHz baseline.

Seg	Codec	Bitrate	CoG	SD	Skew	Kurt	Freq. Peak
		(kbps)	(Hz)	(Hz)	(Hz)	(Hz)	(Hz)
/f/	AMR	12.65	2891	1488	0.65	3.12	2128
/f/	MP3	32	3038	1553	0.65	3.15	2335
/f/	Opus	24	2753	1467	0.84	3.70	1970
/f/	**WAV**	**NA**	**3168**	**1673**	**0.74**	**3.34**	**2445**
[fʲ]	AMR	12.65	2937	1405	0.76	3.55	2161
[fʲ]	MP3	32	3026	1432	0.75	3.57	2320
[fʲ]	Opus	24	2788	1349	0.89	4.15	2128
[fʲ]	**WAV**	**NA**	**3131**	**1544**	**0.87**	**3.84**	**2408**
/s/	AMR	12.65	4505	1091	-0.22	5.77	4383
/s/	MP3	32	4621	1095	-0.35	5.90	4614
/s/	Opus	24	4489	1099	-0.42	6.44	4487
/s/	**WAV**	**NA**	**4789**	**1165**	**-0.09**	**5.71**	**4743**
/z/	AMR	12.65	4054	1215	-0.16	5.78	3486
/z/	MP3	32	4191	1206	-0.42	6.07	3777
/z/	Opus	24	4043	1198	-0.42	6.54	3583
/z/	**WAV**	**NA**	**4357**	**1241**	**-0.13**	**6.17**	**3869**
/ʃ/	AMR	12.65	3270	908	1.70	8.77	2970
/ʃ/	MP3	32	3253	859	1.46	8.28	2993
/ʃ/	Opus	24	3150	776	1.48	9.76	2967
/ʃ/	**WAV**	**NA**	**3271**	**896**	**1.71**	**9.45**	**3002**
/θ/	AMR	12.65	2932	1586	0.53	3.06	1868
/θ/	MP3	32	3141	1676	0.47	3.01	2107
/θ/	Opus	24	2824	1610	0.71	3.52	1712
/θ/	**WAV**	**NA**	**3321**	**1808**	**0.53**	**3.14**	**2300**
/ð/	AMR	12.65	2226	1425	1.04	4.52	1065
/ð/	MP3	32	2458	1521	0.84	3.92	1240
/ð/	Opus	24	2246	1456	1.07	4.85	1047
/ð/	**WAV**	**NA**	**2525**	**1602**	**0.94**	**4.52**	**1284**

Table 3.4. Mean values for all spectral measures for each fricative in each individual codec and 16 kHz WAV baseline.

Despite the frequency peak values being reported here, it should be noted that following analysis, it was clear that due to the upper-cut off and the effect of the codec compression illustrated in figure

13, the measure of frequency peak is to some extent not truly representative in this context. This is because the fricatives often have energy outside the imposed frequency limits, but in the down-sampled files they are constrained be at 8kHz or below despite the true peak potentially being above this limit in the 44.1 kHz files. For this reason, the frequency peak will not be included in the statistical modelling as the number of statistical 0s skew the model and the desired randomness, and distribution of the residuals. This is also noted by Johnson, who explains how this provides the reasoning for implementing CoG more consistently in acoustic research of fricatives (2011).

A similar tendency was observed for kurtosis, where the mean values presented a distribution deviating from the expected normal distribution. This meant that, as with the frequency peak, the kurtosis values could not be analysed with the mixed effects models. Hence, no p-values will be presented for these two measures in the following results. Regarding kurtosis, it should further be noted that the level of the mean values can potentially be explained by the fact that all the investigated fricatives have the energy centred at the higher end of the spectrum. On the other hand, the maximum values spanning between 13 and 132 suggest a number of outliers or issues with the methodological approach to this measure. It is unclear from the present data which of these are the case, but as the other spectral measures behave more according to expectations, the latter is a likely reason for this observation. For this reason, the kurtosis values will only be reported in terms of mean values in the following sections, and should be considered with a level of caution.

Another general tendency observed in the mixed effects modelling relates to the randomness of the predicted value observed in the residual plots for the voiced segments (i.e. /ð/ and /z/). This is again, similarly to the frequency peak results, due to a number of values occurring at the extreme lower range of the measurements. This is expected for the voiced segments because of their acoustic content generally occurring in the lower frequencies, together with the fact that any fully voiced segments will have most energy centred around frequencies near the lower cut-off value. As it is clear why this pattern is present and the remaining residual plots related to distribution and quantiles are following expected patterns, the voiced segments will still be analysed with the mixed effects models. However, this fact should be kept in mind when assessing the implications and reliability of the outputted results.

The magnitude and direction of the changes in mean values are shown in Table 3.5. Some general trends can be observed e.g. the Opus codec generally results in the largest changes to the spectral measures, while the MP3 codec is the least influential. However, the effect of the codec compression was significant for almost all measures regardless of codec type with some segmental uniqueness.

This indicates a consistency in the effect of the codec compression even in smaller changes. The low p-values are to some extent influenced by the size of the dataset.

Seg	Codec	Bitrate (kbps)	CoG (Hz)	CoG (%)	SD (Hz)	SD (%)	Skew (Hz)	Skew (%)	Kurt (Hz)	Kurt (%)	Freq. Peak (Hz)	Freq. Peak (%)
/f/	AMR	12.65	276	-8.73	185	-11.04	0.09	-12.21	0.22	-6.56	317	-12.96
/f/	MP3	32	129	-4.08	120	-7.16	0.09	-12.28	0.19	-5.73	109	-4.47
/f/	Opus	24	415	-13.09	206	-12.33	-0.10	13.09	-0.35	10.62	474	-19.41
[fʲ]	AMR	12.65	194	-6.18	138	-8.97	0.11	-12.75	0.28	-7.33	248	-10.28
[fʲ]	MP3	32	105	-3.34	112	-7.27	0.12	-13.42	0.27	-6.97	89	-3.69
[fʲ]	Opus	24	342	-10.93	195	-12.63	-0.02	2.16	-0.31	8.13	280	-11.64
/s/	AMR	12.65	284	-5.94	74	-6.35	0.14	-144.44	-0.06	1.03	359	-7.58
/s/	MP3	32	168	-3.52	70	-5.98	0.27	-288.89	-0.19	3.30	129	-2.71
/s/	Opus	24	300	-6.27	66	-5.62	0.33	-366.67	-0.73	12.70	256	-5.40
/z/	AMR	12.65	304	-6.97	25	-2.04	-0.03	14.29	0.39	-6.27	383	-9.89
/z/	MP3	32	167	-3.82	35	-2.83	0.30	-207.14	0.10	-1.59	92	-2.38
/z/	Opus	24	315	-7.22	43	-3.46	0.29	-207.14	-0.37	6.05	286	-7.39
/ʃ/	AMR	12.65	0.94	-0.03	-12	1.38	0.02	-1.04	0.68	-7.18	32	-1.06
/ʃ/	MP3	32	18	-0.55	37	-4.11	0.25	-14.81	1.17	-12.42	8	-0.27
/ʃ/	Opus	24	121	-3.69	120	-13.37	0.24	-13.85	-0.31	3.26	34	-1.14
/θ/	AMR	12.65	389	-11.71	223	-12.31	-0.004	0.77	0.09	-2.71	432	-18.79
/θ/	MP3	32	180	-5.43	133	-7.33	0.05	-10.39	0.14	-4.30	193	-8.41
/θ/	Opus	24	498	-14.99	199	-10.99	-0.18	33.69	-0.37	11.89	588	-25.57
/ð/	AMR	12.65	299	-11.83	299	-18.65	-0.09	9.90	0.01	-0.14	218	-17.00
/ð/	MP3	32	67	-2.66	67	4.19	0.11	-11.49	0.61	-13.43	44	-3.45
/ð/	Opus	24	279	-11.06	279	-17.42	-0.13	13.25	-0.32	7.14	237	-18.48

Table 3.5. Differences in mean values between baseline (WAV) and codec compression in Hz and percentage. Colours indicate the direction of the change. (i.e. blue = decrease; yellow = increase)

A final point is essential. As it will be evident from the individual segment analysis and the section on 1 kHz tokens, the voiced fricatives had a tendency to apart from their typical production to be produced as plosives, voiceless fricatives, and approximants (Johnson 2011, 156).

The following sections will as earlier mentioned be specific to each codec. It should be noted here that due to R not recognising the IPA symbols for /θ/, /ʃ/, /ð/, and [fʲ] these will be written as *theta, esh, eth* and *fj* respectively in the R generated figures.

3.4.1 AMR-WB

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the AMR-WB codec. Before these results are presented the linear predictions for each spectral measure and the individual segments can be found below (figures 3.3 to 3.8). These indicate the directionality of the changes imposed by the AMR-WB codec. The graphs presents the results divided into voiced and voiceless segments as this was the grouping made in the linear prediction modelling. The more specific analysis of these plots will be found in the following sections on the individual segments.

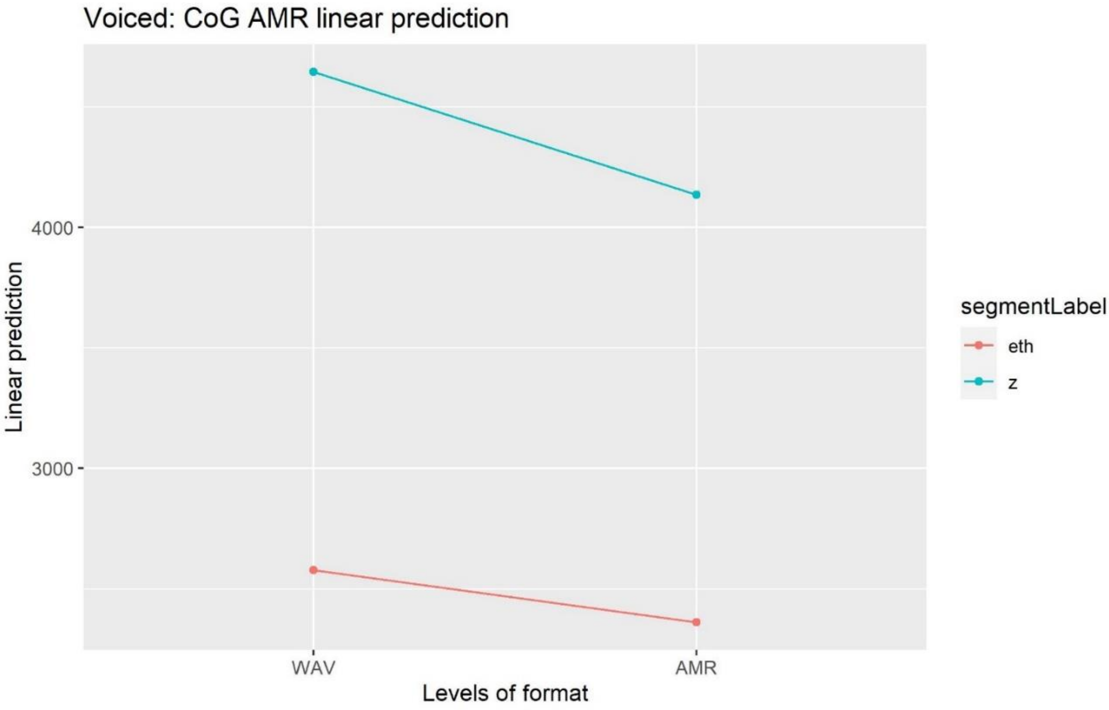


Figure 3.3. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for CoG and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

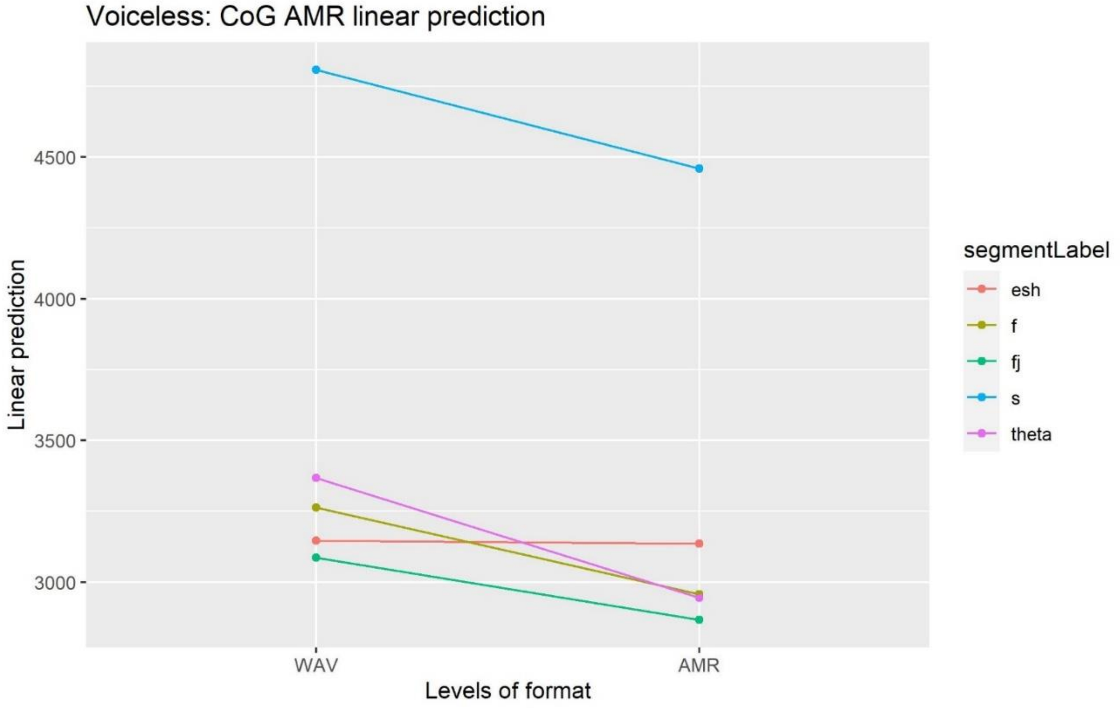


Figure 3.4. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for CoG and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f̠ʲ], and eth = /ð/.

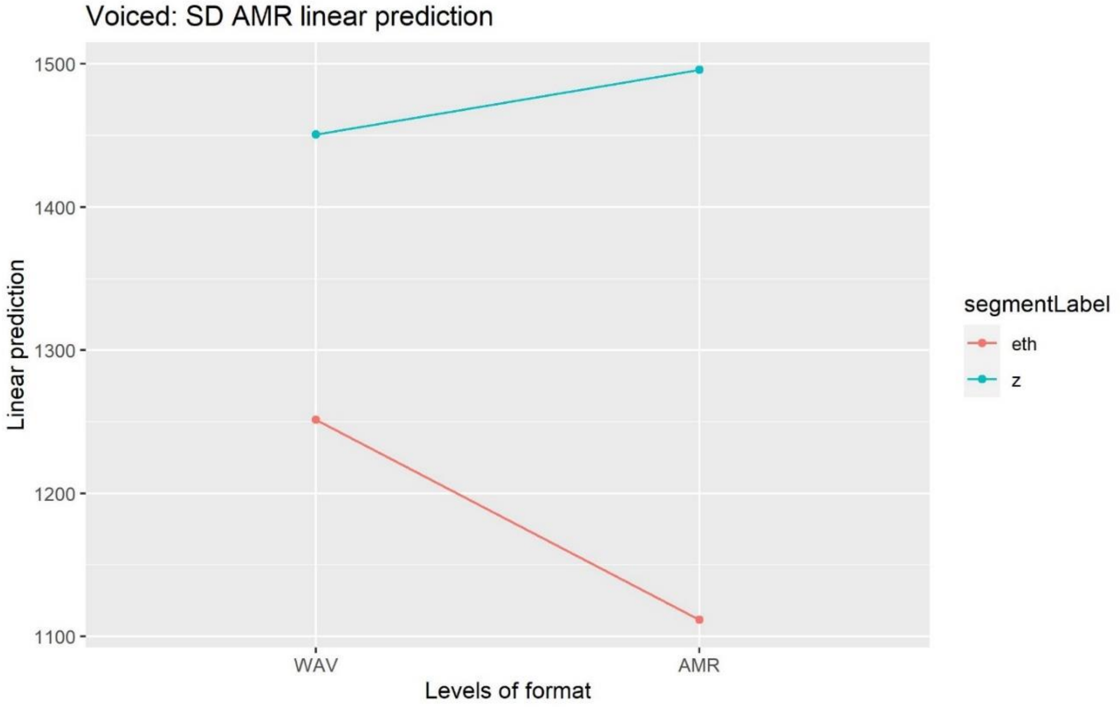


Figure 3.5. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for SD and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f̠ʲ], and eth = /ð/.

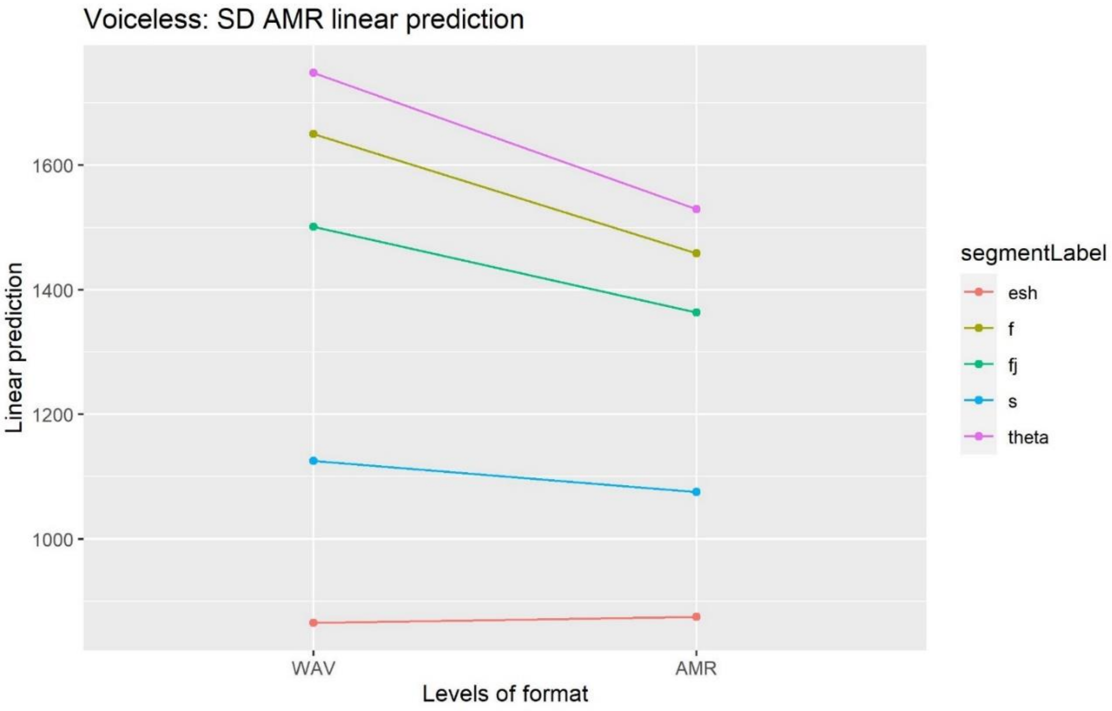


Figure 3.6. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for SD and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [ɸʲ], and eth = /ð/.

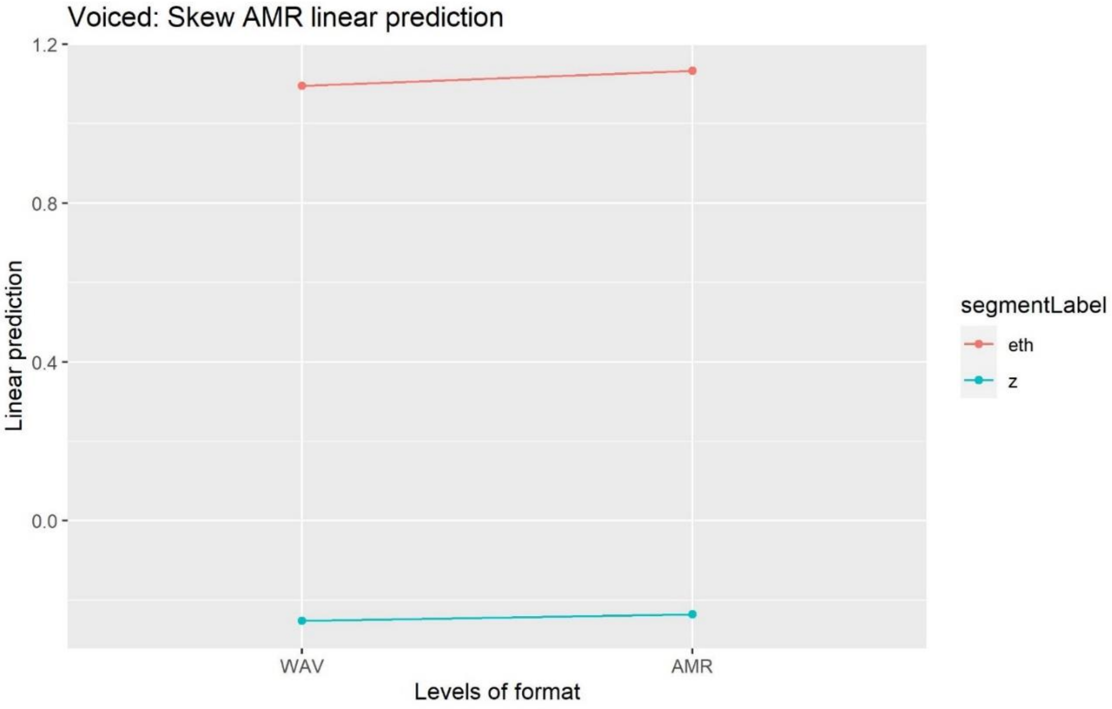


Figure 3.7. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for skewness and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [ɸʲ], and eth = /ð/.

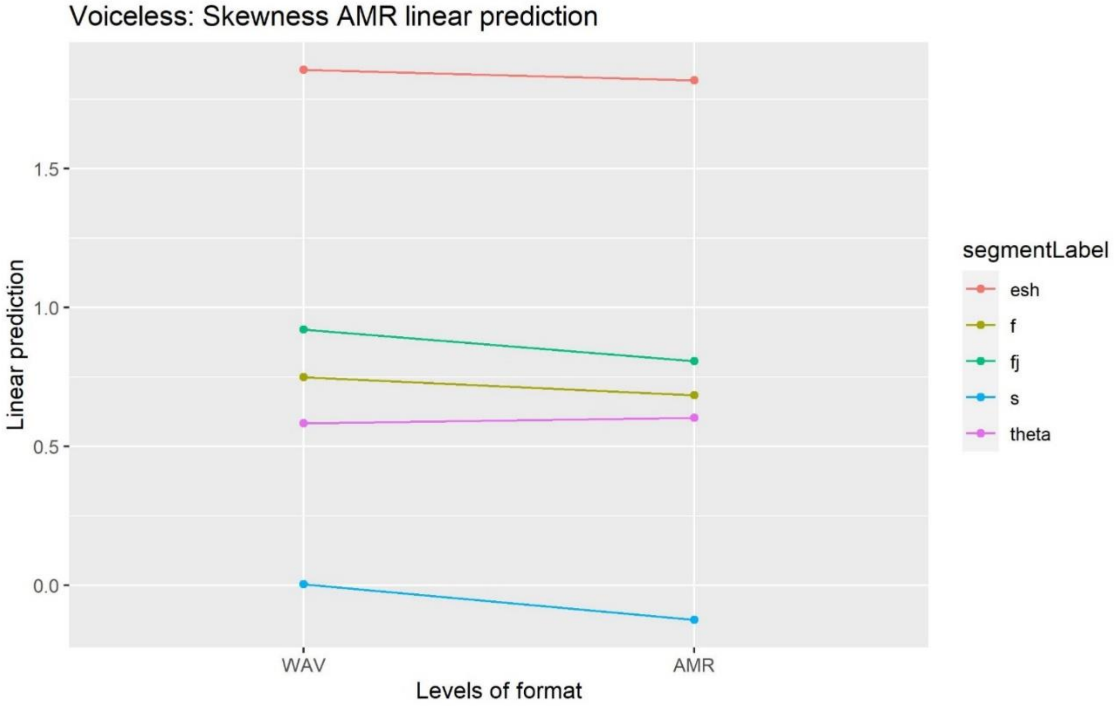


Figure 3.8. Trajectory of the linear predictions in the comparison of WAV and AMR-WB from the mixed effects models for skewness and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

The more general distributions are illustrated below in a set of violin plots for each spectral measure. Again, the specific analysis pertaining to each segment will be found in the relevant sections below.

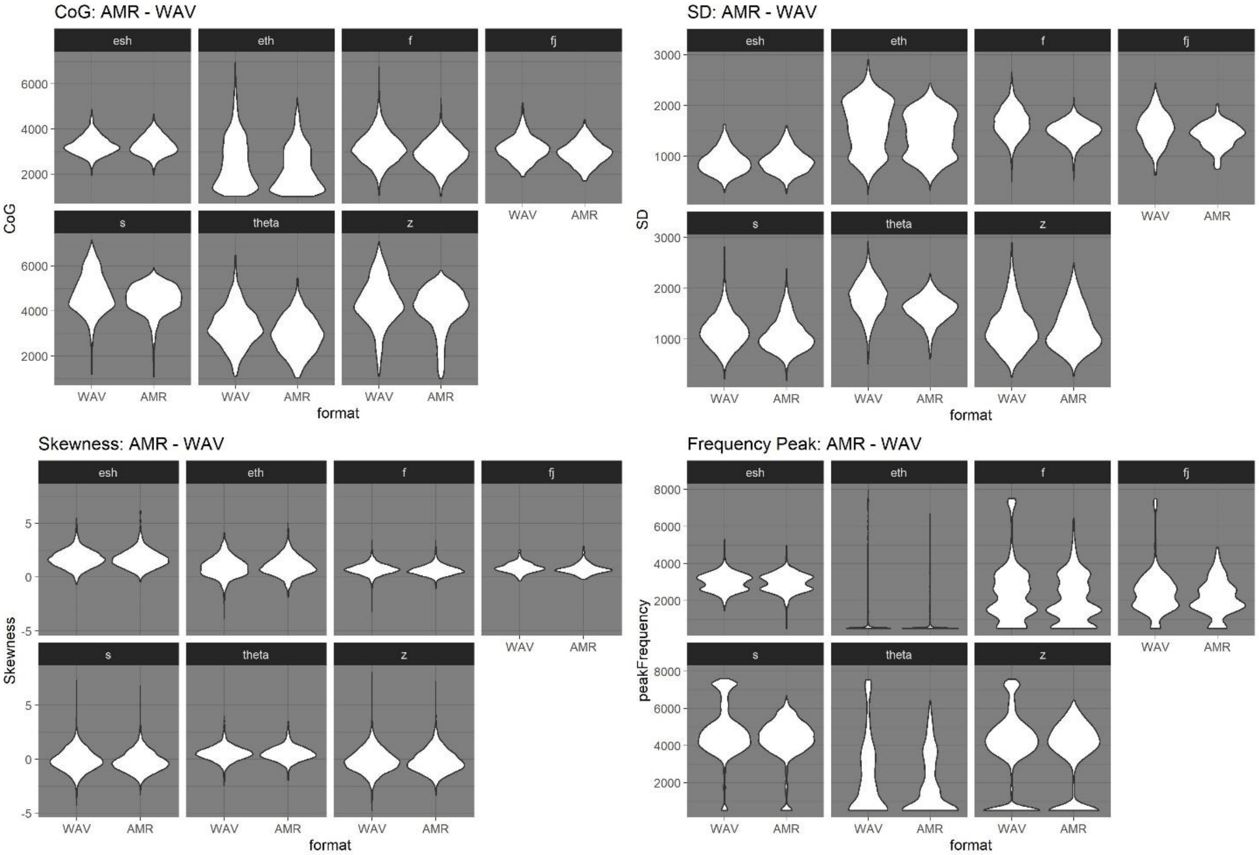


Figure 3.9 Distribution of spectral measure values in WAV baseline and the AMR-WB codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

The overall patterns from these graphs are lower CoG and SD particularly for the voiceless segments, while skewness and the voiced segments have more varied patterns with smaller effects and occasional increases. The violin plots confirm these patterns and show the main effect on frequency peak to be on the frequencies towards the upper-frequency limit.

3.4.1.1 /f/

For /f/ all mean values for the spectral measures are lowered in the AMR-WB compression, which is also evident from the trajectories of the linear predictions. From the distribution plots it is clear that the main effects appear on CoG and SD as well as the higher frequency content. It can further be seen that the AMR-WB compression tends to centre the frequency content around the mean.

Specifically, this is evident for CoG and frequency peak, which both have the mean value lowered with around 300 Hz by the AMR-WB compression (p <.0001). For SD the decrease is just over 190 Hz. In comparison for skewness and kurtosis, the effect of the AMR-codec is far less pronounced. Both are again lowered in terms of mean value, but only with 0.09 for skewness (p<.001) and 0.13 for kurtosis. All p-values and further statistical results can be found in table 3.6 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/f/	276.55	19.45	<.0001
SD	WAV - AMR	/f/	184.71	28.73	<.0001
Skewness	WAV - AMR	/f/	0.09	5.23	<.0001

Table 3.6. Statistical results of the difference between WAV and AMR based on linear prediction models for /f/

All maximum values in the AMR-WB compression confirms the pattern of the codec lowering the spectral measures. For /f/ the biggest lowering of the maximum value in Hz is found for frequency peak, which changed from 7,500 Hz in the WAV baseline to 6,438 Hz in the AMR-WB compression. CoG and SD presented changes in maximum values by around 400 Hz, whereas the skewness maximum value was only lowered with around 0.01.

A typical spectrographic representation of /f/ in the WAV baseline and the AMR-WB compression is presented below in figure 3.10. This show a general reduction in intensity both in the spectrogram and waveform, while the spectrogram also reveal of the frequencies just below 8 kHz are not encoded by the AMR-WB. The waveform further show how the amplitude has been centred around 0dB. The spectrum confirms the lowered intensity, but with more variation across the segment.

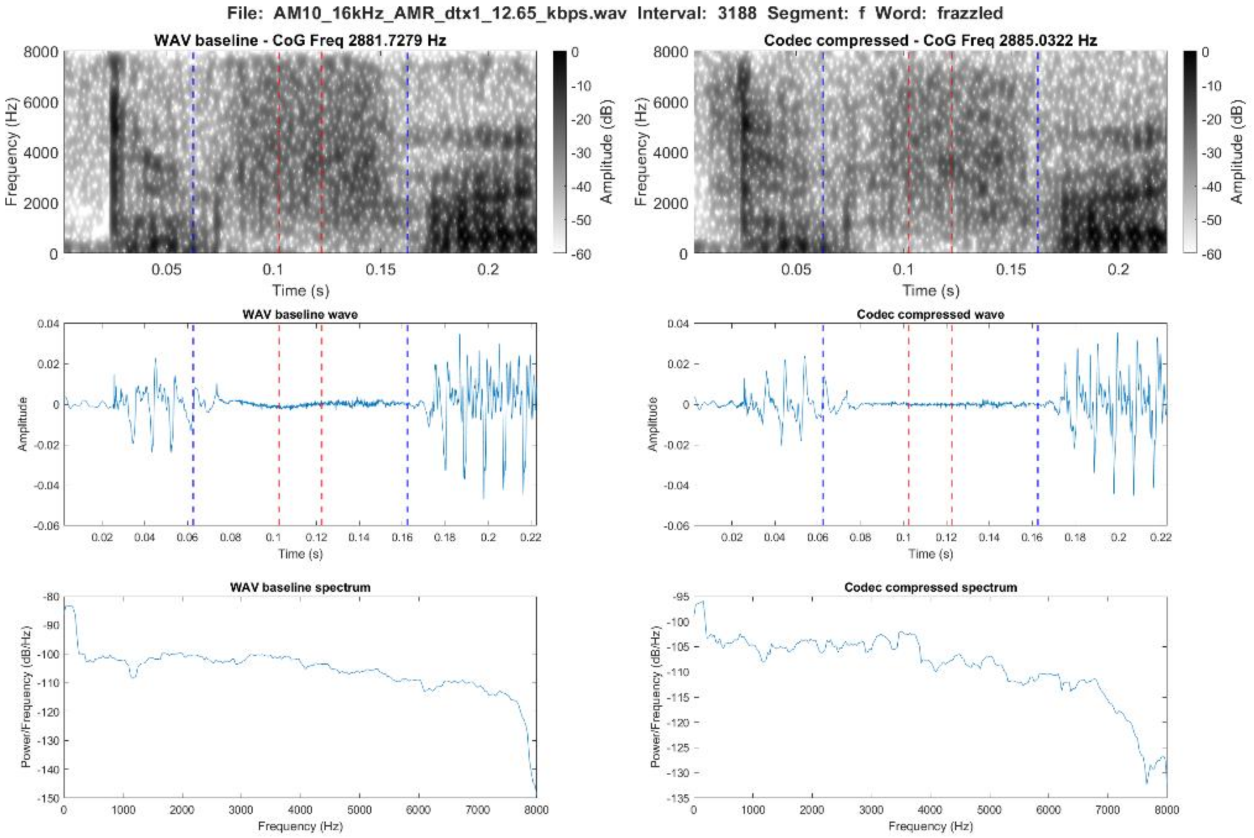


Figure 3.10. Spectrographic comparison of /f/ in the word *frazzled* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.2 /θ/

For /θ/ the shape of the distributions are largely intact. However, again the higher frequencies, CoG, and SD appear to be affected with a substantial change in distribution for SD. The distributional effect on SD is generally very similar to the one observed for /f/. In addition, the AMR-WB has a clear effect on the frequency peak with a greater amount of tokens below 2 kHz.

More specifically, the mean CoG is significantly lowered by the AMR compression by just over 400 Hz ((p = <.0001). A similar lowering is found for frequency peak at just over 400 Hz. SD is also lowered by just over 200 Hz (p <.0001), whereas skewness and kurtosis remain largely unaffected with an increase of 0.004 for skewness and a decrease of 0.09 for kurtosis. For skewness, this minor increase is not significant. Furthermore, the difference in CoG between /f/ and /θ/ is limited in the codec compression and the two sounds become almost identical. This is evident from the mean values as well as the linear prediction of the mixed effects model. The same tendency is observed for skewness, where /f/ and /θ/ again become similar following codec compression. Overall, this follows the observation from the changes in the distribution plots. All p-values and further statistical results can be found in table 3.7 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/θ/	388.89	16.75	<.0001
SD	WAV - AMR	/θ/	222.68	21.21	<.0001
Skewness	WAV - AMR	/θ/	-0.004	-0.14	0.89

Table 3.7: statistical results of the difference between WAV and AMR based on linear prediction models for /θ/

The maximum values confirm the patterns described above. Again, all maximum values are lowered and the most substantial change in Hz is found for frequency peak, which lowers from 7,531 Hz in the WAV baseline to 6,438 Hz in the AMR-WB compression. In that way, the codec compression lowered the frequency peak of /θ/ and /f/ to very similar extents with only a 31 Hz difference in the WAV baseline between the two. By contrast, the AMR-WB compression has lowered the max CoG value of /θ/ with close to 1 kHz and SD with around 600 Hz.

A typical spectrographic representation of /θ/ in the WAV baseline and the AMR-WB compression is presented below in figure 3.11. The spectrogram and waveform as with /f/ shows a reduction in intensity across the frequency range as well as a particular reduction of the frequencies around 8 kHz. Additionally, the final part of the segment is following transmission made almost indistinguishable from the surrounding non-speech sounds. This is especially for the frequencies above 4 kHz. The spectrum confirms this pattern.

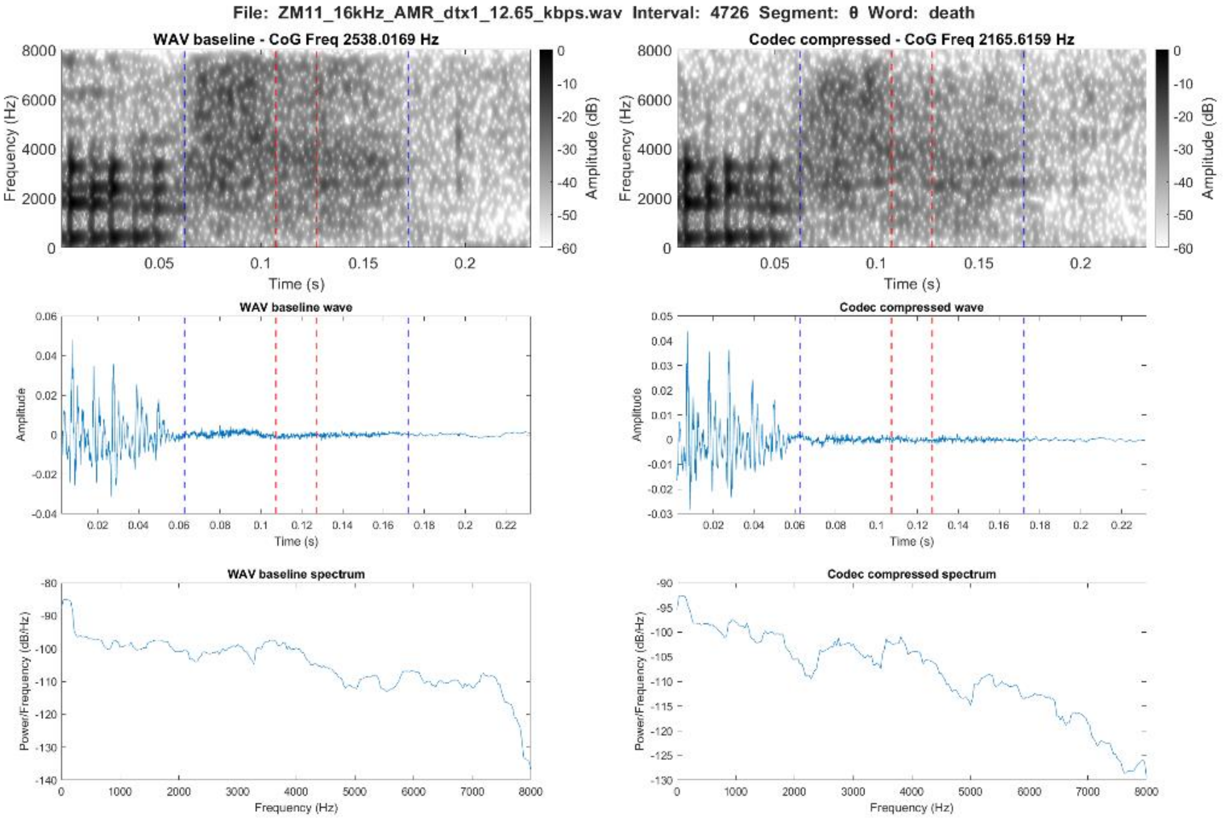


Figure 3.11. Spectrographic comparison of /θ/ in the word *death* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.3 [fʲ]

For [fʲ] all spectral measures are lowered, which is confirmed by the linear prediction from the mixed effect modelling and the distribution plots. Apart from skewness, which appear largely unaffected by the AMR-WB compression, a lowering can be observed in all the distribution plots. This lowering is clearly visible for the frequency peak, which however also appears to increase the number of tokens around the upper-cut off. For CoG the shape of the distribution stays largely unchanged, however for SD a similar pattern to previous segments is found with values being more centred.

The lowering is significant for all measures (p = <.0001) apart from skewness, which presented a lowering in mean value by 0.11. The mean of CoG and frequency peak was lowered by around 200 Hz, whereas SD was lowered by 138 Hz and kurtosis by 0.28. All p-values and further statistical results can be found in table 3.8.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	[fʲ]	193.60	3.62	<.0.001
SD	WAV - AMR	[fʲ]	138.47	5.72	<.0001
Skewness	WAV - AMR	[fʲ]	0.11	1.71	0.09

Table 3.8. Statistical results of the difference between WAV and AMR based on linear prediction models for [fʲ]

In terms of maximum values, it is interesting to note that the frequency peak for [fʲ] dropped from 7,469 Hz in the WAV baseline to 4,875 Hz in the AMR-WB compression. Apart from this, the CoG maximum value showed the biggest effect of the AMR-WB codec with a lowering from 5,150 Hz in WAV baseline to 4,410 Hz. The SD max was also lowered, but to a lesser extent, whereas skewness presented a slight increase of 0.3.

A typical spectrographic representation of [fʲ] in the WAV baseline and the AMR-WB compression is presented below in figure 3.12. The effects on [fʲ] are comparatively less than what has been observed for /f/ and /θ/. A slight decrease in intensity can be observed, but this is mainly to the frequencies around the 8 kHz limit. Both waveform and spectrum confirms this pattern with no substantial changes following the codec compression.

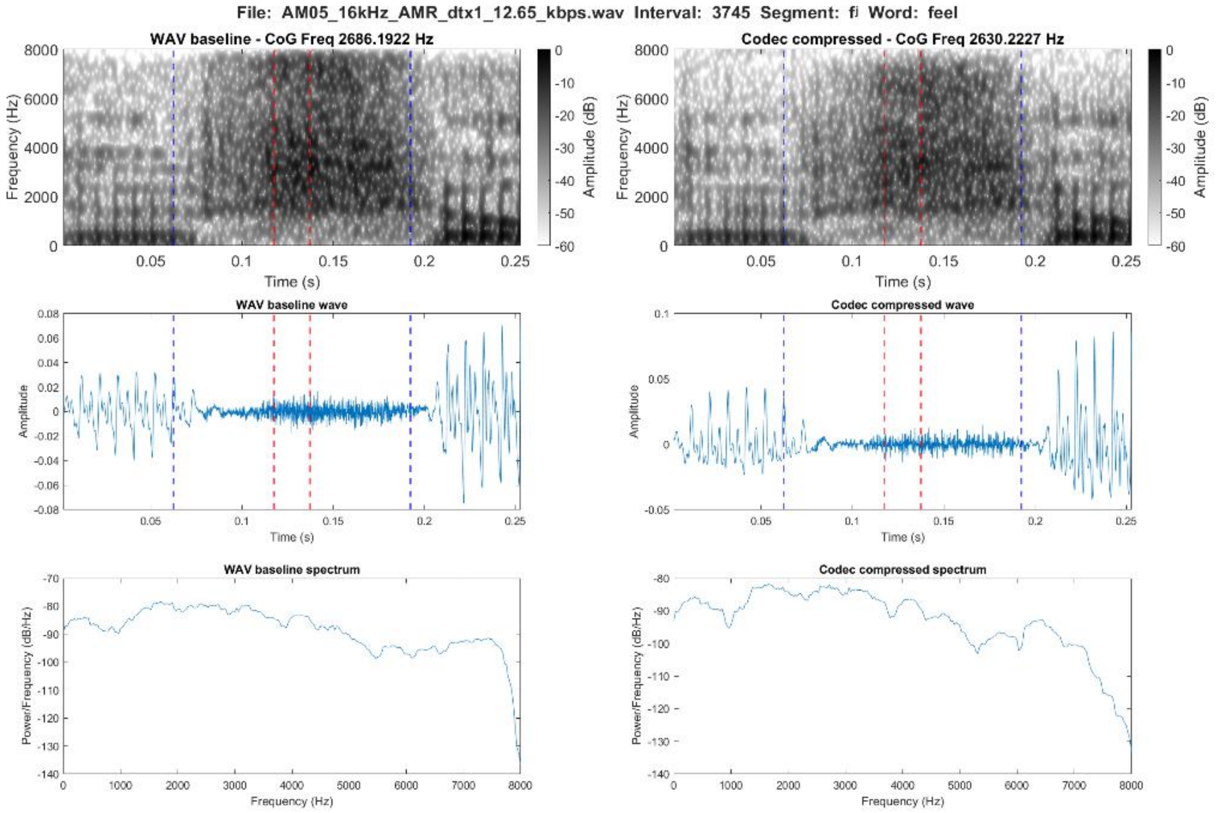


Figure 3.12. Spectrographic comparison of [fʲ] in the word *feel* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.4 /s/

For /s/ all spectral measures were again lowered apart from kurtosis, which showed an increase in mean value of 0.06. The distribution plots reveal the greatest effect of the AMR-WB compression on CoG and frequency peak. It is evident from the distribution that the AMR-WB has particularly affected the frequency content in the higher part of the spectrum of /s/. This can be seen from the distribution of the lower values remaining almost unchanged following compression. For SD a lowering and centring of values can also be observed, whereas the skewness values appear slightly more centred, but in general unchanged.

This limited effect on skewness is substantiated by the mean values, which showed a change of 0.14. Similarly, SD showed only a minor effect of 74 Hz. Nevertheless, the change was significant for both SD and skewness (p <.0001). The lowering was just below 385 Hz for CoG and around a 100 Hz more for frequency peak with a decrease of just below 360 Hz, which is a significant effect for CoG (p<.0001). All p-values and further statistical results can be found in table 3.9 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/s/	284.40	29.36	<.0001
SD	WAV - AMR	/s/	73.98	16.89	<.0001
Skewness	WAV - AMR	/s/	0.14	11.69	<.0001

Table 3.9. Statistical results of the difference between WAV and AMR based on linear prediction models for /s/

Looking at the maximum values, the biggest effect on /s/ was found for CoG with a change from 7,137 Hz in the WAV baseline to 5,934 Hz in the AMR-WB compression. For all the other spectral measures a lowering was found as well by just over 900 Hz for frequency peak and around 450 Hz for SD. Skewness showed a decrease in maximum value of around 0.5.

A typical spectrographic representation of /s/ in the WAV baseline and the AMR-WB compression is presented below in figure 3.13. In contrast to the previous voiceless segments, the codec compression appear to have resulted in an increase in intensity for /s/. From both spectrogram and spectrum this is clear to be particularly around 3-4 kHz and again around 7 kHz.

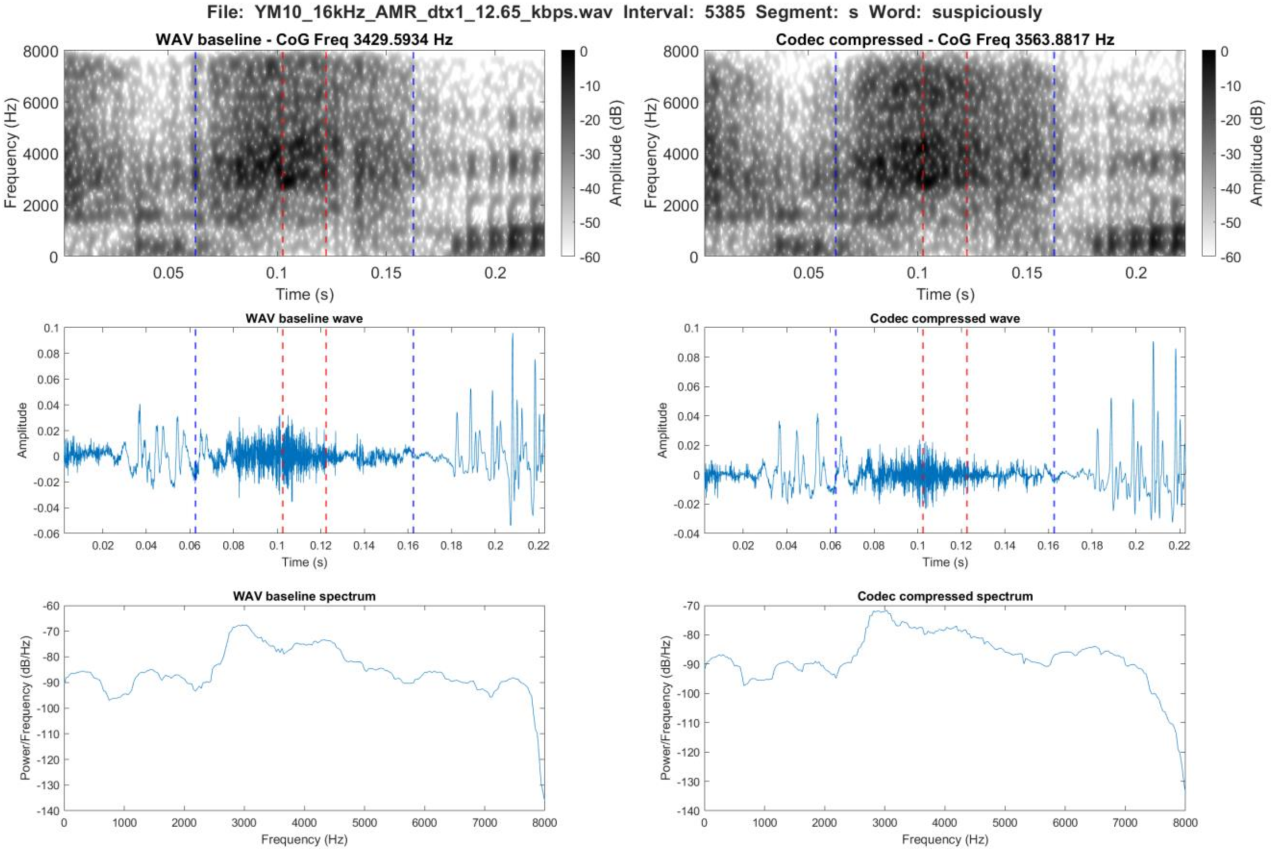


Figure 3.13. Spectrographic comparison of /s/ in the word *suspiciously* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.5 /ʃ/

For /ʃ/ the effect of AMR codec compression is very limited, which can been seen in the distribution plots as well as the spectral measures that stay largely unchanged from the baseline WAV condition. The largest difference in mean value is observed for frequency peak with a 32 Hz lowering, whereas CoG lowered by just below 1 Hz. By contrast, the mean SD increased by around 12 Hz. None of these changes reached significance, however. For skewness, the change in mean value was only at 0.02 and was likewise not significant. For kurtosis, the change was bigger than for some of the other voiceless segments with a lowering of 0.67 in mean value. All p-values and further statistical results can be found in table 3.10 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/ʃ/	0.87	0.04	0.97
SD	WAV - AMR	/ʃ/	-12.31	-1.17	0.24
Skewness	WAV - AMR	/ʃ/	0.02	0.66	0.51

Table 3.10. Statistical results of the difference between WAV and AMR based on linear prediction models for /ʃ/.

The maximum values did not divert from the pattern described above as the biggest change observed was for frequency peak by 313 Hz and only a 17 Hz difference for SD. CoG lowered by 216 Hz and skewness by around 0.6.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the AMR-WB compression is presented below in figure 3.14. Both spectrogram, waveform, and spectrum confirm how the codec compression have little to no effect on /ʃ/.

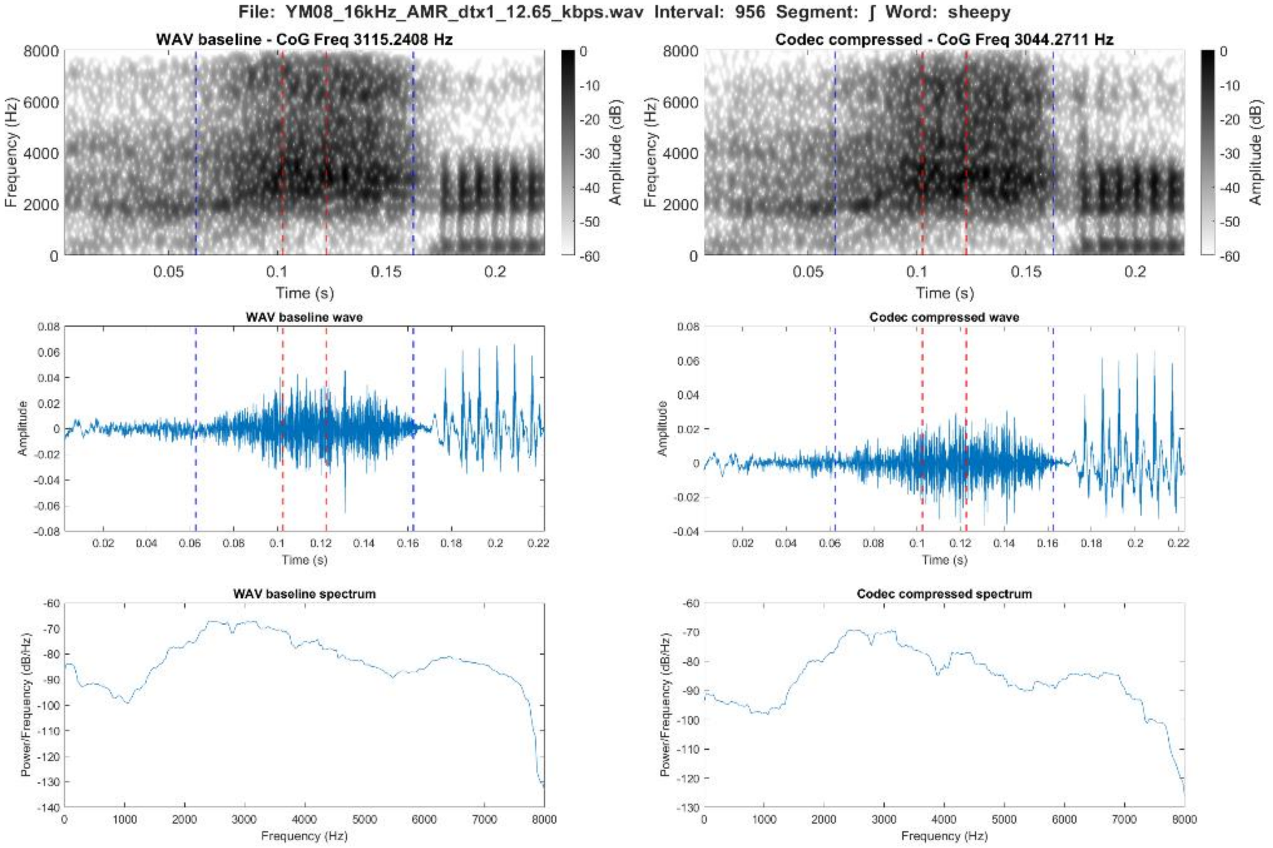


Figure 3.14. Spectrographic comparison of /ʃ/ in the word *sheepy* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.6 /z/

The effect of the AMR codec on /z/ is more varied than the effect observed for the voiceless segments. In terms of distribution, the most apparent change is for CoG. Here it is again the content in the higher frequencies that appear to be affected and centred lower in the spectrum. The distribution of the skewness values appear to be mainly affected in terms of a limitation in outliers both above and below the mean. On the other hand, SD and frequency peak show, in different ways an increased number of tokens in the lower part of the spectrum.

The mean value of CoG was lowered with around 300 Hz (p <.0001) and frequency peak similarly by just above 380 Hz. SD showed a decrease in mean value of 25 Hz, which was again significant (p <.0001). Skewness remained almost unchanged between the baseline WAV and the AMR compression in terms of mean value with a decrease of only 0.03, which did not reach significance. Kurtosis showed a lowering of 0.39. All p-values and further statistical results can be found in table 3.11 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/z/	303.68	16.02	<.0001
SD	WAV - AMR	/z/	25.35	3.00	0.003
Skewness	WAV - AMR	/z/	0.03	1.57	0.12

Table 3.11. Statistical results of the difference between WAV and AMR based on linear prediction models for /z/

The maximum values for frequency peak spanned from a maximum value of 7,563 Hz in the WAV baseline to 6,438 Hz in the AMR-compression. Both CoG and SD showed similarly high values, which lowered by around 120 to 400 Hz. Skewness lowered from just above 8 to just above 7.

Two typical patterns were observed in spectrographic representation of /z/ related to the phonological voicing distinction made by the MFA. This is illustrated in figure 3.15 and 3.16 below with two comparisons of /z/ in the WAV baseline and the AMR-WB compression. Firstly, Figure 3.15 illustrates a phonetically voiced example, while figure 3.16 illustrates the phonological distinction made by the MFA with /z/ phonetically voiceless.

With this in mind, the effect of the codec compression for the voiced /z/ in figure 3.15 appears as a general reduction across the frequency range with a gradual upper cut-off just under 8 kHz. For the voiceless /z/ in figure 3.16, the spectrogram shows how some of the formant structure from the preceding vowel is enhanced in /z/ following the codec compression, while the fricative is overall reduced in intensity. The latter is visible from both waveform and spectrum.

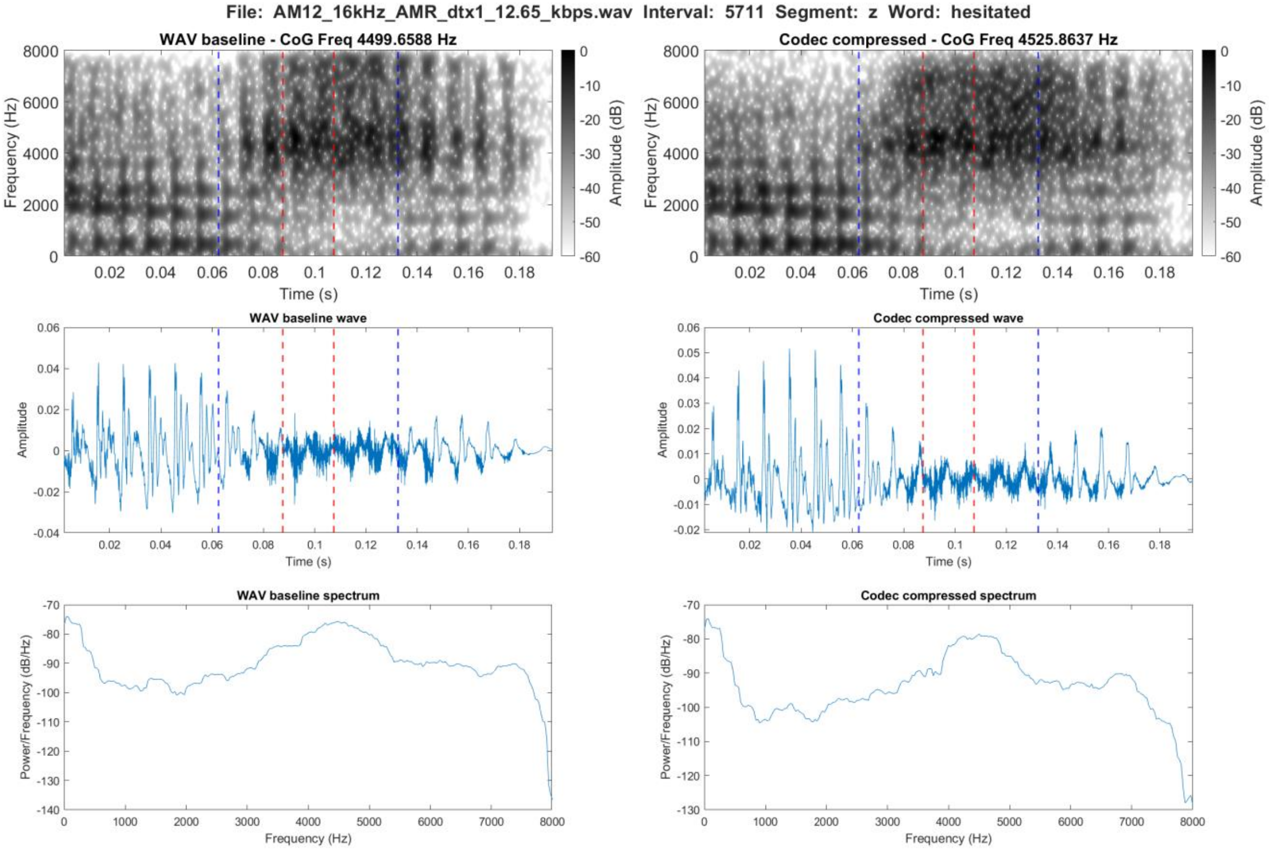


Figure 3.15. Spectrographic comparison of voiced /z/ in the word *hesitated* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

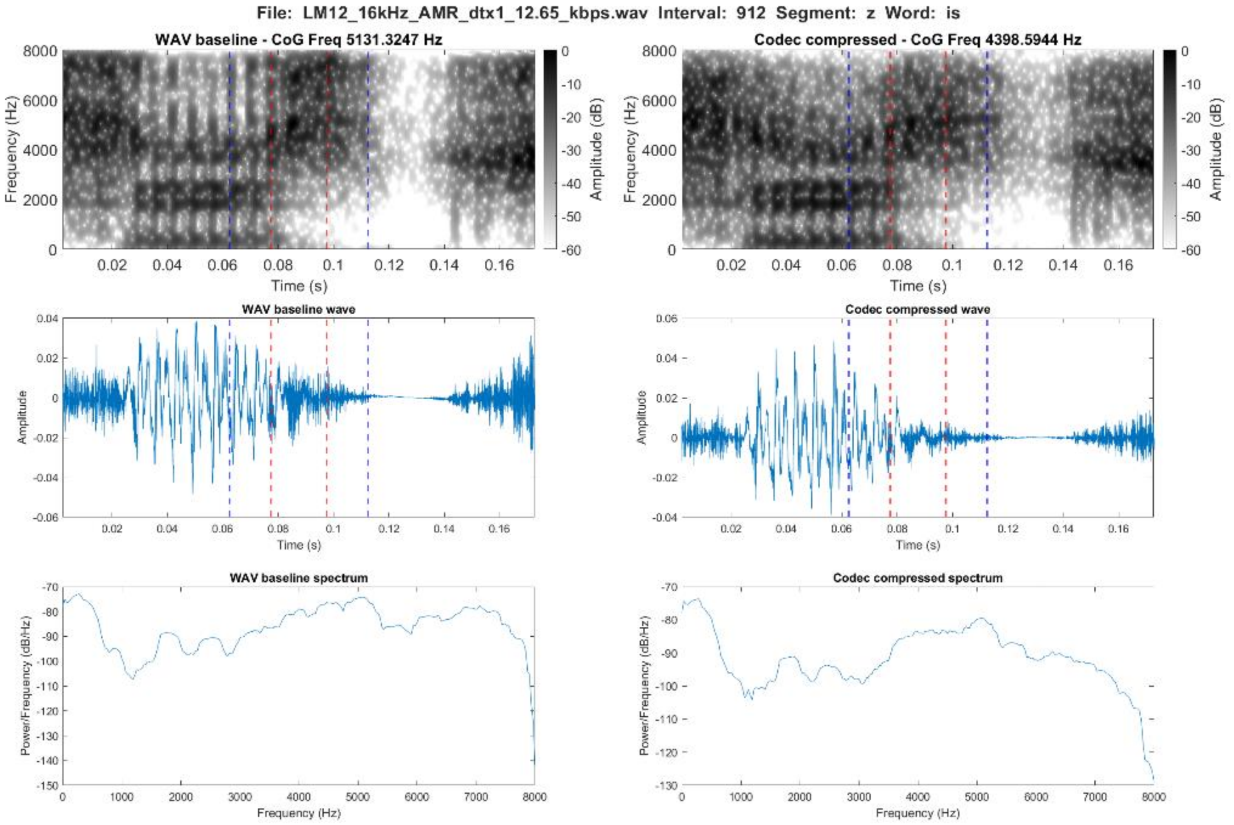


Figure 3.16. Spectrographic comparison of voiceless /z/ in the word *is* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.1.7 /ð/

For /ð/ all measures apart from skewness were lowered, which is substantiated by the distribution plots. For CoG and SD in the distribution plots, this lowering can also be observed. For these two measures, it is also clear from the plots that the values are generally distributed more equally across the spectrum in comparison to the other segments. This is apart from frequency peak, which in both baseline WAV and AMR-WB compression is centred and primarily found around the 1 kHz cut-off. For skewness, /ð/ follows similar patterns to the other segments with the values being more centred around the mean.

For both CoG and SD this was significantly so with a decrease in mean value of 216 Hz for CoG and a little less 139 Hz for SD (p<.0001). For both skewness and kurtosis, the lowering was again very slight. However, the effect on skewness (i.e. increase in mean of 0.04) was significant according to the mixed effects modelling (p <.0001). All p-values and further statistical results can be found in table 3.12.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/ð/	298.57	13.11	<.0001
SD	WAV - AMR	/ð/	176.89	17.42	<.0001
Skewness	WAV - AMR	/ð/	-0.09	-3.79	<.001

Table 3.12. Statistical results of the difference between WAV and AMR based on linear prediction models for /ð/

For /ð/ the frequency peak changed from a maximum value of 7,938 Hz in the WAV baseline to 6,688 Hz in the AMR-compression. SD showed similarly high values, which lowered by around 500 Hz. Skewness likewise lowered by around 0.9.

As with /z/, two typical variations of /ð/ were observed in the spectrographic representations. Here this relates to the fact that /ð/ was at times produced with a stop and thus, had more plosive structure. These two types of /ð/ in the WAV baseline and the AMR-WB compression is presented below in figure 28 and 29.

Similar to /z/, the spectrogram and spectrum for the voiced /ð/ in figure 3.17 shows a general reduction in intensity. Figure 3.18 illustrates an example of the plosive like production of /ð/. A slight decrease in intensity can be observed in both spectrogram, waveform and spectrum. However, the main effect is again a reduction in the frequencies around the upper frequency limit. For /ð/ this reduction is from around 7 kHz.

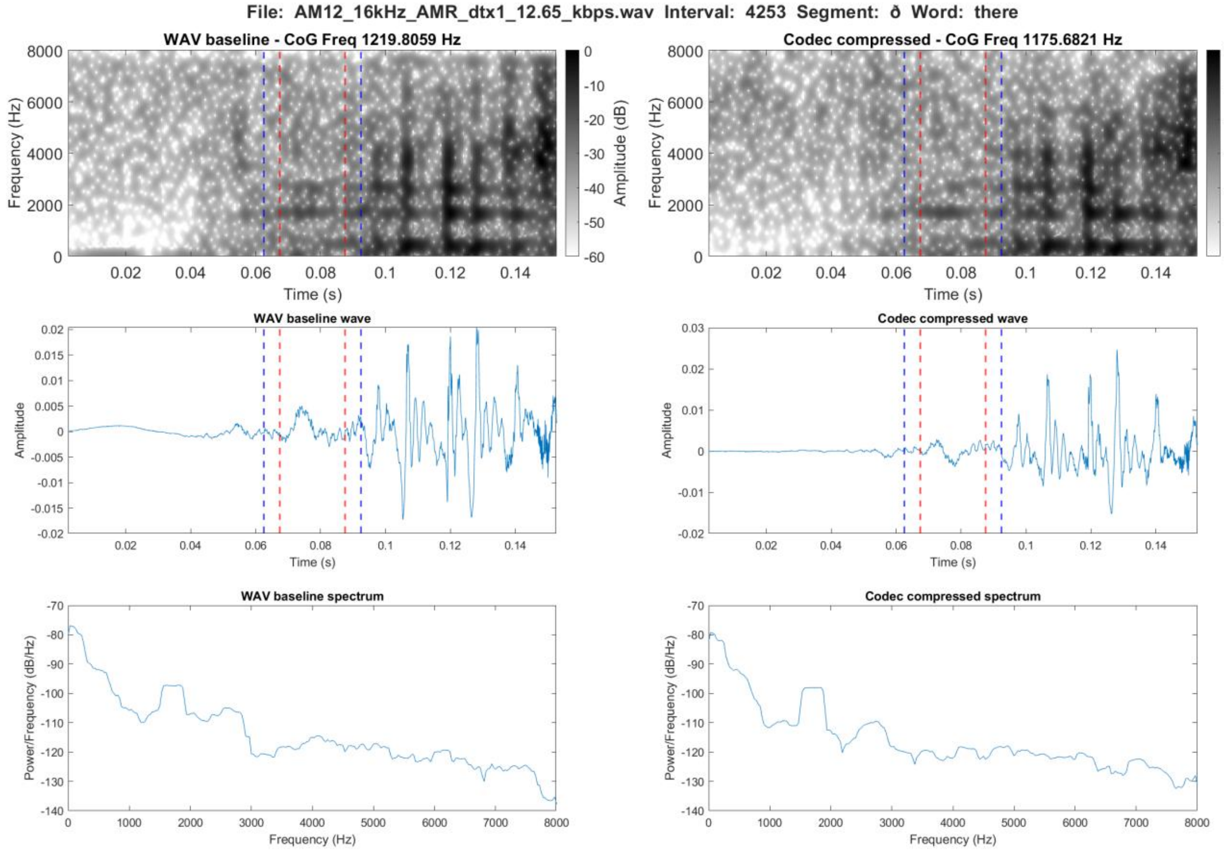


Figure 3.17. Spectrographic comparison of voiced /ð/ in the word *there* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

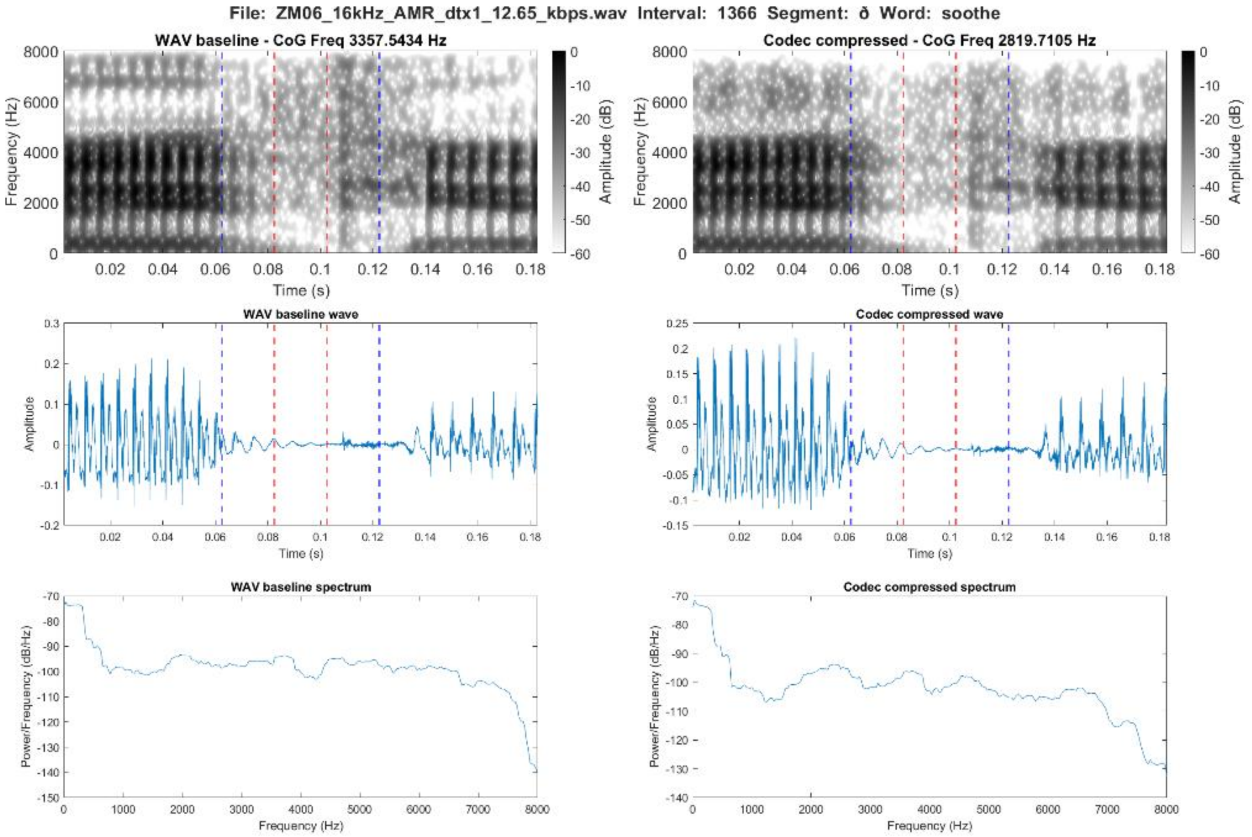


Figure 3.18. Spectrographic comparison of plosive /ð/ in the word *soothe* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

3.4.2 MP3

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the MP3 codec. Before these results are presented the linear predictions for each spectral measure and the individual segments can be found below (figure 3.19 to 3.24). These indicate the directionality of the changes imposed by the MP3 codec. As with the AMR-WB results, the graphs present the results divided into voiced and voiceless segments and the more specific analysis of the plots will be found in the following sections on the individual segments.

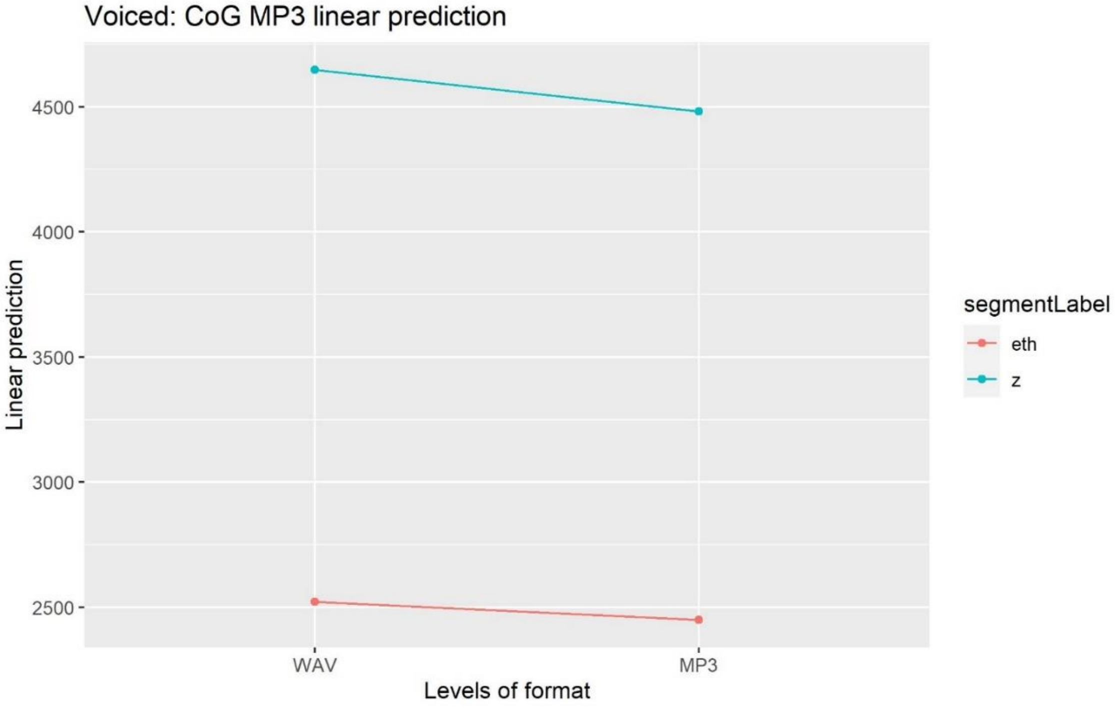


Figure 3.19. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for CoG and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

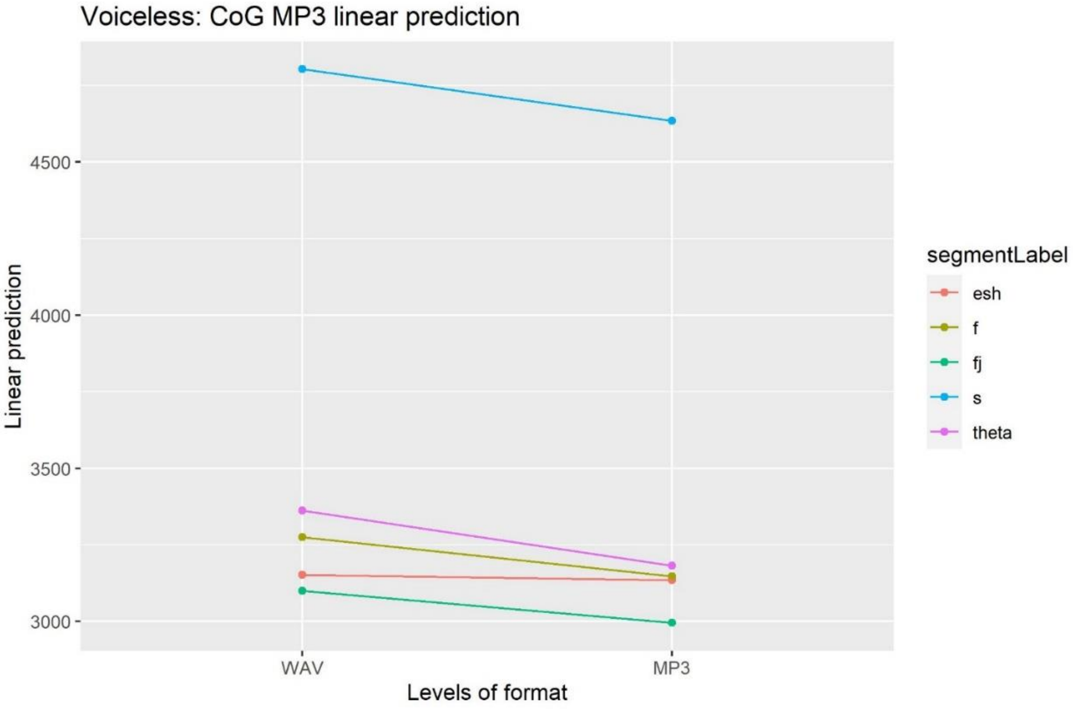


Figure 3.20. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for CoG and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

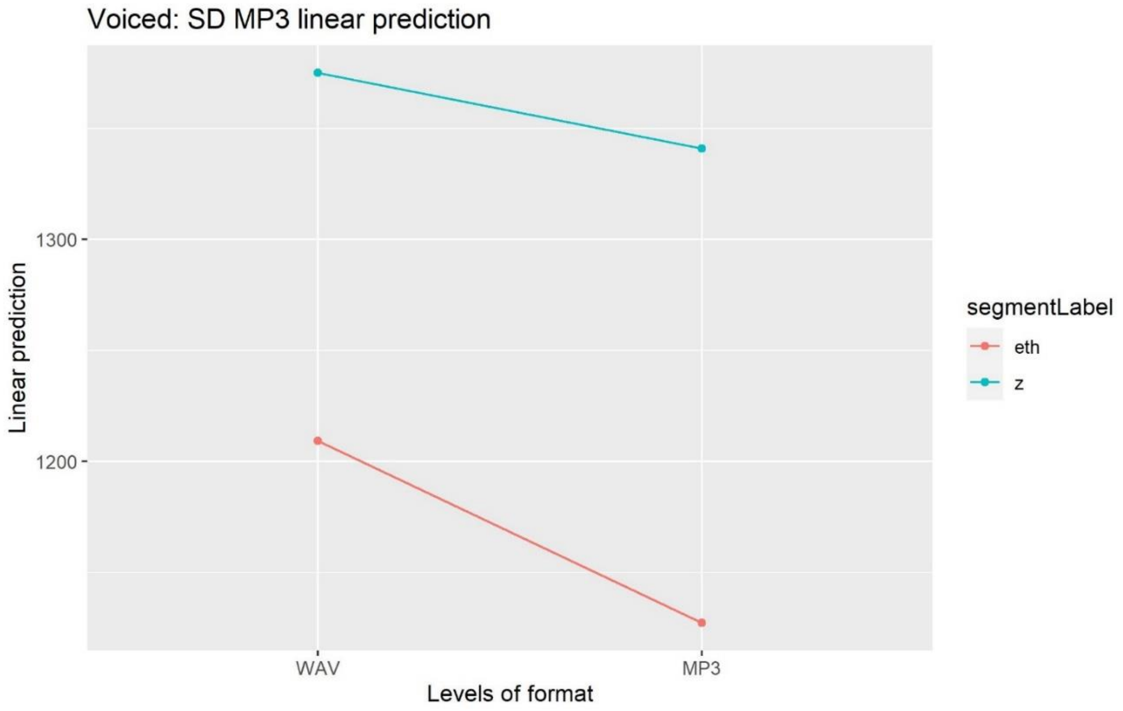


Figure 3.21. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for SD and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

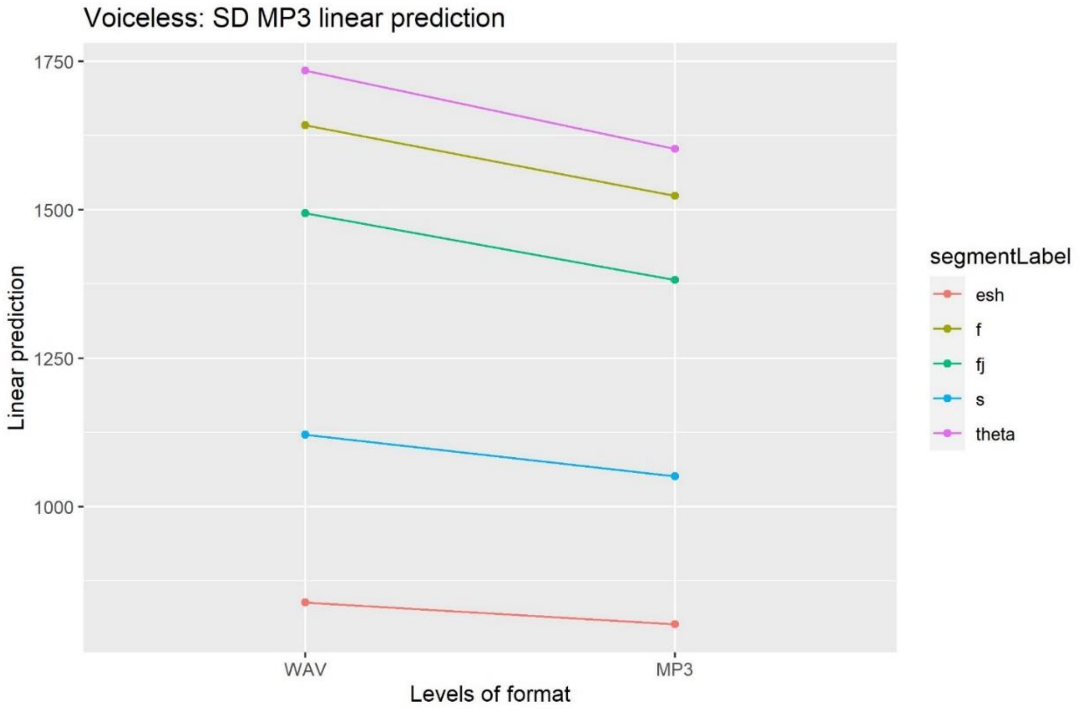


Figure 3.22. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for SD and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

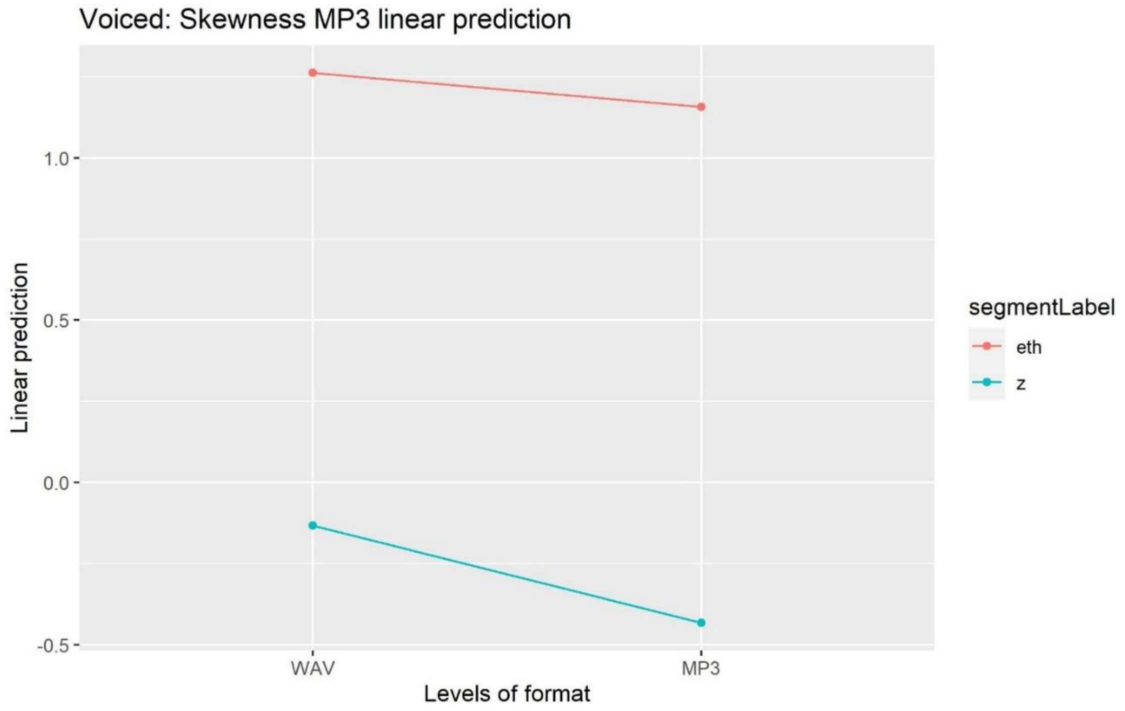


Figure 3.23. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for skewness and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

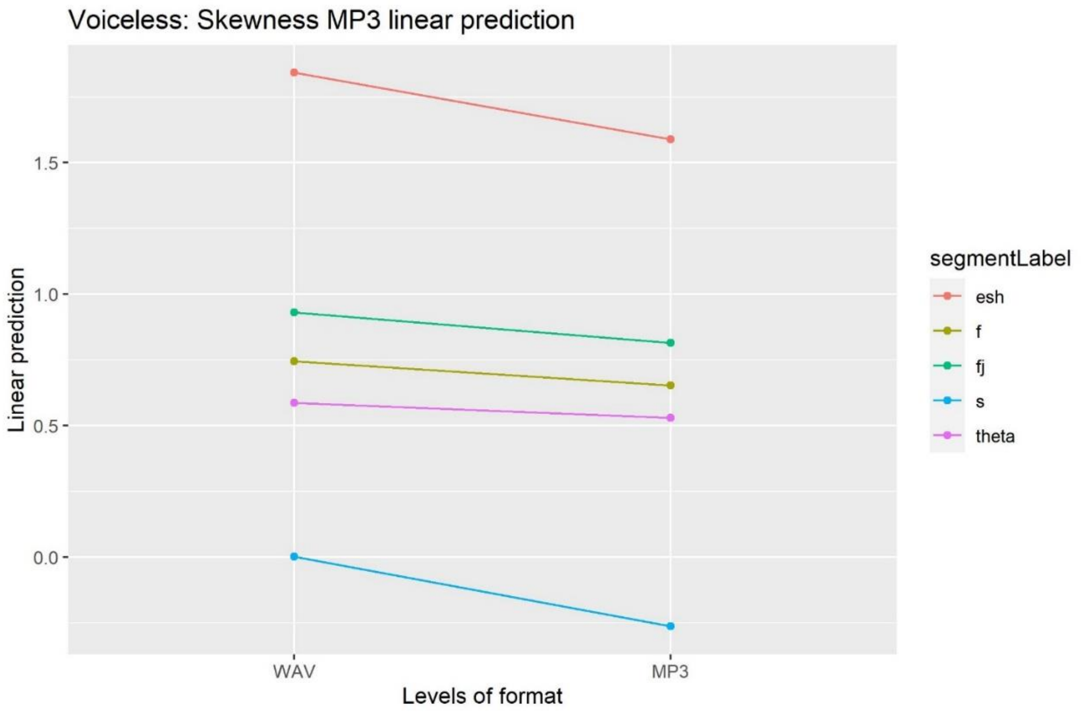


Figure 3.24. Trajectory of the linear predictions in the comparison of WAV and MP3 from the mixed effects models for Skewness and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The general distributions are illustrated below in a set of violin plots for each spectral measure. Again, the specific analysis pertaining to each segment can be found in the relevant sections below.

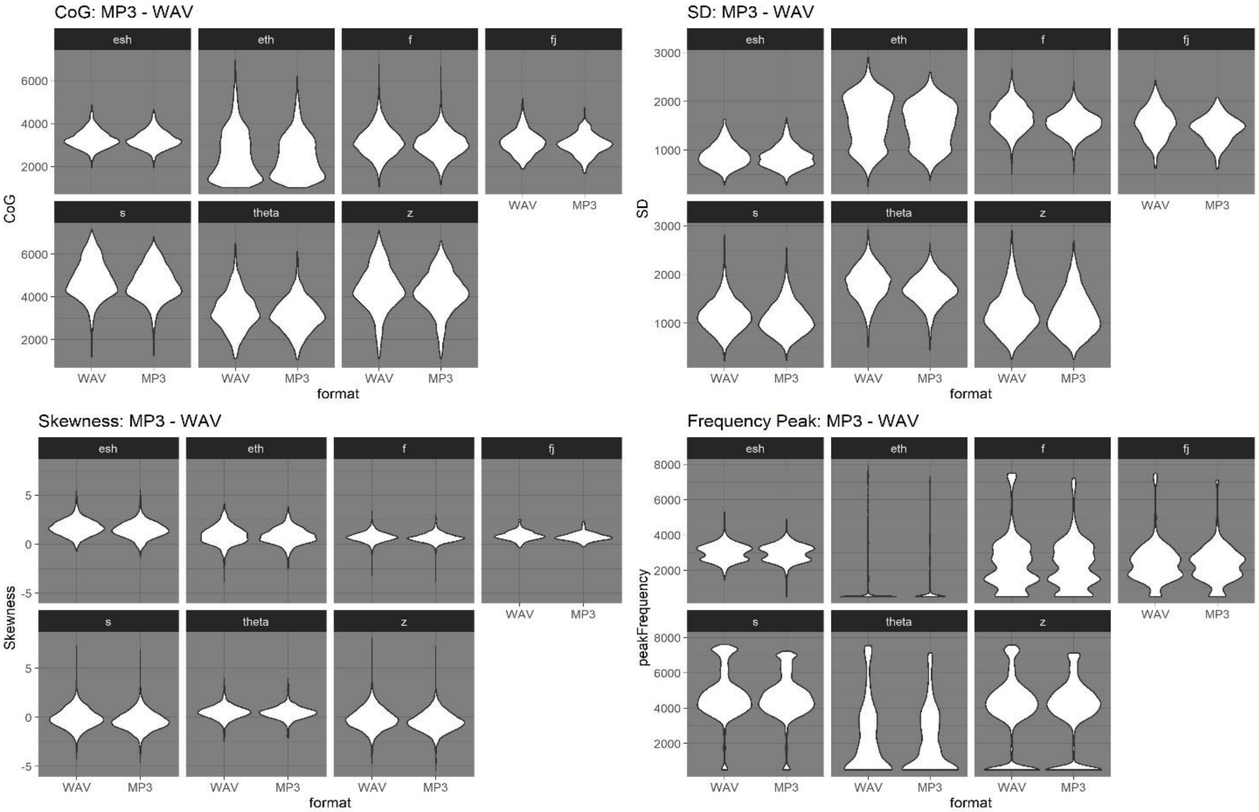


Figure 3.25. Distribution of spectral measure values in WAV baseline and the MP3 codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

The overall pattern following the MP3 compression for all segments and spectral measures, apart from CoG for /ð/, was to decrease.

3.4.2.1 /f/

For /f/, all spectral measures were lowered in the MP3 compression, which is confirmed by the linear prediction plots. These also indicates that the effect was greatest for CoG and SD, which was also seen in the violin distribution plots. Based on these plots, both CoG and SD were lowered and had more values centred around the mean. For skewness, the effect was almost non-observable and the frequency peak distribution remained almost unchanged apart from a lowering of the topmost values.

For CoG and SD the effect was significant (p <.0001) with a lowering of mean value around 300 Hz for CoG and just below 200 Hz for SD. From the linear prediction it is evident that for /f/ the lowering in CoG made /f/ almost identical to /ʃ/. Spectral peak followed a similar pattern with a lowering of the mean value of just above 350 Hz. For skewness, despite the change being only 0.06 it still reached significance (p<.0001). Kurtosis was equally lowered from 3.32 to 3.13. All p-values and further statistical results can be found in table 3.13 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	/f/	128.31	8.96	<.0001
SD	WAV - MP3	/f/	119.38	18.44	<.0001
Skewness	WAV – MP3	/f/	0.09	5.30	<.0001

Table 3.13. Statistical results of the difference between WAV and MP3 based on linear prediction models for /f/

The max values for /f/ substantiate the patterns observed above as all measures were lowered. The biggest change apart from frequency peak was found for SD, which was lowered from 2,664 Hz to 2,422 Hz in the MP3 compression. The rest of the spectral measures showed limited changes in max values. This was seen as a just below 60 Hz lowering of CoG, a 0.3 lowering of skewness value, and finally a lowering from 7,500 Hz for frequency peak in the WAV baseline to a peak at 7,219 Hz in the MP3 compression.

A typical spectrographic representation of /f/ in the WAV baseline and the MP3 compression is presented below in figure 3.26. Little change is observable in the comparison between the spectrogram, waveform and spectrum before and after codec compression. The only notable change appear as a slight change in intensity mainly visible in the spectrogram.

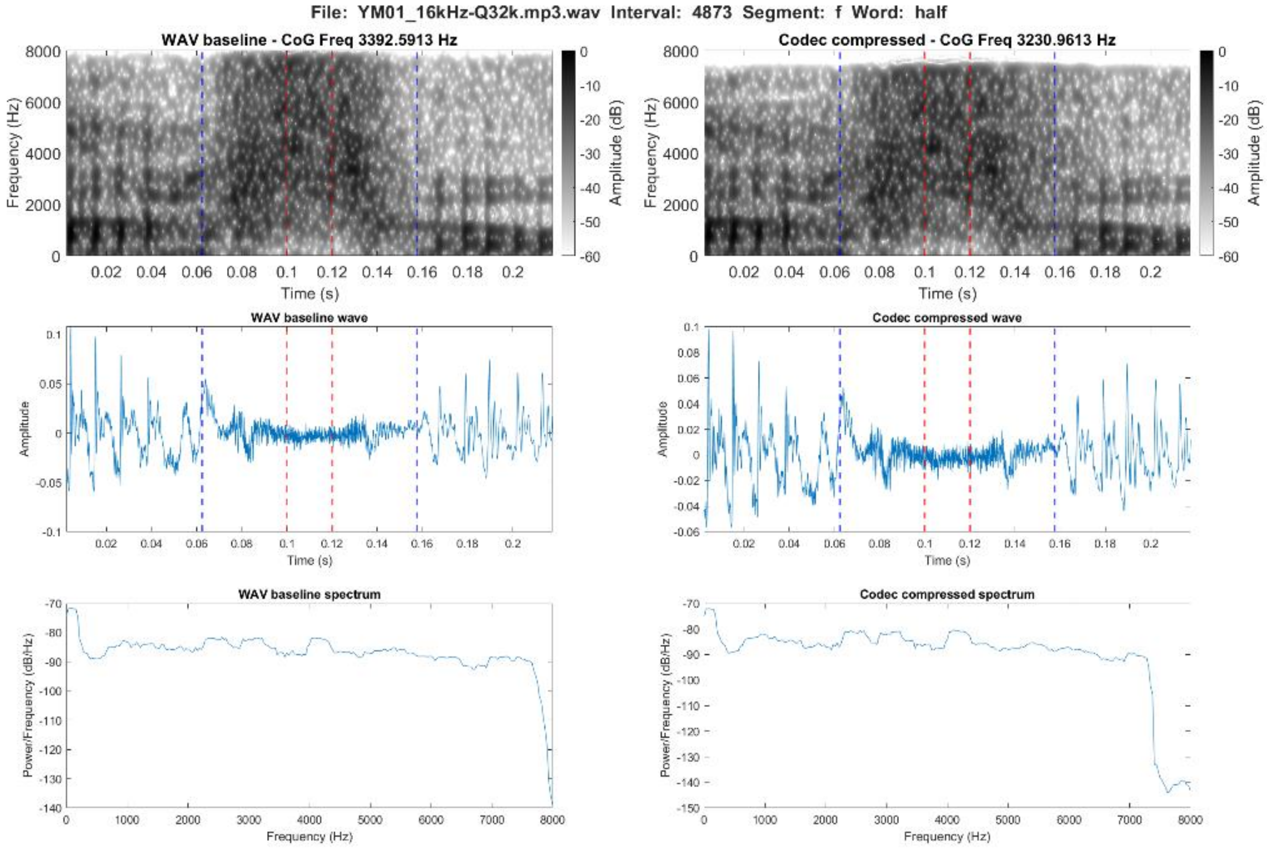


Figure 3.26. Spectrographic comparison of /f/ in the word *half* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.2 /θ/

/θ/ appears to follow the same pattern as the previous voiceless fricatives especially for CoG and skewness. This can be seen in the linear prediction, though the trajectory of skewness indicates only a slight lowering. For SD a more clear decrease of the mean value, and more values centred around this point, are evident from the distribution plots. A lowering and effect on the high frequencies are visible from the frequency peak distributions.

CoG was again significantly lowered (p <.0001) with a change in mean value of 420 Hz from the baseline to the MP3 compression. Similarly SD lowered by 218 Hz (p<.0001), while spectral peak lowered by just under 490 Hz. For skewness the tendency was upwards with a change in mean of 0.02, which only just reached significance (p = 0.05). From the linear prediction from the mixed effect modelling, it is evident that based on CoG, /θ/ becomes almost identical to /f/ following compression with the MP3 codec. All p-values and further statistical results can be found in table 3.14 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	/θ/	0.09	5.69	<.0001
SD	WAV - MP3	/θ/	131.48	12.43	<.0001
Skewness	WAV – MP3	/θ/	0.06	1.97	0.05

Table 3.14. Statistical results of the difference between WAV and MP3 based on linear prediction models for /θ/

Similar to /f/, the biggest change in maximum value for /θ/ was for frequency peak with a lowering from 7,531 Hz to 7,125 Hz. For CoG the effect of the MP3 compression is illustrated by a lowering of 94 Hz in maximum value, while SD lowered from 6,491 Hz to 6,113 Hz. Skewness again showed very limited effects of the compression on the maximum value with a change from 4.02 in the WAV baseline to 3.99 following codec compression.

A typical spectrographic representation of /θ/ in the WAV baseline and the MP3 compression is presented below in figure 3.27. Apart from a clear reduction in the frequencies just below the 8 kHz frequency limit, little change can be observed based on /θ/ on the basis of these plots.

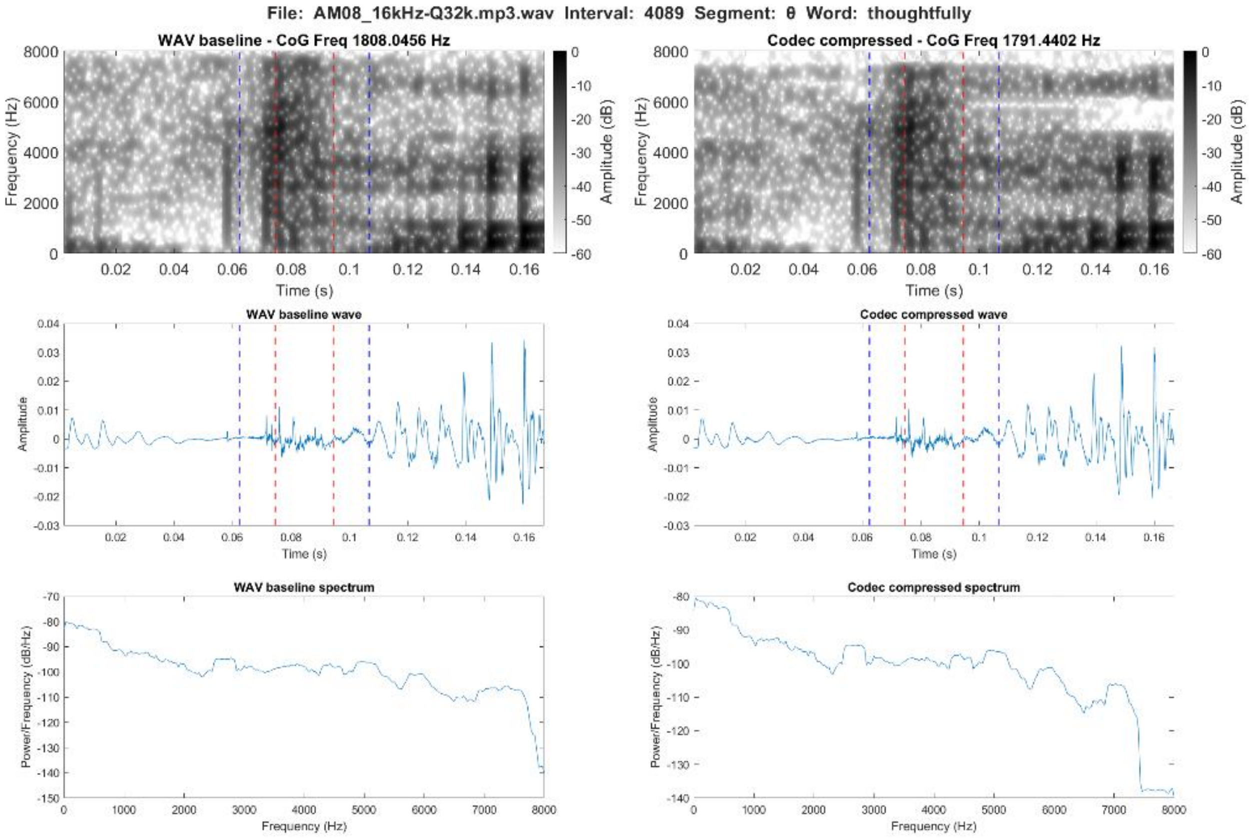


Figure 3.27. Spectrographic comparison of /θ/ in the word *thoughtfully* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.3 [fʲ]

For [fʲ] all spectral measures were lowered with similar distributional changes as what was observed for /f/. This is also evident from the linear predictions. Thus, the biggest changes were observed for CoG and SD following the MP3 compression, while skewness and frequency peak remained more or less unchanged. However, a slight change and tendency to more centred values in the MP3 compression can be observed for all spectral measures.

For CoG and SD, this was with changes in mean values of just around 100 Hz for both segments (p= 0.05 and <.0001). Skewness presented a lowering of mean value of 0.12, which was not significant. CoG presented a lowering of 104 Hz, and SD a lowering of 112 Hz. This downwards trend is confirmed by the linear prediction, where [fʲ] does not markedly become more or less alike any of the other fricatives. The mean kurtosis was lowered by 0.28, and frequency peak by less than 100 Hz. All p-values and further statistical results can be found in table 3.15.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	[fʲ]	104.55	1.95	0.05
SD	WAV - MP3	[fʲ]	112.30	4.62	<.0001
Skewness	WAV – MP3	[fʲ]	0.12	1.80	0.07

Table 3.15. Statistical results of the difference between WAV and MP3 based on linear prediction models for [fʲ]

As with the previous fricatives, the biggest effects in terms of maximum values for [fʲ] was observed for frequency peak with a change from 7,469 Hz in the WAV baseline to 7,094 Hz in the codec compression. The smallest effect was observed for skewness with a change around 0.2, while CoG was lowered by 411 Hz and SD lowered by 375 Hz. Thus, following all the tendencies described above.

A typical spectrographic representation of [fʲ] in the WAV baseline and the MP3 compression is presented below in figure 38. Very limited changes can be observed for [fʲ]. This is again apart from a clear upper frequency cut-off at just below 8 kHz.

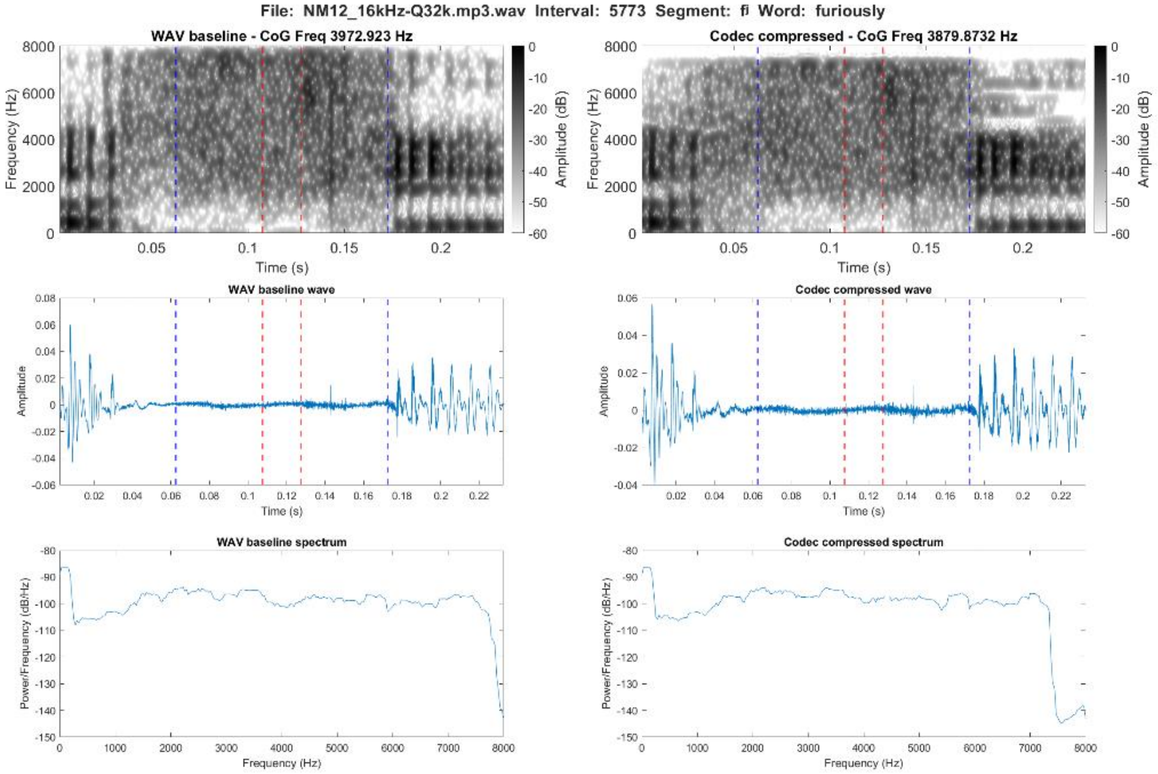


Figure 3.28. Spectrographic comparison of [fʲ] in the word *furiously* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.4 /s/

For /s/, all spectral measures apart from kurtosis were lowered by the MP3 compression. An observation, which is confirmed by the linear predictions. The most substantial effects in terms of distribution was for CoG and SD, where for SD a clear lowering of the mean is visible. For CoG the difference appears mainly in the higher frequency content. The same can be said for frequency peak, where the distributions below the mean appear largely unchanged. The change to the distribution of the skewness values for /s/ caused by the MP3 compression is mainly observable as a slight overall lowering.

In detail, this presented itself as a lowering of just under 170 Hz in mean value for CoG, and a change of around 70 Hz for SD. Nevertheless, both were significant (p <.0001). For skewness, the lowering of the mean value was again significant with a change from -0.08 to -0.35 (p<.0001). On the other hand, kurtosis increased by 0.19, while frequency peak was lowered by just under 130 Hz. All p-values and further statistical results can be found in table 18 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	/s/	168.72	17.34	<.0001
SD	WAV - MP3	/s/	69.69	15.83	<.0001
Skewness	WAV – MP3	/s/	0.27	22.56	<.0001

Table 3.16. Statistical results of the difference between WAV and MP3 based on linear prediction models for /s/

The maximum values were as the mean values lowered by the MP3 compression, but to a greater extent. The most substantial change was for CoG, where the maximum value changed from 5,150 Hz in the WAV baseline to 4,739 Hz in the codec compression file. In comparison, SD decreased by 268 Hz and frequency peak by 379 Hz. Skewness likewise lowered from a max of 7.33 in the baseline WAV to 6.94 in the MP3 compression.

A typical spectrographic representation of /s/ in the WAV baseline and the MP3 compression is presented below in figure 3.29. As with the previous segments, apart from the effect of the upper frequency limit, little to no change is observable in the spectrogram waveform, or spectrum following the codec compression.

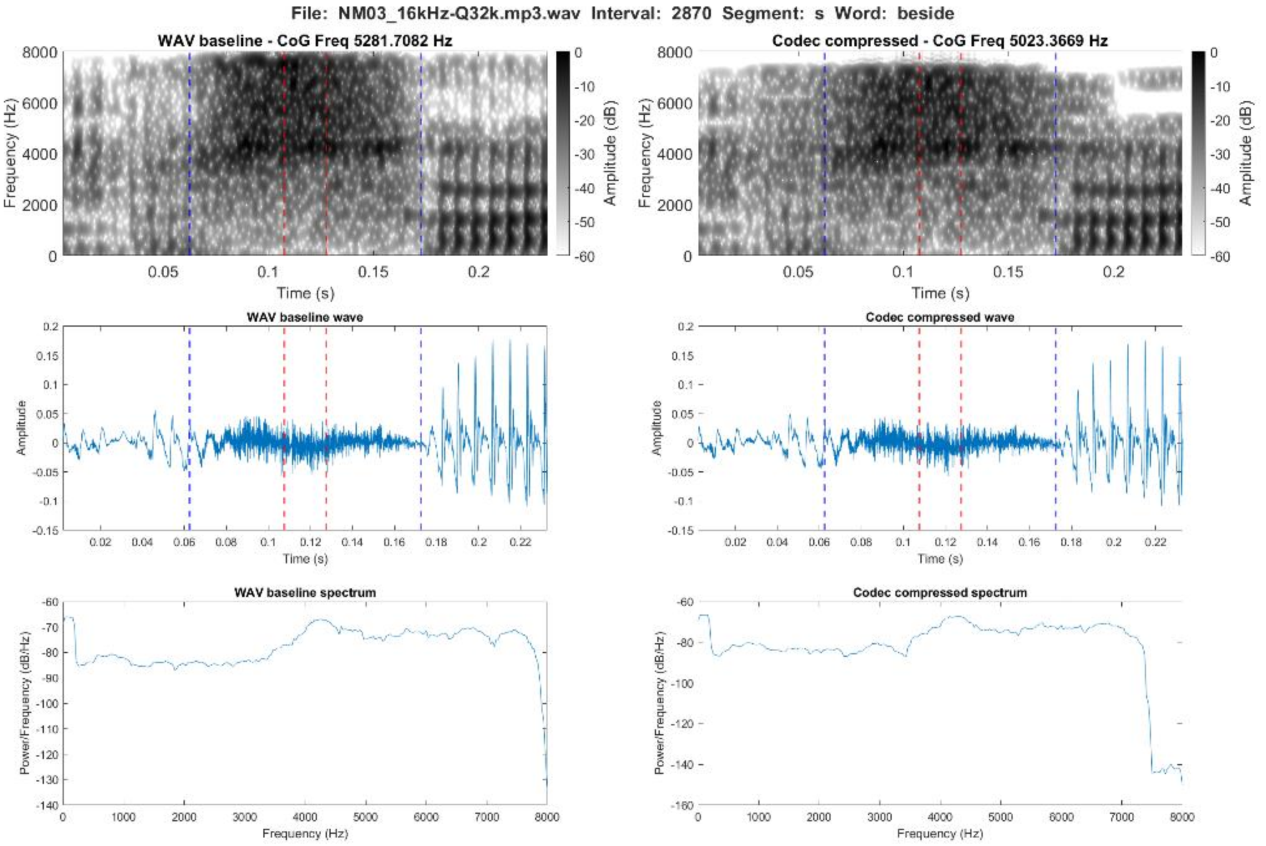


Figure 3.29. Spectrographic comparison of /s/ in the word *beside* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.5 /ʃ/

For /ʃ/ the effect of MP3 codec compression is very limited and the spectral measures again stayed largely unchanged despite a general downwards trend from the baseline WAV condition. This is also evident from the distribution plots, where no substantial changes are observable for CoG and skewness in the comparison of the WAV baseline and the MP3 compression. Despite the distribution plots not revealing any substantial changes to skewness, the linear prediction illustrate a downward trend. SD displays a smooth distribution with no evident disruption in the WAV baseline, whereas the MP3 compressed values appear with more values around the mean, and a division into two sections at this point. For frequency peak, the main thing to observe is the change to the higher frequencies, which were clearly lowered and the distribution made less distinct.

For SD and skewness the effect of the compression was significant (SD: p = .0001, skewness: p <.0001) with changes in mean values of 36 Hz and 0.25 respectively. For CoG the lowering in mean of just under 18 Hz was not significant. The changes to kurtosis amounts to 1.17 and around 8 Hz for frequency peak. All p-values and further statistical results can be found in table 3.17 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	/ʃ/	17.85	0.77	0.44
SD	WAV - MP3	/ʃ/	36.82	3.49	<.001
Skewness	WAV – MP3	/ʃ/	0.25	9.01	<.0001

Table 3.17. Statistical results of the difference between WAV and MP3 based on linear prediction models for /ʃ/

The limited effects of the codec compression on /ʃ/ can also be seen in the maximum values, where CoG was lowered with just around 200 Hz, whereas SD was increased with 26 Hz. Similarly, skewnes increased by 0.1. The biggest change in maximum value was the frequency peak, which lowered from 5,281 Hz in the WAV baseline to 4,875 Hz in the MP3 compression.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the MP3 compression is presented below in figure 3.30. The most notable change can be seen in the spectrogram and spectrum, where an increase in amplitude and intensity is visible in a band between 3 and 4 kHz, while the upper frequency limit is again clearly visible between 7 and 8 kHz.

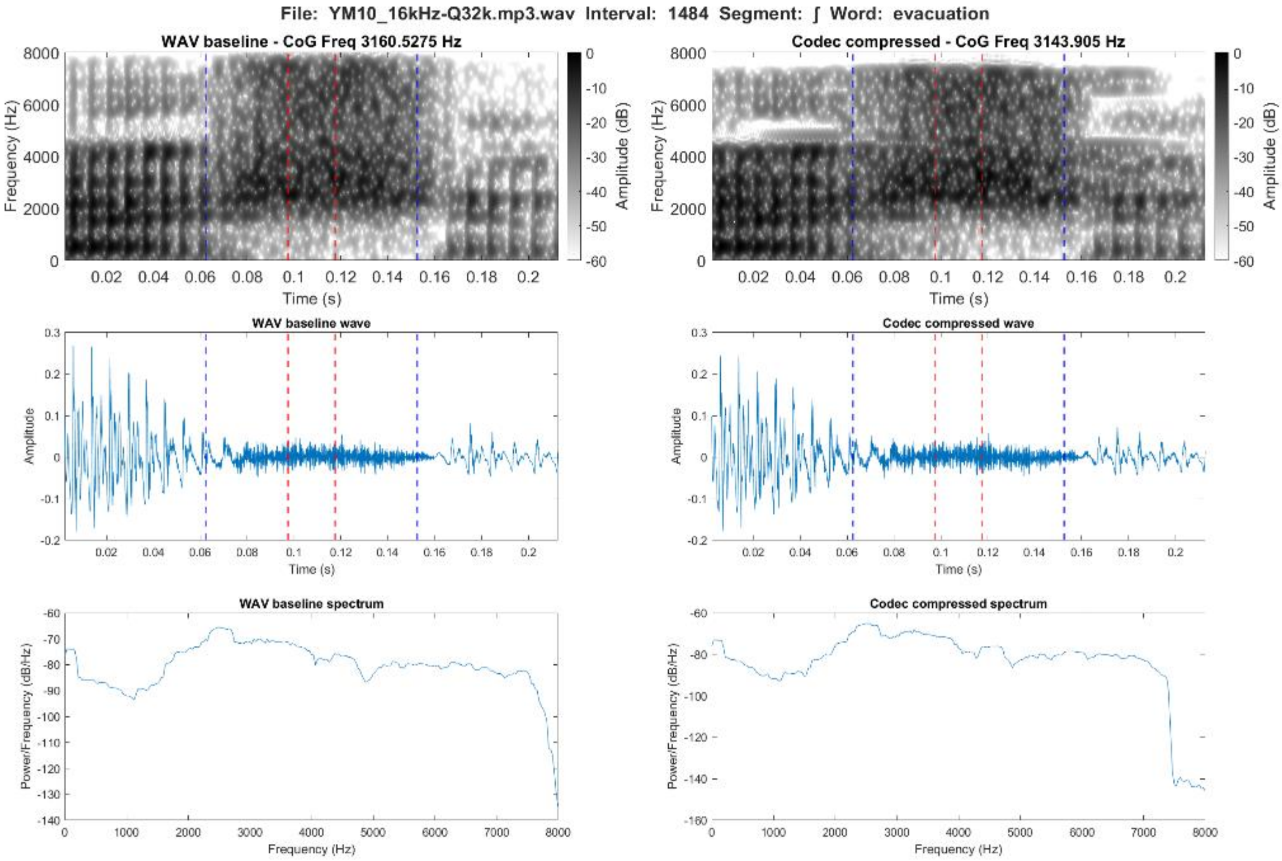


Figure 3.30. Spectrographic comparison of /ʃ/ in the word *evacuation* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.6 /z/

For /z/ all measures were again lowered by the codec compression, which is also clear from the linear predictions, where all spectral measures show a similar downwards trajectory. However, the distributional effects of the MP3 compression were limited. CoG and frequency peak both show a lowering of the values in the higher part of the spectrum. The distribution of SD becomes more smooth and less distinct in the MP3 compression, but are otherwise similar to the WAV baseline. Skewness is like with the previous segments, the least affected measure and adhere to similar tendencies with a slight lowering and more values centred around the mean following codec compression.

This limited effect is also evident from the mean values, where the only measure with a change more than 100 Hz was CoG with a lowering of 166 Hz. Despite this, the mixed effects modelling returned significant p-values for all measures (p <.0001) and showed clear downwards trajectories for all measures. The change in mean value for SD was at 34 Hz, and just above 90 Hz for spectral peak. Skewness lowered by 0.31 and kurtosis by 0.10. All p-values and further statistical results can be found in table 3.10 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - MP3	/z/	166.10	9.02	<.0001
SD	WAV - MP3	/z/	34.19	4.06	<.0001
Skewness	WAV – MP3	/z/	0.30	14.83	<.0001

Table 3.18. Statistical results of the difference between WAV and MP3 based on linear prediction models for /z/

In comparison to the mean values, the maximum values demonstrated greater changes caused by the MP3 compression. This is evident from a change in CoG from 4,862 Hz in the WAV baseline to 6,610 Hz in the MP3 compression, as well as a change in SD from 2,900 Hz to 2,692 Hz. The frequency peak was also lowered, here with 438 Hz, while the maximum value for skewness was lowered from 8.05 to 7.27.

A typical spectrographic representation of /z/ both as phonetically voiced and voiceless in the WAV baseline and the MP3 compression is presented below in figure 3.31 and 3.32. The voiced example illustrates how the frequency content above 7 kHz is clearly reduced, while a band of frequencies between 4 and 5 kHz appear intensified by the transmission.

For the voiceless /z/, a slight reduction in intensity can be observed based on figure 3.31 as well as an increase in non-speech frequency content not present in the original WAV files (see figure 41). In comparison to the voiced /z/ in figure 42, the frequencies around the 7 kHz cut-off is more gradually reduced and the cut-off appear closer to 8 kHz.

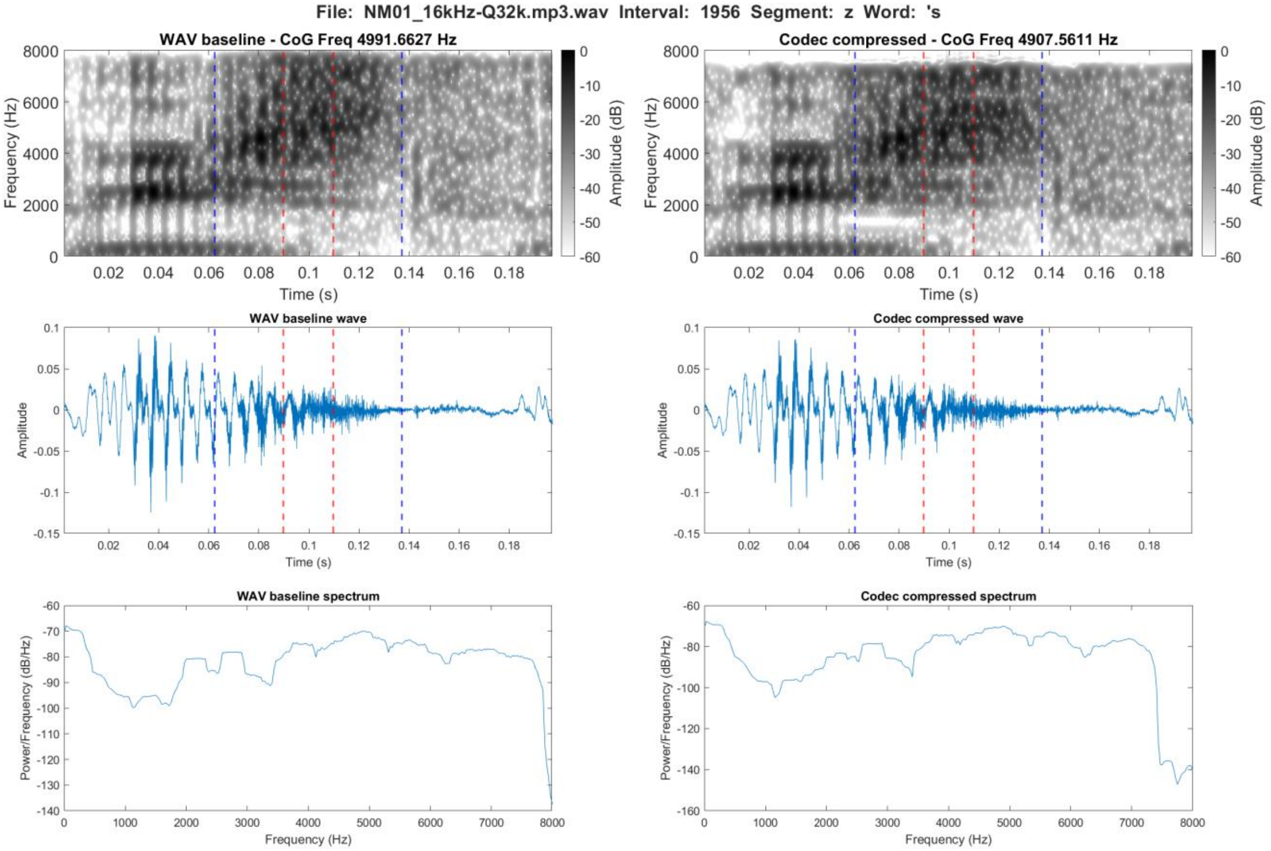


Figure 3.31. Spectrographic comparison of voiced /z/ in the word *animals* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

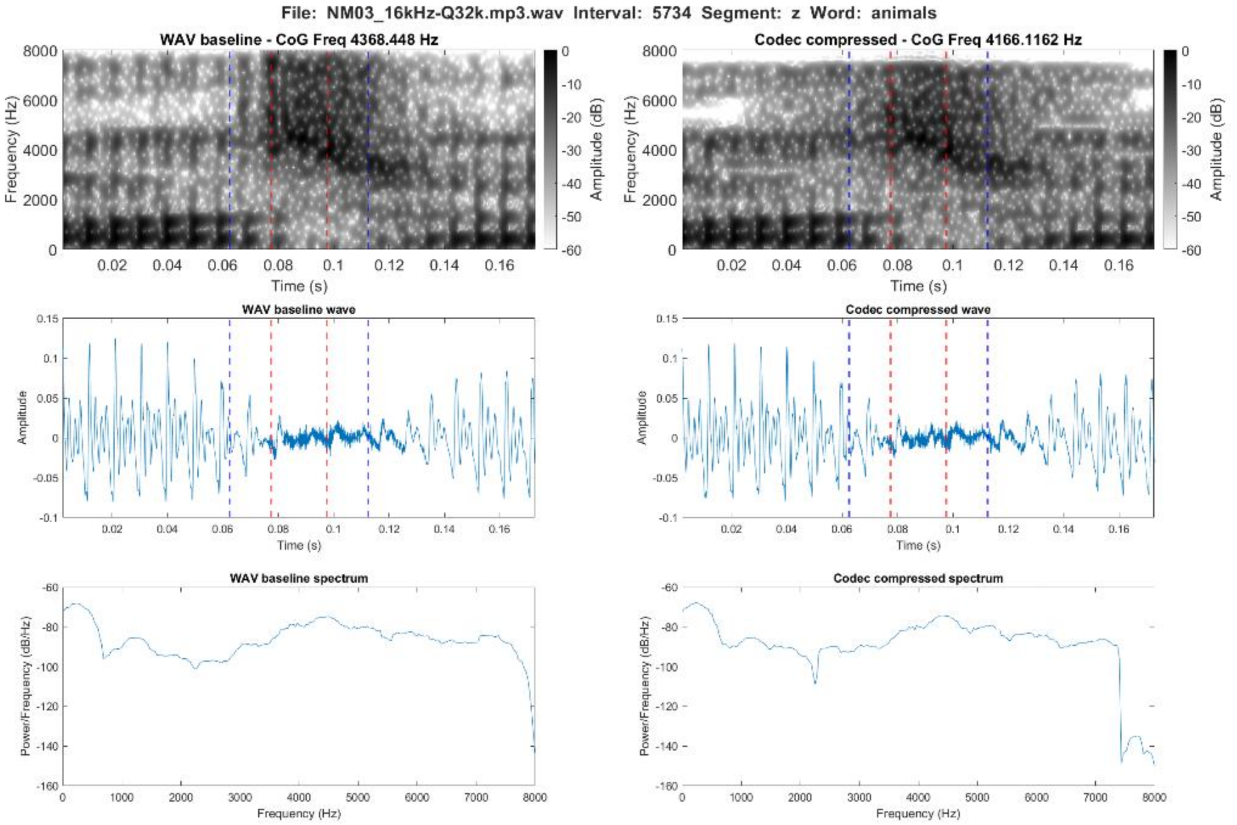


Figure 3.32. Spectrographic comparison of voiceless /z/ in the word *animals* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.2.7 /ð/

For /ð/ like with /z/ all measures were lowered, but less so than what was observed for some of the voiceless segments. The general downwards trend is confirmed by the mixed effects modelling, and the changes were again all significant even though they were all less than 100 Hz (p <=.001). From the distribution plots, /ð/ shows both similar and different patterns to the other fricatives, however, the tendencies across spectral measures are generally more varied as consequence of the MP3 compression. CoG and SD to different degrees show an increase of the lower values and a decrease in the higher values. This is particularly evident for SD. For skewness, the distribution plot again show more values centred around the mean as well as fewer outliers at the lower end of the values following the codec compression. The frequency peak values are already centred around the cut-off value in the WAV baseline and in the MP3 compression, the values appear slightly more compact and a little higher

Specifically, for CoG, this was a lowering in mean value of just above 70 Hz, for SD just below 82 Hz and for skewness 0.11. Kurtosis was lowered by 0.57 and spectral peak by just below 50 Hz. All p-values and further statistical results can be found in table 3.19 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - MP3	/ð/	70.84	3.14	0.002
SD	WAV - MP3	/ð/	81.97	7.96	<.0001
Skewness	WAV – MP3	/ð/	0.11	4.26	<.0001

Table 3.19. Statistical results of the difference between WAV and MP3 based on linear prediction models for /ð/

Despite the minor changes in mean value, the maximum values showed more substantial changes. This was particularly evident for frequency peak with a change from 7,938 Hz in the WAV baseline to 7,281 Hz following codec compression. CoG was also lowered from 6,963 Hz to 6,206 Hz. Similarly, SD demonstrated a lowering of maximum value from 2,907 Hz to 2,692 Hz, while skewness was lowered with 0.78.

A typical spectrographic representations of /ð/ in the WAV baseline and the MP3 compression are presented below in figure 3.33 and 3.34. These again illustrate /ð/ produced as a voiced fricative (figure 3.33), and as a plosive (figure 3.34).

From the spectrogram is it evident that the codec compression has generally reduced the formant structure apparent in the WAV baseline, while two clear drops in amplitude are visible 4 and 5 kHz.

For the second example, frequency content is again inserted by the MP3 compression, but here it is in the initial part of the segment. In addition, it is clear than the compression has intensified the energy of the burst. The upper frequency limit here appear to vary slightly across the segment.

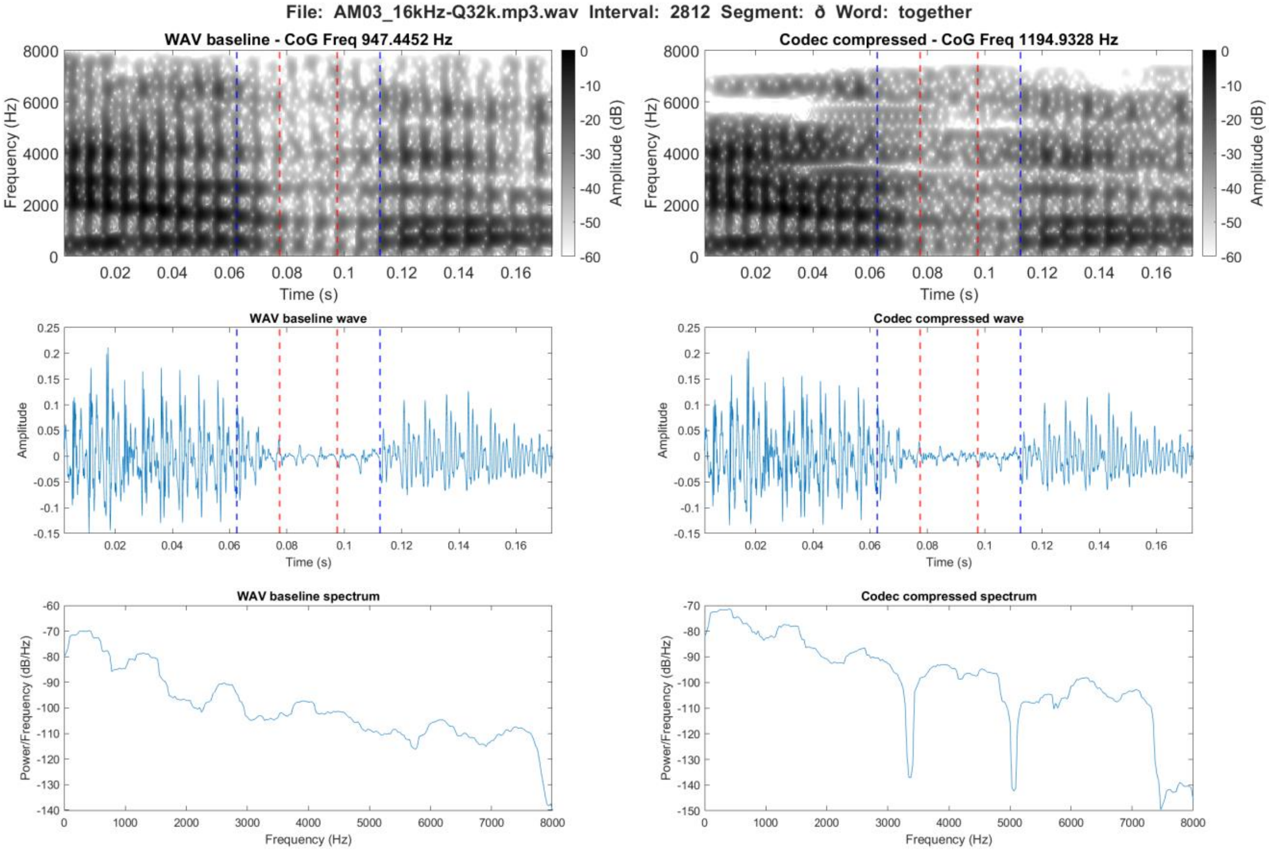


Figure 3.33. Spectrographic comparison of voiced /ð/ in the word *the* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

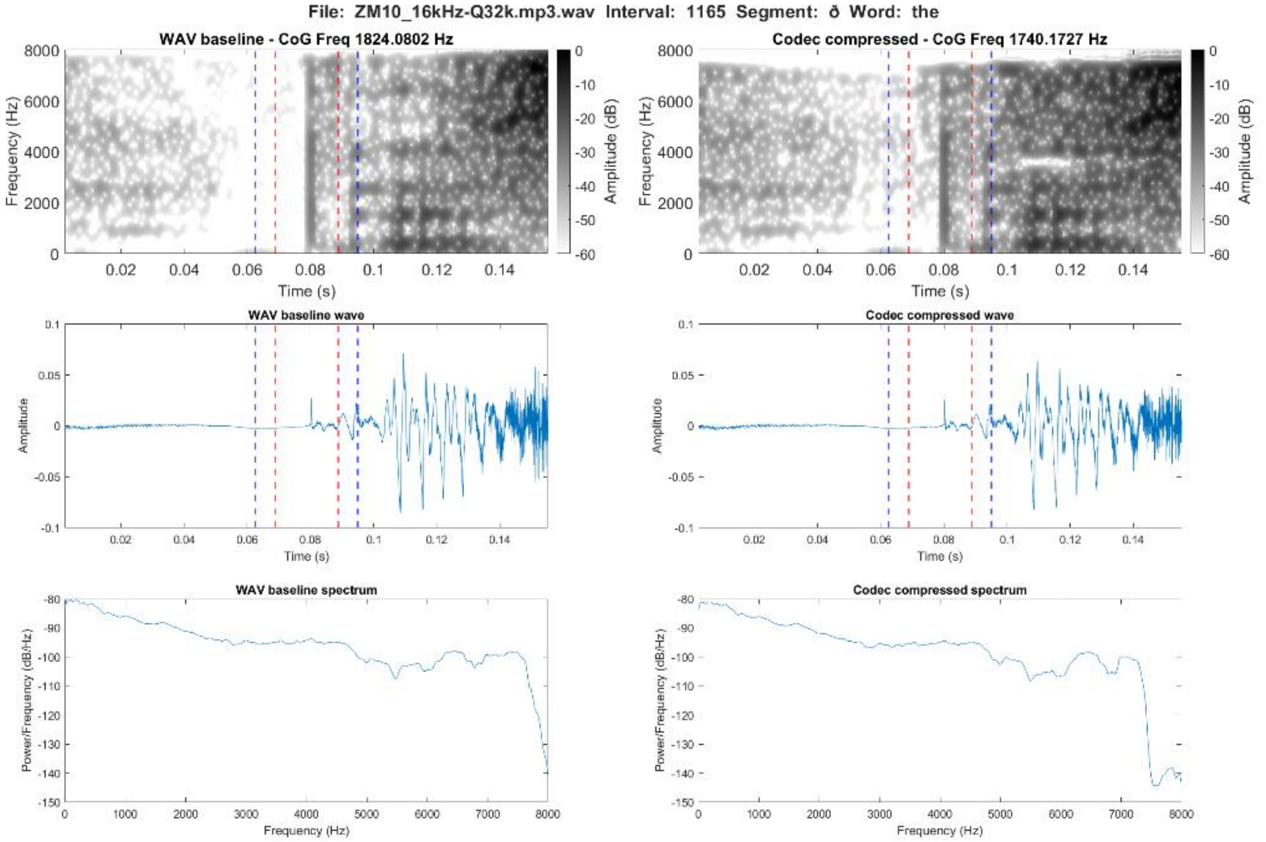


Figure 3.34. Spectrographic comparison of /ð/ in the word *the* in the WAV baseline (left) and the MP3 compression at 32 kbps (right)

3.4.3 Opus

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the Opus codec. As with the previous codecs, the linear predictions for each spectral measure and the individual segments can be found below (figure 3.35 to 3.40). These indicate the trajectory of the changes inferred by the Opus codec. The graphs present the results divided into voiced and voiceless segments and the more specific analysis of the plots can be found in the following sections on the individual segments.

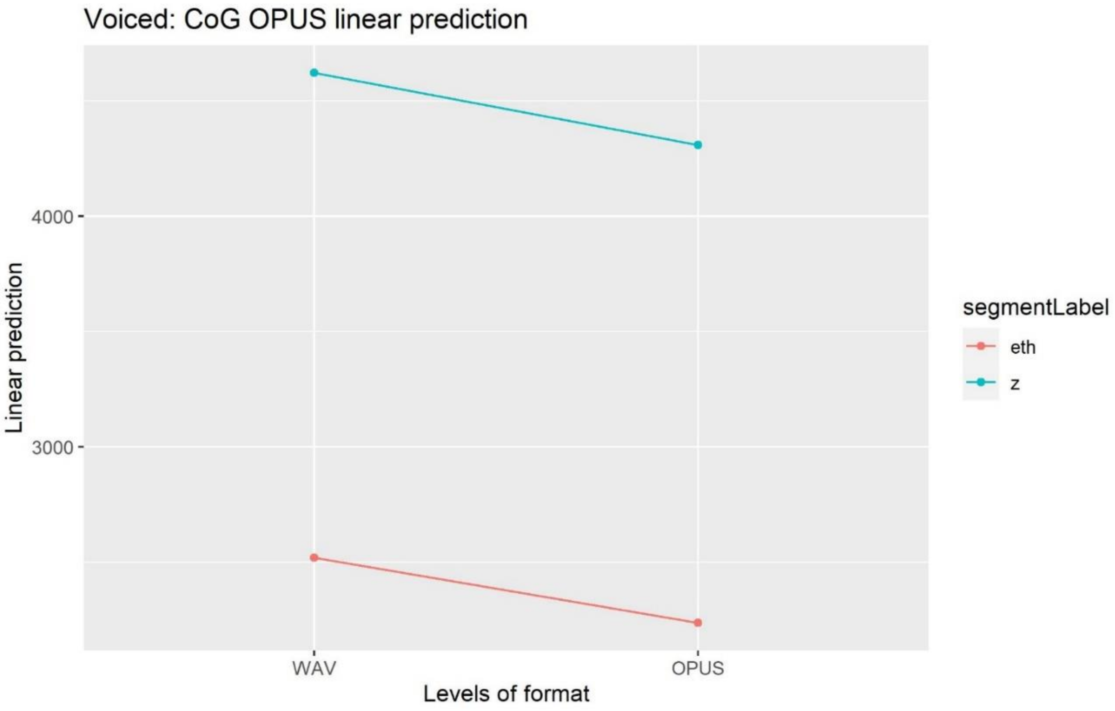


Figure 3.35. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for CoG and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

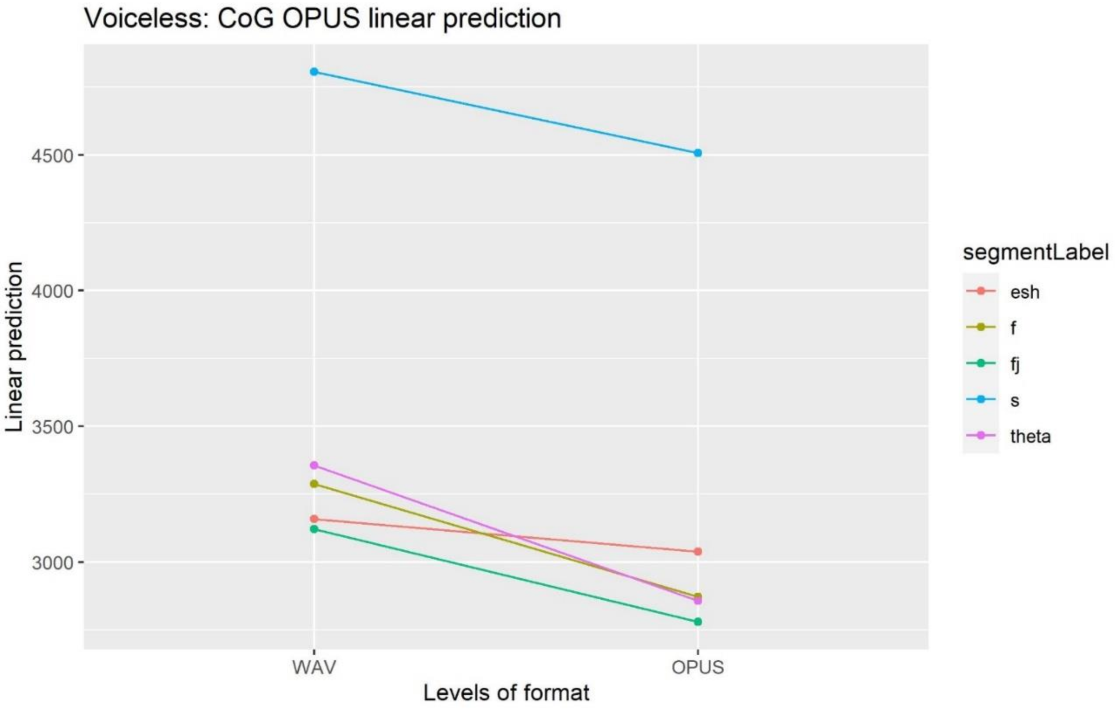


Figure 3.36. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for CoG and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

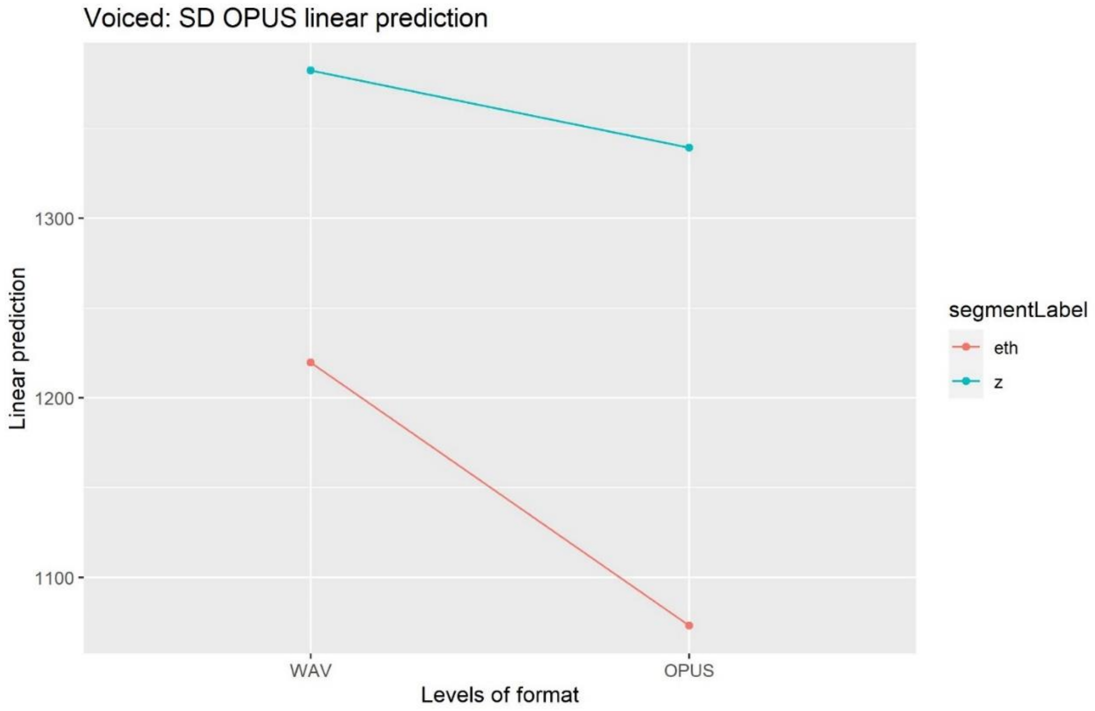


Figure 3.37. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for SD and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

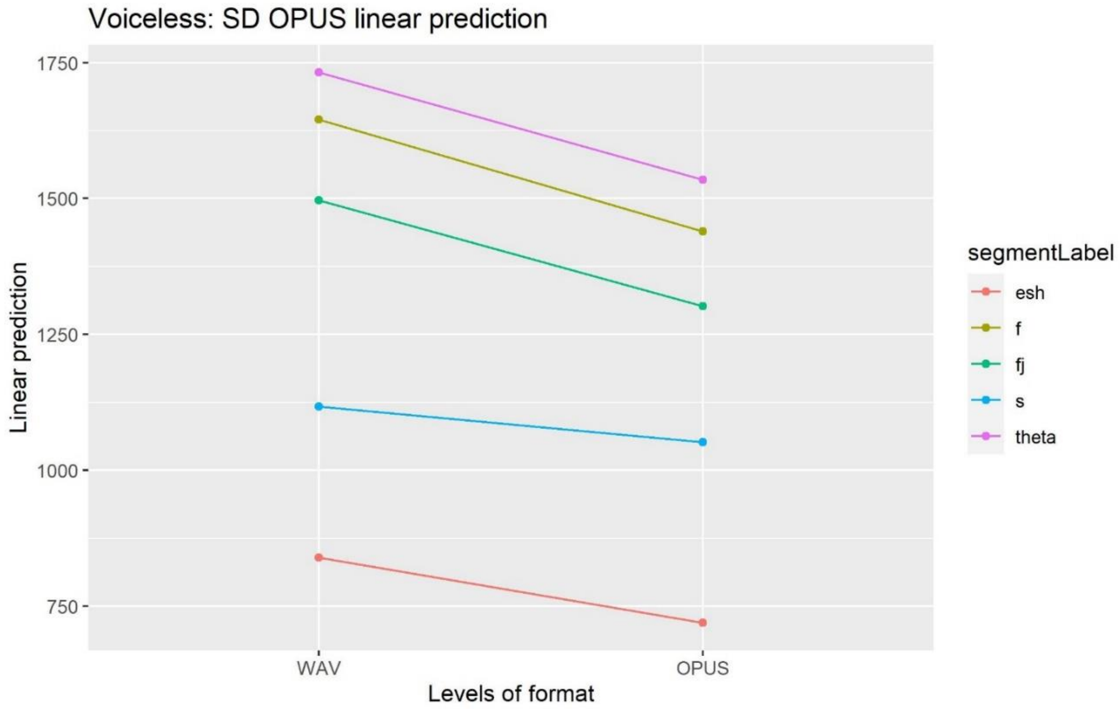


Figure 3.38. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for SD and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

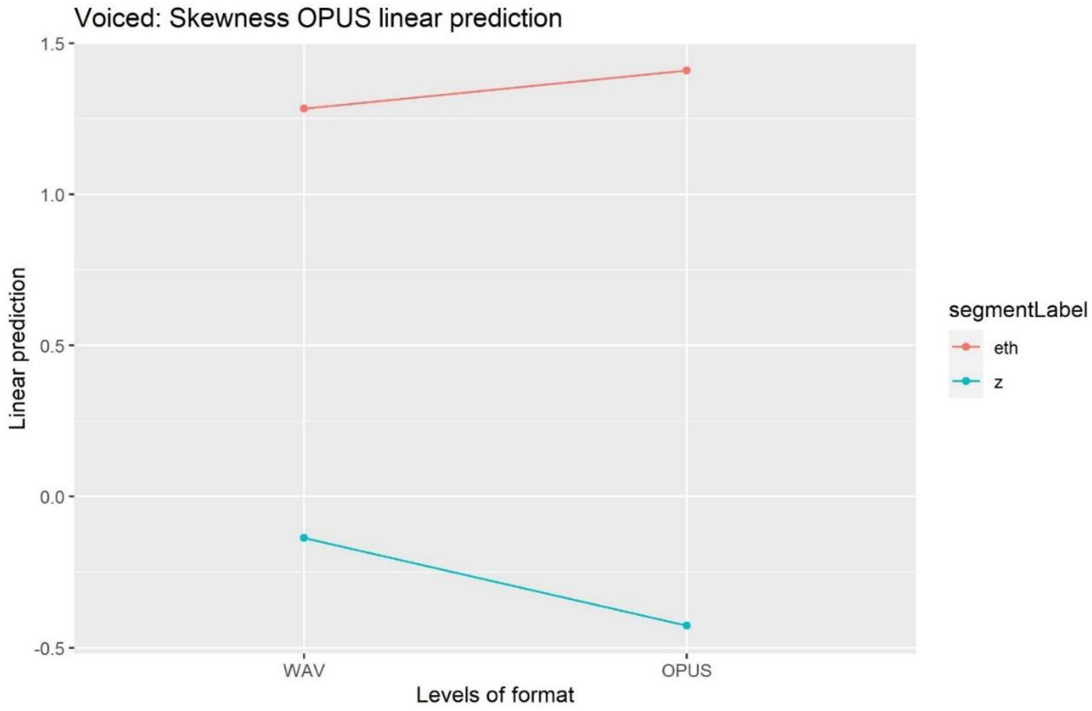


Figure 3.39. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for Skewness and individual voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

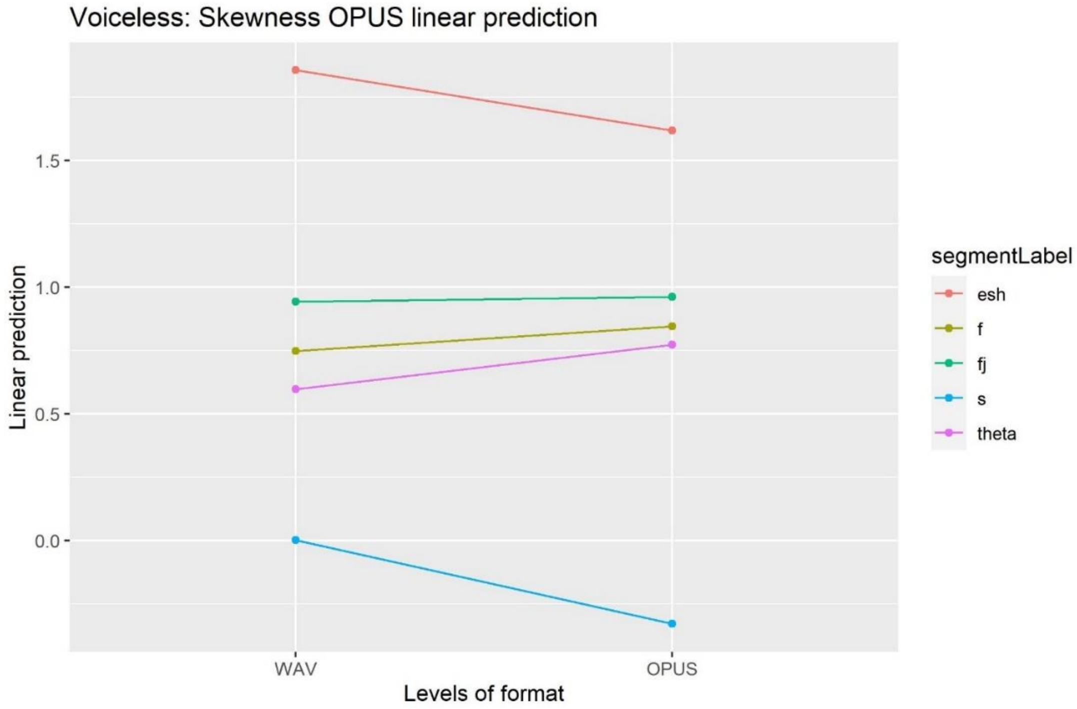


Figure 3.40. Trajectory of the linear predictions in the comparison of WAV and Opus from the mixed effects models for Skewness and individual voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The general distributions are illustrated below in a set of violin plots for each spectral measure. The specific analysis pertaining to each segment can be found in the relevant sections below.

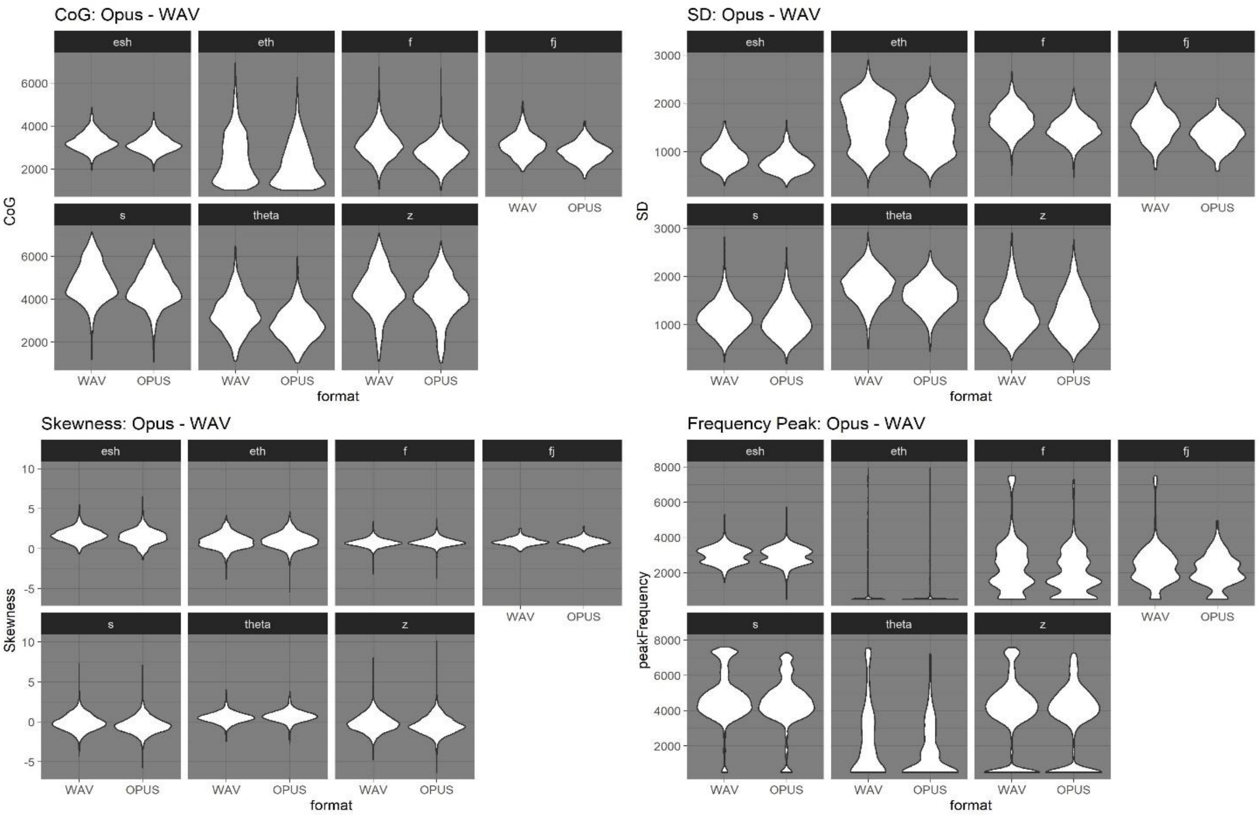


Figure 3.41. Distribution of spectral measure values in WAV baseline and the Opus codec compression grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

Apart from skewness, the general pattern is lowered spectral values following the Opus compression for all segments. For skewness, /f/, /θ/, and /ð/ group together and increase to varying degrees.

3.4.3.1 /f/

For /f/ the distributional plots reveal a general lowering and an effect across the spectrum for CoG and SD. For these two measures, the values again as with previous codecs tend to centre more around the mean. The same tendency can be seen for skewness, which also shows values ranging wider following the Opus compression. The distribution of the frequency peak reveal a limitation of higher frequency content and an increase in number of values around the lower cut-off.

Specifically, the mean values of CoG and SD were significantly lowered by 415 Hz and by just below 206 Hz respectively (p<.0001). Skewness and kurtosis were on the other hand increased by the Opus compression. For skewness, this was with 0.10 (p <.0001). For kurtosis the increase was 0.35. The mean value of frequency peak was similar to CoG lowered. This was with 477 Hz. All of which is confirmed by the linear predictions from the mixed effects models. In addition, the linear prediction shows that for /f/ CoG becomes almost identical to /θ/ in the codec compression. All p-values and further statistical results can be found in table 3.20 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - Opus	/f/	415.16	28.61	<.0001
SD	WAV - Opus	/f/	206.00	30.42	<.0001
Skewness	WAV - Opus	/f/	-0.10	-5.32	<.0001

Table 3.20. Statistical results of the difference between WAV and Opus based on linear prediction models for /f/

The maximum values generally confirm the tendencies observed in the distribution plots. This was by a lowering of CoG from 6,754 Hz to 6,667 Hz and a lowering of SD from 2,664 Hz to 2,333 Hz. The frequency peak max was lowered with 219 Hz from 7,500 Hz to 7,281 Hz. On the other hand, skewness showed an increase in maximum value from 3.4 to 3.7.

A typical spectrographic representation of /f/ in the WAV baseline and the Opus compression is presented below in figure 3.42. Especially based on the spectrogram and spectrum, a reduction in intensity is visible particularly in the frequencies above 4 kHz.

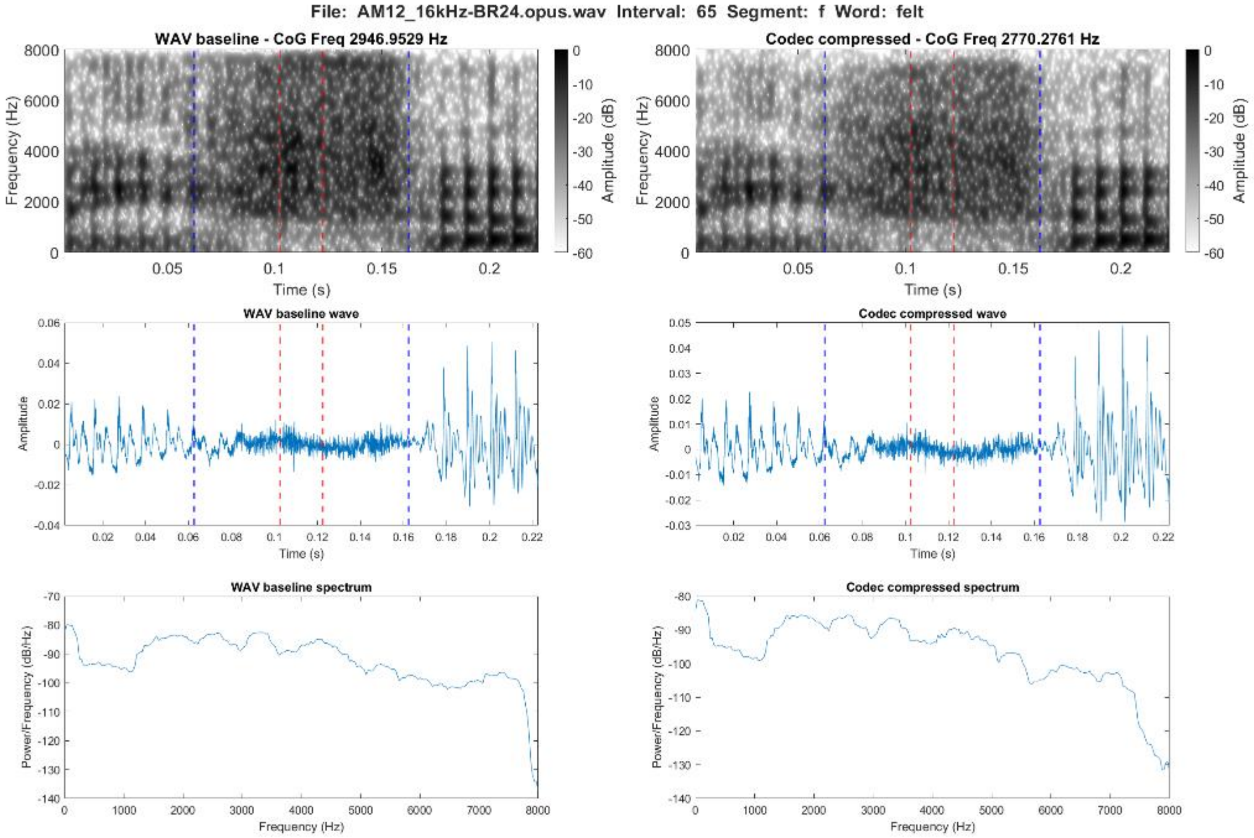


Figure 3.42. Spectrographic comparison of /f/ in the word *felt* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.2 /θ/

/θ/ shows similar tendencies to /f/ in terms of distribution. A general lowering can be seen, together with a greater number of values centred around the mean for CoG and SD. For skewness, a slight increase can be observed with no visible changes to the shape of the distribution. This is all confirmed by the linear predictions, where skewness is further noted to increase more than what is observed for /f/. In comparison, the frequency peak values are clearly lowered, which is evident from a reduction of values in the higher end of the spectrum and an increase around the lower cut-off value following the codec compression.

More specifically, the mean value for CoG was again significantly lowered (p <.0001) by a change in the mean value of just below 500 Hz. SD was significantly lowered by a change of 197 Hz (p <.0001). Skewness was increased following the Opus compression. This was with 0.18 and again significant (p <.0001). This increase combined with the changes seen to /f/ made these two sounds more similar in terms of skewness in the codec compression. The mean kurtosis value was increased with 0.38, while the frequency peak lowered by almost 600 Hz. All p-values and further statistical results can be found in table 3.21 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - Opus	/θ/	497.76	20.98	<.0001
SD	WAV - Opus	/θ/	197.33	17.83	<.0001
Skewness	WAV – Opus	/θ/	-0.18	-4.94	<.0001

Table 3.21. Statistical results of the difference between WAV and Opus based on linear prediction models for /θ/

The maximum values decreased following the Opus compression for all spectral measures. CoG was lowered from 6,491 Hz to 5,981 Hz, while SD was lowered from 2,922 Hz to 2,532 Hz. A slightly smaller change was seen for frequency peak, which changed from 7,531 Hz in the WAV baseline to 7,219 Hz in the Opus compressed files. Finally, skewness likewise presented a decrease from 4.02 to 3.79.

A typical spectrographic representation of /θ/ in the WAV baseline and the Opus compression is presented below in figure 3.43. Again, a slight reduction in intensity can be observed from the spectrogram and spectrum in the frequencies primarily above 4 kHz. The upper frequency limit is again clear between 7 and 8 kHz in the codec compression.

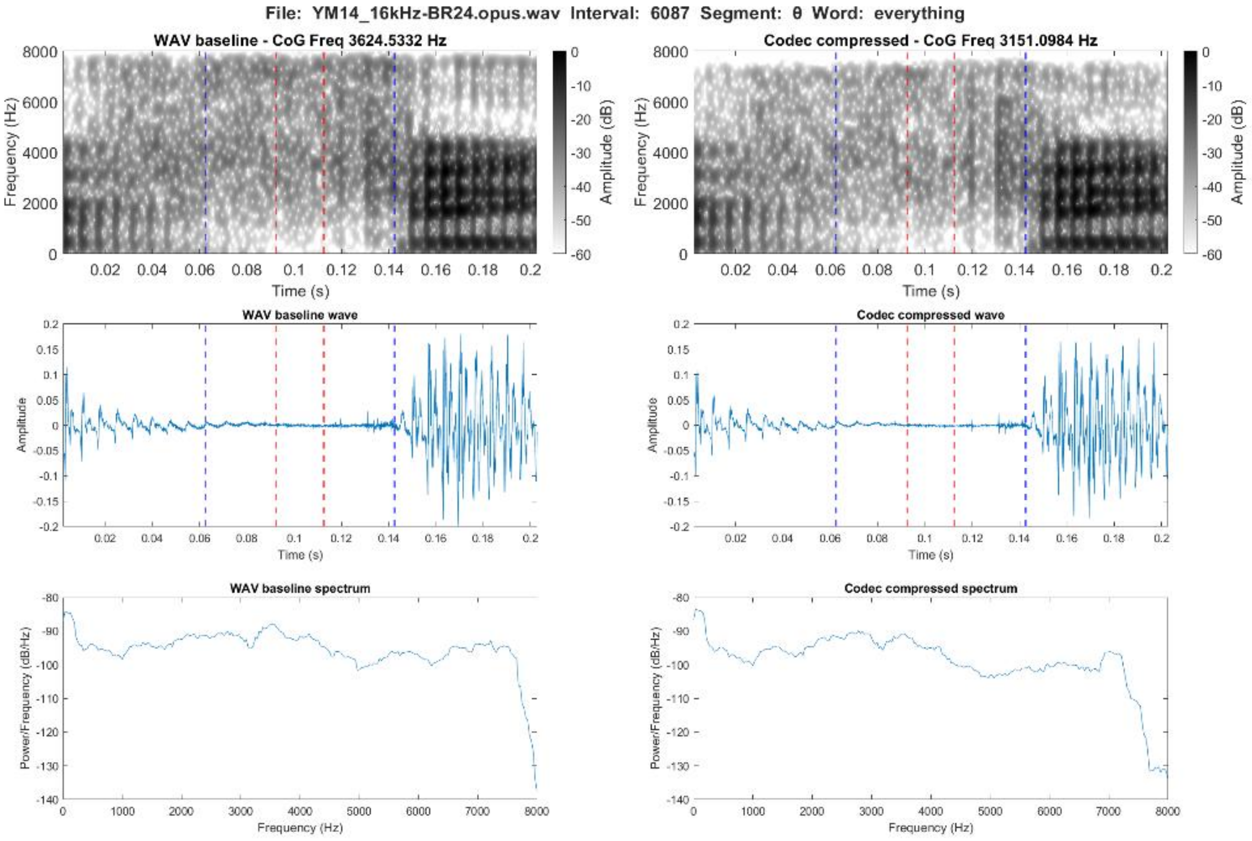


Figure 3.43. Spectrographic comparison of /θ/ in the word *everything* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.3 [fʲ]

[fʲ] follows the pattern of the previous voiceless segments in the Opus compression with a lowering of CoG, SD and frequency peak and an increase of skewness and kurtosis. All the violin plots apart from skewness illustrate this lowering and change in distribution from the baseline WAV caused by the Opus compression. This is especially for the higher frequency content, which is both limited and change shape in distribution. No change is visually observable for skewness in terms of distribution. The linear predictions follow the observations from the distribution plots, however, again with no clearly visible change to skewness.

For CoG the lowering of 242 Hz in mean value was significant (p <.0001) and so was the lowering by just under 195 Hz of SD (p <.0001). The 0.02 increase to the mean value of skewness did not reach significance. Kurtosis increased with 0.31, whereas the mean frequency peak is lowered with 280 Hz. The observed downwards trends as well as the almost unchanged skewness value was confirmed by the linear predictions from the mixed effects modelling. All p-values and further statistical results can be found in table 3.22 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - Opus	[fʲ]	342.28	6.28	<.0001
SD	WAV - Opus	[fʲ]	194.96	7.67	<.0001
Skewness	WAV – Opus	[fʲ]	-0.02	-0.28	0.78

Table 3.22. Statistical results of the difference between WAV and Opus based on linear prediction models for [fʲ]

From the maximum values, the changes observed in the distribution plots were substantiated. This meant a lowering of CoG, SD and frequency peak. The CoG maximum value changed from 5,150 Hz to 4,243 Hz, SD from 2,436 Hz to 2,077 Hz, and frequency peak from 7,469 Hz down to 4,938 Hz. In contrast, but confirming previous observations, the skewness maximum value increased with around 0.3 from 2.52 to 2.79.

A typical spectrographic representation of [fʲ] in the WAV baseline and the Opus compression is presented below in figure 3.44. Very limited changes apart from the reduction related to the upper frequency limit are observable.

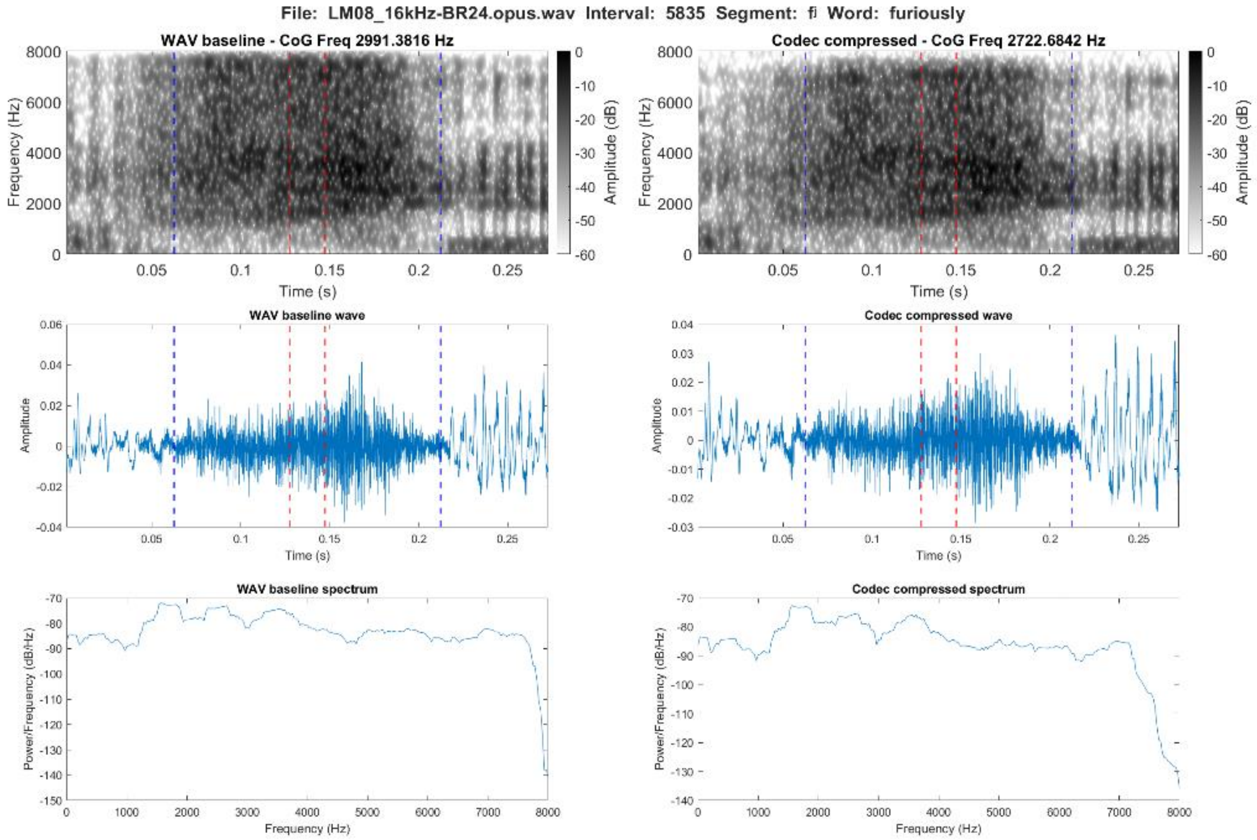


Figure 3.44. Spectrographic comparison of [fʲ] in the word *furiously* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.4 /s/

For /s/ all spectral measures apart from kurtosis were lowered following compression with Opus, which is also evident from the linear predictions. CoG follows the pattern of the previous segments, when looking at the distribution plots. This can be seen as a lowering of the values as well a more values centred around the mean. The same can be observed for skewness. Despite SD also presenting a lowering, the distribution changes so the values are more equally dispersed. For frequency peak, the effect of the Opus codec can mainly be observed as a change to the shape of the distribution of the values in the higher end of the spectrum.

For CoG and SD this were significant (p <.0001) as the mean value for CoG was lowered by 300 Hz, while SD was lowered by just 65 Hz. Skewness was also significantly lowered (p <.0001). This was a change in mean value of 0.33. Kurtosis was increased by 0.72, while the frequency peak was lowered by 256 Hz. These downwards trends follow the linear predictions from the mixed effects modelling. All p-values and further statistical results can be found in table 3.23 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - Opus	/s/	300.09	30.42	<.0001
SD	WAV - Opus	/s/	65.31	14.19	<.0001
Skewness	WAV – Opus	/s/	0.33	26.94	<.0001

Table 3.23. Statistical results of the difference between WAV and Opus based on linear prediction models for /s/

The tendencies with lower values following the Opus compression reported above were also seen in the maximum values. The biggest changes were for CoG and frequency peak, with CoG changing from 7,137 Hz to 6,797 Hz and frequency peak from 7,594 Hz to 7,281 Hz. SD showed a smaller change in maximum value from 2,815 Hz to 2,610 Hz, while skewness lowered with 0.2 from 7.33 to 7.13.

A typical spectrographic representation of /s/ in the WAV baseline and the Opus compression is presented below in figure 3.45. Overall, apart from a slight reduction in the higher frequencies little to no effect can be observed for /s/.

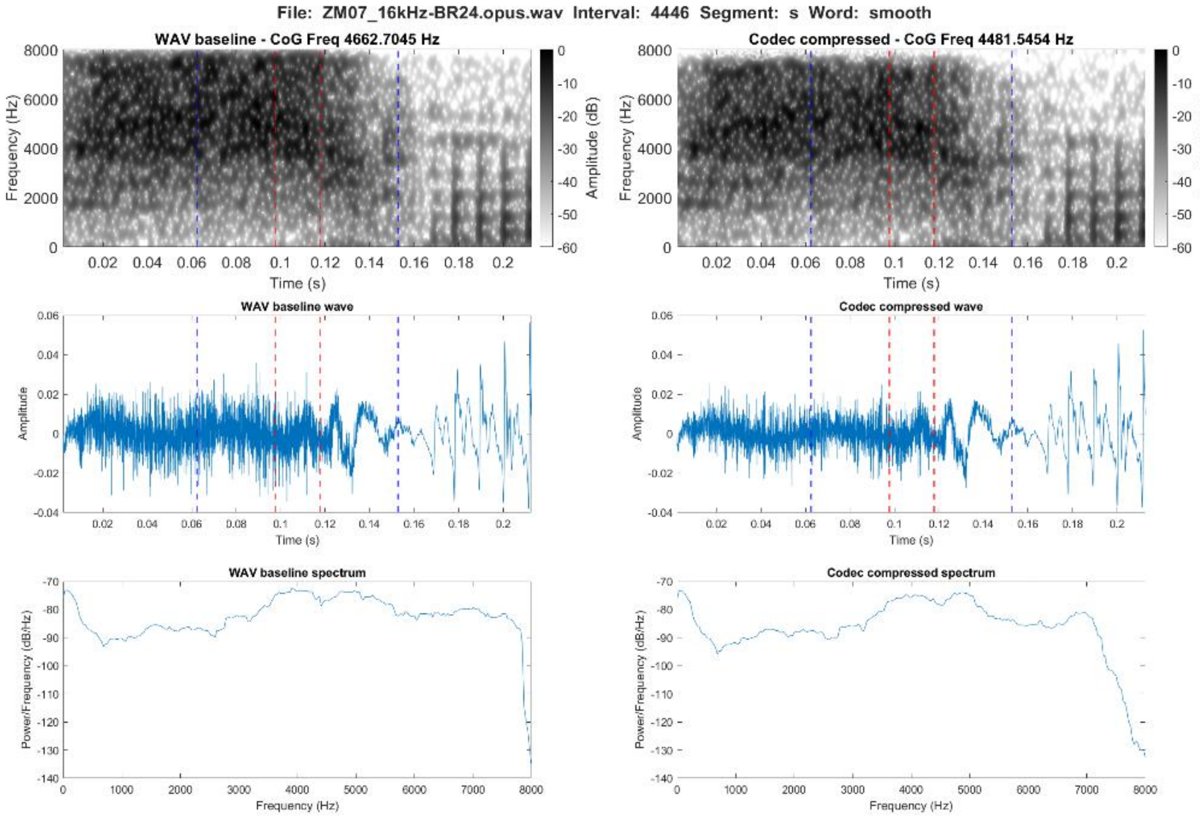


Figure 3.45. Spectrographic comparison of /s/ in the word *smooth* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.5 /ʃ/

For /ʃ/ all spectral measures apart from kurtosis show a downwards trend following codec compression, which is also seen in the distribution plots as well as the linear predictions. Despite this, CoG and skewness present very limited changes in the shape of the distribution, while a centring of values around the mean can again be observed for SD. The distribution of the frequency peak values is overall unaffected by the Opus compression based on the violin plots. However, it is evident that the values have become less centred and reaches further towards the extremes both in the lower and higher part of the spectrum.

In detail, this was a change in mean value of approximately 120 Hz for both CoG and SD (p <.0001). SD was also significantly lowered with a change of mean value by 0.24 (p<.0001). Kurtosis increased by 0.31, and the frequency peak lowered by just 34.14 Hz. All p-values and further statistical results can be found in table 3.24 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - Opus	/ʃ/	120.82	5.11	<.0001
SD	WAV - Opus	/ʃ/	119.81	10.86	<.0001
Skewness	WAV – Opus	/ʃ/	0.24	8.04	<.0001

Table 3.24. Statistical results of the difference between WAV and Opus based on linear prediction models for /ʃ/

The maximum values for /ʃ/ presented a more mixed pattern to the previously reported segments. The Opus compression decreased the maximum value for CoG from 4,862 Hz to 4,634 Hz, while it increased SD with 21 Hz to 1,649 Hz. Similarly, frequency peak increased from 5,281 Hz to 5,719 Hz and skewness from 5.28 to 6.53.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the Opus compression is presented below in figure 3.46. The spectrum show a clear drop in amplitude at 7 kHz, this can also be seen in the spectrogram and waveform, where the frequencies are reduced in intensity.

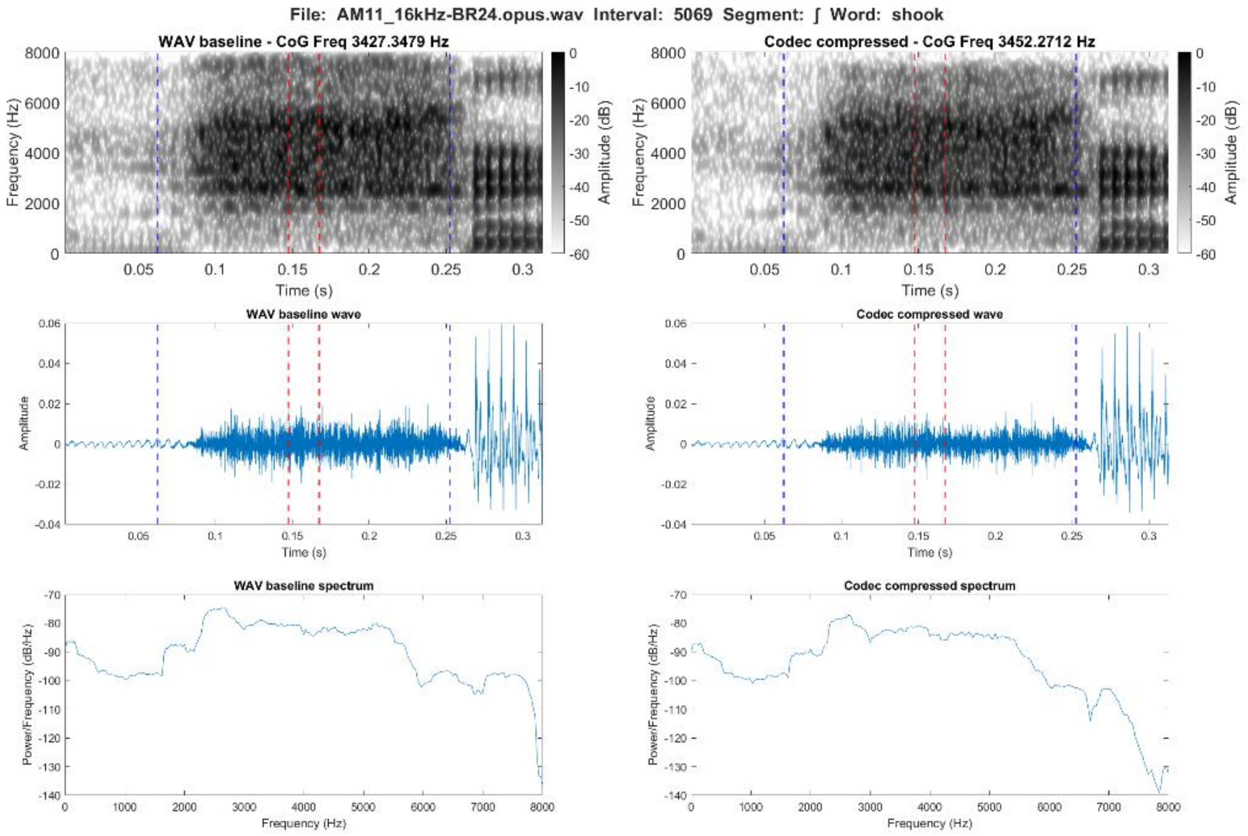


Figure 3.46. Spectrographic comparison of /ʃ/ in the word *shook* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.6 /z/

For /z/ all measures apart from kurtosis were again lowered in the codec compression, which can also be seen in the linear prediction plots, where all the measures present similar downwards trajectories. From the violin plots, the distributional effects of the Opus compression on /z/ is generally very similar to what is observed for /s/ for all spectral measures. This entails a lowering and centring of values for CoG, a change in distribution shape to be more even for SD, a lowering but no substantial change in shape for skewness, and a main effect on the higher values in the spectrum for frequency peak.

For CoG this was a significant change illustrated by a 313 Hz change in mean value (p <.0001). Similarly, the lowering of just below 41 Hz for SD also reached significance (p<.0001). Skewness was lowered by 0.30, which was significant (p <.0001). For kurtosis the Opus compression resulted in an increase by 0.37 in the mean value, while for frequency peak this was as a lowering by 287 Hz. These tendencies are confirmed by the trajectories observed in linear predictions from the mixed effects modelling. All p-values and further statistical results can be found in table 3.25 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - Opus	/z/	313.08	16.41	<.0001
SD	WAV - Opus	/z/	40.94	4.70	<.0001
Skewness	WAV – Opus	/z/	0.30	13.85	<.0001

Table 3.25. Statistical results of the difference between WAV and Opus based on linear prediction models for /z/

Apart from skewness, all maximum values were lowered for /z/ following the codec compression. This was a change in CoG from 7,082 Hz to 6,736 Hz, a change for SD from 2,901 Hz to 2,769 Hz, and for frequency peak from 7,563 Hz to 7,250 Hz. It is worth noting that as with the mean values, these values were in the high end of what was theoretically expected for /z/. In comparison, skewness showed a relatively substantial increase with over 2.0 from 8.05 to 10.11.

Two typical spectrographic representations of /z/ in the WAV baseline and the Opus compression are presented below in figure 3.47 and 3.48.

In the voiced example in figure 3.47 from both spectrogram and spectrum, a general reduction in intensity is visible across the frequency range, while the shape of the waveform again remain almost identical between the two conditions. For the voiceless example in figure 4.48, only a slight reduction in intensity across the frequency range can be observed together with the upper-frequency limit just under 8 kHz.

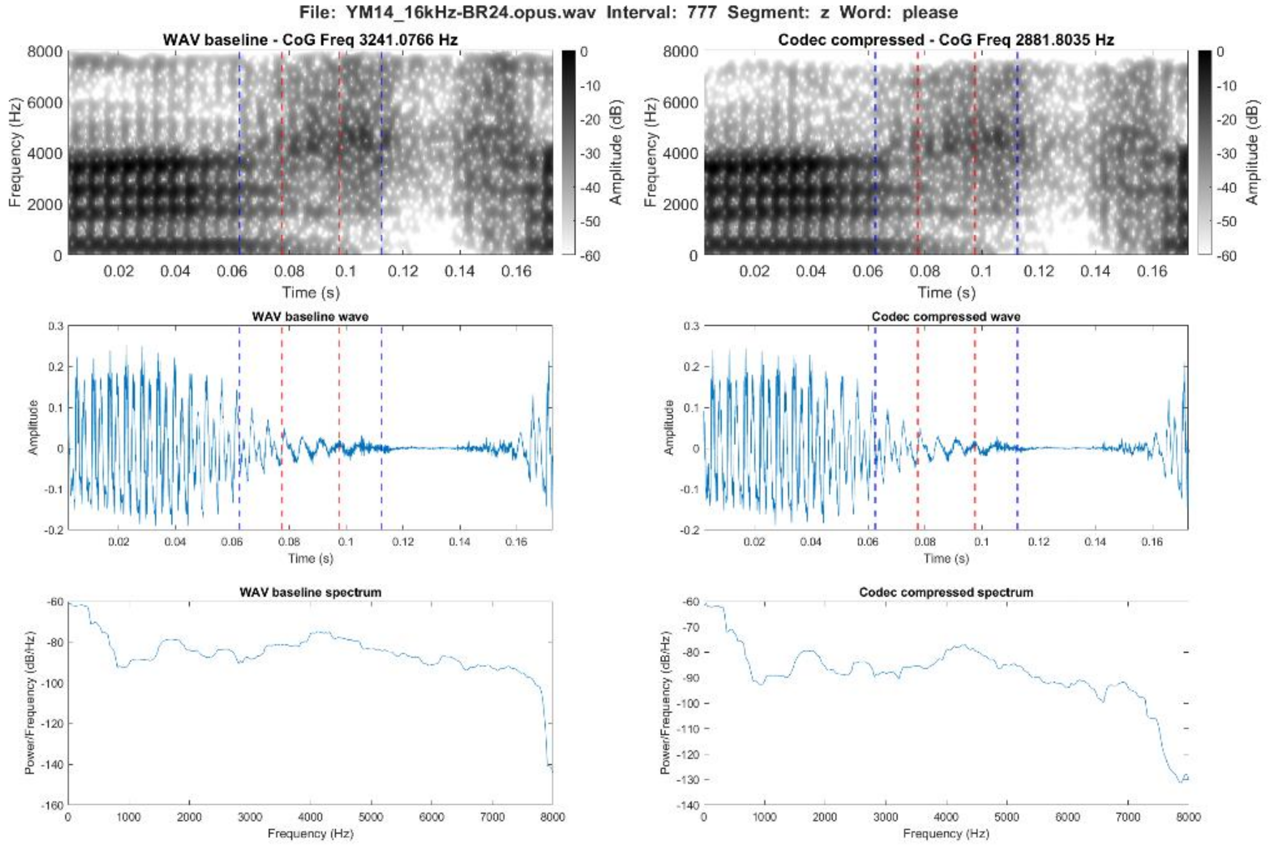


Figure 3.47. Spectrographic comparison of /z/ in the word *please* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

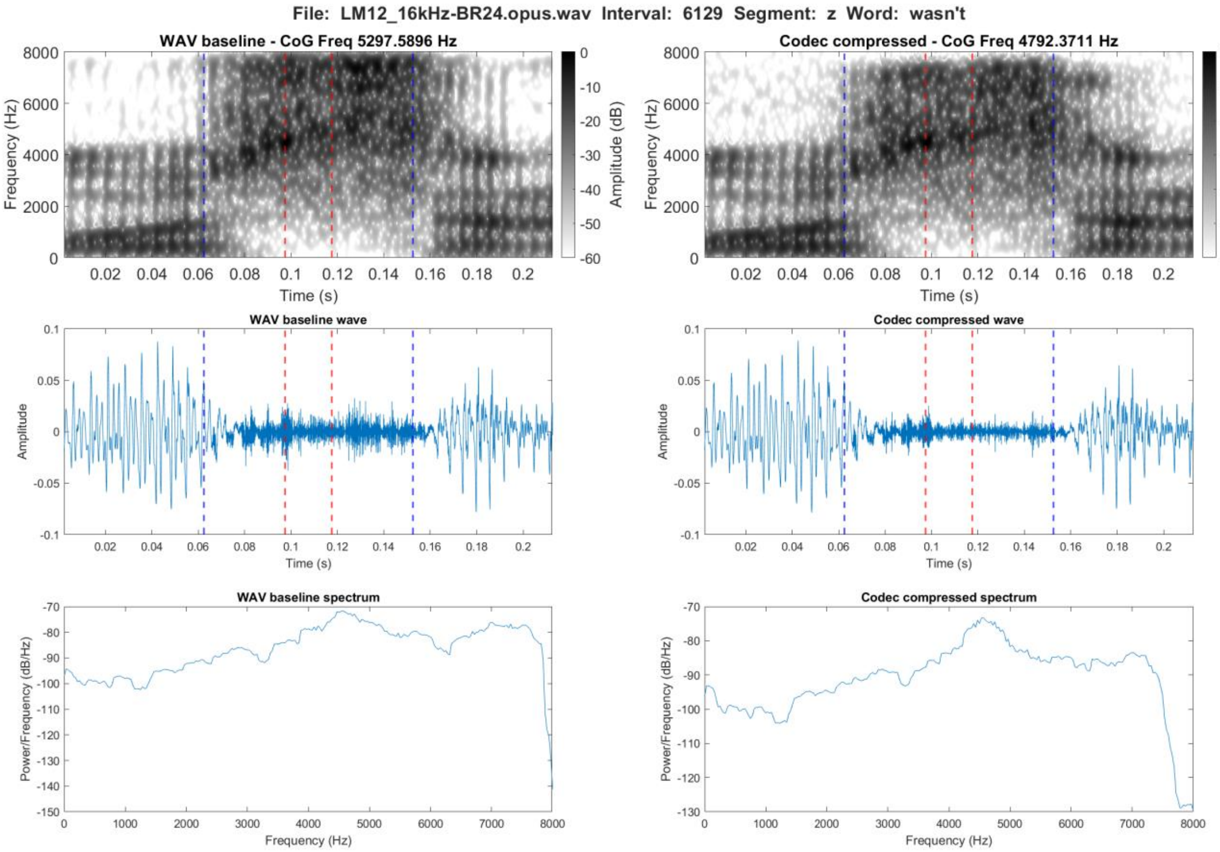


Figure 3.48. Spectrographic comparison of voiceless /z/ in the word *wasn't* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.3.7 /ð/

For /ð/ the codec compression similarly to /ð/ caused a lowering of mean values of CoG, SD and frequency peak. This lowering can also be observed in the violin plots, with a more even distribution of values visible for SD. For skewness, a very slight increase is evident from the violin plot. This is, however, with more sporadic values towards the lower frequency cut-off at 1 kHz. These observations are again confirmed by the linear prediction plots, where it can also be seen that the trajectory of CoG for /ð/ is similar to what can be observed for /z/.

For both CoG and SD this was significant with a change in mean values of 283 Hz and 145 Hz respectively (p<.0001). For skewness and kurtosis the mean values increased. This was significant for skewness with a change of 0.13 (p <.0001). The increase in kurtosis was 0.33, while the mean frequency peak was lowered with 249 Hz. It should be noted here that the linear prediction from the mixed effects models of skewness indicate a clear downwards trend, which does not align with the observed increase in mean value. All p-values and further statistical results can be found in table 3.26 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - Opus	/ð/	283.04	12.13	<.0001
SD	WAV - Opus	/ð/	145.03	13.63	<.0001
Skewness	WAV – Opus	/ð/	-0.13	-7.79	<.0001

Table 3.26. Statistical results of the difference between WAV and Opus based on linear prediction models for /ð/

The maximum values for /ð/ were lowered for CoG and SD. This was from 6,963 Hz to 6,293 Hz for CoG, and from 2,907 Hz to 2,764 Hz for SD. In comparison, skewness followed previously reported tendencies and showed an increase in maximum value from 4.13 to 4.58. Frequency peak was the only maximum value following any codec compression, which remained unchanged at 7,938 Hz in both WAV baseline and Opus compression.

Two typical spectrographic representation of /ð/ in the WAV baseline and the Opus compression are presented below in figure 3.49 and 3.50. In figure 3.49 illustrating the voiced production, the only notable effect of the codec compression, is a change in upper frequency limit between 7 and 8 kHz. For the plosive in figure 3.50, a reduction in intensity of the burst is visible in the spectrogram following codec compression. In addition, a clear drop in amplitude and reduction in frequencies are visible from above just under 7 kHz.

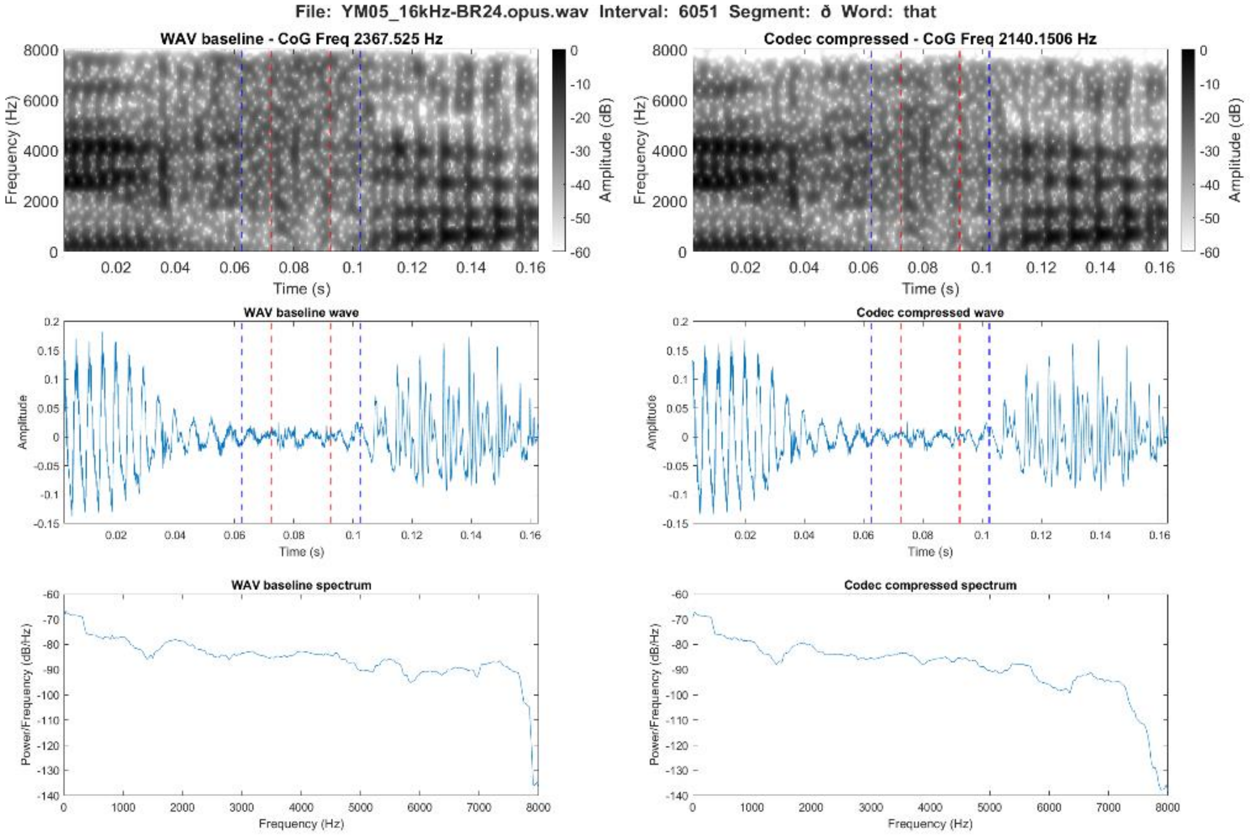


Figure 3.49. Spectrographic comparison of voiced /ð/ in the word *that* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

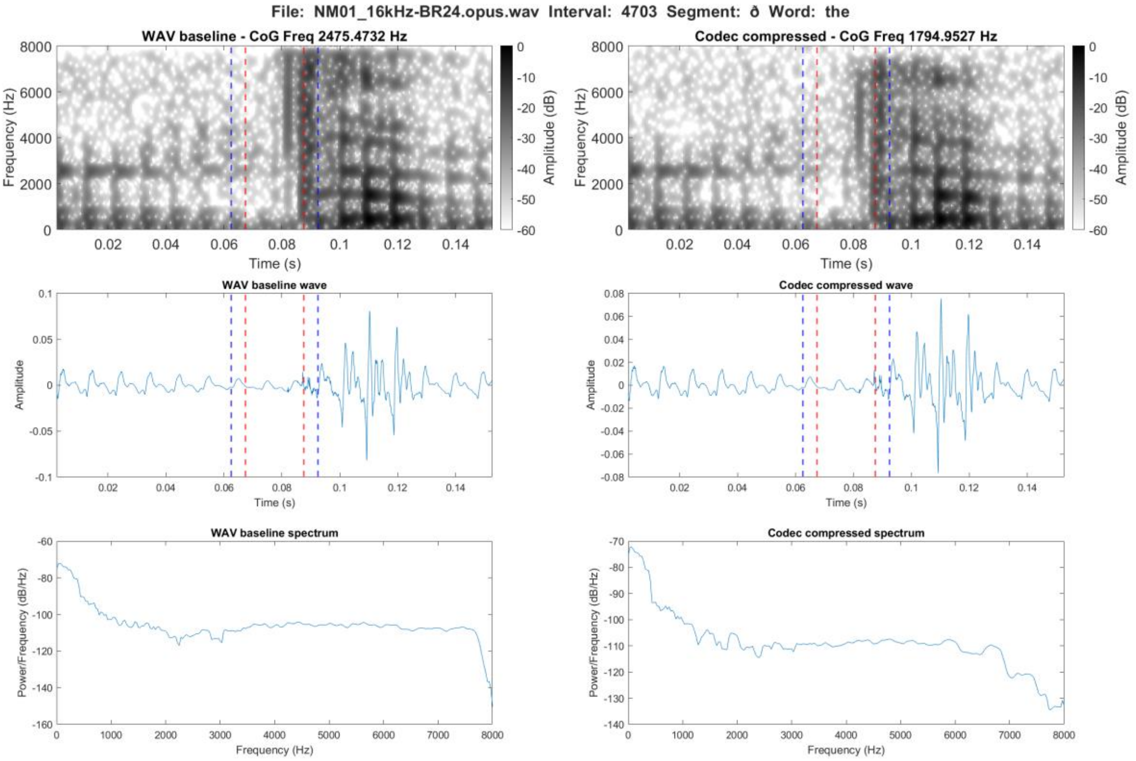


Figure 3.50. Spectrographic comparison of plosive /ð/ in the word *the* in the WAV baseline (left) and the Opus compression at 24 kbps (right)

3.4.4 1 kHz tokens

As mentioned earlier, the 1 kHz tokens are the ones, where one or more tokens in either codec compression or WAV baseline have a CoG below 1 kHz. These are of interest because any movement from above 1 kHz to below 1 kHz due to any of the three codec-compressions potentially suggest substantial effects to the original WAV baseline. All the 1 kHz tokens were paired with their counterparts in the other codecs and baseline (see section 3.3.5 for further details on this).

In total, the number of 1 kHz tokens including all 1 kHz tokens i.e. all groups and those removed not to skew the main dataset, amounted to 6,136 tokens or 7.7 percent of the entire dataset. This equates to 1,534 tokens in each condition. The tokens, which were removed not to skew the main dataset, were excluded from any further analysis.

The exact number of 1 kHz tokens included in the further analysis is given below in table 3.27 for each codec and segment individually. The total number of pairs was 3,590 or 4.5 percent of the entire dataset. None of these pairs were of /ʃ/ or [fʲ].

Segment	**WAV-AMR-WB**	**WAV-MP3**	**WAV-Opus**
	12.65 kbps	**32 kbps**	**24 kbps**
	Group	Group	Group
	A, B & C	A, B & C	A, B & C
/f/	13	9	14
/θ/	12	6	11
/s/	5	3	5
/z/	68	35	49
/ð/	1,191	1,020	1,149

Table 3.27. Number of rejected pairs in groups A, B, and C for each segment in each of the codec compression

It should be noted here that a large part of the pairs of /ð/ are likely belonging to group A (see Table 3.6) due its naturally low CoG.

Due to the limited number of tokens in this dataset mixed effects modelling was not used for the analysis. Hence, the analysis of all the 1 kHz tokens is based on descriptive statistics. The following will focus on the mean values of the 1 kHz tokens as these include all three groups, which means it will be the sum of pairs with all tokens below 1 kHz, pairs with movement from above to below 1 kHz and vice versa.

Firstly, the trajectory of the mean CoG for all codecs and segments revealed that only for the AMR-WB do the mean values move from above to below 1 kHz, and this is only for /θ/ and /z/. The remaining tokens and codecs are all below the 1 kHz limit in both conditions. However all had a downwards trajectory apart from in the MP3 compression, where all but /f/ increased (see figure 3.51 to 3.53)

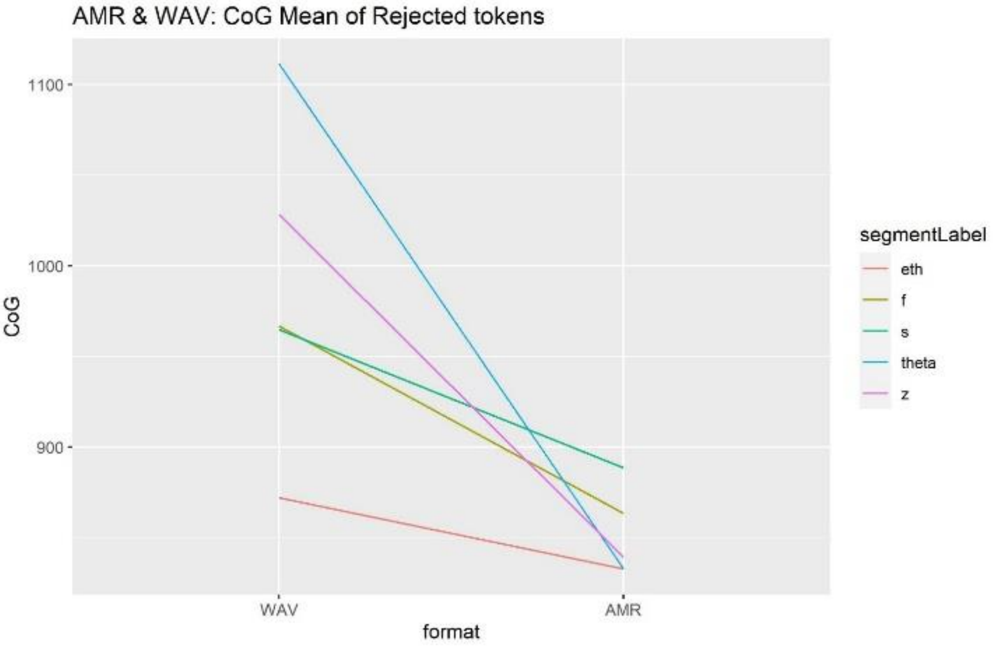


Figure 3.51. Trajectory of mean CoG values in 1 kHz tokens from the baseline WAV to AMR-WB incl. group A, B and C.
The symbols are interpreted as follows *Theta* = /θ/*, esh* = /ʃ/, *fj* = [f̍], and *esh* = /ʃ/. *Rejected tokens* = 1 kHz tokens.

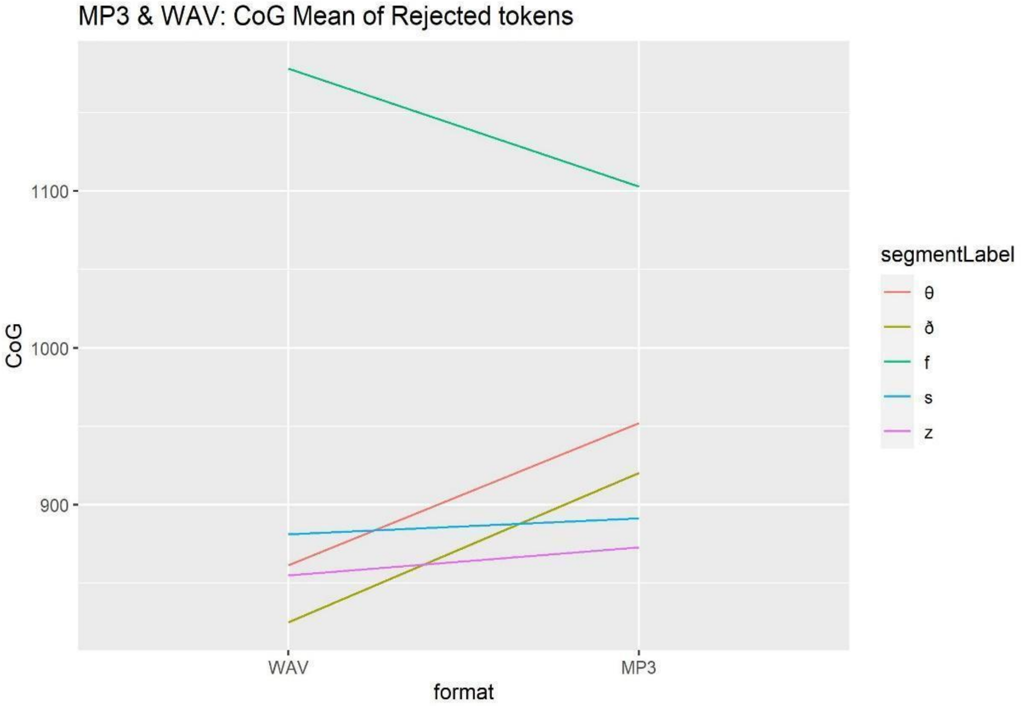


Figure 3.52. Trajectory of mean CoG values in 1 kHz tokens from the baseline WAV to MP3 incl. group A, B and C.
The symbols are interpreted as follows *Theta* = /θ/*, esh* = /ʃ/, *fj* = [f̍], and *esh* = /ʃ/. *Rejected tokens* = 1 kHz tokens.

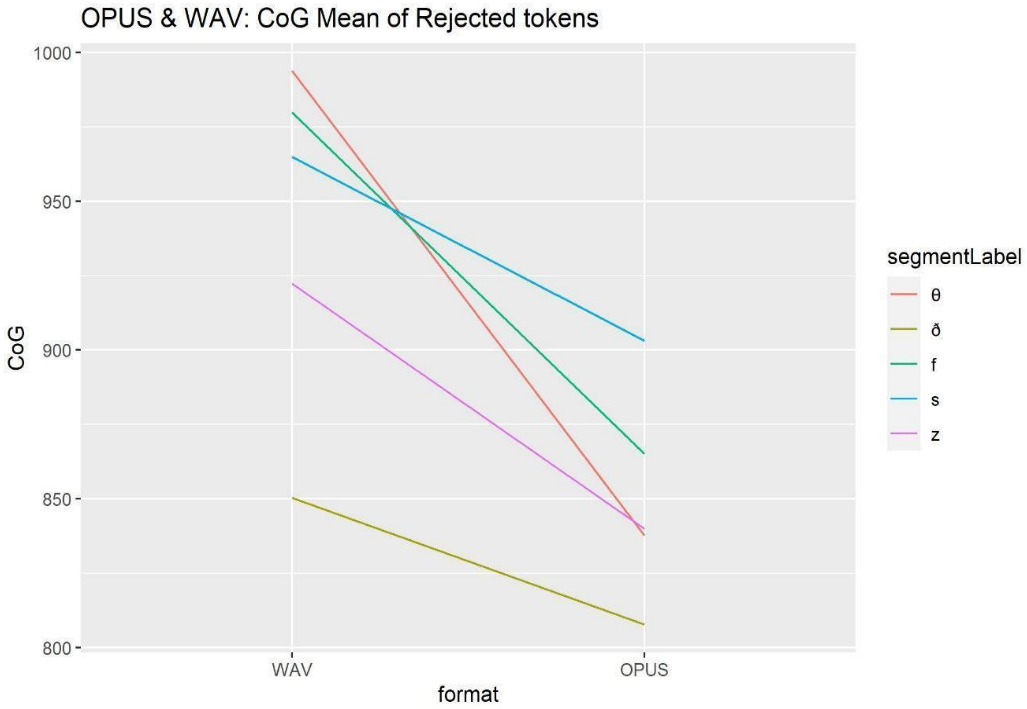


Figure 3.53. Trajectory of mean CoG values in 1 kHz tokens from the baseline WAV to Opus incl. group A, B and C. The symbols are interpreted as follows *Theta* = /θ/*, esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/. *Rejected tokens* = 1 kHz tokens.

The mean values for all the spectral measures including group A, B and C show how all segments in one or more codec compressions both decrease and increase for each spectral measure. Frequency peak overall lowers or stays unchanged as the increase found for /z/ is less than 5 Hz. Moreover, CoG and SD are generally lowered, while for skewness and kurtosis, the main tendency is for an increase following codec compression. All mean values for each codec compression are presented in table 3.28.

Seg	Codec	Bitrate	CoG	SD	Skew	Kurt	Freq. Peak
		(kbps)	(Hz)	(Hz)	(Hz)	(Hz)	(Hz)
/f/	AMR	12.65	879	566	3.59	24.24	572
/f/	MP3	32	1103	742	3.21	23.08	556
/f/	Opus	24	865	610	3.69	23.29	575
/s/	AMR	12.65	917	665	3.46	21.67	544
/s/	MP3	32	891	581	3.85	29.92	542
/s/	Opus	24	903	640	3.62	25.21	519
/z/	AMR	12.65	825	713	4.05	26.15	512
/z/	MP3	32	873	731	3.82	24.72	520
/z/	Opus	24	840	714	4.07	26.17	515
/θ/	AMR	12.65	872	808	3.64	11.89	519
/θ/	MP3	32	952	934	3.36	16.85	516
/θ/	Opus	24	838	781	4.71	32.69	523
/ð/	AMR	12.65	846	640	3.94	28.05	511
/ð/	MP3	32	920	743	3.47	21.64	512
/ð/	Opus	24	808	613	4.27	32.59	510

Table 3.28. Mean values for all spectral measures for the 1 kHz tokens in bitrates divided by segment, and codec compression
(increase from WAV marked in yellow; decrease from WAV marked in blue; no change marked with no shading).

The analysis also included a spectrographic analysis of all the 1 kHz tokens. This analysis was a visual inspection done by the author and is used to give an indication of some potential tendencies. In total, 780 spectrograms were generated for the pairs in group B and C.

This analysis resulted in a division of the rejected pairs into groups as follows: voiced pairs, reductions, increases and errors. The voiced pairs are instances, where the relevant segment occur between voiced segments and are produced without any frication (157). The reductions include more slight reductions mainly of the high frequency content, to more general reduction across the spectrum and in a few instances complete non-encoding (175 pairs). Lastly, increases are pairs where the CoG changes to above 1 kHz (153 pairs), while errors are mistakes in boundaries set by the MFA, which meant the fricatives were not correctly analysed. Examples of the reductions and increases can be found in figure 3.54 and 3.57 below.

None of these examples is of Opus as the effects observed for this codec were limited though similar tendencies could be observed in terms of reduction of the higher frequencies and increases. In line with this, most of the examples are of AMR-WB as this was the codec with the greatest effects and most tokens in this dataset.

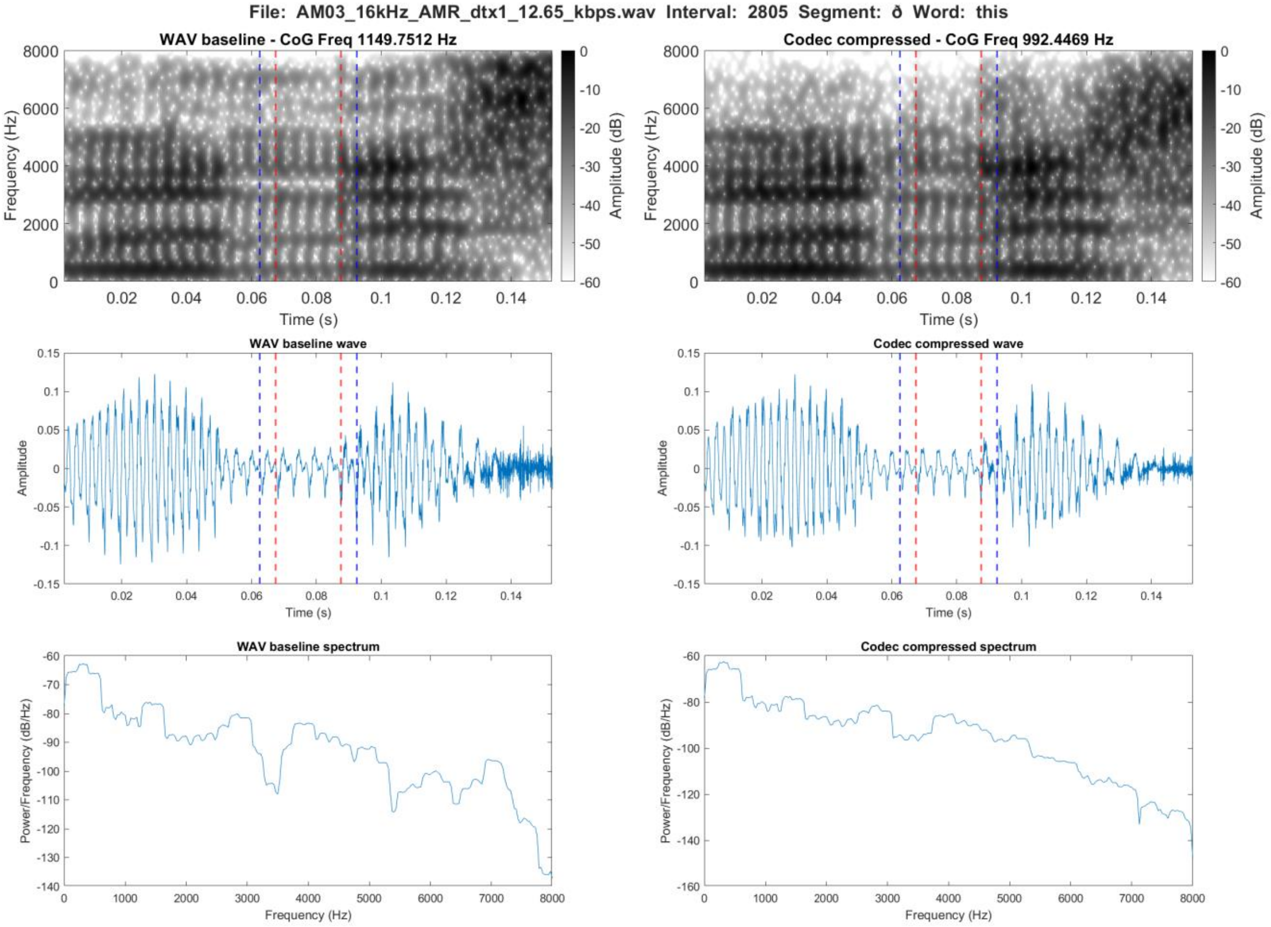


Figure 3.54. Example of fully voiced pair of /ð/ in the word *this* in the WAV baseline (right) and the AMR-WB compression (left)

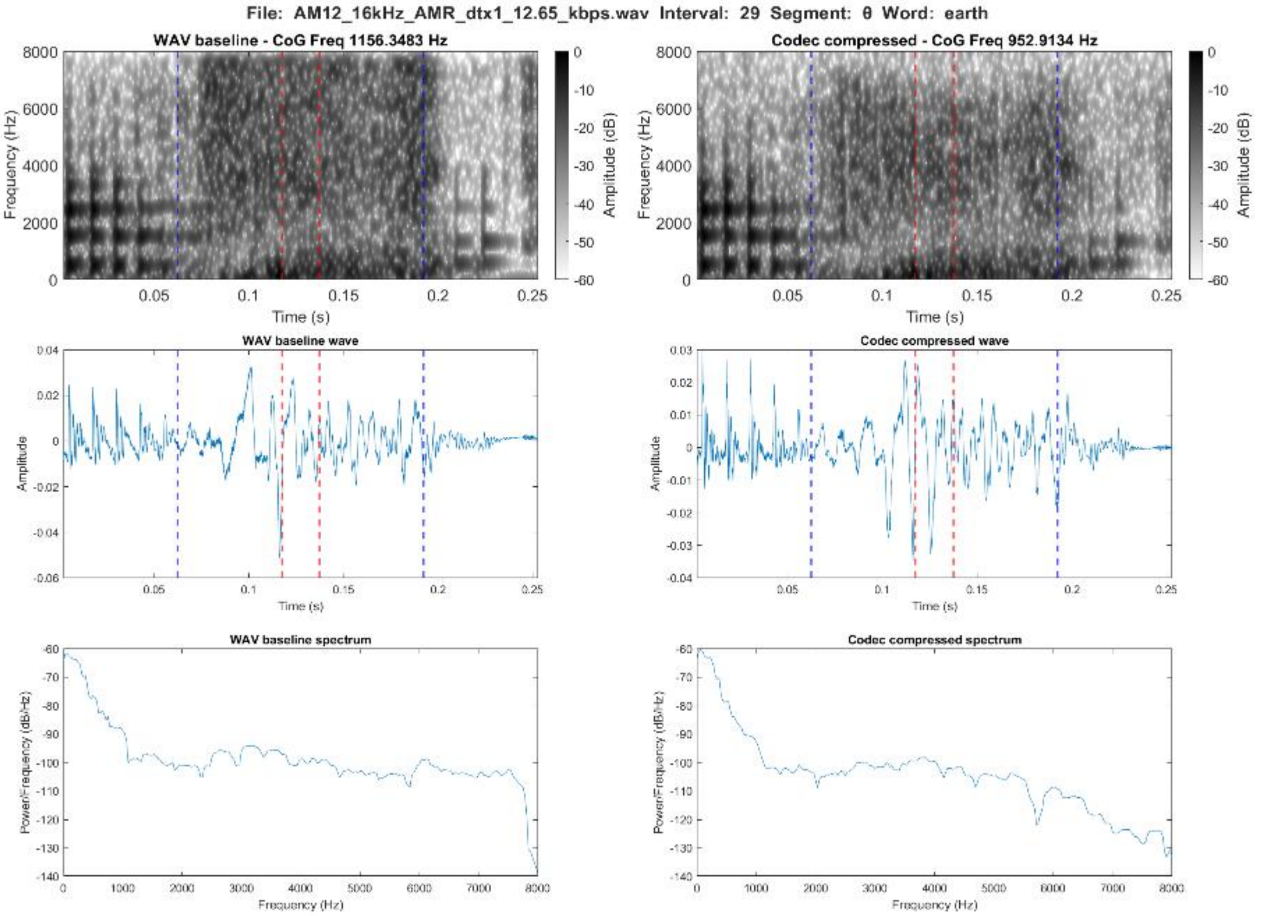


Figure 3.55. Main reduction of higher frequency content illustrated as /θ/ in the word *earth* in the AMR compression

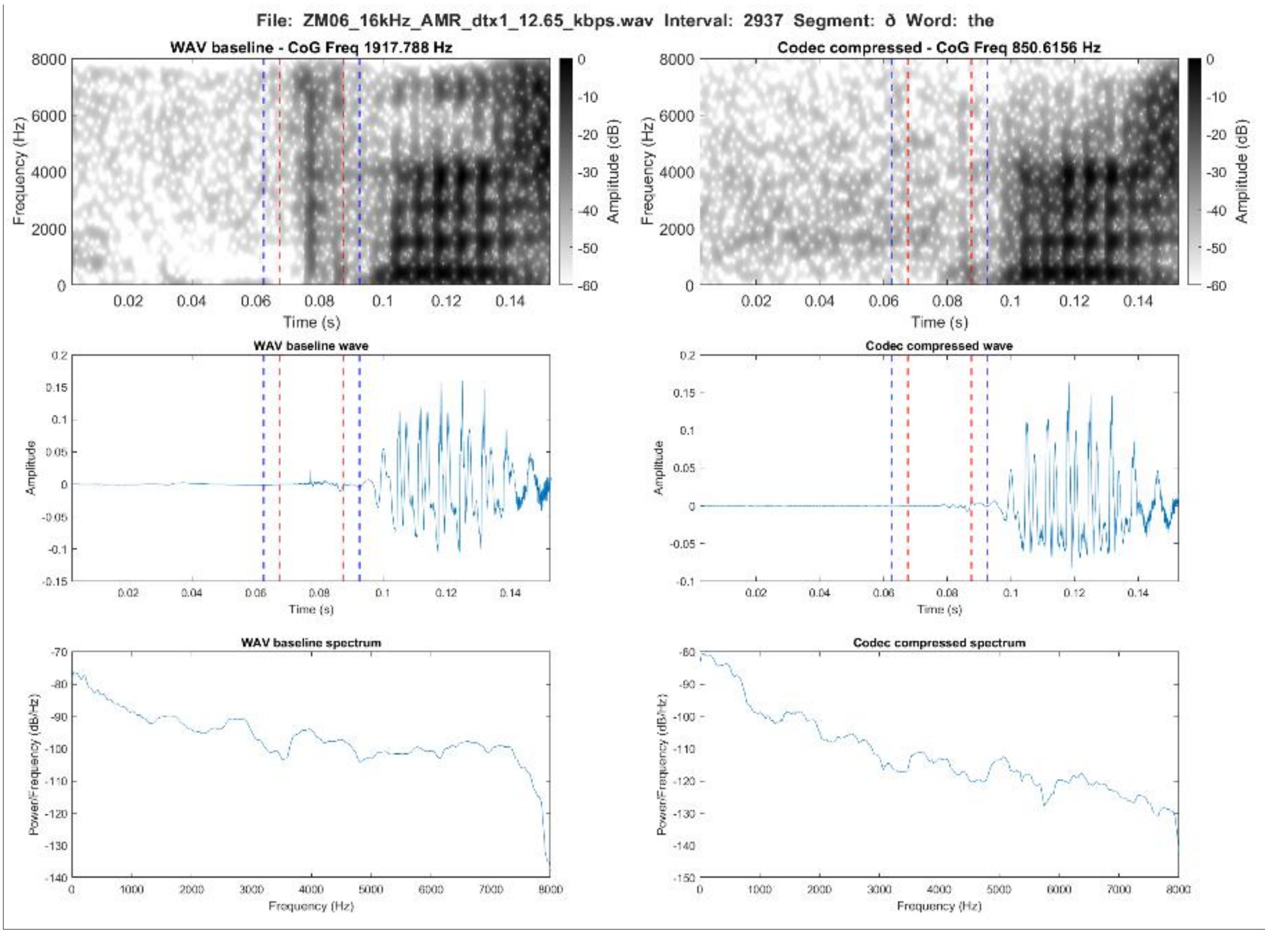


Figure 3.56. Non-encoding with insertion of comfort noise illustrated as plosive produced /ð/ in the word *the* in the AMR-WB compression.

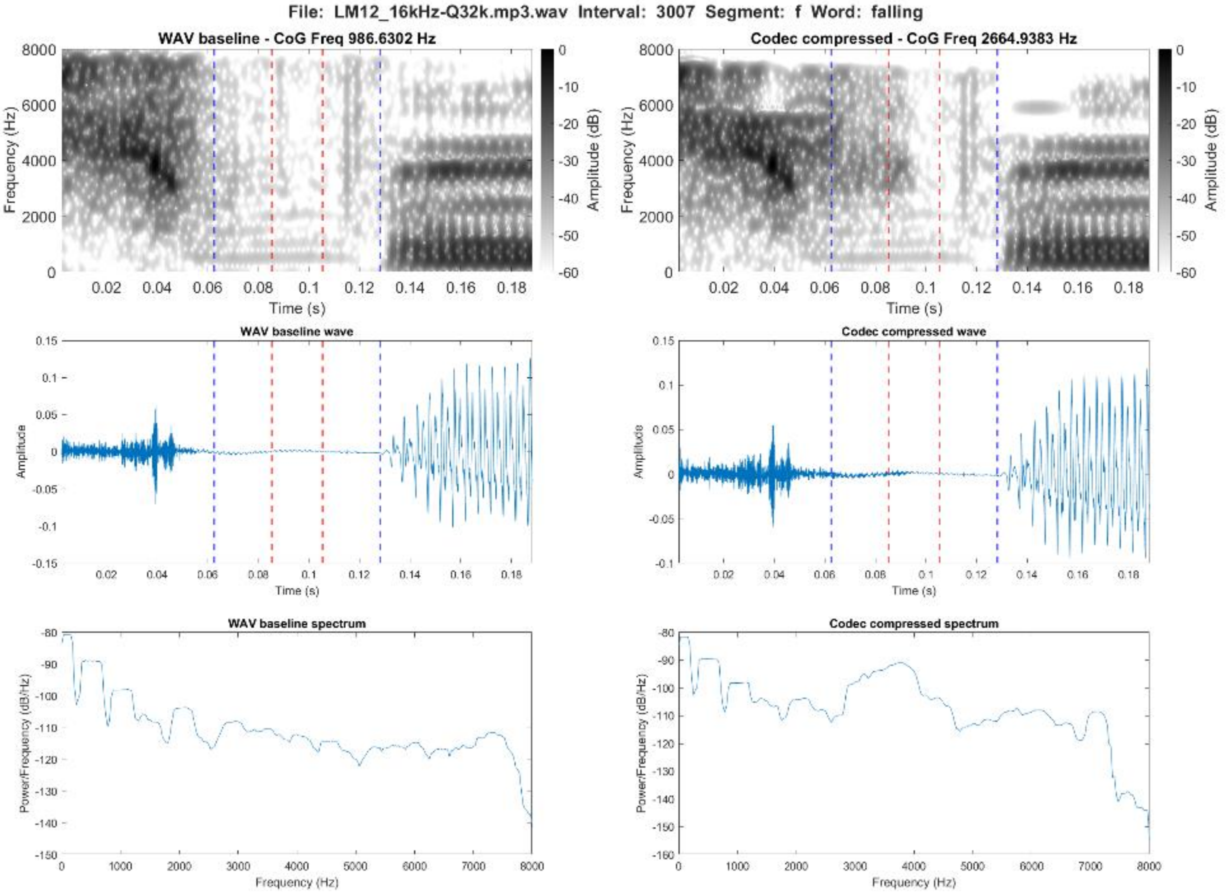


Figure 3.57. Increase caused by insertion of new frequency content by the MP3 compression illustrated as /f/ in the word *falling* in the MP3 compression

In addition, especially for /ð/ and /z/ a large number of the 1 kHz tokens are in fact phonetically voiced segments, which have their mean CoG value just above the 1 kHz boundary in the WAV baseline. Examples of this are presented in figure 3.58 and 3.59 below.

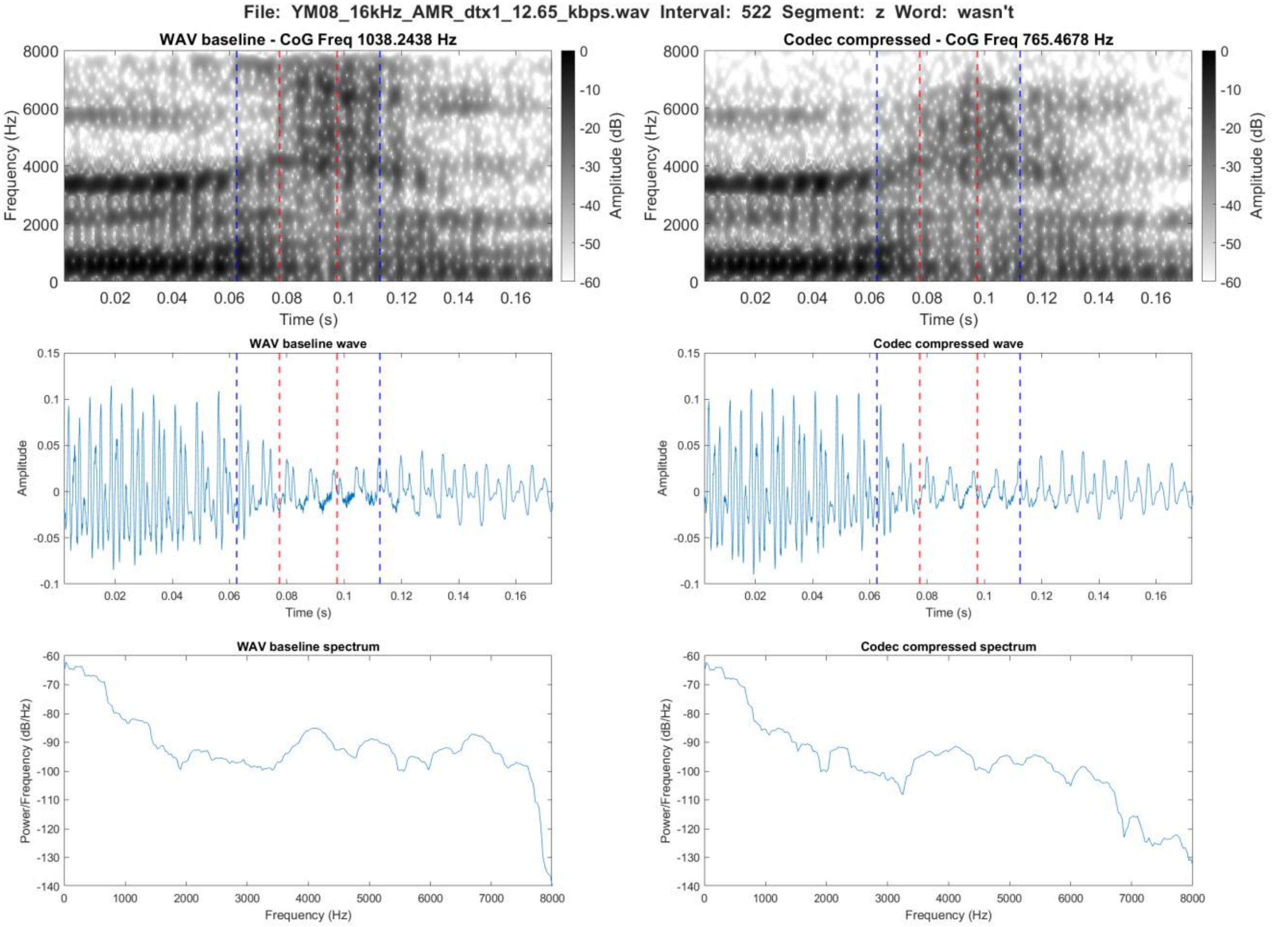


Figure 3.58. Spectrographic comparison of voiced /z/ in the word *wasn't* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

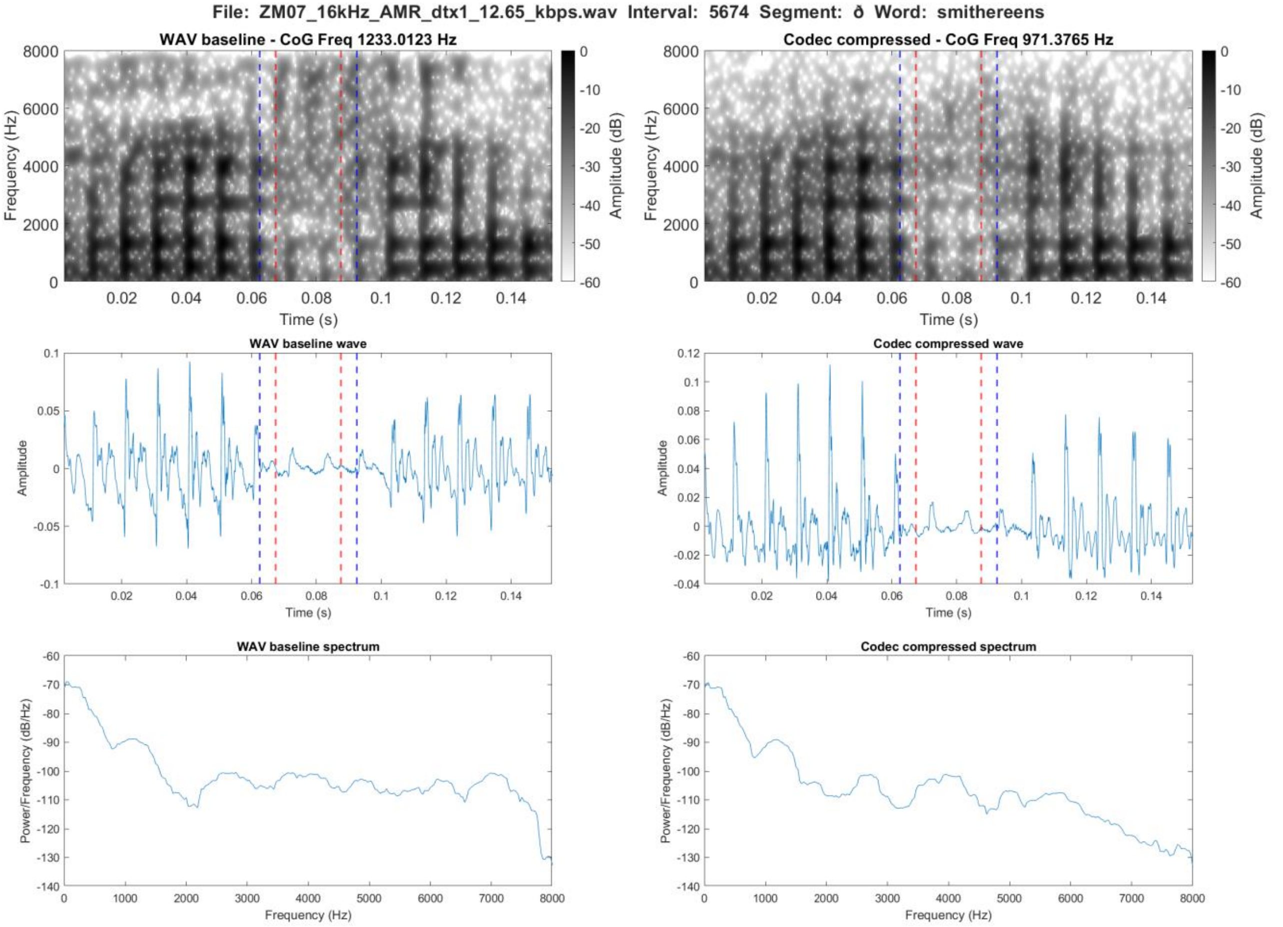


Figure 3.59. Spectrographic comparison of voiced /ð/ in the word *smithereens* in the WAV baseline (left) and the AMR-WB compression at 12.65 kbps (right)

In sum, these effects on the 1 kHz tokens are highly variable across codecs and segments e.g. the non-encoding of the burst in /ð/ is typical for the AMR-WB, while MP3 is the codec, which appear to insert most additional frequency information. For Opus, the main effects are observed as the reduction of the high frequency content. The reductions across the spectrum as well as the increases caused by insertion of frequency content or removal of lower frequency content mainly occur in initial and final position.

3.5 Discussion and Conclusion

This study aimed to investigate the acoustic effects of codec compression with AMR-WB, MP3 and Opus on the fricatives /f/, /s/, /z/, /ʃ/, /ð/, and /θ/. It did so based on three research questions related to: a) the different spectral measures and their relation to the codec compression, b) to what extent the codecs affected the fricatives the same or differently, and c) how the results potentially help create a baseline and understanding of the acoustic consequences of codec compression.

Before addressing the results related to the spectral measures, it is essential to consider the implications of the phonological voicing distinction applied by the MFA as well as the plosive production of /ð/.

It was evident from the spectrographic analysis that the MFA made a phonological voicing distinction whereas this thesis for reasons related to the codecs have applied a phonetic distinction. However, the phonological distinction was not consistent throughout, which is something that will be addressed in the main discussion and conclusion in Chapter 6. For the present study, this meant that both examples of the voiceless as well as voiced productions were presented. For the results this means that some of the similarity in behaviour between /z/ and the voiceless fricatives might be explained by this distinction. However, differences between /z/ and /s/ were observed, which indicate that a number of tokens have indeed been distinct from their voiceless counterparts.

Secondly, for /ð/ particularly in initial position, this was produced like a plosive across the different dialects. Because of the substantial number of instances of /ð/ in initial position e.g. in a word like *the*, this entails that the spectral measures do not solely illustrate the impact of codec compression on the fricative /ð/. For the qualitative analysis, it adds an additional perspective to this Chapter as well as the following Chapter 4 on bitrates, as it gives an indication on the possible effects of codec compression on plosives.

The limitations in bandwidth were introduced both in the initial down-sampling and more importantly by the codecs. First of all, the reason for considering the initial down-sampling is the fact that this already imposes an upper-frequency limit before the codec compression. However, the level of comparison allowed by this down-sampling is preferable to a full bandwidth (i.e. 44,1 kHz) signal where no such comparison between the WAV signal and the codec compressed signals would have been valid.

The white noise signal revealed that the codecs, each to different degrees, limit the frequency band even further. For AMR-WB the codec-compressed signal trails off above 6 kHz, which is a substantial additional limitation to the frequency band. The fact that the codec compressions have not used a band of frequencies between 6 kHz and the 8 kHz limit imposed by the initial down-sampling naturally affects the spectral measures. As mentioned above, the frequency peak will not be a representative measure of the actual acoustic composition of the fricatives, but rather an artefact of the limited frequency band and the 8 kHz cut-off. This was clear from the statistical modelling, which indicated a number of statistical zeros, which made the modelling of this measure invalid based on the plotted residuals. This opens the door for further research into the exact effect of the limited bandwidth relative to a high quality 44,1 kHz file, which would single out the exact effect of the codec compression. CoG is the alternative measure to frequency peak. However, it is possible that despite providing results, which is a more accurate representation of the frequency content, that it is in fact more affected by the down-sampling than the frequency peak. This is because CoG is concerned with the overall energy distribution whereas the frequency peak is only concerned with a single measure point. Thus, CoG will also due to the down-sampling be a relative indicator. Again, further research is needed to establish the exact impact of the down-sampling alone.

The maximum values for the frequency peak for the individual segments and codecs presented a general lowering, but not coherently across each. This suggests that despite the initially down-sampled and thus limited bandwidth, the codec compressions still have segment and codec specific effects e.g. AMR-WB lowering the frequency peak maximum value for all segments and more so than any other codec.

Furthermore, from the white-noise plots and the inferred limitations for each codec in addition to the codec type, there appears to be a relation between the bitrate, the available frequency band and the observed effects of the different codecs. This will be investigated further in Chapter 4, where a number of different bitrates are investigated for each codec.

As predicted, the results show that the included fricatives were all affected by the transmission in comparison to the 16 kHz WAV baseline, and that these effects are segment and codec specific, to different extents. Firstly, the mixed effects modelling revealed that the lossy codec compression regardless of codec type resulted in statistically significant spectral changes. However, a number of segment-dependent effects are observed e.g. in terms of the very small effects down to below 1 percent in mean values and often non-significant effects on /ʃ/ in comparison to changes of up to 25 percent for /θ/ in the Opus compression.

/f/ and [fʲ] were expected to behave similarly in the codec compression. However, this prediction was not borne out as [fʲ] was generally less affected by the codec compression. This might be explained in the statistical analysis by the limited number of [fʲ] compared to /f/, while the spectrographic analysis suggest that the intense frequencies in the lower part of the spectrum for [fʲ] is potentially making it less affected by the codec compression.

From the mean values, it is clear that a general downwards trend is found especially for CoG and SD independent of codec type, but to different degrees depending on the codec type. This is excluding the SD values for /ʃ/ and /z/ in the AMR-WB compression, where slight increases are observed. However, AMR-WB is also the codec with most tokens rejected from the dataset (though marginally so), and the codec with the clearest downwards trajectory of CoG mean values for these 1 kHz tokens. Thus, it is possible that these increases in the main dataset are the result of these tokens being removed, which in turn has limited and evened out the effect in the main dataset.

For skewness and kurtosis, a clear pattern is not immediately evident from the mean values. However, it can be seen the effects appear segment-dependent in terms of magnitude and directionality of the changes. Thus, AMR-WB and Opus both show a segment-dependent pattern with both decreases and increases. For AMR-WB the increases are found for three of the fricatives (i.e. /θ/, /ð/, and /z/), and four out of seven fricatives in the case of Opus (i.e. /θ/, /ð/, /f/, and [fʲ]). MP3, on the other hand, lowered all skewness values and appears to have a slightly bigger effect than AMR-WB. It is interesting to note that the effects on /s/ and /z/ are substantially larger than for any of the other segments with decreases by up to just over 365 percent. These two segments both had skewness values below 0 in the WAV baseline, which suggest that the codec compressions have greater effects on segments with more energy centred below the mean than above. This will be addressed further in the main discussion and conclusion together with the results from the following studies in Chapter 4 and 5.

Finally, the effects of kurtosis are very similar to skewness in terms of magnitude, and indicate a smaller effect of the codec compression across all codec types in comparison to skewness. The effect of Opus and AMR-WB on kurtosis is consistent across segments with Opus increasing all values and AMR-WB decreasing all values. MP3 shows a tendency to lower the values similar to AMR-WB except a slight increase of /s/. Regardless, kurtosis presents some high values both for mean and maximum values already in the WAV, which suggests an unexpected peakedness for these fricatives, or some level of measurement error potentially related to the chosen methodology, which has yet to be identified.

The fact that Opus was the codec with the overall greatest effect on the mean values followed by the AMR-WB, which had the greatest effect on the peak frequencies, implies an effect of the high quality noise suppression and detection on the acoustic characteristics of these fricatives. It is furthermore interesting to note here that MP3 was unexpectedly the codec with the smallest effect on the fricatives. This, however, might be explained by the fact that this was the codec with the highest bitrate. In other words, it was the codec based on this parameter working with the highest level of quality. As mentioned in section 2.5 on digital conversion, newer technology will be able to provide better overall quality at lower bitrates and thus, bitrate is not the sole correlate of quality. Chapter 4 engages more in depth with the effect and influence of higher and lower bitrates to further clarify this aspect.

In that way, the codec compression generally limits the high frequency information and centres the energy lower in the spectrum, which is also evident from the spectrographic representations.

This results in /f/ and /θ/ being more similar and at times almost identical in acoustic composition based on these spectral moments, while it likely limits phonetically distinct high frequency information for e.g. /s/. The CoG for /s/is lowered by the codec compression, while /ʃ/ is often not affected by the compressions. From a sociolinguistic perspective, this might indicate /s/-retraction. In previous research, the evaluation of /s/-retraction is as mentioned in Chapter 2 (section 2.10) primarily based on the relation between the two sounds and whether the distance is significantly decreased rather than exact CoG mean values. The CoG values for /s/, /ʃ/ and /str/found by Baker et al. are generally higher than what is found in the present study (2011). The exact reason for this is beyond the scope of the present study, but is suggested to be related to use of different methods and should be addressed in any future research. With this in mind, despite the decreases to CoG for /s/following codec compression often being significant, the changes were slight, when comparing the CoG mean values of /s/ to /ʃ/, and at no point do the two become identical or /s/lower than /ʃ/ for this measure. In that way, more research is needed to determine whether, these changes are enough to constitute /s/-retraction e.g. by calculation of relation characteristics between the sounds in each condition i.e. WAV and codec compression.

In addition, the fact that /f/ and /θ/ become more alike introduces the potential that the two sounds acoustically as well as perceptually can be interpreted as TH-fronting. However, this needs further investigation to establish to what extent these measures are representative of this dialectal feature.

It is not only /f/ and /θ/, which become more alike in the codec compression, but also in certain cases and especially in the 1 kHz tokens /ʃ/ and /ð/, so that /f/, /θ/, /ʃ/ and/or /ð/ becomes almost identical in terms of CoG. This is key as it means the well established distinction between sibilants and non-sibilants from the spectral moments is potentially limited in codec compressed speech.

Furthermore, the skewness values show how the codec compression largely leaves the distribution unchanged from the baseline values. Thus, the distribution of the energy is similar and the compression created a limited amount of additional extreme values and even at times results in a more normal distribution (i.e. skewness values closer to 0).

As an indicator of peakedness and outliers, it is evident that even in the baseline the kurtosis values are leptokurtic, which means the number of extreme values (outliers) are substantial, the distribution is heavily tailed or from a phonetic perspective that the level of peakedness in the spectrum is substantial. This in itself is surprising working with fricatives and suggests that this specific measure for technical and methodological reasons is likely not suitable. This assumption is substantiated by the behaviour of the kurtosis values in the mixed effects modelling, where the predicted values indicated a high number of statistical zeros. Thus, a better understanding of skewness and especially kurtosis is needed both methodologically and in relation to the effect of codec compression to make any assumptions about these measures. The only conclusion to be drawn is from the comparison of the main dataset tokens and the 1 kHz tokens, where the kurtosis values are even higher than in the main dataset. This suggests that the codec compression affects not only CoG in these tokens, but also both skewness and kurtosis. The reason for this is still to be determined.

Thus, the codec compressions have generally, as expected, had the greatest effect on /θ/. This is not in the sense that /θ/ had the most tokens rejected from the dataset. It is rather seen in the light of the CoG of /θ/ substantially decreasing in both AMR-WB and Opus, while MP3 showed mixed and highly segment-dependent patterns.

It is of further interest that the mean CoG in the AMR-WB compression becomes almost identical for /z/, /θ/, and /ð/, and the same can be seen for /θ/ and /z/ in the Opus compression, while MP3 has very little effect on this measure. This shows an effect across phonetic voicing, which for /ð/ and /z/ is potentially related to the plosive like production of the former, and the phonetically voiceless realisation of /z/. Nevertheless, the effects are codec dependent and again suggest an effect of the bitrate on the level of influence of the codec compressions.

The last part of the study looked at 1 kHz tokens based on one or more CoG values of a segment being above or below the 1 kHz threshold in one or more codec compressions or baseline. This threshold was set as CoG values below this value could suggest substantial effects of the codec compression, e.g. the segment not being recognised as speech and in that way not encoded by the codecs. The patterns observed here must be taken with caution as the number of tokens is limited and vary substantially from segment to segment. Therefore, any predictions or patterns can only be valued as indications of a potential more general trend.

/ð/ was the most rejected token, which is likely related to its already low mean CoG value in the WAV baseline. However, it is of interest that an effect is still observed on this segment i.e. in group A, B and C, but that this is codec dependent. AMR-WB as with the other segments lowers the mean CoG value of /ð/, but to a lesser degree than the other segments, MP3 by contrast increased /ð/, while Opus lowered the CoG value with a similar trajectory as the other segments. This might suggest that AMR-WB has a greater effect in the higher frequencies, while Opus affects the segments equally across the spectrum. The effect of the MP3 compression on the other hand is not as clear. All tokens with a mean value below 900 Hz appear to be increased in CoG mean value by the codec compression, but the effect is segment-dependent (i.e. both voiced and voiceless sounds increased similarly, while others are not). This suggests that MP3 affects the segments with most energy centred in the lower part of the spectrum differently from the higher frequencies. The reason for this is still to be determined.

As a large amount of the tokens of /ð/ and particularly /z/ were phonetically voiced in the 1 kHz tokens, this suggests that the codecs are as suggested sensitive to phonetic voicing, and that the effects are influenced by bitrate.

The overall findings follow the initial predictions. This is specifically in terms of the observed lowering of both CoG and SD. In addition, as expected, both skewness and kurtosis are affected, but to varying degrees. Moreover, it was expected that the effects would be greatest in initial position. This is borne out when observing the 1 kHz tokens, where the majority of tokens occur in this position. This is keeping in mind that a large number of these tokens are of /ð/ in the word *the* and thus, the effect is mainly attributed to this specific context. The reason for this is likely as mentioned above found in the fact that these tend not to be phonetic fricatives, but rather plosives or approximants. This hypothesis is again supported by the visual inspection of the spectra.

The fact that the fewest of the 1 kHz tokens occur in medial position, however, more generally supports the hypothesis that the way the codec activates and deactivates, depending on the detected presence of speech, affects initial and final sounds more than medial ones, where the codec is already activated.

Despite a few tokens being markedly reduced and non-transmitted and the trajectories of the mean CoG reveal increases, it is clear from the spectrographic representations that the reported mean values for these 1 kHz tokens hide a varied set of effects. The spectrographic representations showed both reductions and increases of these pairs. These are only indicative findings, but the fact that reductions occur and even non-encodings e.g. of plosive /ð/ are key knowledge for further research and understanding of digital transmission of speech.

In cases, where only a limited amount of data are available, a reduction or non-encodings might be key to the interpretation or misinterpretation of a given utterance. This supports the initial hypothesis that the codec compression misinterprets a number of these segments as noise and on occasion inserts comfort noise to compensate for the silence. The insertion of frequency information is important, because this acoustic information is an artefact of the codec compression, and thus is not a representation of the actual speech production. The general reduction of the CoG of these tokens due to the limited bandwidth is also evident from these pairs as most of them are smaller changes from just above to just below 1 kHz.

As predicted, the effect of the codec compression is not entirely random across segments and codec type. This finding supports the initial claim that the inconsistencies in results found in previous research is to some extent caused by flawed and varied methodological setups. In addition, the fact that previous research has suggested that the effect of codec compression is greater on female speakers, suggest the results here could be even clearer for female speakers (e.g. Niebuhr and Siegert, 2021).

Finally, segment duration and position were found to improve the accuracy of the statistical modelling. Duration is a distinctive feature for the investigated fricatives, which is the likely reason for the increased accuracy of the models with this as a fixed effect. Position is relevant for similar reasons, while it was also evident from some of the spectrographic examples that the formant structure of the preceding and following sounds influenced the way the fricative was affected by the codec compression. In addition, duration and word position might play a role as the codecs as previously mentioned activates and deactivates in the beginning and end of detected speech. In that way, it becomes more likely to make misinterpretations, and misrepresents sounds in initial and final position. These effects will be discussed further in the main discussion and conclusion in Chapter 6. However, as previously mentioned a more in depth analysis of these features is beyond the scope of this thesis, but is encourage as topics for future research.

These results have implications for any acoustic or auditory analysis of codec compressed and digitally transmitted speech, but especially so for sociolinguistics and forensic phonetics.

For sociolinguistics, TH-fronting and /s/-retraction were previously mentioned, another potential perspective relates to the use of MP3 as file format when downloading and analysing material from online sound files and films. The results from this study have shown how MP3 significantly affect the spectral measures, and from the qualitative analysis, reductions in intensity were seen. The effects are for the spectral measures often limited and the smallest in the comparison between the three codecs. However, the spectrographic analysis revealed how MP3 at times insert frequency information not present in the baseline WAV, which is problematic if doing acoustic analysis. Thus, based on this study, sociolinguists are encourage to consider the implications of the MP3 compression on the files in question, and inspect them qualitatively. The following Chapter as well as the final main discussion and conclusion will further consider this perspective.

In a forensic phonetic perspective, in speaker comparison and profiling the alterations or lack of spectral information mean that sounds might become less distinct within a recording or might not be directly comparable to a higher quality reference recording. Moreover, in cases involving content determination a non-transmission or marked reduction in energy may force the forensic phonetician to rely more (consciously or unconsciously) on context, which thus, increases the influence from priming and potential bias. If such data is used it is essential to be aware of the basic workings of digital transmission, the codec used as well as the equipment and hardware in order to compensate for this, and never make direct acoustic or phonetic comparisons between high quality and codec compressed recordings or between different codec compression technologies. In that way, in terms of data-collection and using digitally transmitted and codec-compressed speech, this study follows previous research and urges caution especially if the data is used for acoustic or segmental purposes (e.g. Sanker et al. 2021, Leemann et al. 2020, Siegert and Niebuhr 2021).

The coming Chapters will investigate the observed effects in more detail and introduce further variables i.e. different bitrates (Chapter 4) as well as live transmitted speech (Chapter 5), where background noise and network access are likely to increase the effect of the codec compression.

Chapter 4 : The effect of bitrate on spectral implications of codec compression

4.1 Introduction

The previous study established a baseline of spectral implications of codec compression with AMR-WB, MP3 and Opus at one set bitrate for each codec. However, as bitrate is key to the quality level i.e. how much acoustic information can be captured in the codec-compressed signal, it is relevant to look at the same three codecs, but compare the spectral implications as a factor of different bitrates.

As described in section 2.5 on digital conversion, the bitrate is a way to express the relation between sampling rate and bit depth in PCM WAV, while representing the amount of data used to encode the signal per second in the digital transmission. In that way, this Chapter takes a step further into the technical aspects of the codec compression in order to better understand, the various variables potentially affecting speech during digital transmission.

In linguistics and phonetics, research is yet to be conducted on the precise effects of different bitrates and codecs on fricatives. From a live-transmission and thus, everyday perspective, the study of different bitrates is also relevant as the available bitrate will not always be the same, and not always be the best quality option due to geographical location, amount of traffic on the network and other technical complications (See section 2.5).

In sum, this motivates this second study, which will look at the same three codecs as in the previous study, but here the codec compression will be made at two additional bitrates for each codec i.e. one low quality and one high quality. As with the previous study the chosen bitrates were based on quality estimates and their corresponding bitrates (3GPP 2022; Valin, Vos, and Terriberry 2012; Triton 2022). The same seven fricatives are investigated (i.e. /s, z, f, θ, ð, ʃ/and [fʲ]) with no additional measures added in comparison to the baseline study. Thus, this study works with the static spectral measures CoG, SD, skewness, kurtosis, and frequency peak. These are the first four spectral moments and frequency peak.

The results from this study, as well as a comparison between these and the baseline results, will be used with the aim to answer the following three research questions:

BitrateRQ1: What do the included measures indicate about the effect of different bitrates on fricatives for each codec compression?

BitrateRQ2: To what extent are the observed effects codec dependent? And, what is the interaction between the codecs and bitrates based on these effects?

BitrateRQ3: In what way does this inform potential phonetic implications of codec compression and digital transmission on fricatives in view of varying bitrates?

4.2 Predictions

As very little research has been done on the acoustic effects of different bitrates using different codec compressions, the predictions for this study will primarily be based on knowledge of the technical setup of the codecs. This also means that the predictions will be less detailed than for the previous baseline study. As such, the predictions are in essence the same as for the previous study, but with effects expected to be more prevalent for the lower bitrates and less for the higher bitrates.

In brief, this means that the less intense voiceless fricatives i.e. /f/ and /θ/ are expected to be most affected and reduced in intensity. For the lower bitrates, this is predicted to lead to more tokens moved from the main dataset to the 1 kHz tokens due to changes from above to below 1 kHz or vice versa from WAV to the codec compression. [fʲ] was expected to behave similar to /f/ in the previous study, but showed to be less affected. Thus, this pattern is predicted to carry through to the present study.

The effects on /s/ is as with the other fricatives expected to be exaggerated in the lower bitrates. This means a lower CoG due to limited frequency information in the higher end of the spectrum and a decreased SD due compression of the acoustic information in a more limited bandwidth and with less data available to represent the signal. On the other hand, as the effects observed in the baseline study were often minor for /s/, the high bitrates are predicted to leave /s/ with lesser changes to the spectral measures. The latter prediction is also valid for /ʃ/, which showed little to no effect of any of the codec compression already in the baseline.

For /z/ the changes observed for the spectral measures were at times limited with effects down to a only few percentages. Therefore, /z/ is as /s/ predicted to be almost intact in the higher bitrates, and slightly more affected in the lower bitrate conditions. However, the effects of the lower bitrates are predicted to be less than for /s/ due to the inherently lower CoG of /z/.

Lastly, the baseline study revealed how a number of /ð/ tokens, were not in fact true fricatives and e.g. produced as plosives, voiceless fricatives or approximants. This limits the possible conclusions and prediction to be made about this segment, however, MP3 particularly in the baseline had a tendency to increase the CoG from below 1 kHz in the baseline to above in the codec compression. This effect is expected to be seen in the present study as well. This is especially in the lower bitrates as the increase was often followed by a reduction in high frequency content.

Overall, on the basis of the results from the baseline study, AMR-WB and Opus are generally expected to have the biggest impact on the spectral measures, and the effects observed on the high frequency content are predicted to be more prevalent in the lower bitrates for all codecs.

4.3 Methodology

The methodology for this study was overall identical to the baseline study as both work with the same dataset and measures. Therefore, this section will only contain brief overviews of the information presented in the methodology for the baseline study (see section 3.3). The reader will be referred to more detailed information from the previous section with section references throughout. Philip Harrison, as in the previous study in Chapter 3, assisted with the spectral analysis and generation of spectrograms.

4.3.1 Corpus and Participants

The *You Came to Die?! corpus* (Best et al., 2012-2015) was used for this study. In total, the corpus consists of thirty male and thirty female speakers, all native speakers of English and aged between 18 and 41. The participants spoke five different accents of English with six speakers of each accent. These were Australian (AUS), New Zealand (NZL), London (LON), Newcastle (NCL), and York (YRK) English. See section 3.3.1 for further details.

4.3.2 Materials

The dataset elicits the following seven fricatives /s, z, f, θ, ð, ʃ/and [fʲ] in read speech from an approximately 10 minuteute reading of the *Chicken Little* story in varying phonetic context and in word-initial, word-medial, and word-final position (e.g. <fear>, <painful>, <safe>; see full transcript in appendix 1).

The 16 kHz resampled files were compressed with the three codecs of interest (see section 3.3.3 on sound files for details) at three different bitrates for each codec (see table 4.1 for bitrates). The *average* bitrate is the bitrate used in the baseline study from which the results were presented in the previous Chapter.

Quality	**Codec**	**Bitrate (kbps)**
Low	MP3	16
Average	MP3	32
High	MP3	48
Low	AMR-WB	6.6
Average	AMR-WB	12.65
High	AMR-WB	23.85
Low	Opus	12
Average	Opus	24
High	Opus	64

Table 4.1. Overview of different qualities and corresponding bitrates for each codec compression

This resulted in two additional datasets, with all the low bitrates in one, and all the high quality bitrates in the other with the same number of tokens in each as in the baseline study. To recap, this means a total of 85,280 fricatives in each additional dataset with 21,320 in each codec compression and baseline. The number of individual segments in total including all codecs and baseline were 5,104 tokens of /ʃ/, 5,156 tokens of /θ/, 15,572 tokens of /ð/, 13,668 tokens of /f/, 960 tokens of [fʲ], 29,336 tokens of /s/, and 15,484 tokens of /z/. For additional details on the number of tokens, see section 3.3.2.

As in the baseline study, a separate dataset (1 kHz tokens) was created for both the low and the high quality bitrate datasets with all tokens with a CoG below 1 kHz and their counterparts in any other codec compression or baseline. These datasets were analysed separately and the results presented in section 5.4. These tokens were analysed separately to avoid skewing the main dataset, while also investigating the tokens with the potentially most apparent effect of the codec compression (see details in section 3.3.5 on data extraction and measurements).

In sum, this meant that the low bitrate quality dataset consisted of 78,940 tokens with 19,735 per codec compression, while the 1 kHz tokens amounted to 6,340 in total. For the high bitrate quality dataset, the total number of tokens were 80,588 with 20,147 per codec compression, and 4,692 tokens rejected. This is illustrated in the two tables below, where information position of the segments within the word are added to illustrate the distribution.

Segment	**Total** (without rejected)	**Rejected** (total)	**Initial position**	**Medial position**	**Final position**
/f/	13,584	84	8,972	2,536	2,156
[fʲ]	960	0	600	360	0
/s/	29,300	36	15,624	6,580	7,128
/z/	15,032	452	252	2,392	12,800
/ʃ/	5,104	0	3,336	1,396	912
/θ/	5,084	72	1,096	1,880	2,180
/ð/	9,876	5,696	12,916	1,396	912

Table 4.2. Number of tokens per segment with different criteria including all codecs and bitrates without the 1 kHz tokens for the **low** quality bitrates.

Segment	**Total** (without rejected)	**Rejected** (total)	**Initial position**	**Medial position**	**Final position**
/f/	13620	48	8,900	2,528	2,132
[fʲ]	960	0	600	360	0
/s/	29320	16	15,624	6,580	7,116
/z/	15288	196	248	2,316	12,724
/ʃ/	5104	0	3,336	1,168	600
/θ/	5112	44	1,084	,	2,164
/ð/	11184	4,288	10,016	516	652

Table 4.3. Number of tokens per segment with different criteria including all codecs and bitrates without the 1 kHz tokens for the **high** quality bitrates.

The exact number of 1 kHz tokens for each codec and the individual bitrates is part of the results in section 4.4.

4.3.3 Sound files

The sound files were down-sampled in the same manner as for the baseline study, which means the files were down-sampled to 16 kHz.

The only change to the analysis of the sound files in comparison to the baseline study was the addition of two more bitrates per codec. This meant that all files were codec compressed twice more, using the same software, with each additional set bitrate.

With the addition of these 6 bitrate conditions, the spectra for each segment in each condition was extracted using the same MATLAB script as the baseline study (Harrison 2022; MathWorks Inc. 2010). See section page section 3.3.3 for further details on the sound files and codec compression.

4.3.4 Segmentation

This study used the same forced aligned and corrected TextGrids as the baseline study. Thus, the reader is referred to 3.3.4 for further details.

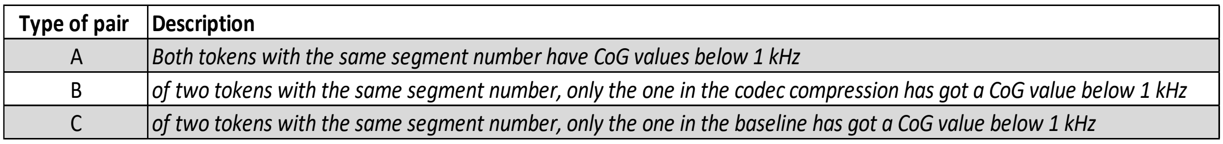
4.3.5 Data extraction and measurements

The measurements were as mentioned the same as for the baseline study, and the method was in large part identical. The main difference was that the analysis included two additional runs for each codec compression with the added bitrates.

As in the baseline study (see Chapter 3), a white noise signal was generated and subjected to codec-compression at each of the new bitrates. This was done to estimate the actual upper-frequency limit of the codec compressions and to illustrate whether the different bitrates affect this.

The spectral measures were extracted using the same MATLAB script written by Philip Harrison (Harrison 2022; MathWorks Inc. 2010) in the 20 ms central frame. All of this was done using multitaper analysis.

Again, any tokens in each compression with a CoG below 1 kHz were moved from the main dataset and put into a separate dataset, which was subsequently analysed. These 1 kHz tokens consisted of three categories, which are summarised in table 4.4 below.



Type of pair	Description
A	*Both tokens with the same segment number have CoG values below 1 kHz*
B	*of two tokens with the same segment number, only the one in the codec compression has got a CoG value below 1 kHz*
C	*of two tokens with the same segment number, only the one in the baseline has got a CoG value below 1 kHz*

Table 4.4. Description of types of pairs in the dataset including only the 1 kHz tokens.

The final part of the analysis was the generation of spectrograms for the 1 kHz tokens in categories B and C. However, as with the previous study the number of tokens in the 1 kHz dataset for this study ruled out visual inspection of each spectrogram individually. Thus, a random subset was inspected to give an indication of more general tendencies. For further details on the process see section 3.3.5.

4.3.6 Statistical analysis

Firstly, summary statistics were generated for each measure for each codec and new bitrate and the differences in mean values between the baseline WAV and the low and high quality bitrates were calculated. The statistical analysis presented in section 3.3.6 in the previous Chapter was repeated again using R for each new dataset (i.e. for each bitrate in each codec compression) (RStudio Team 2019; RCore Team 2020). This resulted in the following variables in the mixed effects models with each spectral measure as the dependent variable:

A range of models were compared for each codec and spectral measure including the following independent variables: format (with 4 levels (2 per model): baseline 16 kHz, and one of the following AMR-WB, MP3, or Opus), speaker (with 30 levels: individual speakers), word position (with three levels: initial, medial, and final), segment duration, and preceding segment (with 21 levels) and following segment (with 60 levels).

The best-fitted models turned out to be the same as those used for the baseline study and the average bitrate. Thus, the model had format, segment, preceding segment, following segment, and duration as fixed effects, and speaker and word as random effects. For additional and more exact description of the statistical analysis and procedure see section 3.3.6.

4.4 Results

This section will be divided into two main sections based on the two bitrate qualities. Hence, the first main section will present the results of the low quality bitrate dataset for each codec, while the second main section will present the results from the high bitrate dataset again for each codec. The comparison between the two and between these results and the previous baseline study will be done in section 4.4.1 as well as in the discussion and conclusion of this Chapter in section 4.5).

Firstly, to indicate the actual upper-frequency limit of the codec compressions for each bitrate, a white noise signal was codec compressed. The result of this is illustrated in figure 4.1 to 4.3 below. The following values are not obtained with the usual -3 dB definition for the cut-off, but are estimated from the graphs and hence, should be seen as indicative. The cut-off is interpreted as the point where the values trail off and no longer follow the trajectory of the white noise.

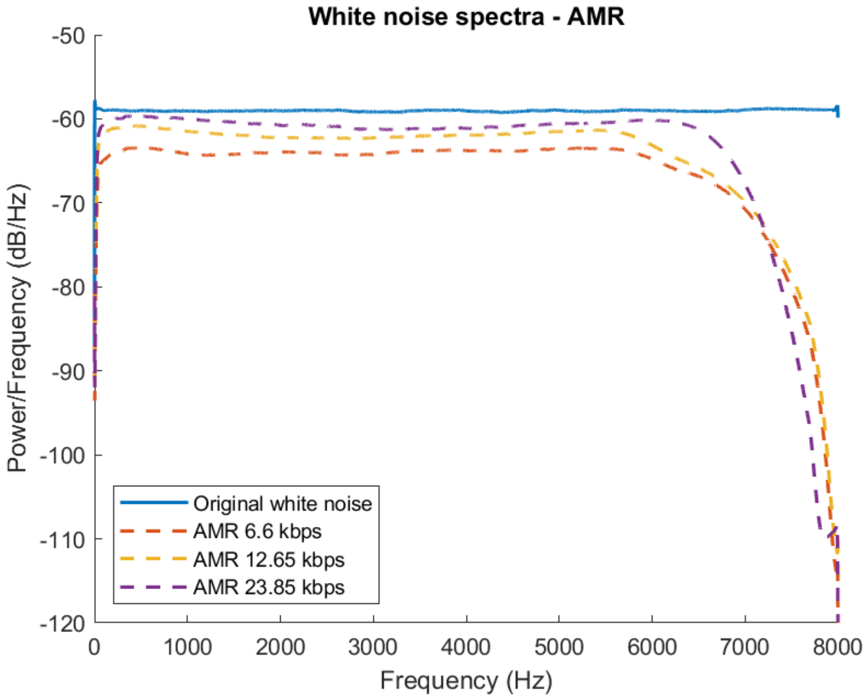


Figure 4.1. Frequency spectra of original white noise signal and codec compressed versions for AMR-WB at every bitrate

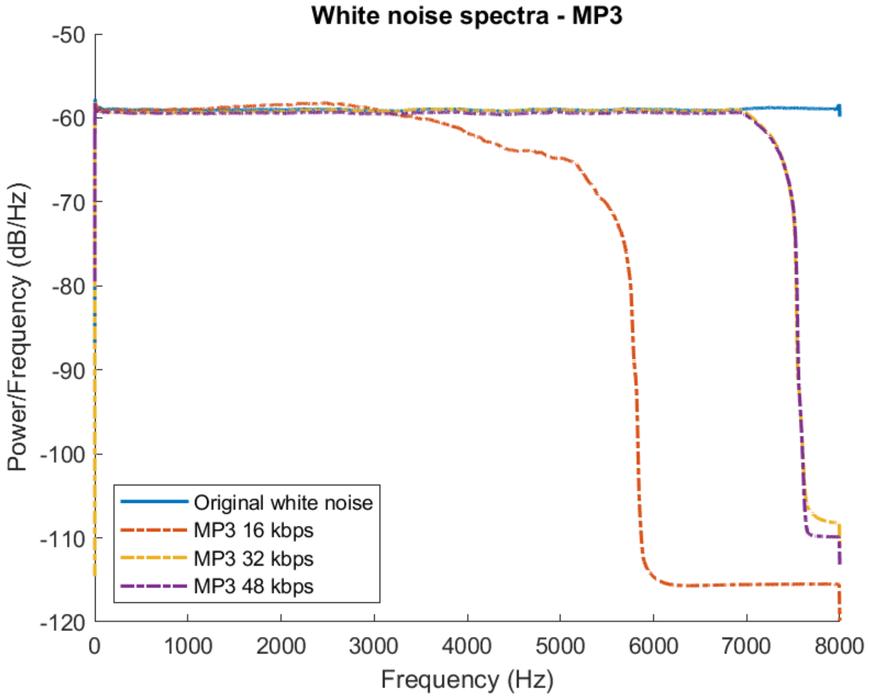


Figure 4.2. Frequency spectra of original white noise signal and codec compressed versions for each codec at every bitrate

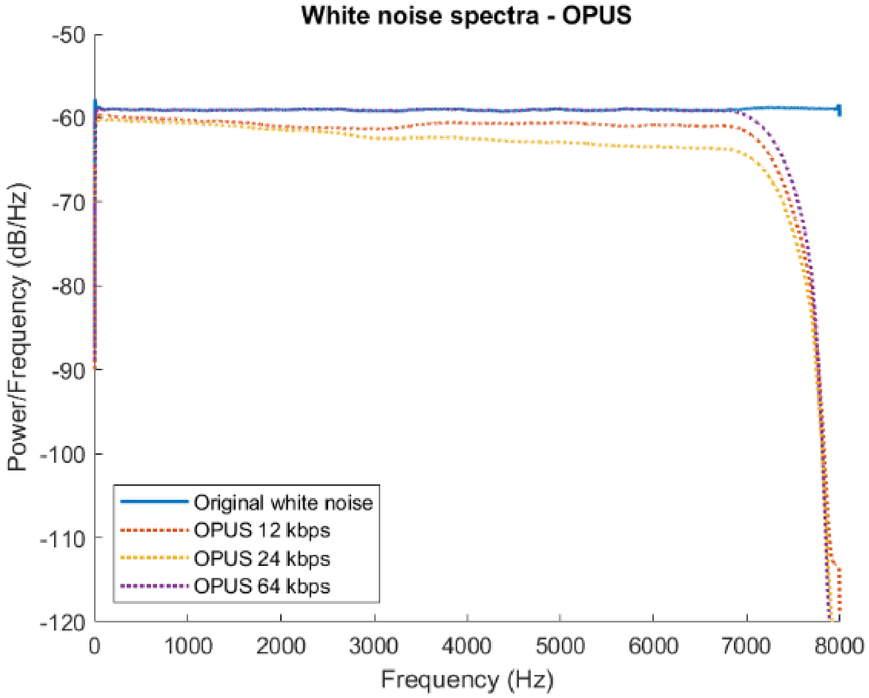


Figure 4.3. Frequency spectra of original white noise signal and codec compressed versions for each codec at every bitrate

The plots show how for AMR-WB the high bitrate provided a higher cut-off frequency than the average and low bitrate, while all three trailed off gradually towards the 8 kHz limit. For MP3, the effect of bitrate was substantial particularly for the low bitrate. Both the average and high bitrate trailed off around 7 kHz, while the low bitrate resulted in a trail off already between 3 and 4 kHz up until 6 kHz. For Opus, the signal trails off for all bitrates around 7 kHz.

The following sections will be specific to each codec and bitrate i.e. low or high. The previous Chapter presented spectrographic examples of the various productions of the voiced fricatives, however, due to the amount of data presented in this Chapter, only one spectrographic example will be provided for each segment.

Lastly, it should again be noted here that due to R not recognising the IPA symbols for /θ/, /ʃ/, [fʲ], and /ʃ/ these will be written as *theta, esh, fj* and *esh* in the R generated figures.

4.1.1 Low Quality bitrates

All results reported in the following are from the low quality bitrate dataset excluding all the 1 kHz tokens. These tokens will be addressed and analysed separately in section 4.1.3. This section will also provide details on the exact number of 1 kHz tokens for each codec.

As expected, the mean values in Table 4.5 reveal that all segments are affected by the codec compression, and that these effects are codec as well as segment-dependent. As in the baseline study, the frequency peak is to a certain extent not truly representative in this context due to the limited bandwidth and sampling rate. However, when seen together with the CoG, the measure can still give an indication of the effect on the higher frequency content. Regardless, the frequency peak will not be included in the statistical modelling as the number of statistical 0s skew the model and the desired randomness, and distribution of the residuals. Again, the same tendency was observed for kurtosis, which for this reason was also not included in the statistical modelling. Consequently, no p-values will be presented for these measures, and thus, these values will only be reported from descriptive statistics.

The voiced fricatives (i.e. /ð/ and /z/) deviated slightly from the desired randomness in the residual plots predicted values. However, this is to be expected from the acoustic structure of these sounds often with energy centred low in the spectrum towards the lower cut-off (see section 3.3.6 and

3.4 for further details). In that way, as it is clear why this pattern is found and the remaining residual plots i.e. distribution and quantiles are following expected patterns, the voiced segments are still included in the mixed effects modelling. However, this should be kept in mind when assessing the implications and reliability of the outputted results.

Seg	**Codec**	**Bitrate (kbps)**	**CoG (Hz)**	**SD (Hz)**	**Skew**	**Kurt**	**Freq. Peak (Hz)**
/f/	AMR	6.6	2895	1440	0.61	3.19	2210
/f/	MP3	16	2472	1111	0.31	2.66	2149
/f/	Opus	12	3031	1564	0.65	3.23	2293
/f/	**WAV**	**NA**	**3159**	**1669**	**0.75**	**3.43**	**2437**
[fʲ]	AMR	6.6	2946	1349	0.69	3.48	2306
[fʲ]	MP3	16	2567	1024	0.31	2.95	2337
[fʲ]	Opus	12	3068	1456	0.71	3.52	2390
[fʲ]	**WAV**	**NA**	**3131**	**1544**	**0.87**	**3.84**	**2408**
/s/	AMR	6.6	4457	1041	-0.42	5.91	4408
/s/	MP3	16	4075	865	-1.50	8.51	4271
/s/	Opus	12	4632	1097	-0.33	5.71	4593
/s/	**WAV**	**NA**	**4787**	**1165**	**-0.08**	**5.72**	**4741**
/z/	AMR	6.6	3872	1190	-0.20	6.39	3264
/z/	MP3	16	3654	1019	-1.22	7.68	3420
/z/	Opus	12	4103	1201	-0.25	6.16	3618
/z/	**WAV**	**NA**	**4306**	**1236**	**-0.07**	**6.38**	**3819**
/ʃ/	AMR	6.6	3274	873	1.43	7.79	3011
/ʃ/	MP3	16	3115	625	0.45	5.39	2995
/ʃ/	Opus	12	3274	874	1.43	8.10	3011
/ʃ/	**WAV**	**NA**	**3271**	**896**	**1.71**	**9.45**	**3002**
/θ/	AMR	6.6	2949	1552	0.49	3.12	1979
/θ/	MP3	16	2437	1204	0.34	2.85	1798
/θ/	Opus	12	3095	1680	0.53	3.17	2042
/θ/	**WAV**	**NA**	**3299**	**1802**	**0.56**	**3.30**	**2282**
/ð/	AMR	6.6	1767	1158	2.07	13.69	881
/ð/	MP3	16	1687	1007	1.25	6.36	909
/ð/	Opus	12	1887	1241	1.95	12.59	960
/ð/	**WAV**	**NA**	**2021**	**1336**	**1.88**	**11.91**	**1048**

Table 4.5. **Low** dataset mean values for all spectral measures for each fricative in each individual codec using the **low** quality bitrate and 16 kHz WAV baseline

The magnitude and direction of the changes in mean values are shown in Table 4.6. Some general trends can be observed e.g. the MP3 codec generally results in the largest changes to the spectral measures, while the Opus codec is the least influential.

Seg	Codec	Bitrate	CoG	CoG	SD	SD	Skew	Skew	Kurt	Kurt	Freq.	Freq.
		(kbps)	(Hz)	(%)	(Hz)	(%)	(Hz)	(%)	(Hz)	(%)	Peak	Peak
											(Hz)	(%)
/f/	AMR	6.6	263	-8.34	229	-8.34	0.14	-19.15	0.24	-6.99	227	-9.32
/f/	MP3	16	687	-21.75	558	-21.75	0.45	-59.25	0.77	-22.37	289	-11.84
/f/	Opus	12	128	-4.04	105	-4.04	0.10	-13.48	0.20	-5.88	145	-5.93
[fʲ]	AMR	6.6	185	-5.91	194	-5.91	0.18	-20.56	0.36	-9.35	102	-4.25
[fʲ]	MP3	16	564	-18.02	520	-18.02	0.56	-64.14	0.88	-23.06	71	-2.96
[fʲ]	Opus	12	63	-2.00	88	-2.00	0.16	-18.23	0.31	-8.18	18	-0.76
/s/	AMR	6.6	329	-6.88	124	-10.65	0.34	-287.60	-0.19	3.25	333	-7.02
/s/	MP3	16	712	-14.88	300	-25.75	1.50	-1010.3	-2.79	48.73	469	-9.90
/s/	Opus	12	155	-3.24	67	-5.79	0.82	-192.11	0.01	-0.19	148	-3.12
/z/	AMR	6.6	434	-10.09	46	-3.71	0.12	-142.49	-0.01	0.20	554	-14.52
/z/	MP3	16	652	-15.14	217	-17.56	1.15	-665.54	-1.30	20.46	398	-10.42
/z/	Opus	12	203	-4.72	36	-2.88	0.18	-116.98	0.22	-3.42	201	-5.26
/ʃ/	AMR	6.6	-3	0.09	23	-2.53	0.29	-16.66	1.66	-17.53	-9	0.30
/ʃ/	MP3	16	156	-4.76	271	-30.26	1.27	-73.98	4.06	-42.99	7	-0.23
/ʃ/	Opus	12	-3	0.10	21	-2.40	0.28	-16.35	1.35	-14.33	-10	0.32
/θ/	AMR	6.6	350	-10.60	251	-13.91	0.07	-12.09	0.18	-5.46	303	-13.29
/θ/	MP3	16	862	-26.14	598	-33.20	0.22	-38.66	0.46	-13.83	484	-21.22
/θ/	Opus	12	204	-6.20	122	-6.77	0.03	-5.26	0.13	-3.91	240	-10.52
/ð/	AMR	6.6	254	-12.57	178	-13.33	-0.19	10.02	-1.79	15.03	167	-15.91
/ð/	MP3	16	334	-16.53	329	-24.66	0.63	-33.67	5.55	-46.60	139	-13.29
/ð/	Opus	12	134	-6.64	95	-7.14	-0.06	3.44	-0.68	5.72	88	-8.37

Table 4.6. Differences in mean values between baseline (WAV) and codec compression using the **low** quality bitrate in Hz and percentage.
Colours indicate the direction of the change. (i.e. blue = decrease; yellow = increase)

Overall, the low bitrate decreased all the mean values across the spectral measures apart from a few cases of skewness and kurtosis, and for CoG and frequency peak for /ʃ/. The latter was, however, minor increases by less than 1 percent.

The following sections will be specific to each codec and bit rate i.e. low or high. As in the previous study, it should be noted here that due to R not recognising the IPA symbols for /θ/, /ʃ/, [fʲ], and /ʃ/ these will be written as *theta, esh, fj* and *esh* respectively in the R generated figures.

4.1.1.1 Low bitrate: AMR-WB

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the AMR-WB codec in the low quality bitrate dataset.

First, the linear predictions for each spectral measure and the individual segments can be found below (figure 4.4 to 4.9). These indicate the directionality of the changes produced by the AMR-WB codec using the low bitrate. The graphs present the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these plots are in the following sections on the individual segments.

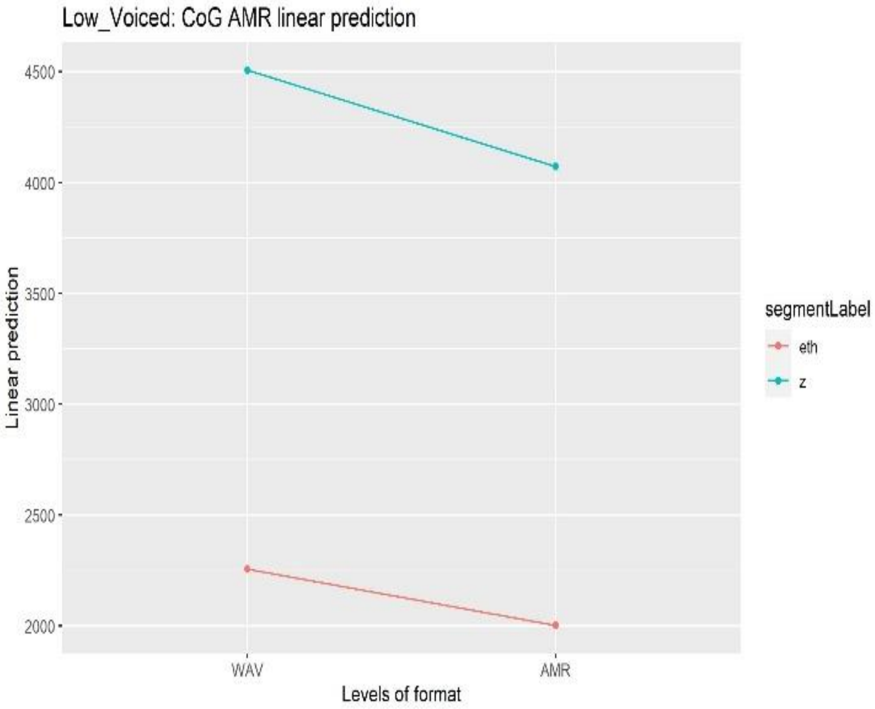


Figure 4.4. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

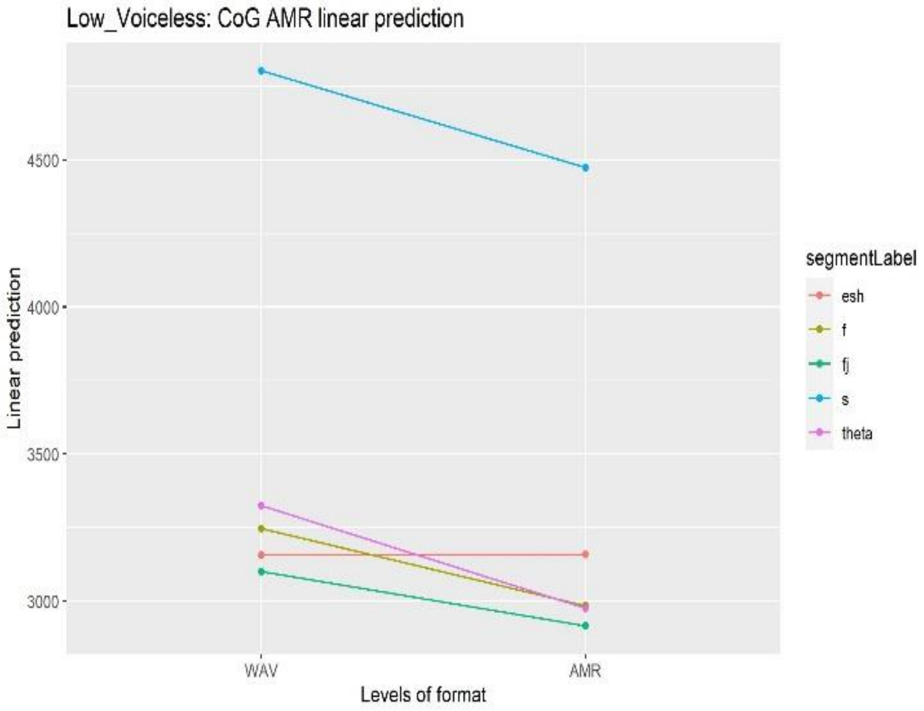


Figure 4.5. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

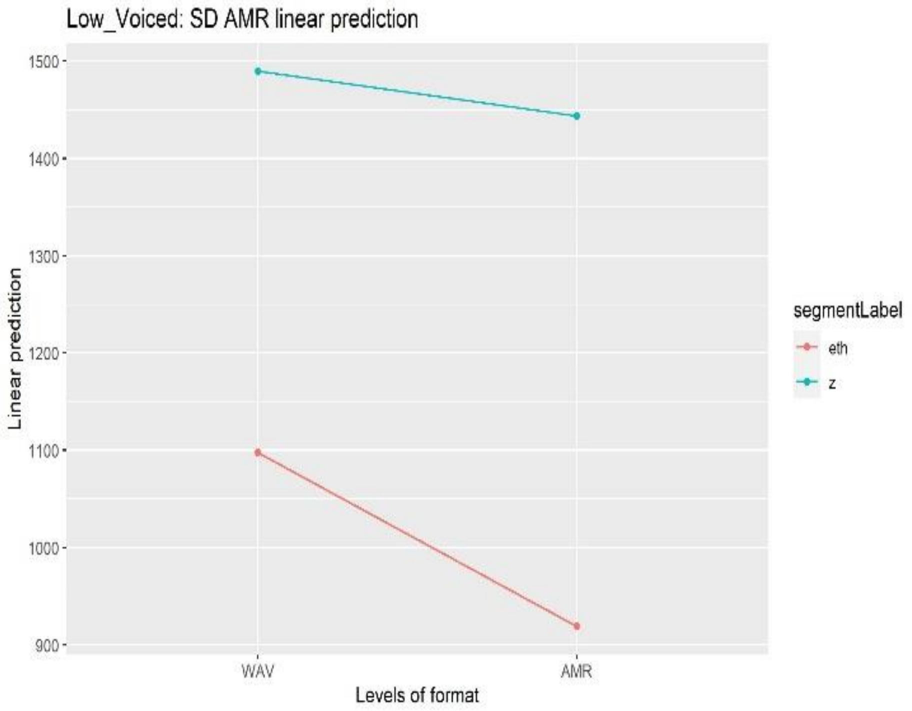


Figure 4.6. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

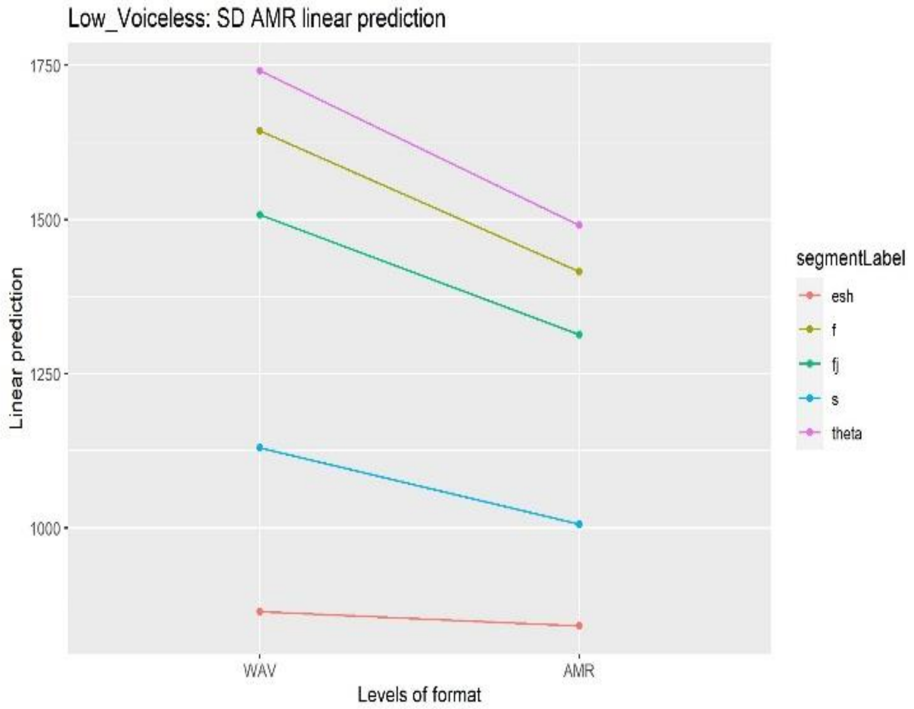


Figure 4.7. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

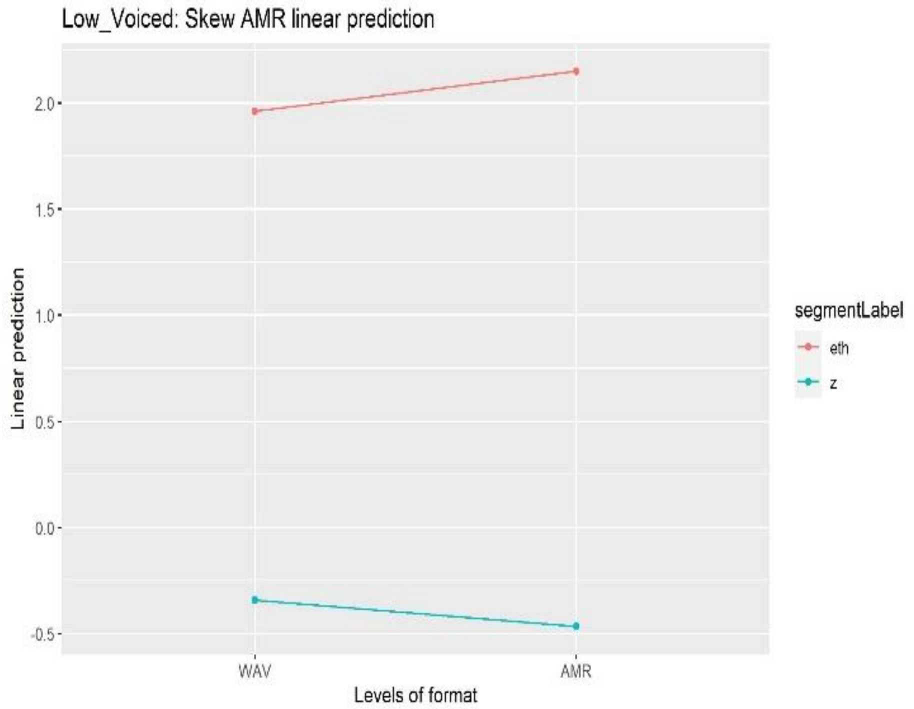


Figure 4.8. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

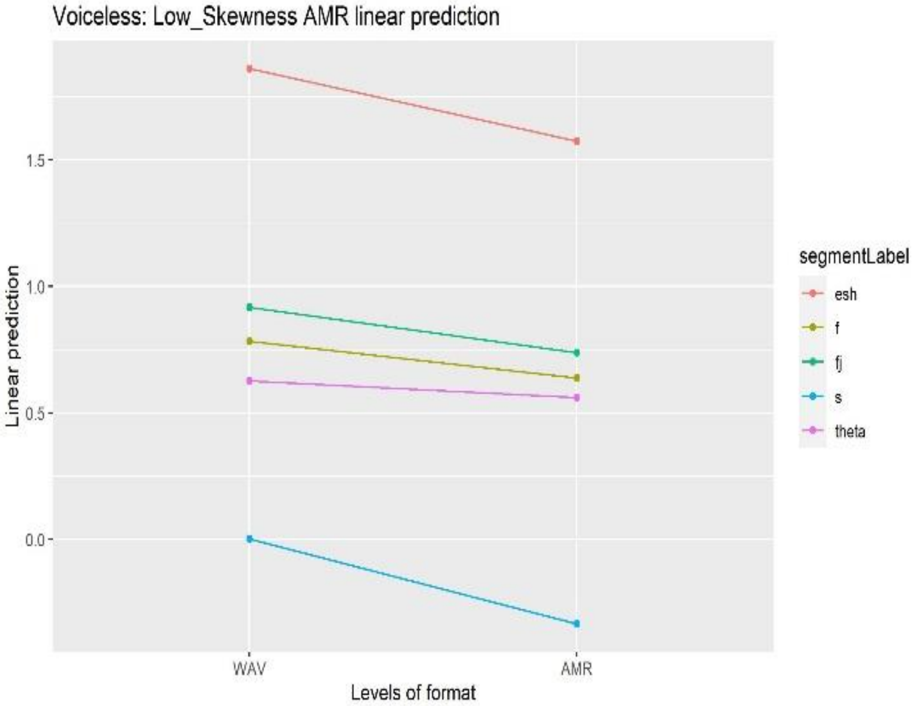


Figure 4.9. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate AMR-WB from the mixed effects models for skewness and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The distributions are illustrated below in a set of violin plots for each spectral measure. Again, the specific analysis pertaining to each segment will be found in the following sections.

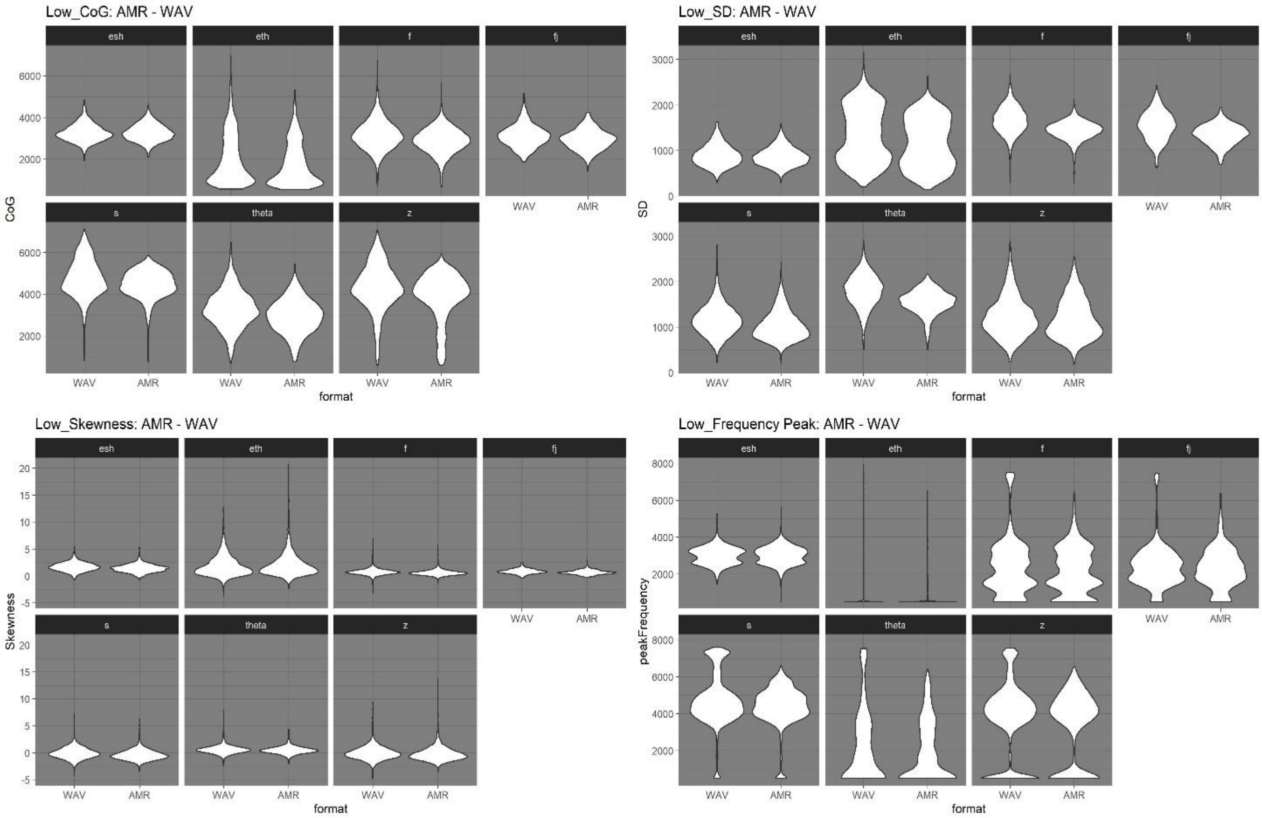


Figure 4.10. Distribution of spectral measure values in WAV baseline and the **low** bitrate AMR-WB codec compression grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

In brief, the plots show how apart from /ʃ/ the spectral measures are lowered for all the segments, while the most prominent changes in distribution are observed for SD.

4.1.1.1.1 Low bitrate: /f/

For /f/ all mean values for the spectral measures are lowered in the AMR-WB compression. This is also evident from the trajectories of the linear predictions, where it can also be seen how /f/ and /θ/ become identical in terms of CoG and more similar for skewness. From the violin plots, it can be observed that the main effects appear for CoG and especially SD as well as the higher frequency content. In addition, the AMR-WB compression tend to centre the frequency content around the mean.

Specifically, the mean values for CoG and frequency peak, lowered by 263 Hz and 227 Hz or around 8 to 9 percent following the AMR.WB compression (p <.0001). Despite the change in distribution, the change for SD was a similar decrease of 8 percent and 229 Hz (p<.0001). A similar percentage change was found for kurtosis with a decrease of 0.24, while skewness lowered by 0.14, which amounts to a change of almost 20 percent (p<.0001). All p-values and further statistical results can be found in table 4.7 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/f/	263.60	14.20	<.0001
SD	WAV - AMR	/f/	228.58	6.38	<.0001
Skewness	WAV - AMR	/f/	329.38	9.69	<.0001

Table 4.7. Statistical results of the difference between WAV and AMR based on linear prediction models for /f/

All maximum values in the AMR-WB compression follow the same pattern as the previous results with all spectral measures lowered. The biggest decrease was found for SD, which changed from 2,664 Hz in the WAV baseline to 2,119 Hz in the AMR-WB compression. CoG and frequency peak presented changes in maximum values of around 400 to 500 Hz, whereas the skewness maximum was lowered by around 1.

A typical spectrographic representation of /f/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.11. This shows a reduction in intensity and frication across the frequency range and particularly around the upper frequency limit, where the reduction is not equal across the segment.

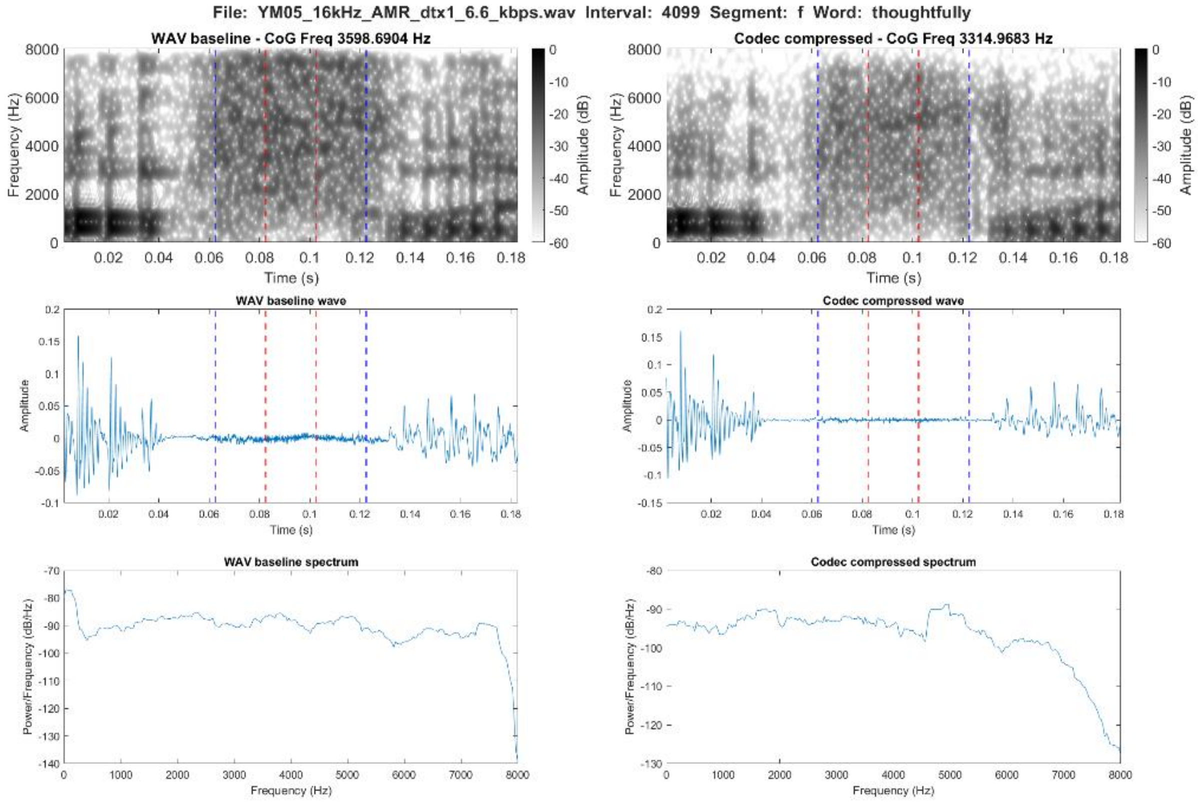


Figure 4.11. Spectrographic comparison of /f/ in the word *thoughtfully* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.2 Low bitrate: /θ/

Similar to /f/ the main effects for /θ/ are found in the distributions of CoG and SD as well as the high frequency content evident from the content removed in the comparison between the two frequency peak distributions. Again, more content is centred around the mean for CoG and SD, while little effects can be observed for skewness.

Specifically, all spectral measures lowered for /θ/ following the AMR-WB compression, which is also evident from the linear predictions despite the change for skewness being limited. This is roughly for all measures by 11 to 14 percent. In detail, this was a change of 350 Hz for CoG (p<.0001), and 251 Hz for SD (p=0.03). For skewness the change was only 0.07, while kurtosis presented a lowering of 5.46. Lastly, the frequency peak lowered from 2,282 Hz to 1,979 Hz. All p-values and further statistical results can be found in table 4.8 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/θ/	349.69	23.12	<.0001
SD	WAV - AMR	/θ/	22.76	10.44	0.03
Skewness	WAV - AMR	/θ/	228.58	6.38	<.0001

Table 4.8. Statistical results of the difference between WAV and AMR based on linear prediction models for /θ/

All maximum values lowered by the AMR-WB compression. This was from 6,754 Hz to 5,686 Hz for CoG, and from 2,664 Hz to 2,119 Hz for SD. The maximum value for the frequency peak was lowered by exactly 1 kHz to 6,500 Hz, while skewness decreased by 1.02 from 6.89 to 5.87.

A typical spectrographic representation of /θ/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.12. The spectrogram and waveform reveal a substantial decrease in intensity and frication across the frequency range with an upper cut-off limit just under 7 kHz. In addition, in the final part of the segment a plosive like structure appears following the codec compression.

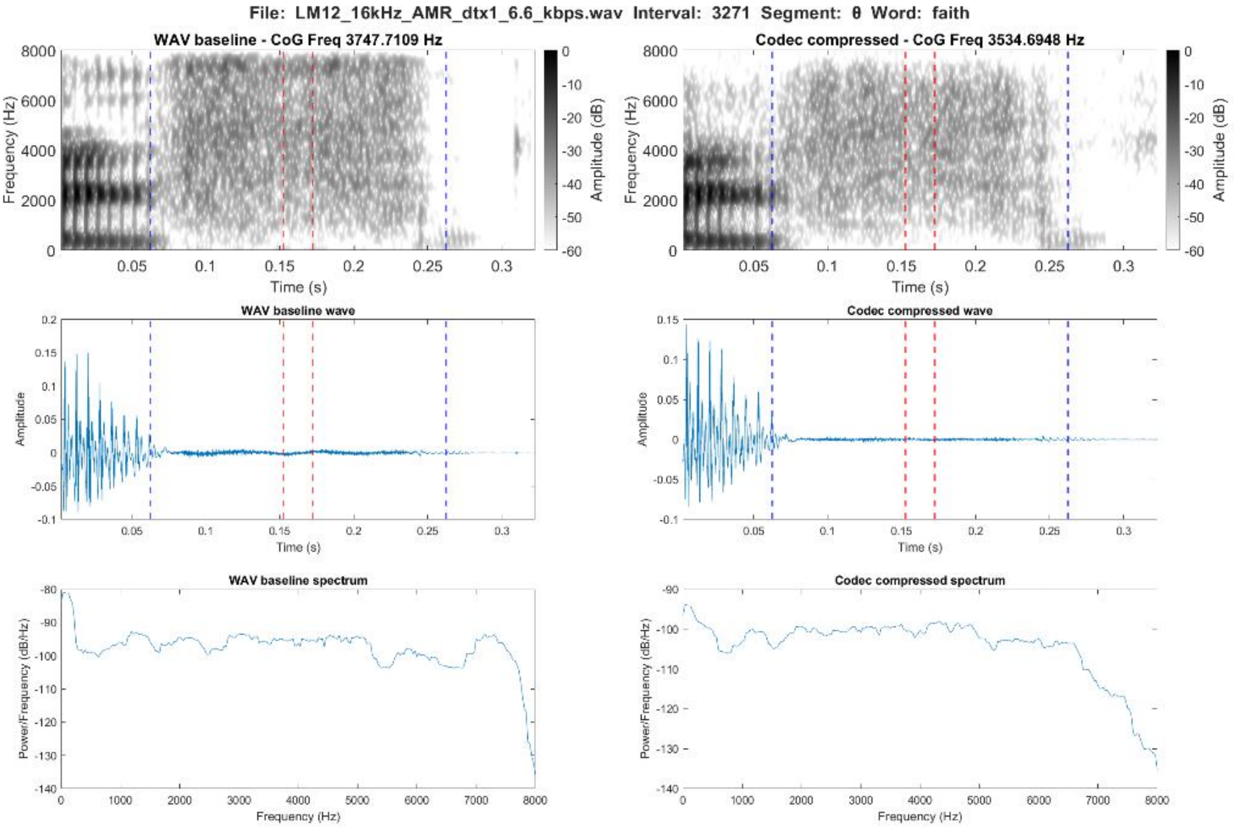


Figure 4.12. Spectrographic comparison of /θ/ in the word *faith* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.3 Low bitrate: [fʲ]

For [fʲ] the distribution of CoG and SD can be seen to lower with more content below the mean, and more values centred around the mean. This is very similar to what is observed for /f/ and /θ/, which is also true for skewness, where very little change is observable. For frequency peak, some of the higher frequency content is again removed, while the distribution is smoothed in shape and more values appear at the lower end of the spectrum.

The linear prediction plots from the mixed effects models show a downwards trajectory for both CoG, SD, and skewness. This is most clear for SD. All spectral measures were lowered for [fʲ] in terms of mean values. This can be seen as changes to CoG and SD by just below 6 percent (p<.001). This meant changes of 185 Hz and 194 Hz respectively. Skewness showed a slightly bigger effect by around 20 percent or 0.18 (p<.01), while kurtosis decreased by 0.36 or just above 9 percent. For frequency peak, the change was limited to only 4 percent, which was a change from 2,408 Hz to 2,306 Hz. All p-values and further statistical results can be found in table 4.9 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	[fʲ]	185.07	53.57	0.0005
SD	WAV - AMR	[fʲ]	194.41	24.06	<.0001
Skewness	WAV - AMR	[fʲ]	0.18	0.06	0.006

Table 4.9. Statistical results of the difference between WAV and AMR based on linear prediction models for /θ/

Again, the maximum values largely followed the pattern of the mean values and lowered following the codec compression. This was from 5,150 Hz to 4,242 Hz for CoG, from 2,436 Hz to 1,950 Hz for SD, and from 7,469 Hz to 6,375 Hz for frequency peak. The only increase was found for skewness, which increased by 0.5 to 3.01.

A typical spectrographic representation of [fʲ] in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.13. Again, a reduction in intensity can be observed across the segment, but especially in the initial and final part. The relatively more intense and formant like structured frequencies between 2 and 4 kHz are maintained in the codec compression.

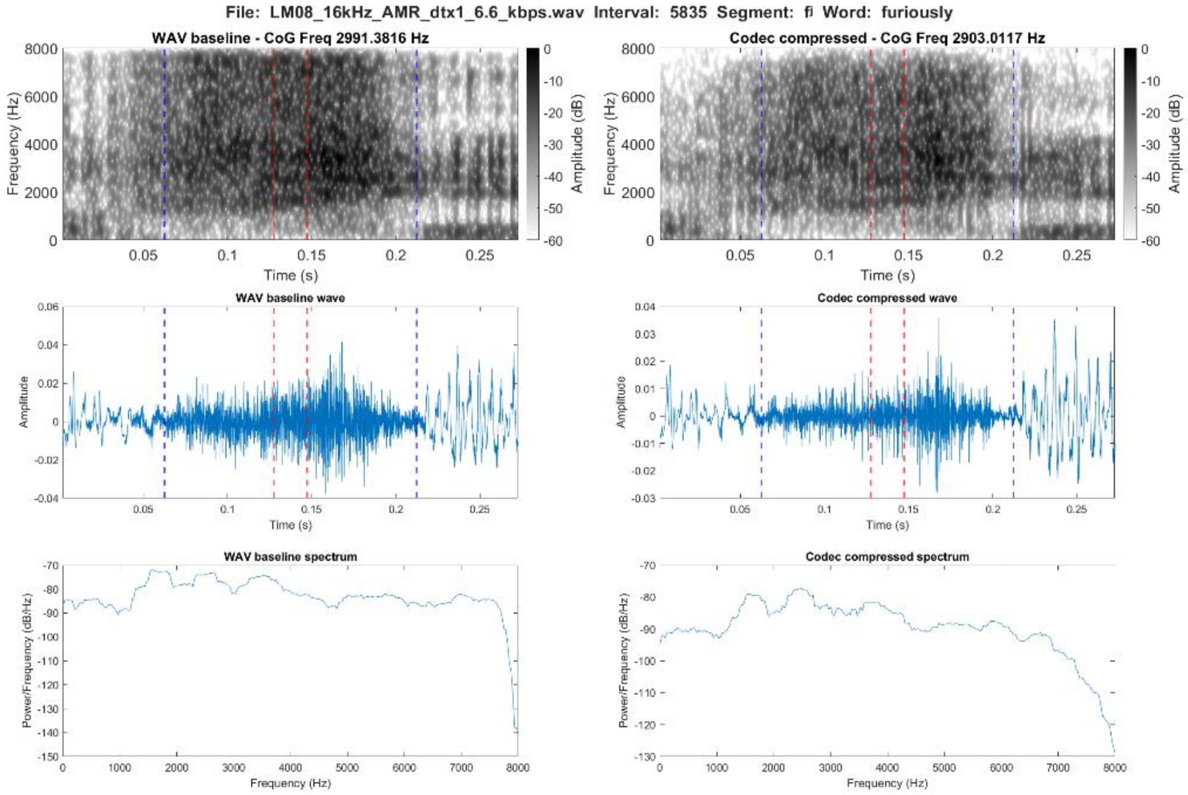


Figure 4.13. Spectrographic comparison of [fʲ] in the word *furiously* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.4 Low bitrate: /s/

The main effects on the distribution of /s/ following the AMR-WB compression is in the higher frequency content above the mean, which appears to be moved lower in the spectrum. In addition, a clear lowering can be seen for SD, where the values centre lower than in the WAV baseline. For skewness the effect is again very limited. The distribution of the frequency peak values are again lowered, but the primary changes happen above the mean with little evident changes observable below 4 kHz.

This decrease was also observed from the linear predictions, where all measures show similar downwards trajectories. More specifically, the mean values revealed decreases for all measures apart from kurtosis, which presented a slight increase of 0.19 or just above 3 percent. For CoG, SD and frequency peak, the changes caused by the AMR-WB compression were around 7 percent to just above 10 percent (p<.0001). For CoG and frequency peak this was a change around of 330 Hz, while it only appeared as a decrease in mean value for SD of 124 Hz. Skewness decreased by 0.34 (p=0.03). All p-values and further statistical results can be found in table 4.10 below

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/s/	329.38	9.69	<.0001
SD	WAV - AMR	/s/	349.69	23.12	<.0001
Skewness	WAV - AMR	/s/	22.76	10.44	0.03

Table 4.10. Statistical results of the difference between WAV and AMR based on linear prediction models for /s/

For /s/ all maximum values decreased following the AMR-WB compression. This was especially true for CoG, which lowered from 7,137 Hz in the WAV baseline to 5,857 Hz in the codec compression. Slightly smaller decreases were found for SD and frequency peak by changes of a little less than 500 Hz for SD to 2,437 Hz, and a change of 31 Hz for frequency peak to 5,625 Hz.

A typical spectrographic representation of /s/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.14. As for previous segments, a reduction in intensity can be observed across the segment, but particularly for /s/ in the final part. The frequency towards the upper frequency limit appear less reduced than what has been observed for the other voiceless segments.

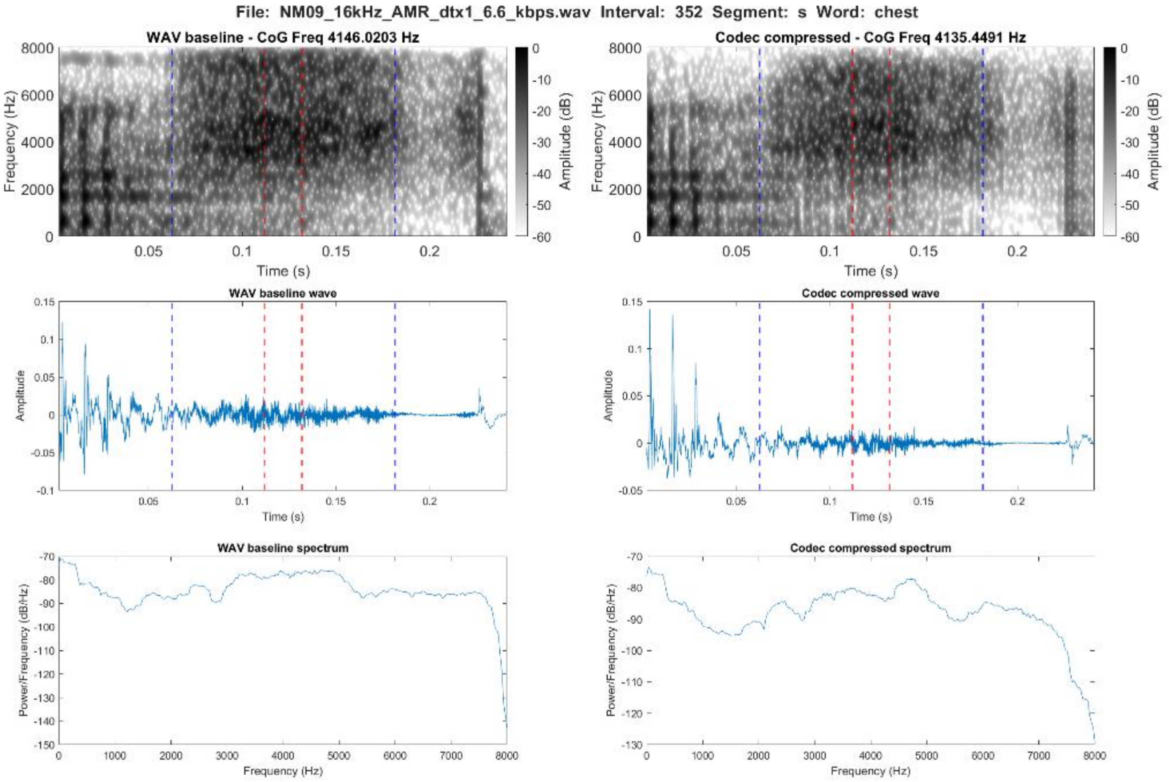


Figure 4.14. Spectrographic comparison of /s/ in the word *chest* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.5 Low bitrate: /ʃ/

Overall, the distributional effects on /ʃ/ are very limited. Small changes was observed e.g. minor decreases for all measures, except for an increase in the more extreme values for frequency peak. This observation is substantiated by the linear prediction plots, where the changes to /ʃ/ following the codec compression is very slight and e.g. SD remains largely unchanged.

The mean values and the changes to these showed that /ʃ/ was the only segment that presented increases of CoG and frequency peak. However, these increases were not significant at only 3 Hz and less than 1 percent for CoG, and 9 Hz for frequency peak or 0.3 percent. This limited effect of the AMR-WB compression can also be observed for SD, which decreased significantly by 23 Hz or 2.5 percent (p=0.03), while skewness followed previous findings and decreased by just below 17 percent (p<.0001). This was a change from 9.45 to 7.79. All p-values and further statistical results can be found in table 4.11.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/ʃ/	-2.77	23.24	0.9
SD	WAV - AMR	/ʃ/	22.76	10.44	0.03
Skewness	WAV - AMR	/ʃ/	0.29	0.03	<.0001

Table 4.11. Statistical results of the difference between WAV and AMR based on linear prediction models for /ʃ/

Despite the slight increases found for the mean values, all maximum values decreased for /ʃ/ in the AMR-WB compression. These decreases were all less than 400 Hz and for skewness only 0.07. Specifically, CoG lowered from 4,862 Hz to 4,678 Hz, SD from 2815 Hz to 2,437 Hz, and lastly the frequency peak lowered from 5,281 Hz to 5,625 Hz.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.15. The spectrogram and waveform reveal a reduction in intensity and frication across the segment and frequency range. The most intense frequency band around 3 kHz is also reduced and loses its distinctive shape in the codec compression.

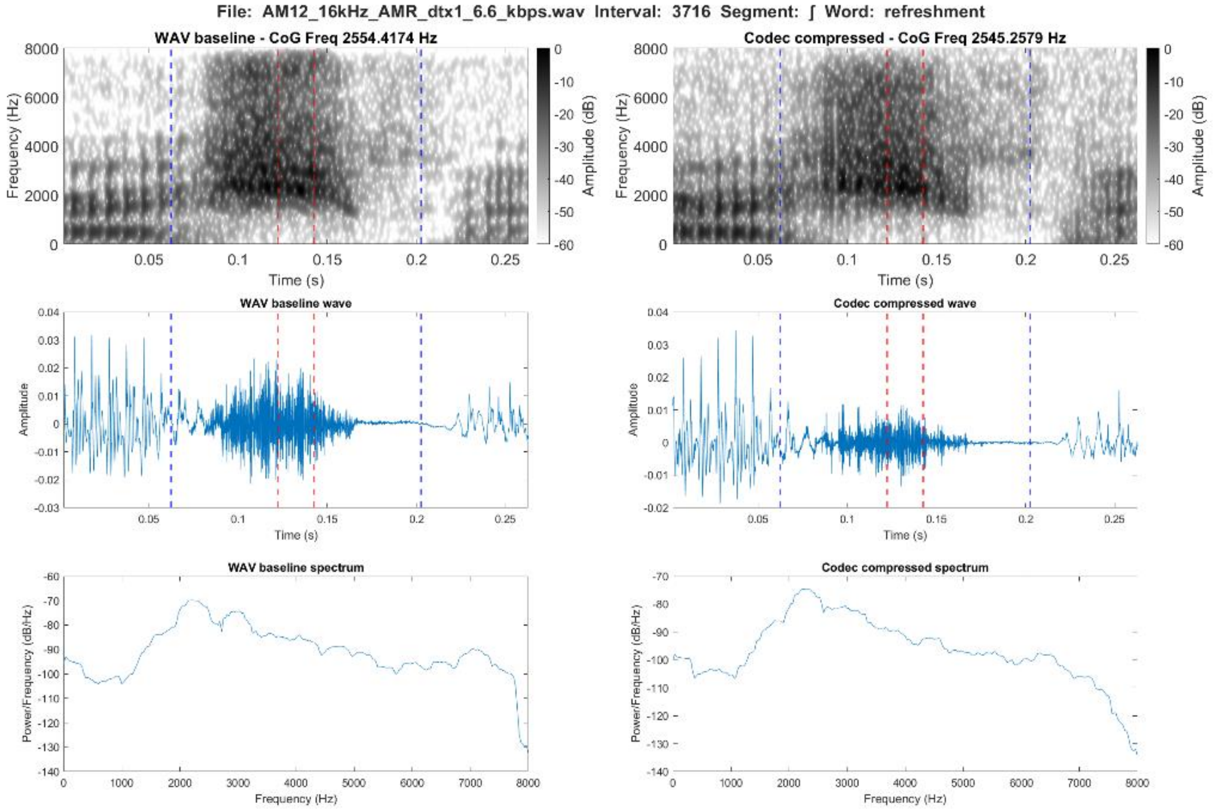


Figure 4.15. Spectrographic comparison of /ʃ/ in the word *refreshment* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.6 Low bitrate: /ð/

A decrease of CoG and SD can be observed from the distribution plots, where both measures have more values appear in the lower part of the spectrum. Opposite previous segments, an increase is visible for skewness, where higher values above 15 appear following the codec compression. The distribution of frequency peak is difficult to interpret as most values were centred around the lower frequency limit. However, a decrease in the higher values and an increase in the lower values can be observed. Again, the linear predictions follow these observations with downwards trends for CoG and especially SD, while skewness presents a slight upwards trajectory

In more detail, CoG, SD and frequency peak decreased by just below 13 percent to just below 16 percent. This was a change of CoG by 254 Hz, 178 Hz for SD, and 167 Hz for frequency peak. Skewness and kurtosis both increased. Skewness increased by 0.19, which equals a 10 percent change (p<.0001). For kurtosis the change was 15 percent, or from 11.91 to 13.69. All p-values and further statistical results can be found in table 4.12 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/ð/	253.79	19.73	<.0001
SD	WAV - AMR	/ð/	178.11	9.12	<.0001
Skewness	WAV - AMR	/ð/	-0.19	0.03	<.0001

Table 4.12. Statistical results of the difference between WAV and AMR based on linear prediction models for /ð/

The maximum values showed a similar pattern to the mean values with decreases of CoG, SD and peak frequency, but an increase of skewness. Specifically, this was decreases of CoG from 6,963 Hz to 5,321 Hz, of SD from 3,161 Hz to 2,638 Hz, and of frequency peak from 7,938 Hz to 6,531 Hz. Skewness showed the biggest change observed yet in maximum values of almost 7, which meant a value of 20.70 in the AMR-WB compression.

A typical spectrographic representation of /ð/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.16. The illustration here is of the plosive production as this was the most prevalent in the randomly generated examples. The spectrogram show a substantial reduction in the primary burst at the final part of the segment, while the frequency information are inserted towards the lower part of the spectrum, where the initial part of the burst occur. In addition, the frequencies ranging from 6 to 8 kHz are also substantially reduced in intensity.

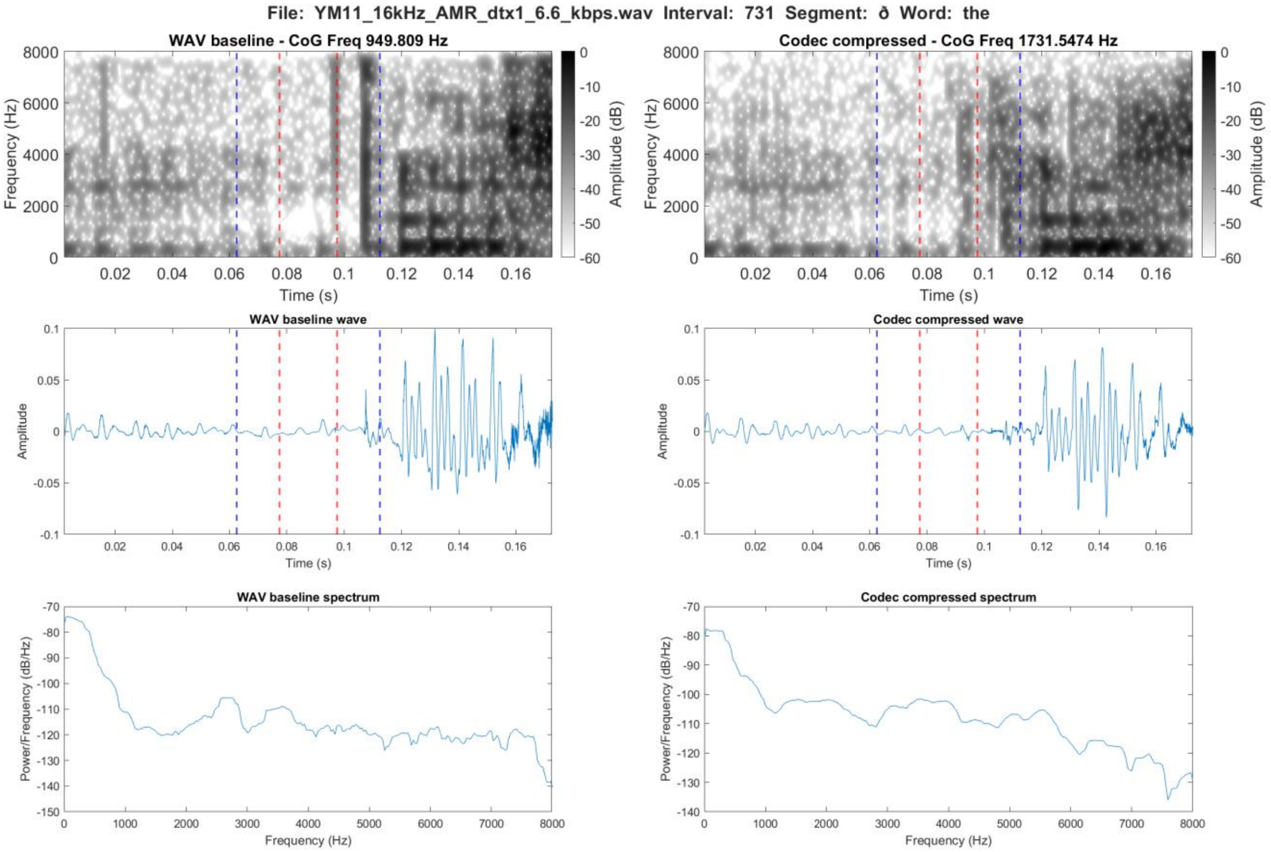


Figure 4.16. Spectrographic comparison of plosive /ð/ in the word *with* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.1.7 Low bitrate: /z/

Decreases can be observed for all spectral measures in the distribution of /z/ in the AMR-WB compression. For CoG and frequency peak this effect does not appear to change the mean substantially, while a clear lowering of the main portion of values can be seen to be lowered for SD. For skewness a number of higher values appear following the codec compression, while a slight lowering can be observed for the remaining values. The linear predictions also indicate the general downwards trajectory observed in the distribution plots. However, only a slight change is observed for SD and skewness in comparison to CoG.

The mean values showed decreases of all spectral measures apart from kurtosis, which presented a very slight increase of 0.01, which was a change of less than 1 percent. The decrease for SD was by 46 Hz or just below 4 percent, while skewness lowered by 142 percent or 0.12 (p<.0001). The mean values for frequency peak and CoG were more in line with previous segments with a change of almost 15 percent from 3,264 Hz to 3,819 Hz for frequency peak and a decrease by 434 Hz for CoG (p<.0001). All p-values and further statistical results can be found in table 4.13 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/z/	434.39	19.59	2.96E-107
SD	WAV - AMR	/z/	178.11	9.12	4.13E-07
Skewness	WAV - AMR	/z/	45.86	9.05	0.0001

Table 4.13. Statistical results of the difference between WAV and AMR based on linear prediction models for /z/

Similar to /ð/, decreases in maximum values were found for /z/ for all measures apart from skewness, which increased by 4 to 13.81. The decreases were prominent for CoG and frequency peak, where the former lowered from 7,082 Hz to 5,932 Hz and the latter from 7,563 Hz to 6,531 Hz. SD presented a decrease of just over 500 Hz to a maximum value of 2,638 Hz.

A typical spectrographic representation of voiced /z/ in the WAV baseline and the AMR-WB compression using the low bitrate is presented below in figure 4.17. The spectrogram and waveform show a reduction in intensity particularly in the voiceless frequencies between 4 and 7 kHz, while the voiced frequencies at the lower part of the spectrum are kept intact following transmission.

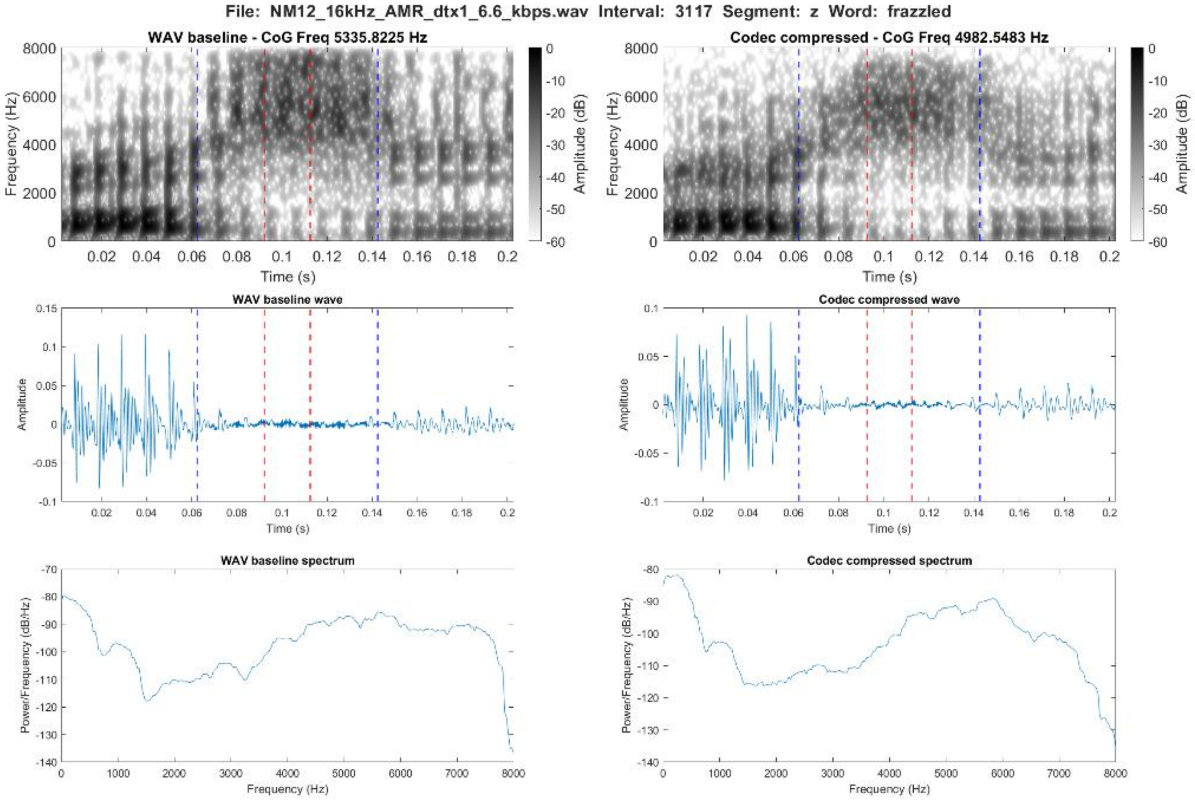


Figure 4.17. Spectrographic comparison of voiced /z/ in the word *with* in the WAV baseline (left) and the AMR-WB compression at 6.6 kbps (right)

4.1.1.2 Low bitrate: MP3

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the MP3 codec in the low quality bitrate dataset.

Firstly, the linear predictions for each spectral measure and the individual segments are presented below (figure 4.18 to 4.23). These indicate the directionality of the changes caused by the MP3 codec using the low bitrate. The graphs illustrate the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these plots are in the following sections for the individual segments.

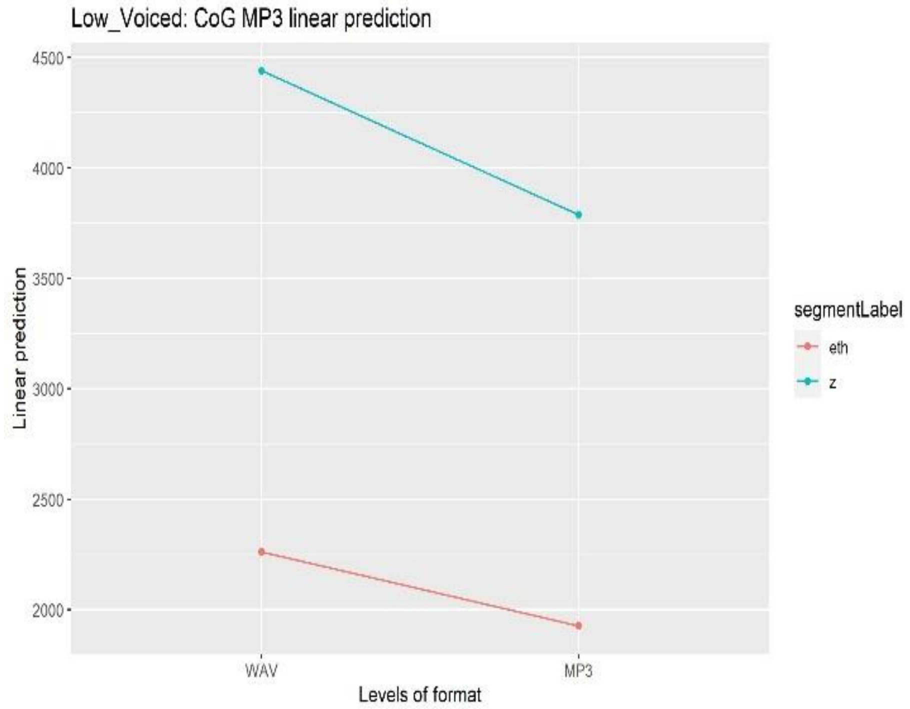


Figure 4.18. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

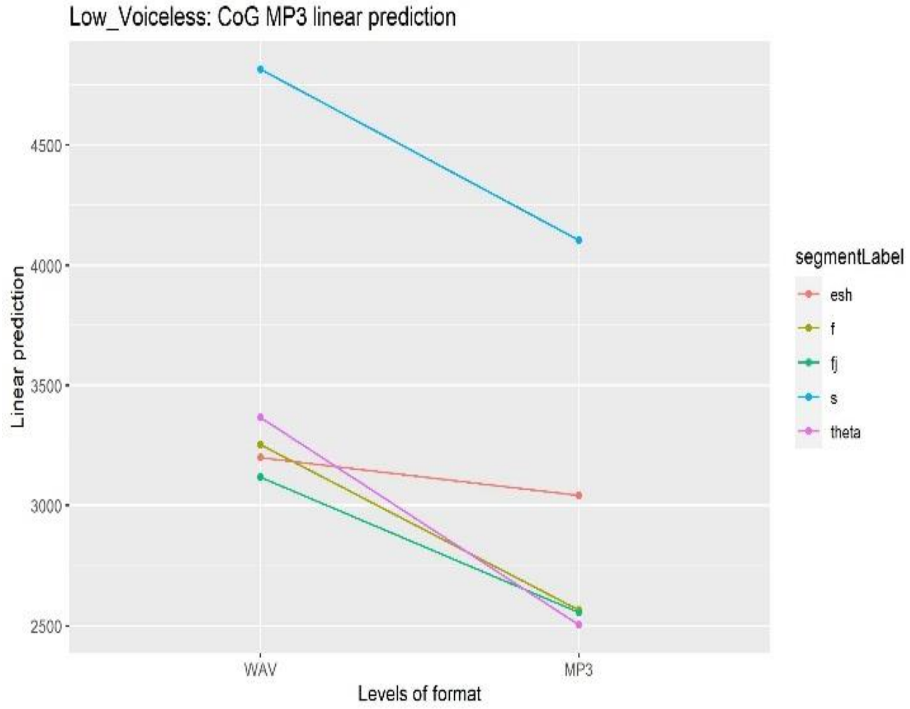


Figure 4.19. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

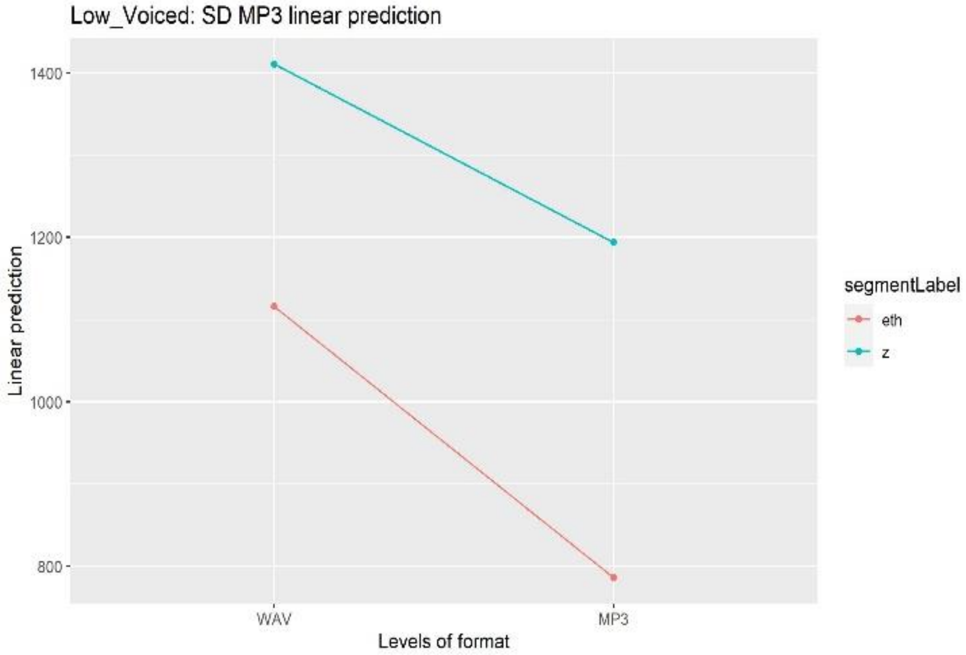


Figure 4.20. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

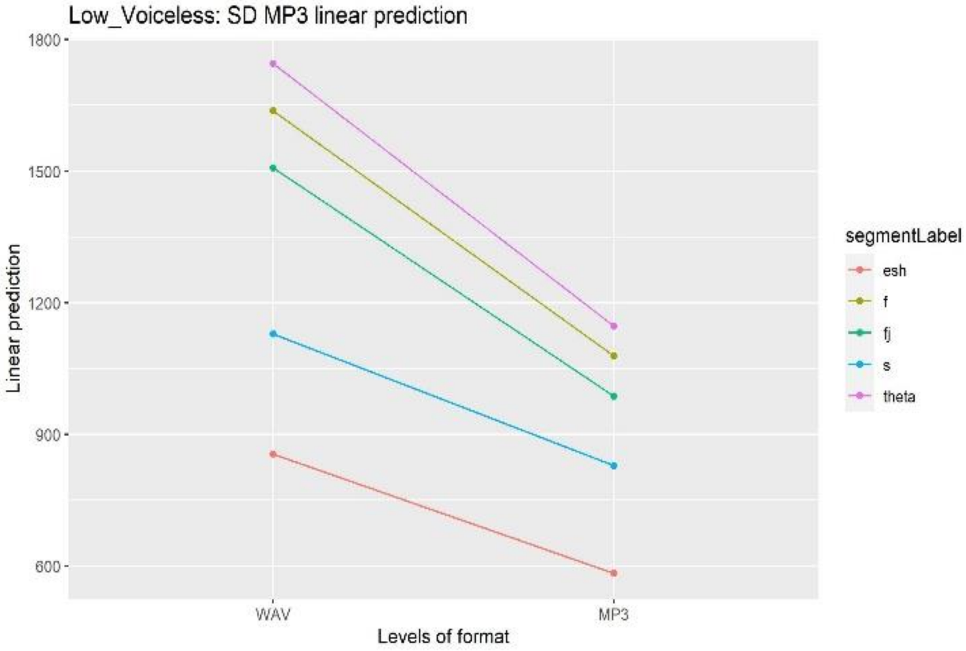


Figure 4.21. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

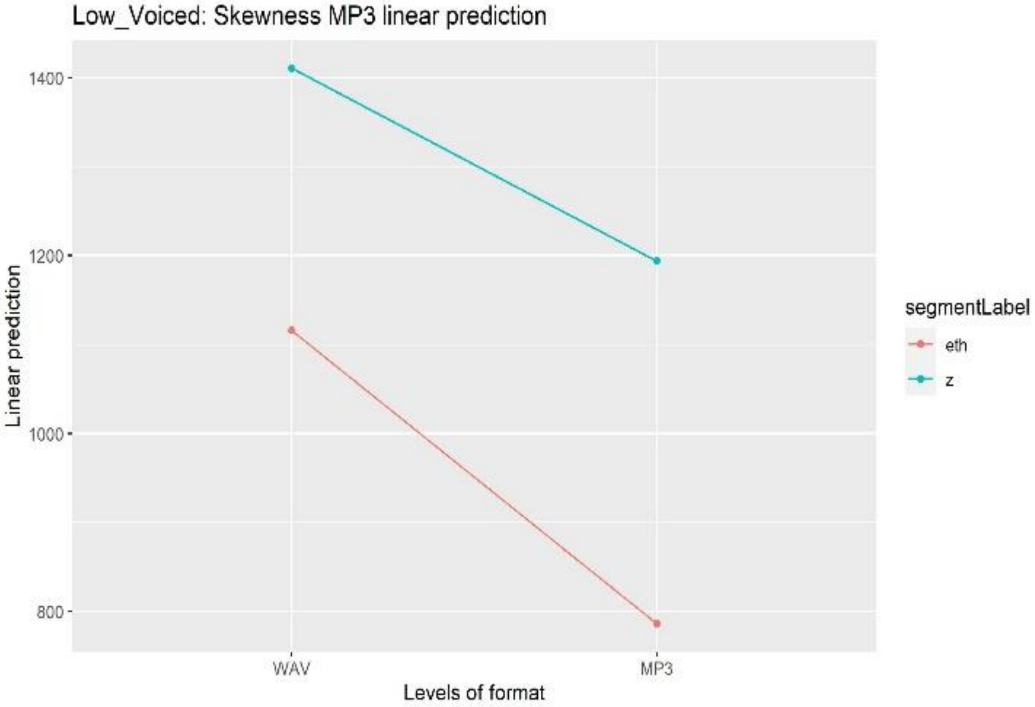


Figure 4.22. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

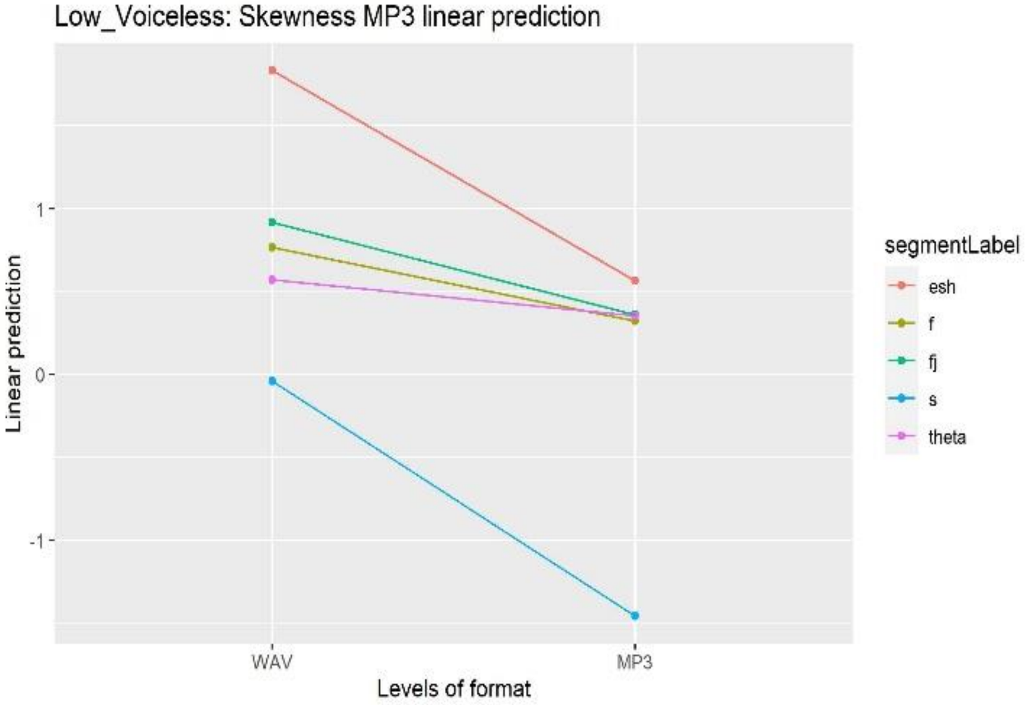


Figure 4.23. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate MP3 from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The distributions for each spectral measure and segment are illustrated below in a set of violin plots. Again, the specific analysis pertaining to each segment will be found in the following sections.

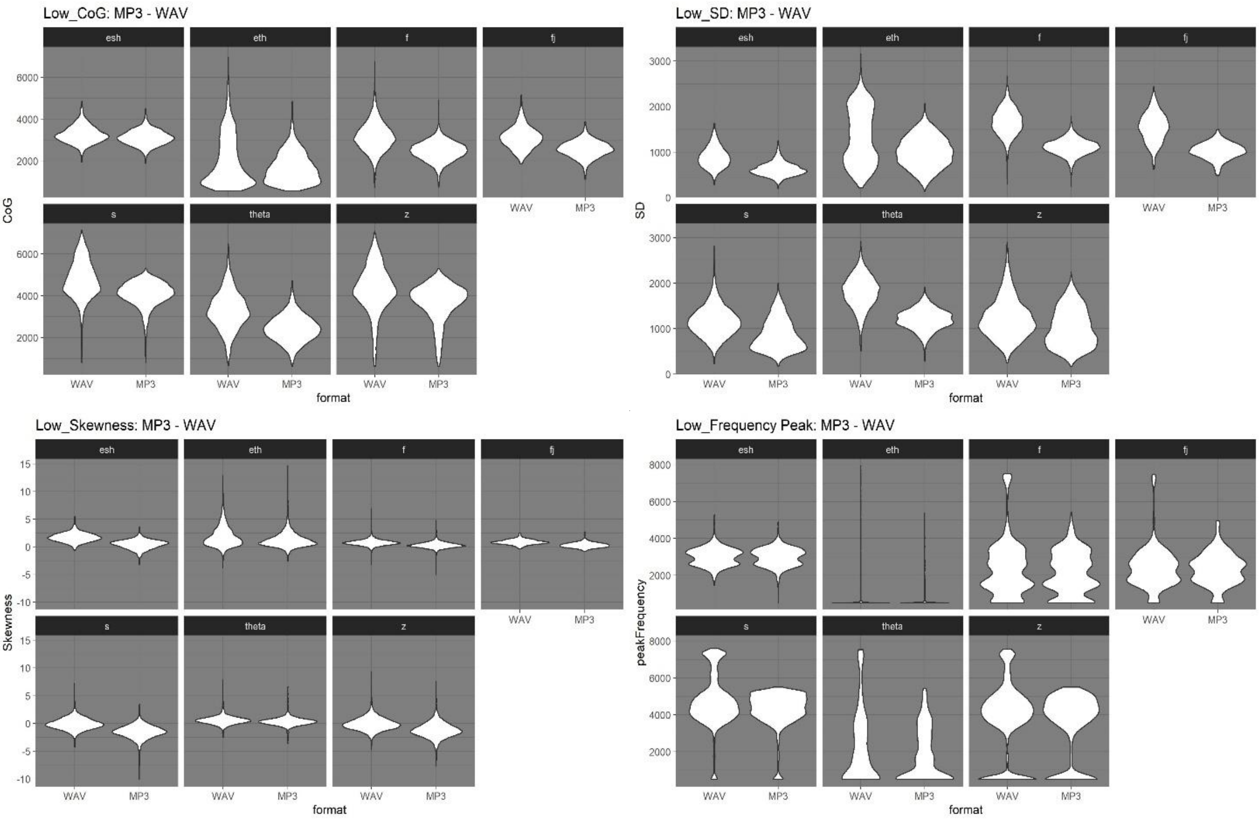


Figure 4.24. Distribution of spectral measure values in WAV baseline and the **low** bitrate MP3 codec compression grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

In sum, the linear predictions and distributions show how all spectral measures were lowered by the MP3 condition, and how particularly SD was affected.

Lastly, all changes to the spectral measures produced by the MP3 codec proved significant with p-values <.0001. Hence, these will not be reported in-text in this section.

4.1.1.2.1 Low bitrate: /f/

For /f/ all mean values for the spectral measures were lowered in the MP3 compression. This was particularly for CoG and SD. For these two measures, /f/ and [fʲ] appeared to become identical following the MP3 compression based on the linear predictions. In addition, for skewness both /f/, [fʲ], and /θ/ appeared to obtain largely the same value in the MP3 compression. From the distribution plots, it was also clear that the MP3 codec compression lowered and altered the distribution of especially CoG and SD. These two measures both became less spread across the spectrum and d around the mean. A decrease was also observed for skewness, while the higher values for the frequency peak were mainly affected by a cut-off at 5-6 kHz.

In more detail, the MP3 codec lowers all mean values for each spectral measure, and this was with up to just over 59 percent. This large change was observed for skewness and amounts to a change of 0.45. For CoG and SD the change was identical in terms of percentage with a change just below 22 percent, or 687 Hz for CoG and 558 Hz for SD. A similar change was found for kurtosis, which decreased by 0.77 or just over 22 percent. The frequency peak was less affected and showed a change of 289 Hz or just below 12 percent. All p-values and further statistical results can be found in table 4.14 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/f/	687.21	48.85	<.0001
SD	WAV – MP3	/f/	558.38	86.24	<.0001
Skewness	WAV – MP3	/f/	0.45	21.87	<.0001

Table 4.14. Statistical results of the difference between WAV and MP3 based on linear prediction models for /f/

As with the mean values, all maximum values for /f/ decreased following the codec compression. This left CoG with a maximum value of 4,920 Hz, SD of 1,789 Hz and frequency peak of 5,406 Hz, which was around 2 kHz lower for CoG and the frequency peak and just under 1 kHz lower for SD. Skewness was likewise lowered from the baseline 6.89 to 4.78.

A typical spectrographic representation of /f/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.25. The spectrogram as well as spectrum show a close to complete non-encoding of frequencies above 6 kHz, while the remaining frequencies appear intensified following the codec compression.

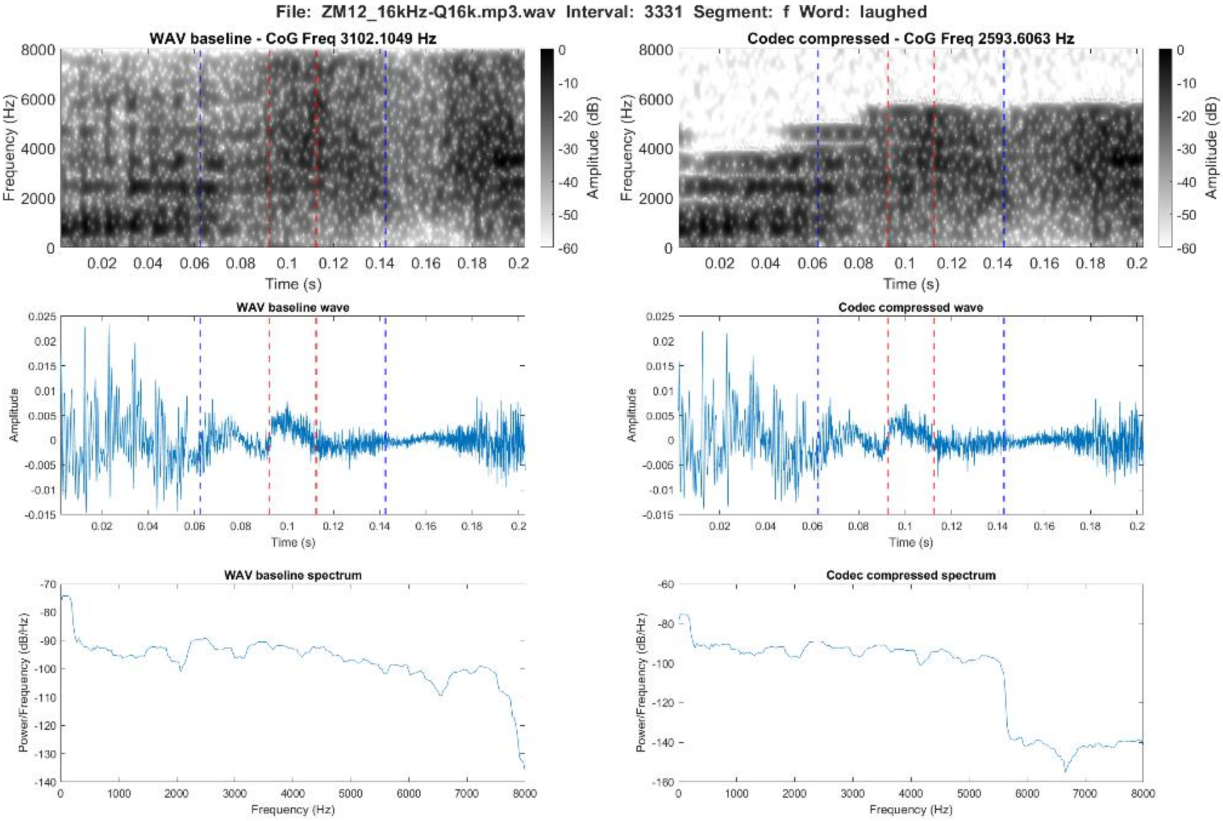


Figure 4.25. Spectrographic comparison of /f/ in the word *laughed* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.2 Low bitrate: [fʲ]

The MP3 compression lowered all spectral measures in comparison to the WAV baseline. This is confirmed from the linear predictions. For [fʲ] the effects on CoG, SD and skewness in terms of distribution is very similar to, what was observed for /f/. The frequency peak is again affected by the cut-off, but here the cut-off is slightly lower and the distribution is more centred in comparison to /f/.

As with /f/ all measures are lowered in terms of mean values. The biggest decrease was again found for skewness, which lowered by 0.56 or just above 64 percent. As with /f/ the effect on CoG and SD were the same percentage, here just above 18 percent. For CoG this was a change of 564 Hz, while for SD it amounted to a change of 520 Hz. Kurtosis lowered by 0.88 or just above 23 percent, while frequency peak stayed almost unaffected with a lowering of just 71 Hz or just below 3 percent. All p-values and further statistical results can be found in table 4.15 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	[fʲ]	564.09	10.63	<.0001
SD	WAV – MP3	[fʲ]	520.18	21.30	<.0001
Skewness	WAV – MP3	[fʲ]	0.56	7.26	<.0001

Table 4.15. Statistical results of the difference between WAV and MP3 based on linear prediction models for [fʲ]

Apart from skewness, all maximum values were also lowered for [fʲ]. The biggest change are seen for CoG and frequency peak, which decreased from 5,150 Hz to 3,878 Hz for the former and from 7,469 Hz to 4,938 Hz for the latter. SD lowered with just under 1 kHz to 1,491 Hz, while skewness increase to 2.68.

A typical spectrographic representation of [fʲ] in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.26. As observed for /f/, all frequencies above 6 kHz was not encoded by the MP3 at this bitrate, but here the final part of the segment presents and even lower cut-off around 5 kHz. No substantial changes can be observed for the encoded frequencies.

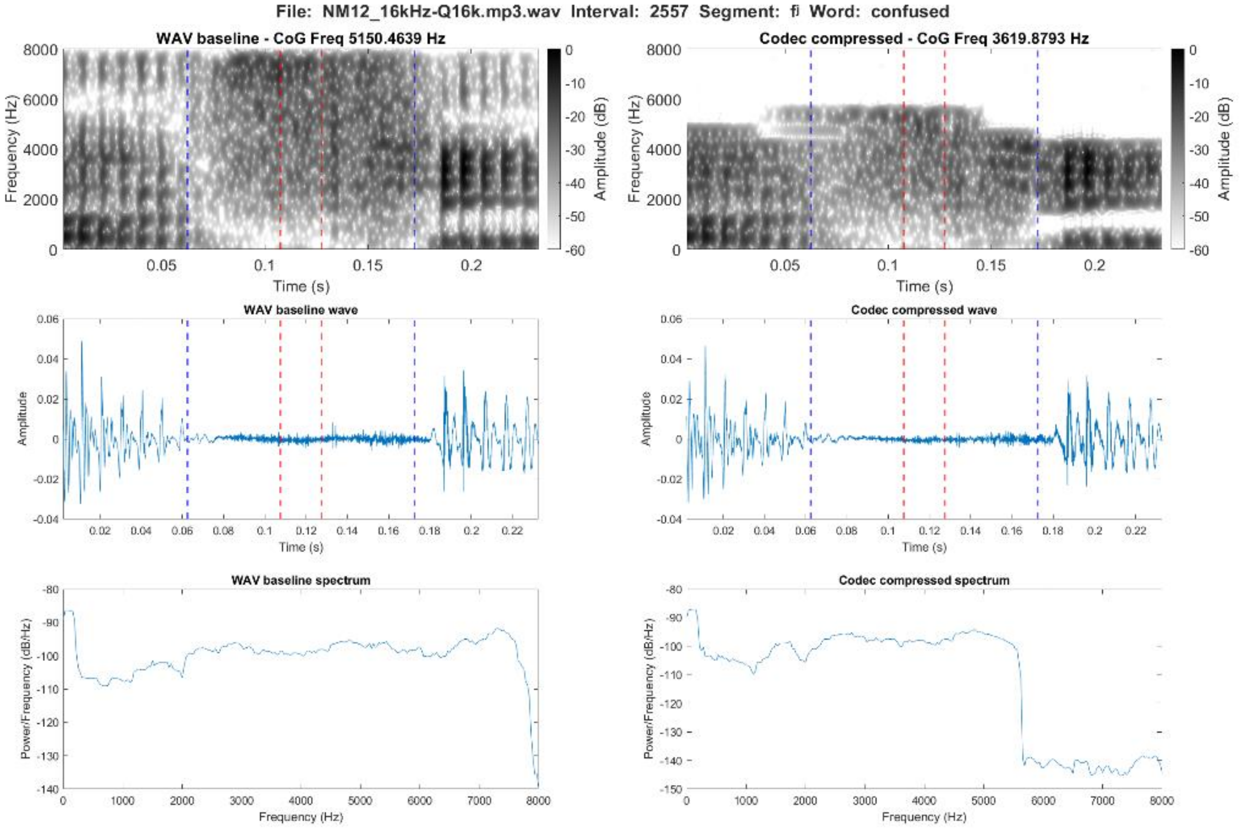


Figure 4.26. Spectrographic comparison of [fʲ] in the word *confused* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.3 Low bitrate: /θ/

For /θ/ a clear lowering and change in distribution can be observed from the violin plots for all spectral measures apart from skewness. This is confirmed by the linear predictions, where the main effects are seen for CoG and SD, while little change is seen for skewness. As with previous segments, the MP3 compression appears to lower the mean and centre the values across a smaller range of values. The higher values above 5-6 kHz in the WAV baseline are again removed, and more values can be sees to appear around the lower cut-off.

Again, the mean values for all spectral measures were lowered by the MP3 compression. For CoG this was by just over 26 percent or 863 Hz, while a slightly bigger effect was seen for SD, which decreased by just over 33 percent or 598 Hz. Skewness presented a lowering of just below 39 percent, while kurtosis changed by just under 14 percent or 0.46. /θ/ showed the greatest effect on frequency peak caused by the MP3 compression with a decrease in mean value by 484 Hz or just over 21 percent. All p-values and further statistical results can be found in table 4.16 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/θ/	862.55	37.66	<.0001
SD	WAV – MP3	/θ/	598.28	56.75	<.0001
Skewness	WAV – MP3	/θ/	0.22	6.50	<.0001

Table 4.16. Statistical results of the difference between WAV and MP3 based on linear prediction models for /θ/

The biggest effect on maximum value for /θ/ was seen for frequency peak, which lowered by just over 2 kHz to 5,438 Hz. CoG lowered from 6,491 Hz to 4,718 Hz, while SD lowered by just over 1 kHz to 1,910 Hz. Lastly, the maximum value for skewness changed from 7.78 to 6.60.

A typical spectrographic representation of /θ/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.27. The spectrographic representation and spectrum show a clear drop in intensity and non-encoding of frequencies above 6 kHz. In addition, drops in amplitude result in bands of frequencies with clear gaps between, this is particular evident between 4 and 5 kHz. These frequencies appear intensified by the codec compression.

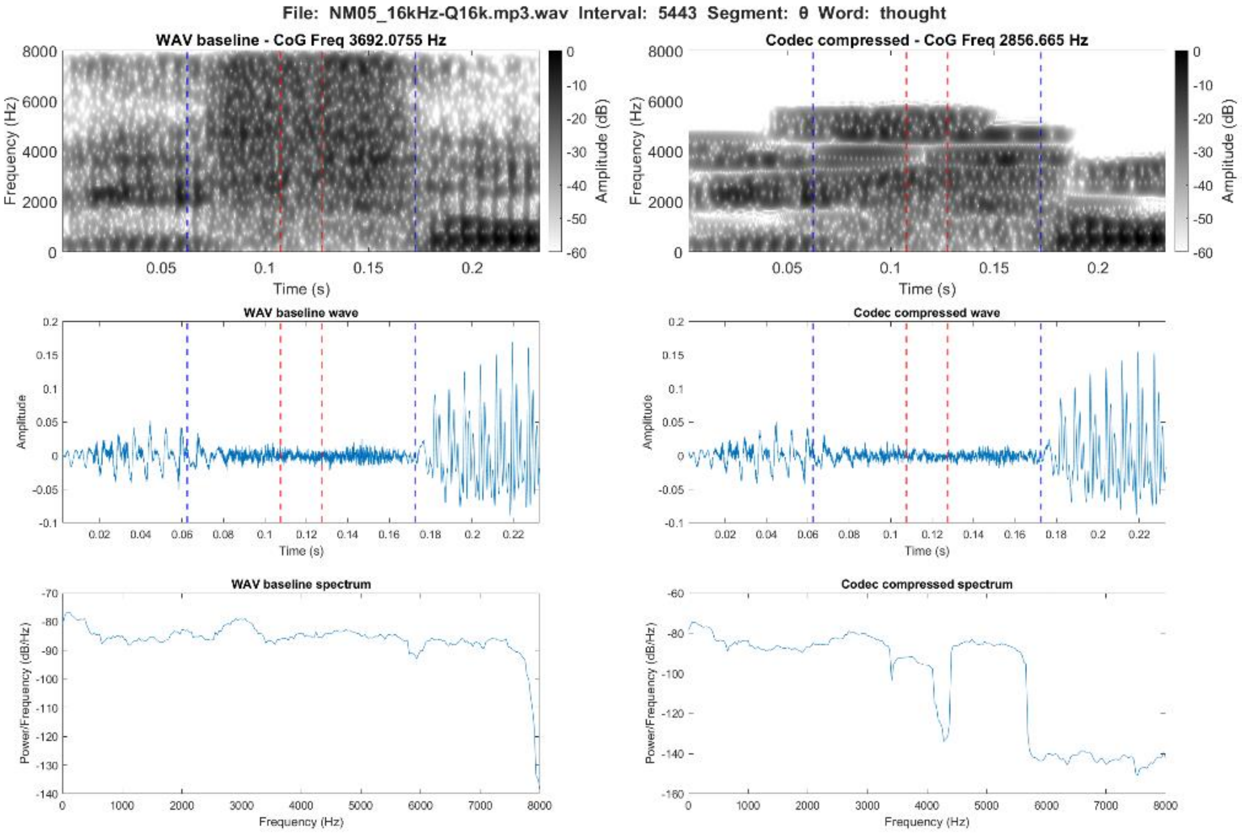


Figure 4.27. Spectrographic comparison of /θ/ in the word *path* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.4 Low bitrate: /s/

For /s/ clear changes to the distribution can be observed from the violin plots. All of which are accompanied by an overall decrease in values. This is especially true for CoG, SD and frequency peak, where CoG behave similar to previous segments, while SD is substantially lowered and presents a distribution more heavily centred in the lower part of the spectrum. For frequency peak, the cut-off is evident, and the plot suggests that a number of frequencies appear just below the cut-off. Skewness is lowered, but in large maintain its distributional shape apart from more tokens appearing at the lower extreme. The linear prediction plots confirm these observations, as all spectral measures show downwards trajectories.

For /s/ the mean values for CoG, SD, skewness and frequency peak all lowered following the codec compression, while kurtosis presented an increase of 2.79 or just below 49 percent. CoG decreased by just below 15 percent, which was a change of 712 Hz. SD followed the findings for the previous segments and showed a decrease by just below 26 percent or 520 Hz, while frequency peak again showed smaller effects. Here a lowering by just below 10 percent. All p-values and further statistical results can be found in table 4.17 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/s/	712.87	74.19	<.0001
SD	WAV – MP3	/s/	299.82	67.85	<.0001
Skewness	WAV – MP3	/s/	1.42	102.02	<.0001

Table 4.17. Statistical results of the difference between WAV and MP3 based on linear prediction models for /s/

Clear decreases in maximum values were again found for /s/. This was by around 2 kHz for CoG and frequency peak, which in the MP3 compression had maximum values of respectively 5,316 Hz and 5,500 Hz. SD lowered by just over 800 Hz to 2,006 Hz, while skewness in the MP3 compression got a value of 3.45 in comparison to 7.33 in the baseline.

A typical spectrographic representation of /s/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.28. As with the previous segments, the upper frequency limit appear just around 6 kHz, while the frequencies in a band just below this limit around 5 kHz is increased in intensity following the codec compression.

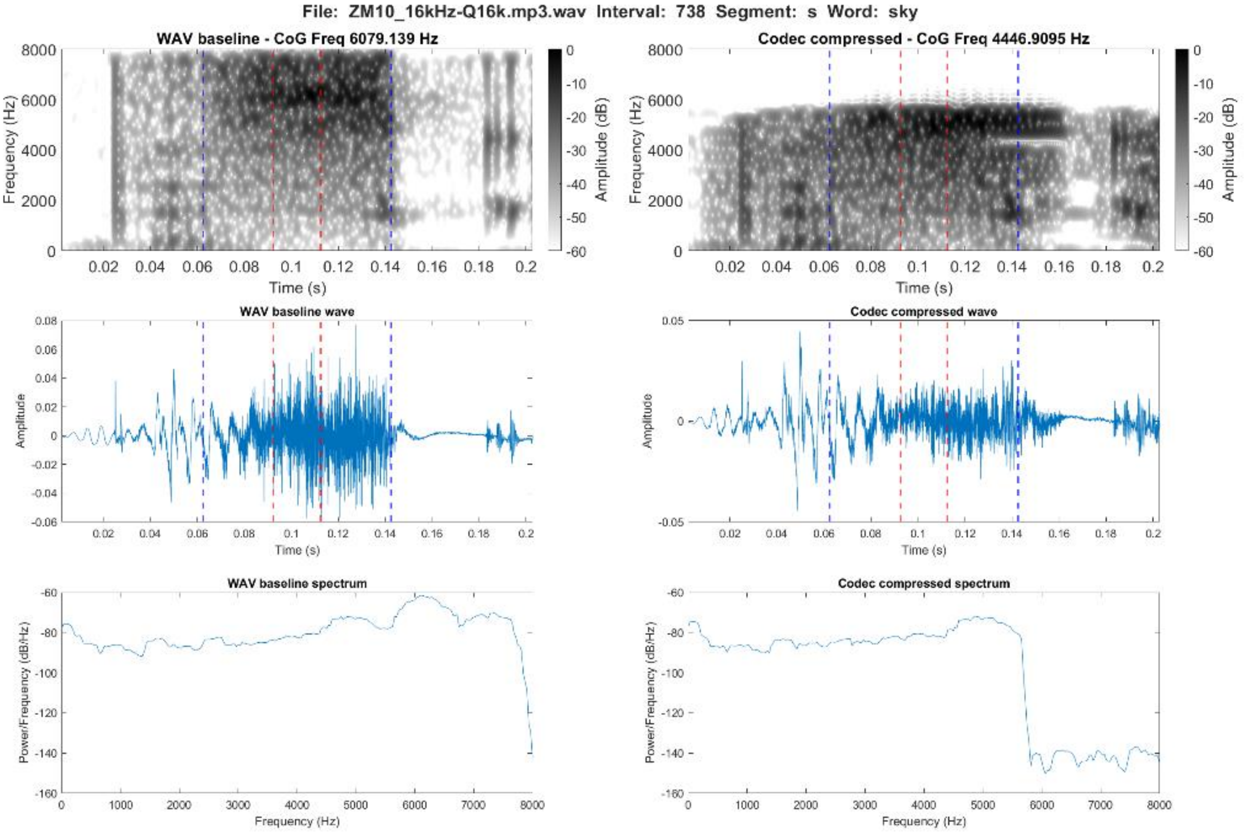


Figure 4.28. Spectrographic comparison of /s/ in the word *sky* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.5 Low bitrate: /ʃ/

The distribution plots indicate the biggest effect of the MP3 codec compression on SD, which is both lowered and change distribution similarly to /f/, [fʲ], and /θ/. For CoG and frequency peak a slight change in the shape of the distribution can be observed with more values below the mean and for frequency peak towards the lower extreme. Skewness presents a clear decrease as well as a slight change to the distribution with more values around the mean. These observations conquer with the linear predictions from the mixed effects models, where SD and skewness can be seen with a clear downwards trajectory, while only a relatively minor decrease can be observed for CoG.

For /ʃ/ CoG and frequency peak showed very limited changes following the MP3 compression. This meant a change of just under 5 percent for CoG or 156 Hz, and a change of just 0.23 percent or 7 Hz for frequency peak. In comparison, skewness showed the biggest effect found for any of the segments with a decrease of almost 74 percent. SD lowered with 271 Hz or just over 30 percent, while kurtosis changed by just under 43 percent. All p-values and further statistical results can be found in table 4.18 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/ʃ/	155.87	6.77	<.0001
SD	WAV – MP3	/ʃ/	271.27	25.60	<.0001
Skewness	WAV – MP3	/ʃ/	1.27	38.03	<.0001

Table 4.18.Statistical results of the difference between WAV and MP3 based on linear prediction models for /ʃ/

Despite all maximum values being lowered, the effects were slightly smaller than for the previous segments. CoG decreased from 4,862 Hz to 4,508 Hz, frequency peak from 5,281 Hz to 4,875 Hz, and SD from 1,628 Hz to 1,251 Hz. Skewness showed a change of just below 2, which meant a maximum value of 3.60.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.29. The spectrogram and spectrum show a drop in frequencies and amplitude just under 6 kHz, while in the final part of the segment the drop happens already around 5 kHz. In contrast, in the initial part of the segment a band of frequencies around 1 kHz is not encoded. The frequencies aligning with the formants in the following vowel appear more intense following the codec compression.

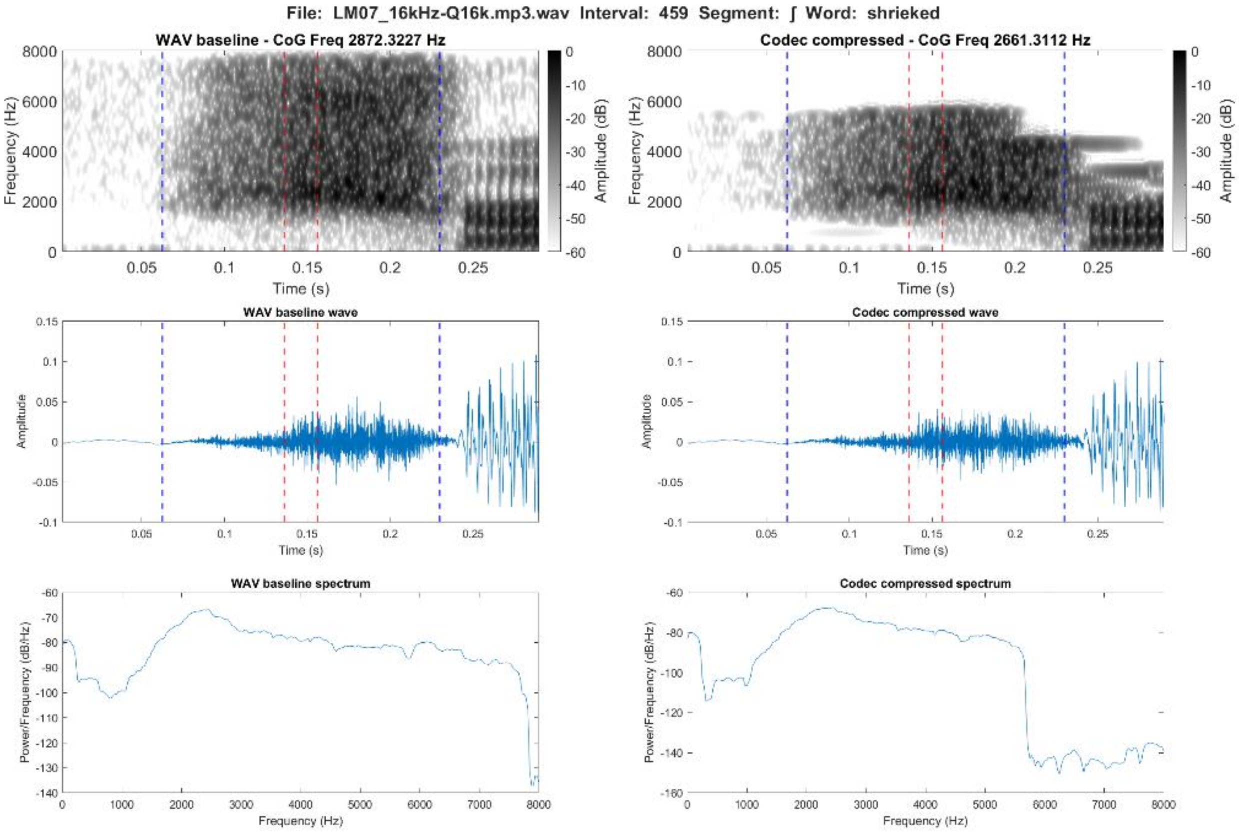


Figure 4.29. Spectrographic comparison of /ʃ/ in the word *shrieked* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.6 Low bitrate: /z/

The main effect for CoG and frequency peak on /z/ in terms of distribution appears to be caused by the upper cut-off inferred by the MP3 compression. For SD and skewness, a lowering can be observed. SD shows some changes in distribution with values more equally distributed, where as the shape of the distribution for skewness shows changes similar to /s/, however, with a an increase in values towards the higher extreme. The downwards trend is confirmed by the linear predictions.

The mean values for /z/ showed relatively smaller changes in comparison the voiceless segments. Apart from kurtosis, all spectral measures again lowered by between just over 10 to just under 18 percent, while kurtosis showed an increase of 1.30 or just over 20 percent. Specifically, CoG increased by just over 15 percent or 652 Hz, while SD lowered with 217 Hz or just under 18 percent. Skewness decreased with 1.15, and frequency peak again showed a change around 10 percent or 398 Hz. All p-values and further statistical results can be found in table 4.19 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/z/	651.99	35.61	<.0001
SD	WAV – MP3	/z/	217.05	25.28	<.0001
Skewness	WAV – MP3	/z/	1.14	38.30	<.0001

Table 4.19.Statistical results of the difference between WAV and MP3 based on linear prediction models for /z/

Similar to some of the voiceless segments, the maximum value for CoG and frequency peak changed with around 2 kHz, which meant maximum values at 5,301 Hz for CoG and 5,500 Hz for frequency peak. SD showed a smaller decrease from 2,901 Hz to 2,241 Hz, while skewness decreased from 9.36 to 7.64 in maximum value.

A typical spectrographic representation of voiceless /z/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.30. From the spectrogram and spectrum, /z/ is almost unrecognisable in the comparison between the WAV file and the MP3 compressed file. This is again because of an upper-frequency limit just under 6 kHz, as well as a drop in amplitude in the initial and mid sections of the segment just under 2 kHz. The frequencies encoded by the MP3 are intensified.

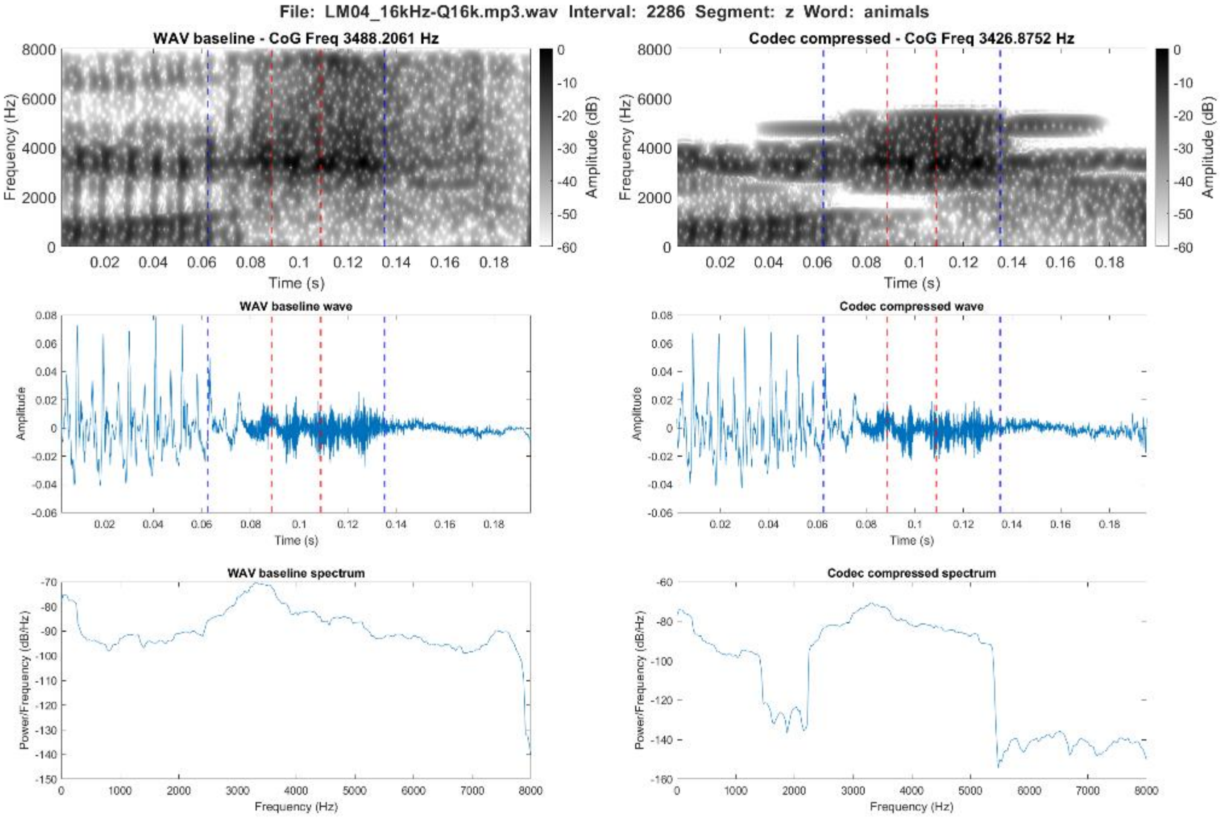


Figure 4.30. Spectrographic comparison of /z/ in the word *animals* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.2.7 Low bitrate: /ð/

Based in the violin plots, some of the clearest changes in distribution is found for /ð/. This is particularly for CoG, SD and skewness, where CoG and SD show decreases and the values spread over a more limited frequency range. For skewness, more values are centred around the mean, but more values also appear at the higher extremes indicating a potential increase following the MP3 compression. Lastly, the main observable effect on frequency peak is caused by the upper cut-off. Overall, the linear predictions indicate that the changes are mainly a downwards trend with the most prevalent changes for SD and skewness.

In more detail, all mean values for /ð/ lowered as a consequence of the MP3 codec compression. This was by around 330 Hz for both CoG and SD or just under 17 percent for the former and just under 25 percent for the latter. Skewness lowered by 0.63 or just under 34 percent, while kurtosis lowered by 5.55 or just under 47 percent. As with previous segments, the effect on frequency peak was less than for the other measures. Here this was a change of 139 Hz or just over 13 percent. All p-values and further statistical results can be found in table 4.20 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – MP3	/ð/	334.26	18.12	<.0001
SD	WAV – MP3	/ð/	329.44	38.08	<.0001
Skewness	WAV – MP3	/ð/	0.63	21.05	<.0001

Table 4.20. Statistical results of the difference between WAV and MP3 based on linear prediction models for /ð/

For /ð/ skewness increased in maximum value, while the remaining spectral measures all decreased. Specifically, skewness increased from 12.87 to 14.64. In comparison, CoG and frequency peak decreased by around 2 kHz to 4,845 Hz for the former and 5,375 Hz for the latter. Lastly, SD lowered from 3,161 Hz to 2,065 Hz.

A typical spectrographic representation of voiced approximant /ð/ in the WAV baseline and the MP3 compression using the low bitrate is presented below in figure 4.31. Again, the upper frequency limit is just under 6 kHz, while a drop in amplitude again occur just over 4 kHz. The frequencies below 4 kHz appear more intense, but with less format structure than in the WAV file.

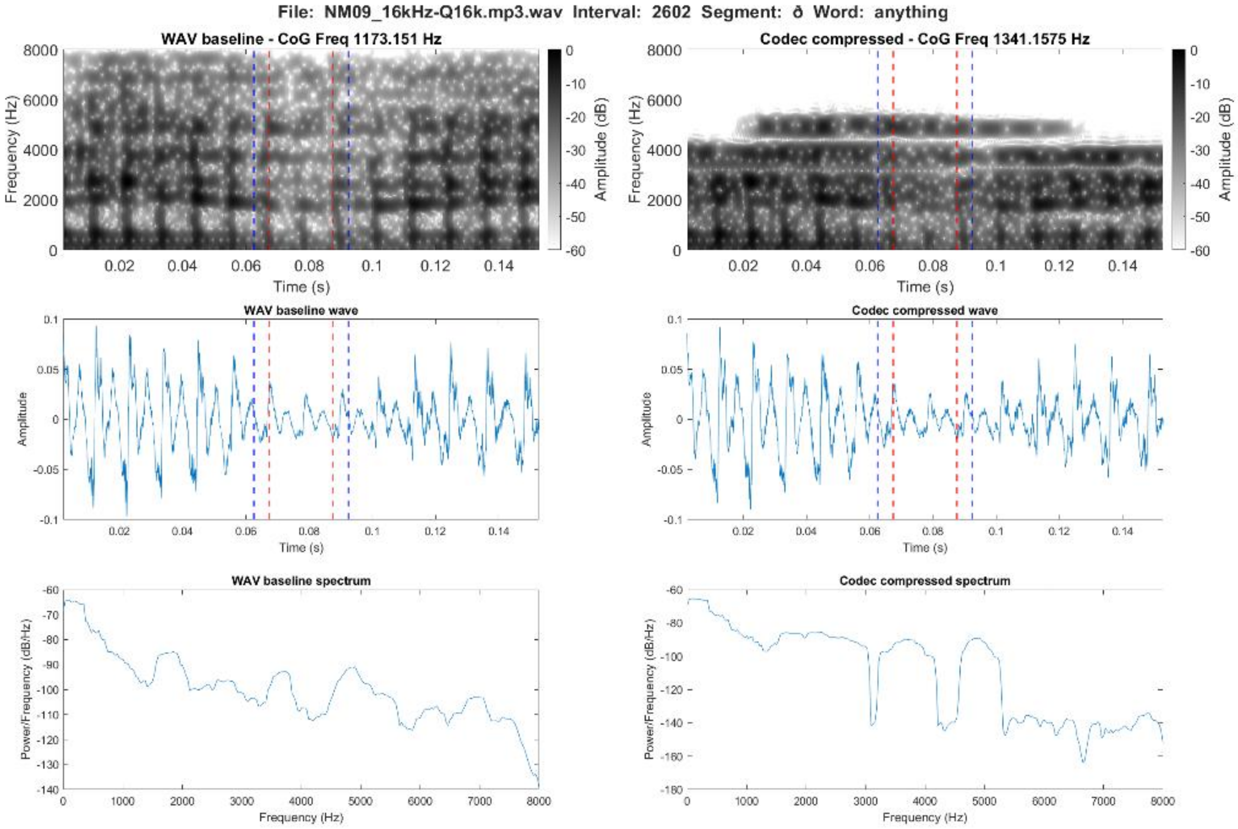


Figure 4.31. Spectrographic comparison of voiced approximant /ð/ in the word *anything* in the WAV baseline (left) and the MP3 compression at 16 kbps (right)

4.1.1.3 Low bitrate: Opus

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the Opus codec in the low quality bitrate dataset.

First, the linear predictions for each spectral measure and the individual segments are presented in figures 4.32 to 4.37 below. These indicate the directionality of the changes inferred by the Opus codec using the low bitrate. The plots illustrate the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these are in the following sections on the individual segments.

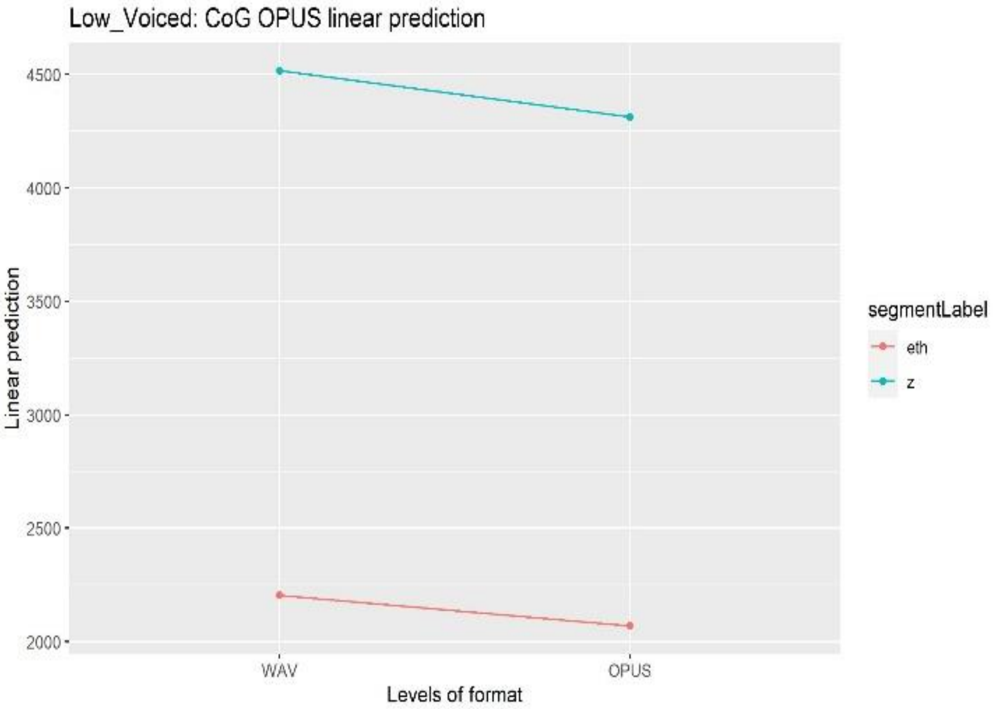


Figure 4.32. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

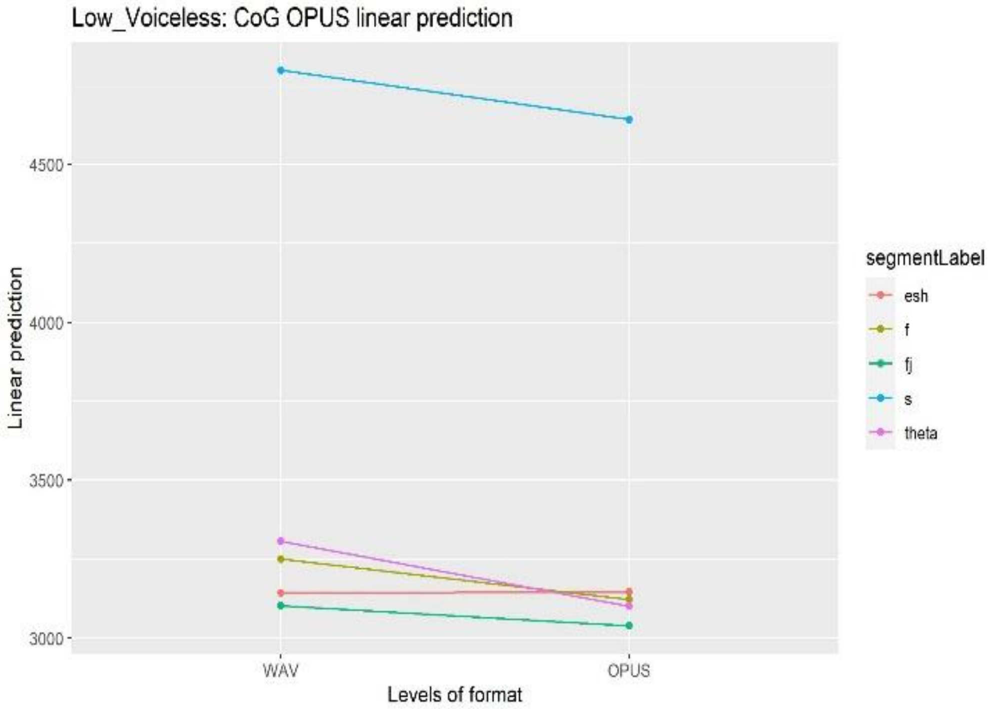


Figure 4.33. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

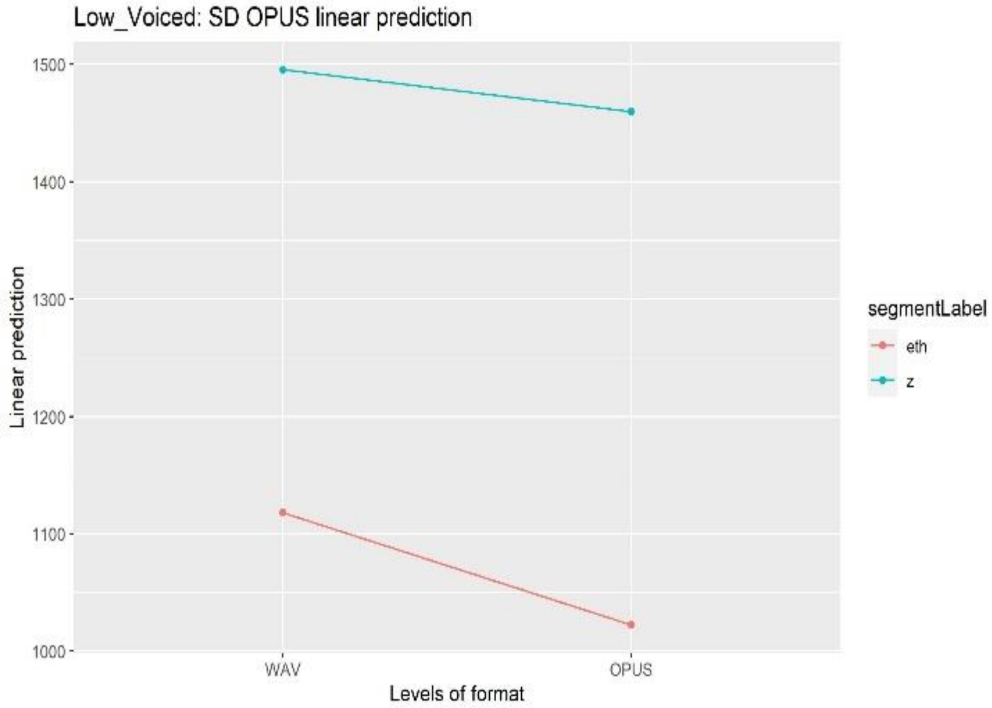


Figure 4.34. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

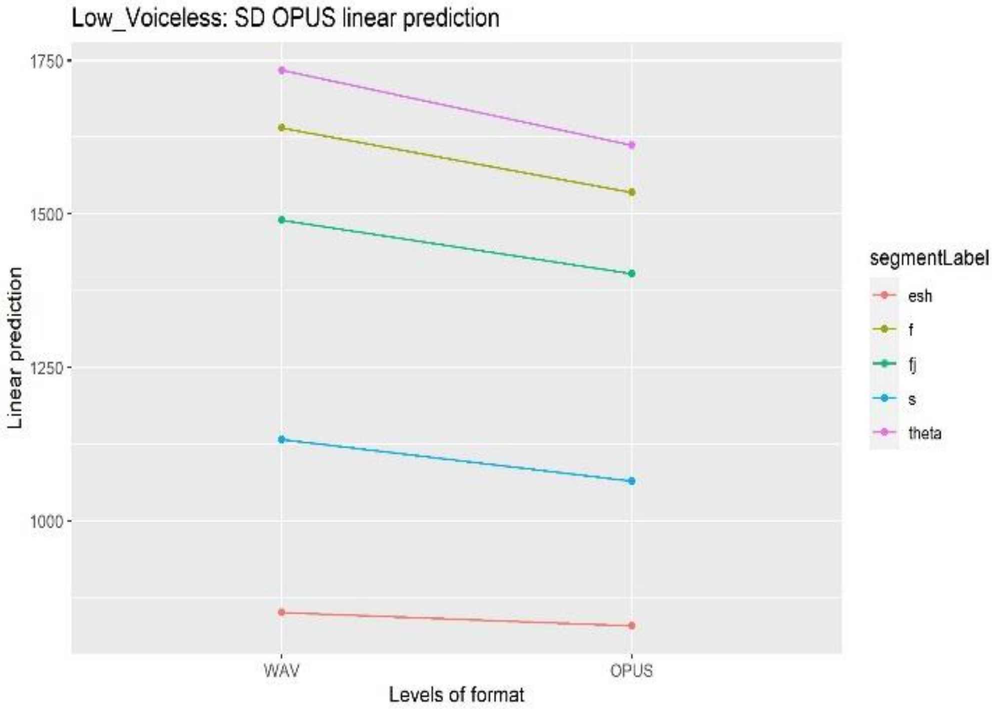


Figure 4.35. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

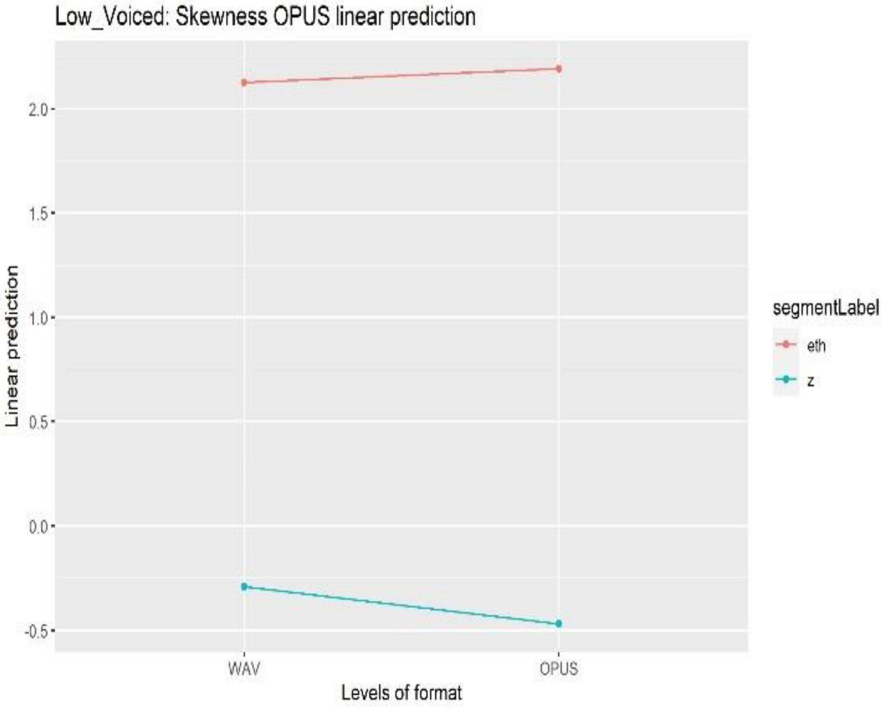


Figure 4.36. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

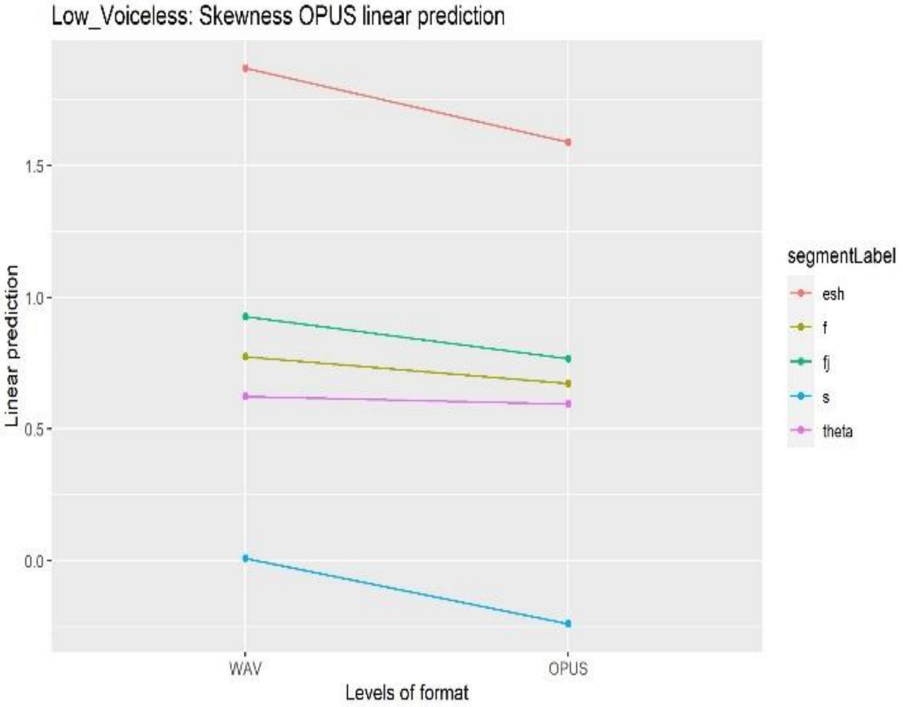


Figure 4.37. Trajectory of the linear predictions in the comparison of WAV and **low** bitrate Opus from the mixed effects models for skewness and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The distributions for each spectral measure and segment are illustrated below in a set of violin plots. Again, the specific analysis pertaining to each segment will be found in the following sections.

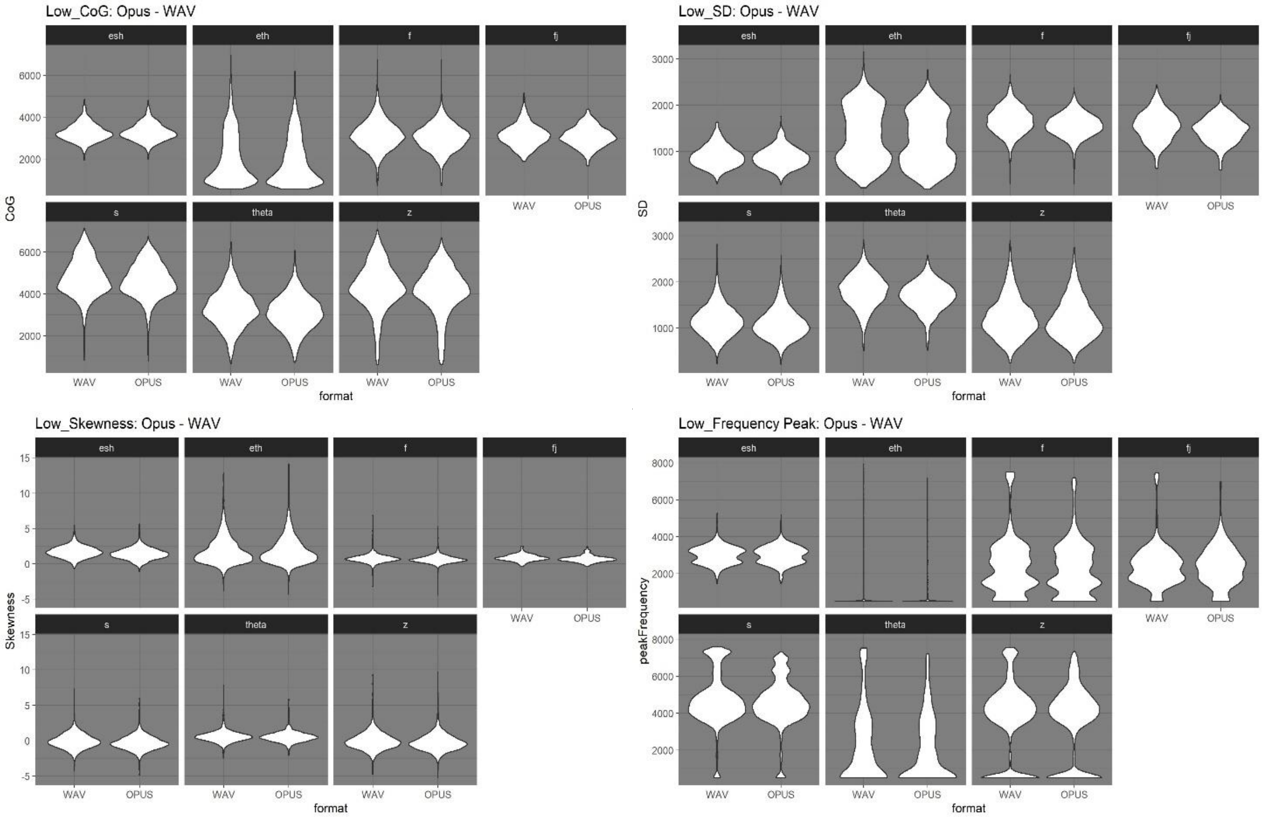


Figure 4.38. Distribution of spectral measure values in WAV baseline and the **low** bitrate Opus codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

In brief, the figures show how the Opus compression generally lowered or kept the spectral measures stable from the original WAV files. The same minor effects appear for the distribution.

4.1.1.3.1 Low bitrate: /f/

In general, the Opus compression has limited effect on the distribution of the spectral measures for /f/. The only notable change is to SD, which like previously presents a slight lowering as well as the value spanning across a more limited set of frequencies. This is also evident from the linear predictions. Here only very slight decreases can be observed, while /f/, /ʃ/, and /θ/ become more alike in terms of CoG.

Regardless, the Opus compression significantly lowered the mean values for all spectral measures for /f/. However, apart from skewness, this was only with between just over 4 and just under 6 percent. This was a lowering of 128 Hz for CoG, and 105 Hz for SD, while the frequency peak was lowered with 145 Hz. Kurtosis was decreased by 0.20. Lastly, skewness was the only measure showing a decrease over 10 percent with a lowering of 0.10. All p-values and further statistical results can be found in table 4.21 below

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/f/	-127.78	-8.66	<.0001
SD	WAV – Opus	/f/	-104.98	-16.04	<.0001
Skewness	WAV – Opus	/f/	0.10	5.76	<.0001

Table 4.21. Statistical results of the difference between WAV and Opus based on linear prediction models for /f/

The maximum values confirmed the pattern for the mean values as all maximum values were likewise lowered following the Opus compression. For CoG this was, however, with less than 10 Hz which gave a maximum value of 6,767 Hz in the Opus compression. SD was lowered from 2,664 Hz to 2,380 Hz, while the frequency peak is lowered by around 300 Hz from 7,500 Hz to 7,188 Hz. Lastly, skewness present a change from 6.89 to 5.30.

A typical spectrographic representation of /f/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.39. Only a slight decrease in intensity can be observed in both spectrogram, waveform and spectrum.

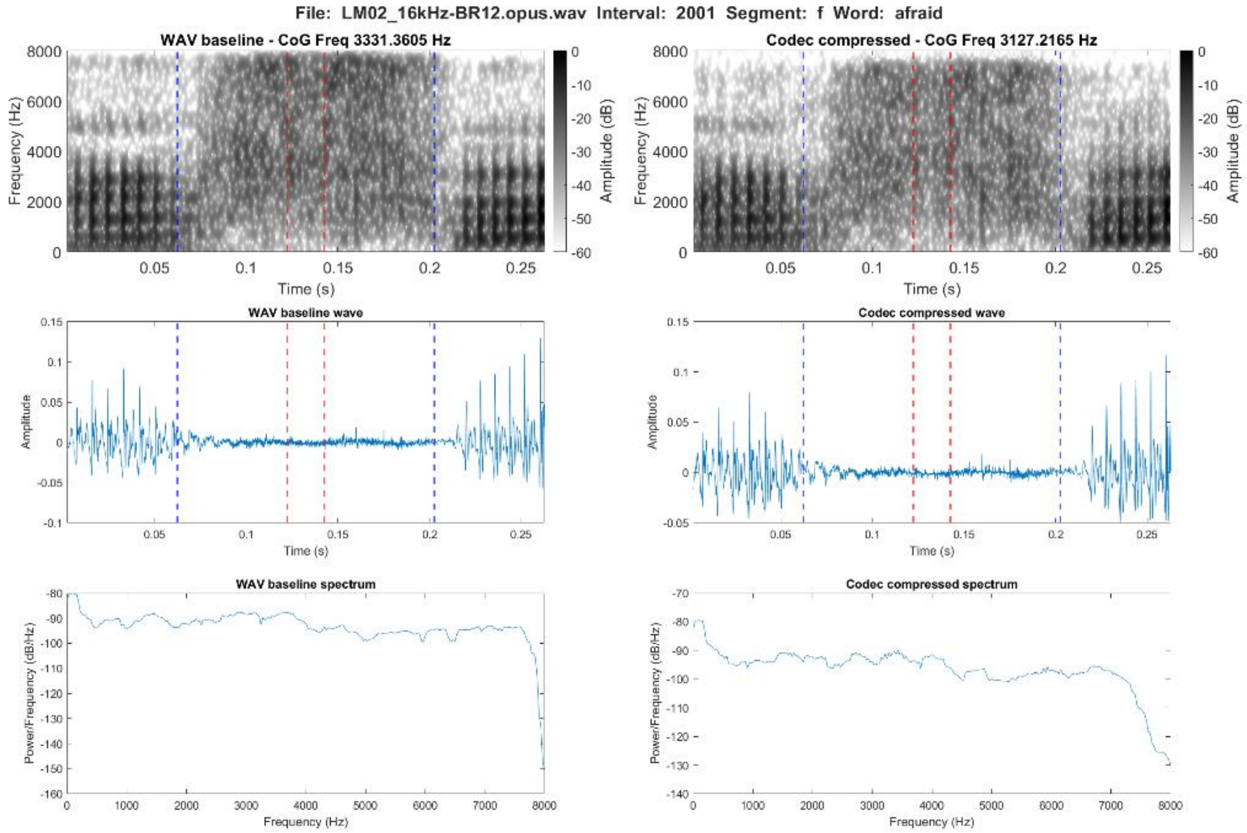


Figure 4.39. Spectrographic comparison of /f/ in the word *afraid* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.2 Low bitrate: [fʲ]

The distribution plots reveal no clear decreases or changes to the distribution shapes for any of the spectral measures. The main noticeable effect on these caused by the Opus compression is to the upper frequencies, which are affected by the upper frequency limit. Again, this is also confirmed by the linear predictions, where only very slight and for CoG almost no changes are observable.

As with /f/ all mean values for the spectral measures were lowered for [fʲ], but not significantly for CoG. For CoG and SD, the change was only 63 HZ for the former, and 88 Hz for the latter, which meant a 2 percent change for both measures. Skewness was again affected comparatively more with a change by just over 18 percent or 0.16, while kurtosis decreased by just over 8 percent or 0.31. The frequency peak stayed almost unchanged by the codec compression with a change of only 18 Hz or less than 1 percent. All p-values and further statistical results can be found in table 4.22 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	[fʲ]	-62.58	-1.12	0.26
SD	WAV – Opus	[fʲ]	-87.76	-3.55	<.0001
Skewness	WAV – Opus	[fʲ]	0.16	2.39	0.02

Table 4.22. Statistical results of the difference between WAV and Opus based on linear prediction models for [fʲ]

All maximum values, were like the mean values, lowered in the Opus compression. The greatest changes were seen for CoG and frequency peak. CoG decreased from 5,150 Hz to 4,382 Hz and frequency peak decreased from 7,469 Hz to 6,969 Hz. SD was lowered by just over 200 Hz to 2,228 Hz, while skewness only changed by 0.05 to 2.47.

A typical spectrographic representation of [fʲ] in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.40. A general reduction in intensity can be observed across the segment as well as a slightly lower upper frequency limit just over 7 kHz.

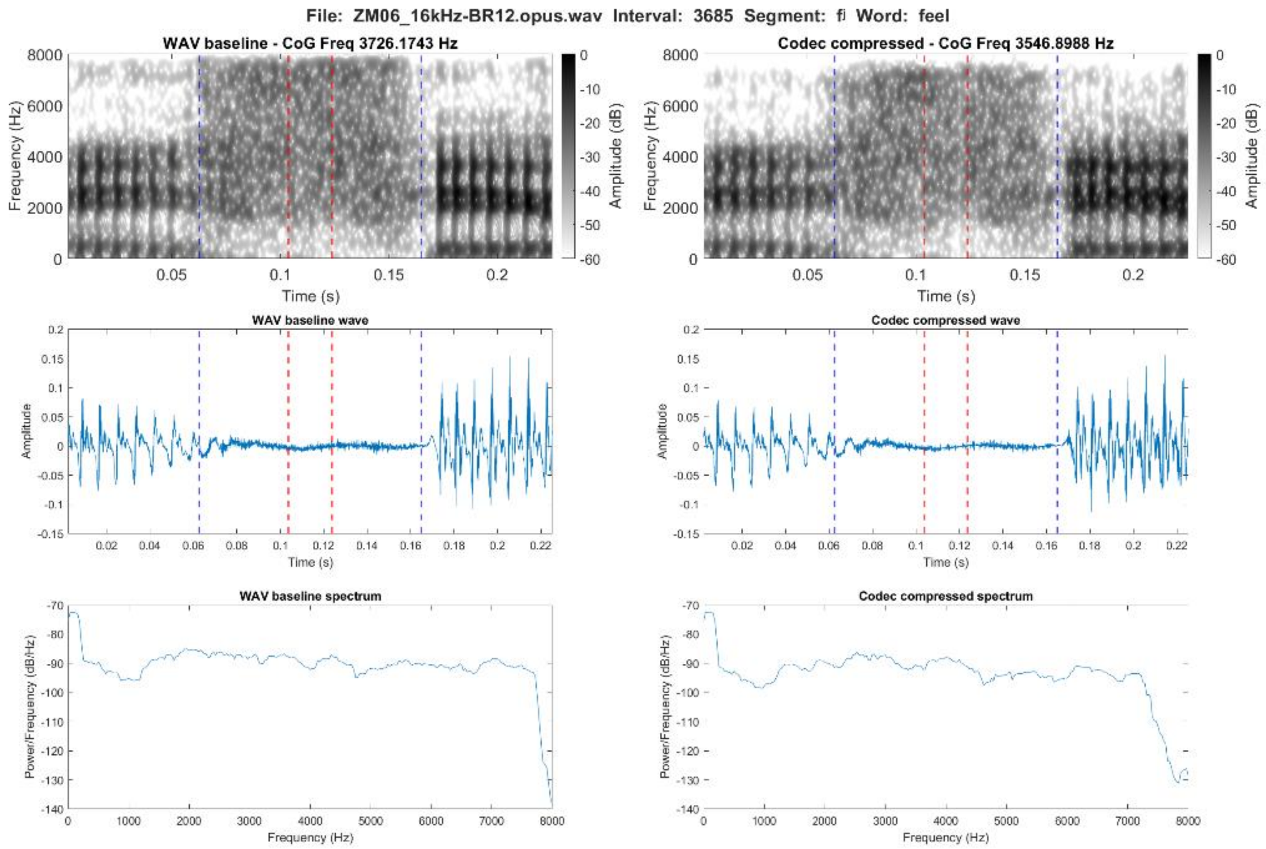


Figure 4.40. Spectrographic comparison of [fʲ] in the word *feel* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.3 Low bitrate: /θ/

The violin plots shows tendencies similar to what have been described for the other codecs and bitrates. This includes a slight lowering of CoG. SD also show a centring of frequencies, which means the values are spread across a comparatively smaller range of frequencies. There is no observable effect on skewness, while a very slight decrease of the upper frequencies and a slight increase in number of values closer to the lower limit is observable for the frequency peak. This is also evident from the linear predictions, where CoG and SD lower, while skewness stays stable.

/θ/ showed decreases in mean values for the spectral measures between just under 4 percent for kurtosis and up just under 11 percent for frequency peak. This was a change of 0.13 for kurtosis and 240 Hz for frequency peak. CoG and SD were both lowered with 6-7 percent, which was a decrease of 104 Hz for CoG and 122 Hz for SD. Here skewness followed a similar pattern of change as the previous measures with a change of 0.03 or just over 5 percent, which was not significant. All p-values and further statistical results can be found in table 4.23 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	/θ/	-204.60	-8.52	<.0001
SD	WAV – Opus	/θ/	-122.08	-11.45	<.0001
Skewness	WAV – Opus	/θ/	0.03	1.02	0.31

Table 4.23. Statistical results of the difference between WAV and Opus based on linear prediction models for /θ/

The maximum values again followed the pattern of the mean values, and were all lowered. This with around 400 Hz for CoG and SD, which gave a CoG maximum of 6,079 Hz and a SD maximum of 2,568 Hz. Frequency peak presented a slightly smaller decrease from 7,531 Hz to 7,219 Hz. Skewness changed by just under 2 to 5.89 in the codec compression.

A typical spectrographic representation of /θ/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.41. As with the two previous voiceless segments, the main effect of the codec compression is a general reduction in intensity across the segment.

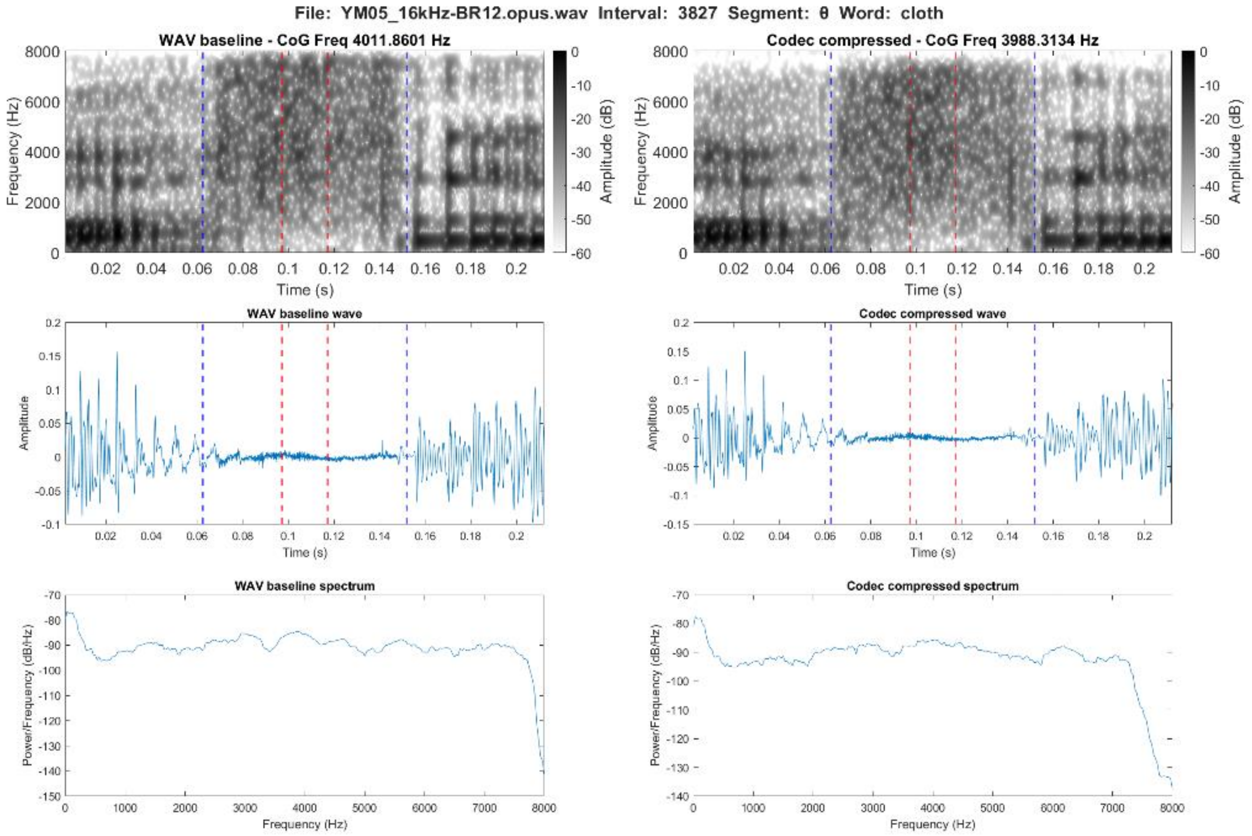


Figure 4.41. Spectrographic comparison of /θ/ in the word *cloth* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.4 Low bitrate: /s/

The effects on the distribution of the spectral measures for /s/ is again slight. For CoG and SD a lowering can be observed, though slight for CoG. Skewness shows no notable change and the effect on frequency peak is mainly limited to the top most frequencies. In contrast, the linear predictions reveal a clear downwards trajectory and especially so for skewness.

The mean values for the spectral measures all shows a decrease following the Opus compression. /s/ follows the pattern seen for the previous segments. CoG lowered by just over 3 percent or 155 Hz, SD with only 67 Hz or just below 6 percent, while the frequency peak decreased with 148 Hz or again just over 3 percent. For skewness the change was again more noticeable than for the other measures, for /s/ this was a decrease of 0.82 or 192 percent. Kurtosis was almost stable across the conditions, and only lowered by 0.19 percent. All p-values and further statistical results can be found in table 4.24 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	/s/	-154.96	-15.39	<.0001
SD	WAV – Opus	/s/	-67.36	-15.08	<.0001
Skewness	WAV – Opus	/s/	0.25	20.60	<.0001

Table 4.24. Statistical results of the difference between WAV and Opus based on linear prediction models for /s/

All maximum values were lowered for /s/. This was from 7,137 Hz to 6,749 Hz for CoG, and from 2,815 Hz to 2,584 Hz for SD. The frequency peak showed a decrease by around 300 Hz to 7,313 Hz, while SD lowered from 7.37 to 6.01.

A typical spectrographic representation of /s/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.42. A general reduction in intensity can again be observed across the segment particularly in the final part.

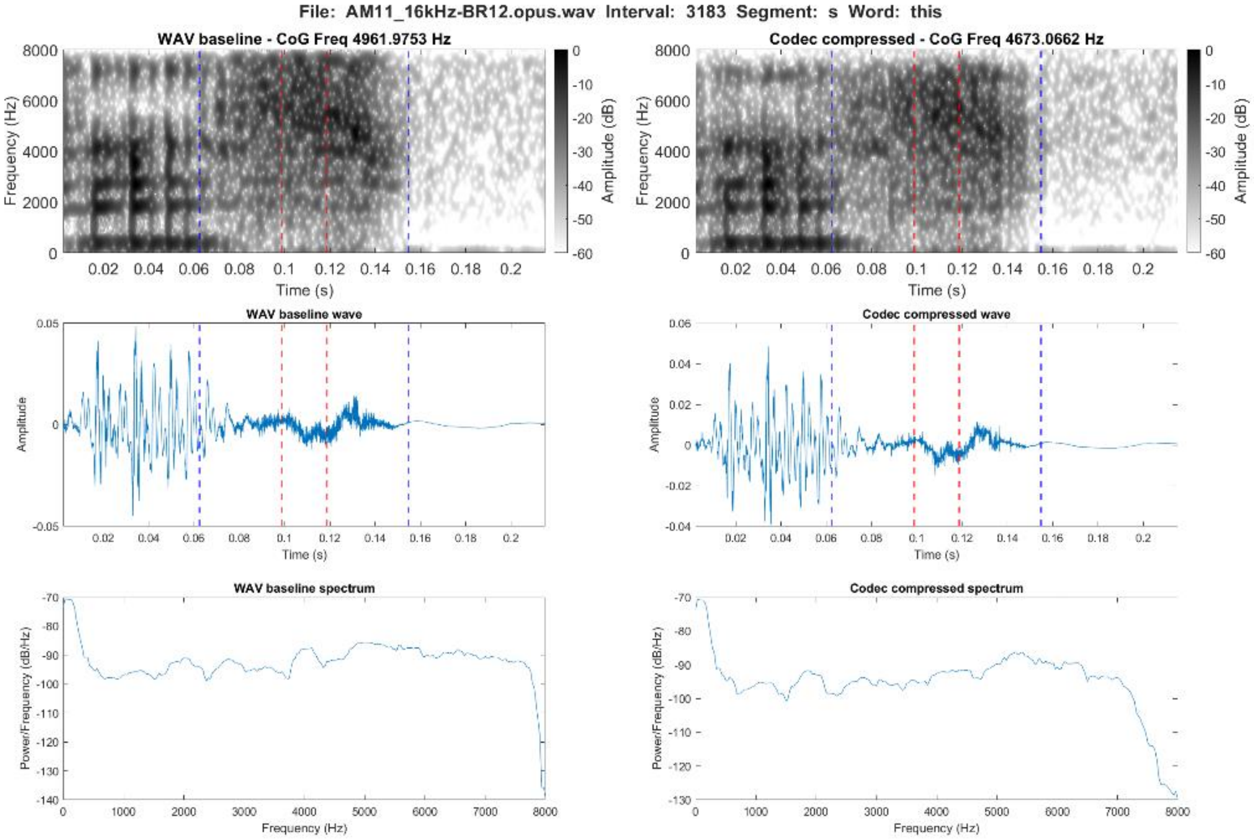


Figure 4.42. Spectrographic comparison of /s/ in the word *this* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.5 Low bitrate: /ʃ/

The distribution plots reveal no notable effect of the Opus compression on /ʃ/. Only slight downwards trajectories are observable for CoG and SD based on the linear predictions, while skewness presents a more clear downwards trend.

The mean values for /ʃ/ presented very slight increases of both CoG and frequency peak following the Opus compression. The increases were both by under 1 percent, which was an increase of 3 Hz for CoG and 10 Hz for frequency peak. This change was not significant for CoG. The remaining spectral measures decreased in mean values. For SD this was by 21 Hz or just over 2 percent, while skewness and kurtosis both showed decreases around 15 percent or for skewness 0.28 and kurtosis 1.35. All p-values and further statistical results can be found in table 4.25 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/ʃ/	3.12	0.13	0.90
SD	WAV – Opus	/ʃ/	-21.46	-2.00	0.05
Skewness	WAV – Opus	/ʃ/	0.28	9.74	<.0001

Table 4.25.Statistical results of the difference between WAV and Opus based on linear prediction models for /s/

As with the mean values, not all maximum values are lowered. However, here the increases were found for SD, which increased from 1,628 Hz to 1,767 Hz, and for skewness, which increased from 5.48 to 5.66. CoG presented a slight decrease of 28 Hz to 4,834 Hz, while frequency peak decreased by around 100 Hz to 5,188 Hz.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.43. A reduction in intensity is visible for /ʃ/ particularly in the initial and final part of the segment, and in a band between 2 and 4 kHz.

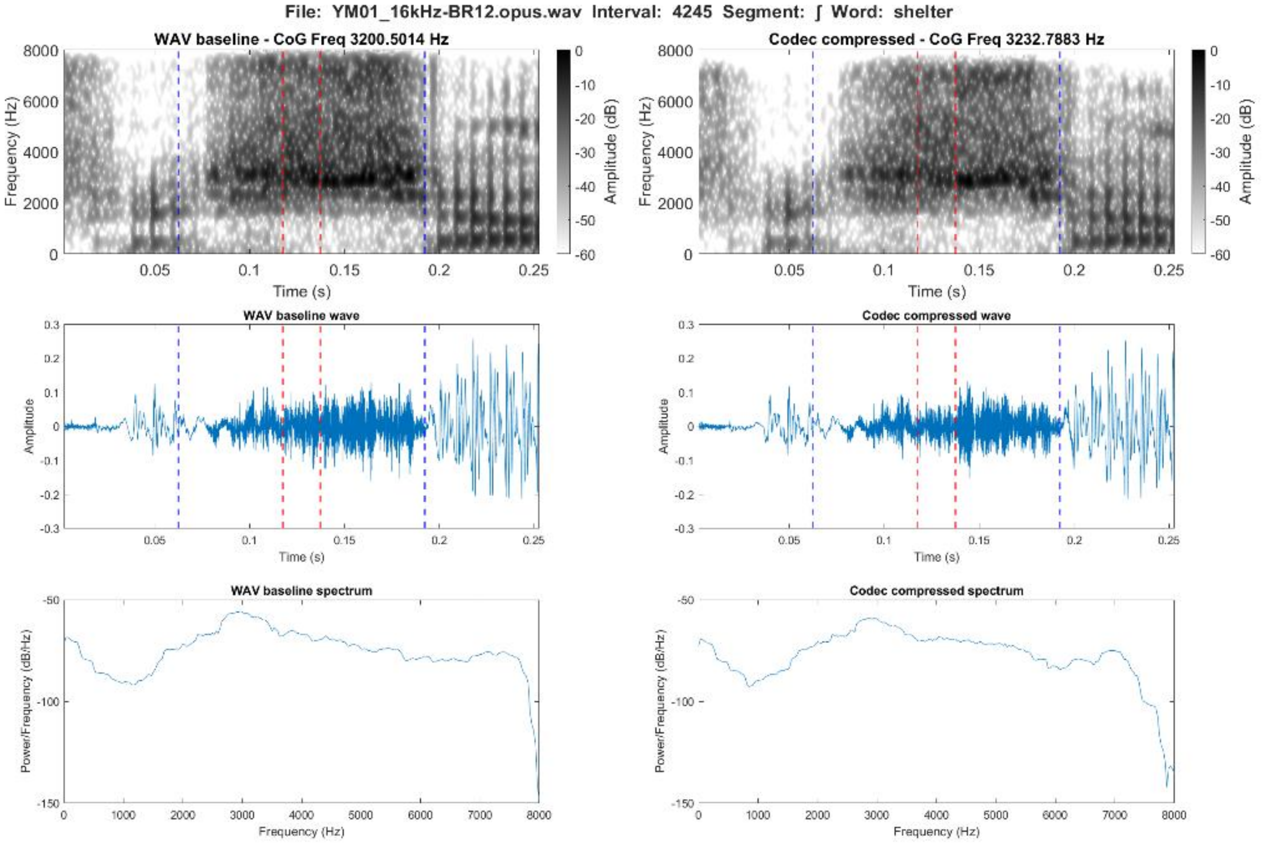


Figure 4.43. Spectrographic comparison of /ʃ/ in the word *shelter* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.6 Low bitrate: /z/

Apart from a slight lowering of SD and the upper values for frequency peak, no substantial effects are observed on the distribution of the spectral measures for /z/ following the Opus compression. The linear predictions reveal slight downwards tendencies at similar trajectories for both CoG, SD, and skewness.

In terms of mean values, as with the most of the voiceless segments, for /z/ all the spectral measures again lowered. This was by just under 3 percent for SD and up to just over 5 percent for frequency peak, which translated into a change of 36 Hz for SD and 201 Hz for frequency peak. The largest effect in Hz was seen for CoG, which lowered with 203 Hz or just under 5 percent. Skewness and kurtosis both decreased with around 0.20. All p-values and further statistical results can be found in table 4.26 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	/z/	134.16	6.77	<.0001
SD	WAV – Opus	/z/	35.54	3.86	<.001
Skewness	WAV – Opus	/z/	0.18	5.72	<.0001

Table 4.26. Statistical results of the difference between WAV and Opus based on linear prediction models for /z/

Apart from skewness, all maximum values lowered following the Opus compression. This was by just under 400 Hz for CoG to 6,697 Hz, and around 150 Hz for SD to 2,755 Hz. Frequency peak lowered from 7,563 Hz to 7,344 Hz, while skewness as mentioned increased from 9.36 to 9.69.

A typical spectrographic representation of partly voiced /z/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.44. A reduction in intensity is particularly prevalent in the initial and final part of the segments, while the frequencies below 4 kHz remain largely intact following the Opus compression.

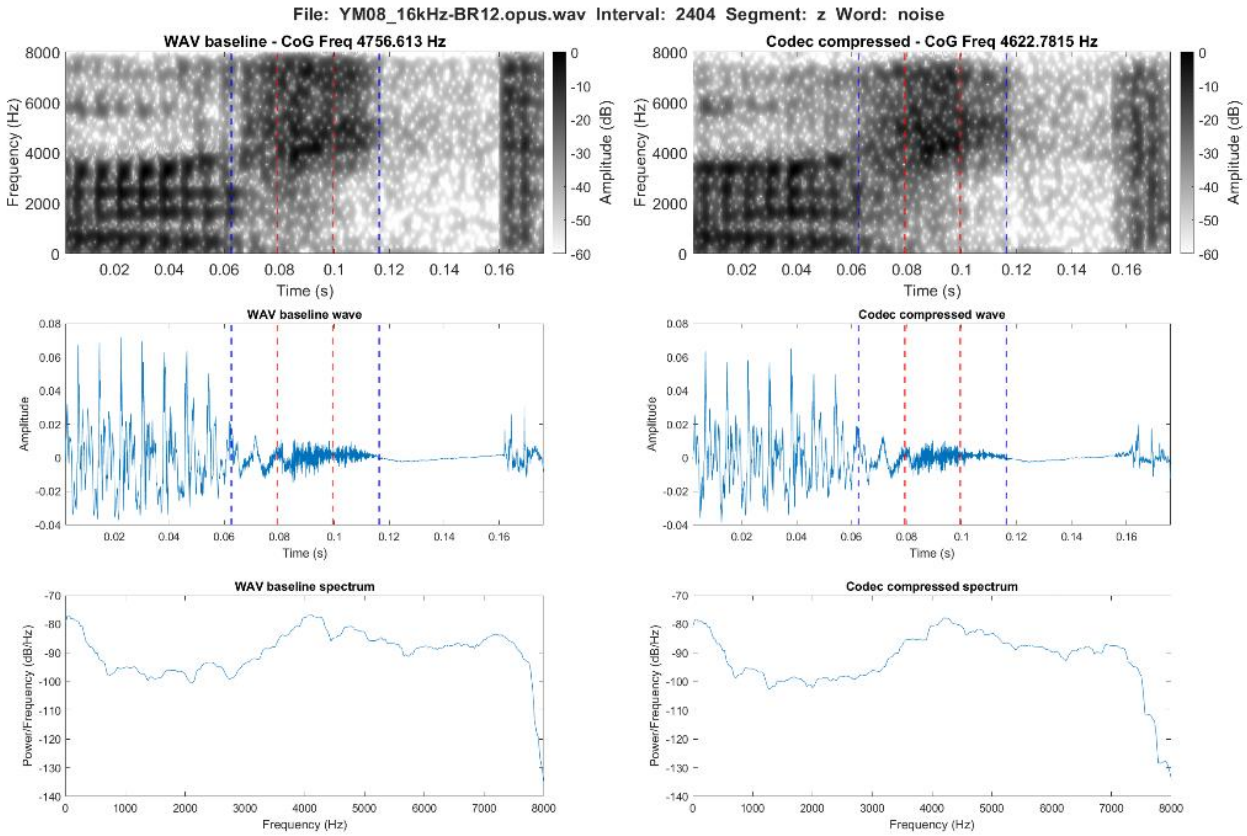


Figure 4.44. Spectrographic comparison of partly voiced /z/ in the word *noise* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.1.3.7 Low bitrate: /ð/

The distribution plots shows slightly more values for CoG towards the lower cut-off following the Opus compression. For SD a slight decrease is also observable, while skewness present a slight overall increase as well as in values towards the higher extreme. Apart from the effect of the upper frequency cut-off, no observable effect is noted for the frequency peak values. Similarly, the linear predictions reveal a slight decrease for CoG, a clear decrease for SD, and a slight increase of skewness.

All effects of the Opus compression on /ð/ were found to be significant (see table 4.27). The mean values for both skewness and kurtosis increased following the Opus compression. This was however, with less than 1 percent or 3.44 for skewness and 5.72 for kurtosis. CoG was lowered by just over 6 percent or 134 Hz, while SD was lowered with just over 7 percent or 95 Hz. Lastly, the frequency peak was found to lower by 88 Hz or just over 8 percent. All p-values and further statistical results can be found in table 4.27 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/ð/	134.16	6.77	<.0001
SD	WAV – Opus	/ð/	95.41	10.29	<.0001
Skewness	WAV – Opus	/ð/	-0.07	-2.09	0.04

Table 4.27. Statistical results of the difference between WAV and Opus based on linear prediction models for /ð/

As with /z/ all maximum values apart from skewness decreased. This was from 6,963 Hz to 6,202 Hz for CoG, and from 3,161 Hz to 2,773 Hz for SD. Frequency peak lowered by just under 700 Hz to 7,188 Hz, while skewness increase from 12.87 to 14.13.

A typical spectrographic representation of voiced /ð/ in the WAV baseline and the Opus compression using the low bitrate is presented below in figure 4.45. The main observable effect on /ð/ is a general reduction in intensity across the segment.

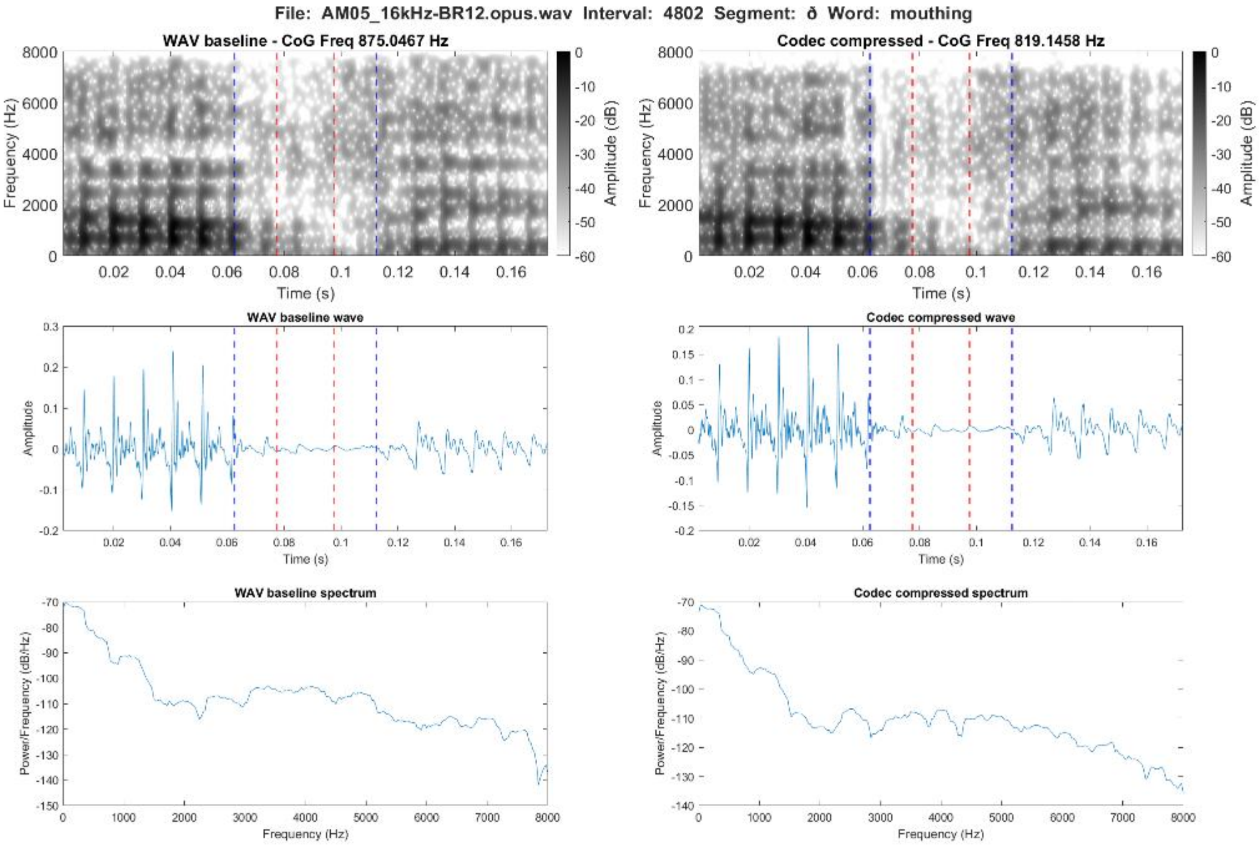


Figure 4.45. Spectrographic comparison of /ð/ in the word *mouthing* in the WAV baseline (left) and the Opus compression at 12 kbps (right)

4.1.2 High Quality bitrates

All results reported in the following sections are from the high quality bitrate dataset excluding all the 1 kHz tokens. These tokens will be analysed separately in section 4.1.3. This section will also provide details on the exact number of 1 kHz tokens e.g. for each codec.

As expected, the mean values in table 4.28 show that all segments were affected by the codec compression, and that these effects were codec as well as segment-dependent. As in the baseline study and the low bitrate section, the frequency peak was to a certain extent not truly representative in this context due to the limited bandwidth and sampling rate. However, when seen together with the CoG, the measure can still give an indication of the effect on the higher frequency content. Regardless, the frequency peak will not be included in the statistical modelling as the number of statistical 0s skew the model and the desired randomness, and distribution of the residuals.

Again, the same tendency was observed for kurtosis, which for this reason was also not included in the statistical modelling. In consequence, no p-values will be presented for these measures, and thus, these values will only be reported from descriptive statistics.

Seg	**Codec**	**Bitrate (kbps)**	**CoG (Hz)**	**SD (Hz)**	**Skew (Hz)**	**Kurt (Hz)**	**Freq. Peak (Hz)**
/f/	AMR	23.85	2885	1493	0.64	3.08	2142
/f/	MP3	48	3023	1554	0.67	3.21	2320
/f/	Opus	64	3027	1547	0.68	3.26	2327
/f/	**WAV**	**NA**	**3166**	**1672**	**0.74**	**3.35**	**2443**
[fʲ]	AMR	23.85	2917	1385	0.71	3.41	2240
[fʲ]	MP3	48	3024	1441	0.78	3.63	2358
[fʲ]	Opus	64	3025	1434	0.78	3.66	2370
[fʲ]	**WAV**	**NA**	**3131**	**1544**	**0.87**	**3.84**	**2408**
/s/	AMR	23.85	4504	1029	-0.56	6.39	4471
/s/	MP3	48	4628	1084	-0.35	5.95	4628
/s/	Opus	64	4641	1076	-0.31	6.00	4620
/s/	**WAV**	**NA**	**4789**	**1165**	**-0.09**	**5.71**	**4742**
/z/	AMR	23.85	4100	1143	-0.54	6.50	3644
/z/	MP3	48	4205	1185	-0.38	6.20	3771
/z/	Opus	64	4225	1169	-0.34	6.30	3771
/z/	**WAV**	**NA**	**4342**	**1242**	**-0.12**	**6.17**	**3852**
/ʃ/	AMR	23.85	3206	813	1.41	8.33	2979
/ʃ/	MP3	48	3246	853	1.49	8.47	2996
/ʃ/	Opus	64	3246	846	1.51	8.60	2997
/ʃ/	**WAV**	**NA**	**3271**	**896**	**1.71**	**9.45**	**3002**
/θ/	AMR	23.85	2993	1626	0.47	2.98	1964
/θ/	MP3	48	3142	1687	0.49	3.04	2150
/θ/	Opus	64	3152	1684	0.49	3.10	2149
/θ/	**WAV**	**NA**	**3318**	**1808**	**0.53**	**3.16**	**2298**
/ð/	AMR	23.85	2267	1471	1.05	4.72	1108
/ð/	MP3	48	2387	1513	0.95	4.37	1189
/ð/	Opus	64	2378	1510	0.98	4.62	1170
/ð/	**WAV**	**NA**	**2451**	**1575**	**1.02**	**4.83**	**1242**

Table 4.28. **High** dataset mean values for all spectral measures for each fricative in each individual codec and 16 kHz WAV baseline

The voiced fricatives (i.e. /ð/ and /z/) deviated slightly from the desired randomness in the residual plots predicted values. However, this is to be expected from the acoustic structure of these sounds (see Chapter 2, section 2.2 and 2.3 for further considerations). In that way, as it is clear why this pattern was found and the remaining residual plots i.e. distribution and quantiles were following expected patterns, the voiced segments were still included in the mixed effects modelling. However, this should be kept in mind when assessing the implications and reliability of the outputted results.

The magnitude and direction of the changes in mean values are shown in Table 4.29. Some general trends can be observed. Firstly, how almost all spectral measures under this quality bitrate were lowered. Secondly, Opus and MP3 appear to affect these measures to similar extents, while AMR-WB had a comparatively larger effect.

Seg	Codec	Bitrate (kbps)	CoG (Hz)	CoG (%)	SD (Hz)	SD (%)	Skew (Hz)	Skew (%)	Kurt (Hz)	Kurt (%)	Freq. Peak (Hz)	Freq. Peak (%)
/f/	AMR	23.85	281	-8.87	180	-10.74	0.10	-14.05	0.27	-8.14	302	-12.35
/f/	MP3	48	143	-4.51	118	-7.05	0.07	-9.14	0.14	-4.27	123	-5.04
/f/	Opus	64	139	-4.39	126	-7.52	0.06	-8.45	0.10	-2.91	117	-4.78
[fʲ]	AMR	23.85	214	-6.84	159	-10.29	0.16	-18.62	0.42	-11.00	168	-6.97
[fʲ]	MP3	48	107	-3.41	102	-6.64	0.09	-10.39	0.21	-5.44	50	-2.07
[fʲ]	Opus	64	106	-3.39	110	-7.13	0.09	-10.33	0.18	-4.69	39	-1.61
/s/	AMR	23.85	284	-5.94	136	-11.67	0.47	550.27	-0.67	11.81	272	-5.73
/s/	MP3	48	161	-3.36	81	-6.97	0.26	308.71	-0.24	4.17	114	-2.41
/s/	Opus	64	148	-3.09	89	-7.68	0.23	263.50	-0.29	5.00	122	-2.57
/z/	AMR	23.85	242	-5.58	99	-7.99	0.43	367.91	-0.33	5.37	208	-5.40
/z/	MP3	48	137	-3.15	57	-4.60	0.26	226.37	-0.03	0.56	81	-2.11
/z/	Opus	64	117	-2.69	72	-5.83	0.23	193.75	-0.13	2.17	81	-2.11
/ʃ/	AMR	23.85	65	-1.99	83	-9.26	0.31	-17.98	1.12	-11.89	22	-0.75
/ʃ/	MP3	48	25	-0.76	43	-4.77	0.22	-12.94	0.98	-10.37	6	-0.20
/ʃ/	Opus	64	25	-0.77	50	-5.54	0.21	-12.04	0.85	-9.03	5	-0.16
/θ/	AMR	23.85	325	-9.79	181	-10.03	0.06	-12.08	0.17	-5.42	334	-14.54
/θ/	MP3	48	176	-5.30	121	-6.67	0.04	-8.19	0.11	-3.60	147	-6.40
/θ/	Opus	64	166	-5.01	123	-6.83	0.04	-7.27	0.05	-1.70	149	-6.47
/ð/	AMR	23.85	183	-7.48	104	-6.62	-0.03	2.54	0.12	-2.38	134	-10.80
/ð/	MP3	48	63	-2.59	62	-3.96	0.07	-7.09	0.47	-9.65	53	-4.26
/ð/	Opus	64	73	-2.97	65	-4.13	0.04	-3.77	0.21	-4.35	72	-5.83

Table 4.29. Differences in mean values between baseline (WAV) and **High** quality codec compression in Hz and percentage. Colours indicate the direction of the change. (i.e. blue = decrease; yellow = increase)

In sum, the tables show how the high bitrates across codecs and segments lowered the spectral measures. This was, however, generally by smaller percentages than what was observed for the low bitrates. The exception was the skewness values for /s/ and /z/, which changed by up to just over 550 percent.

4.1.2.1 High bitrate: AMR-WB

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the AMR-WB codec in the high quality bitrate dataset.

First, the linear predictions for each spectral measure and the individual segments can be found below (figure 4.46 to 4.51). These indicate the directionality of the changes imposed by the AMR-WB codec using the high bitrate. The graphs present the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these plots are in the following sections on the individual segments.

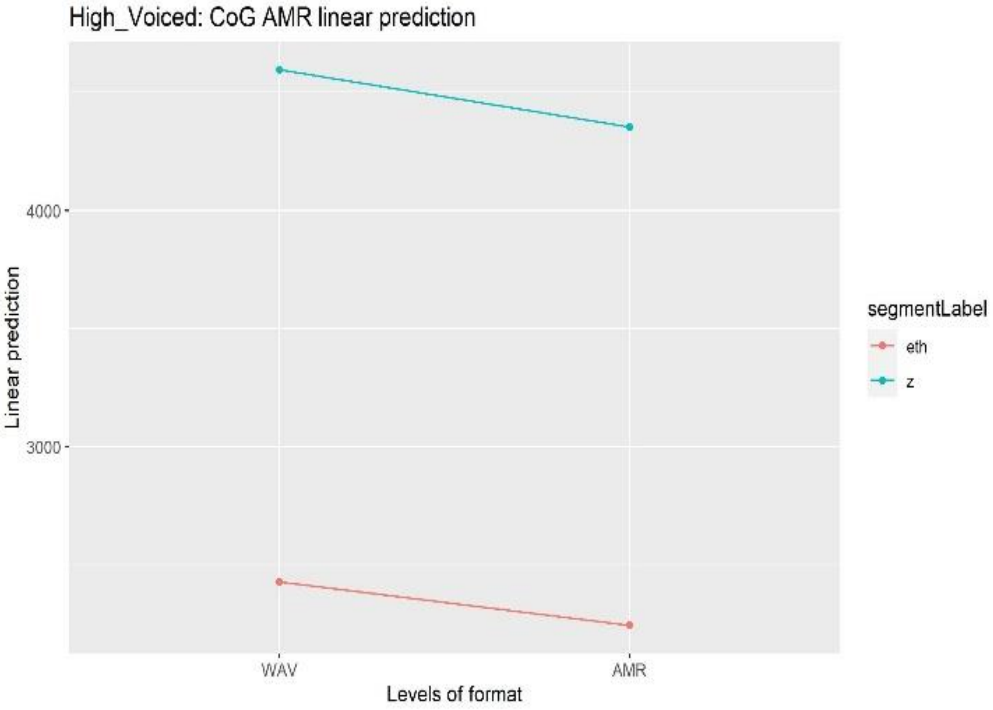


Figure 4.46. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

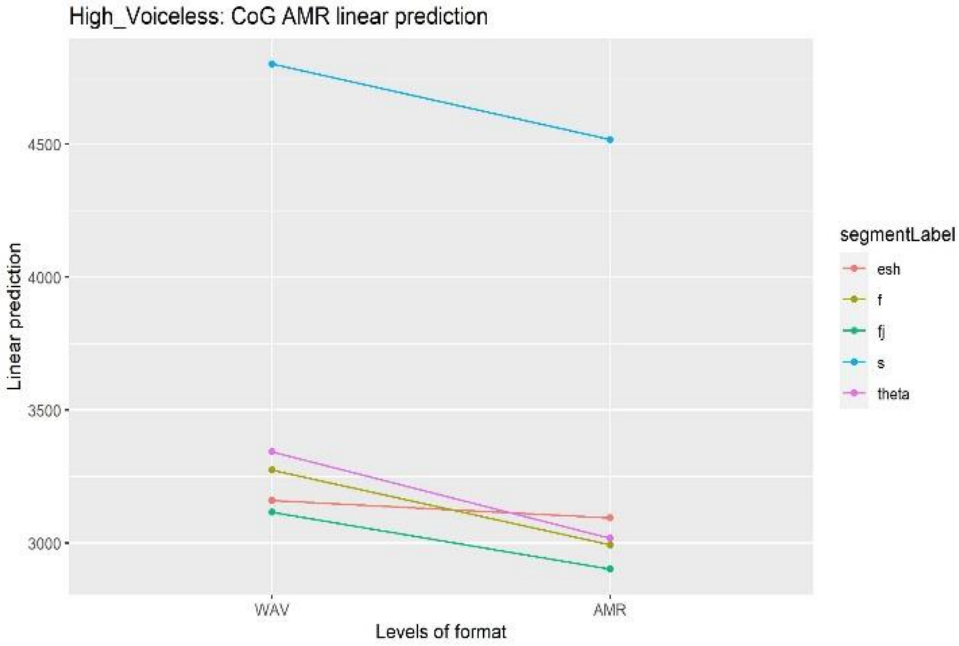


Figure 4.47. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

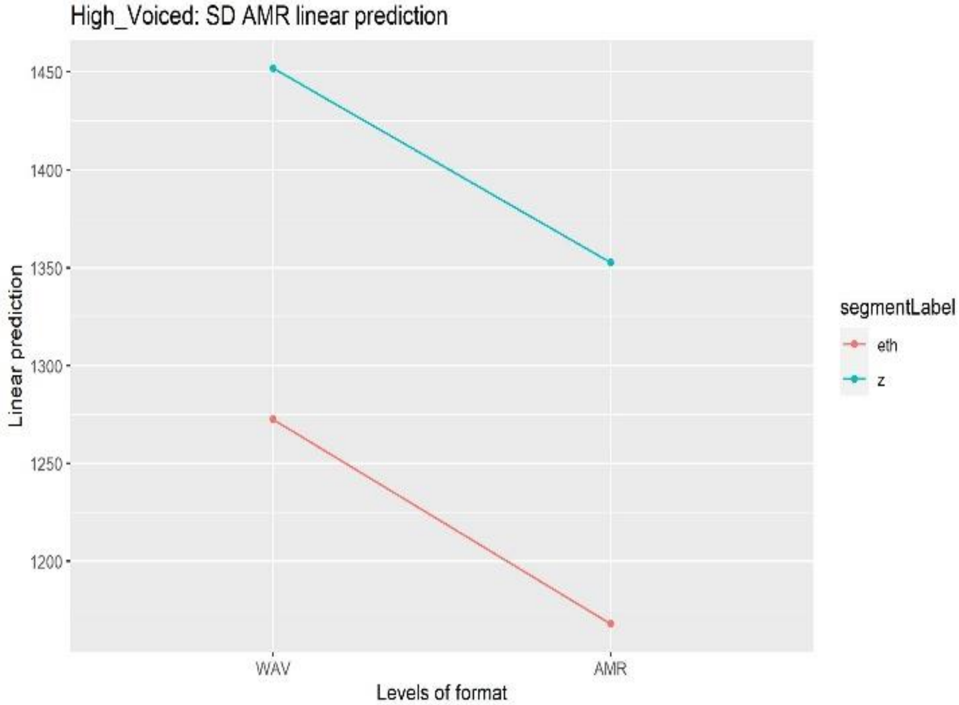


Figure 4.48. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

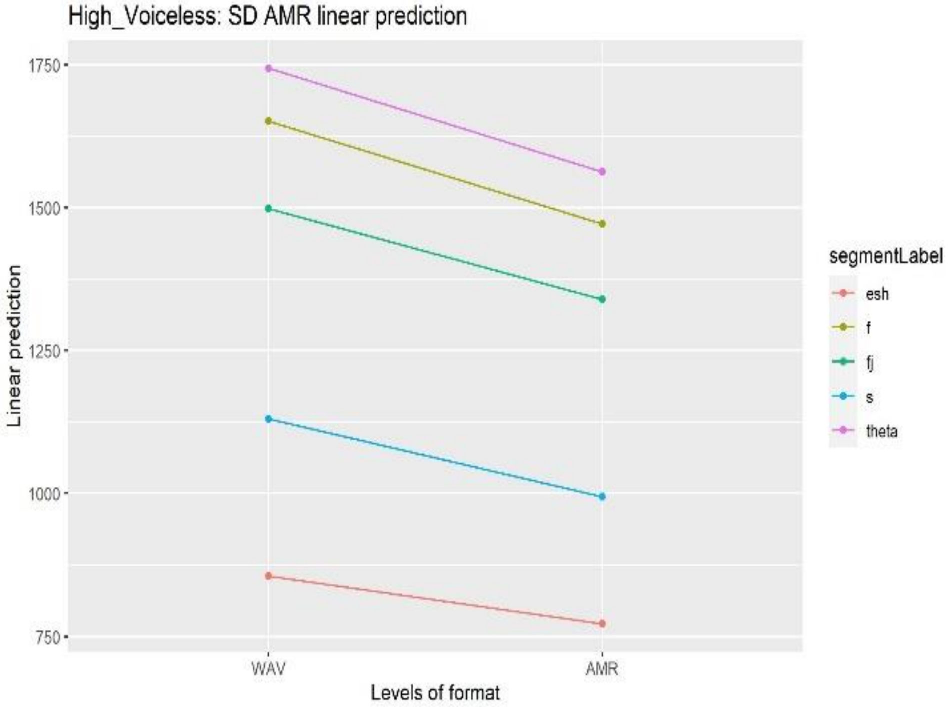


Figure 4.49. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

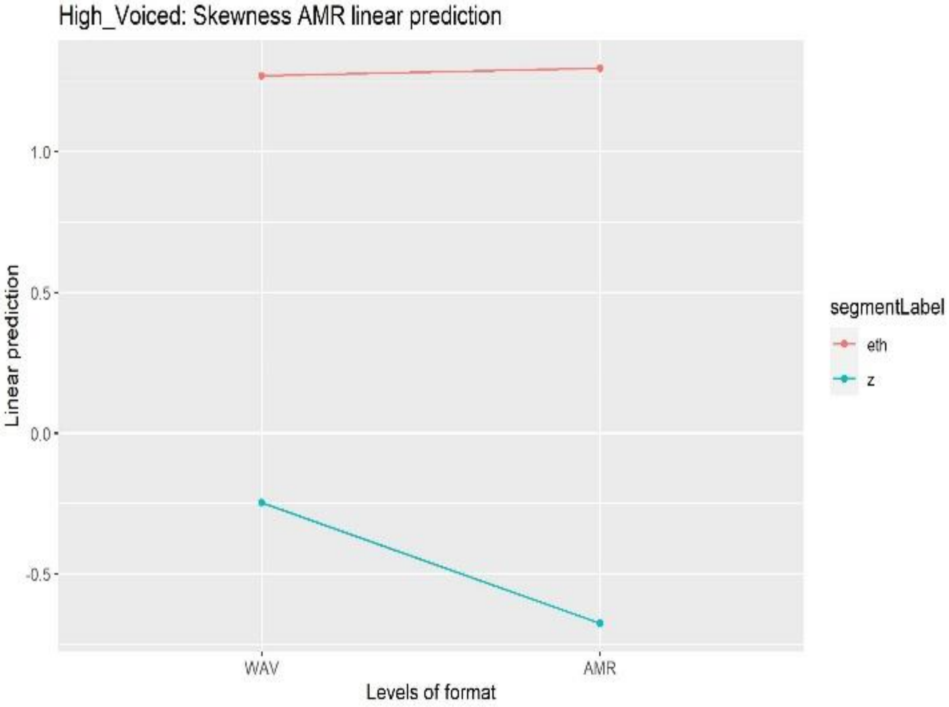


Figure 4.50. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

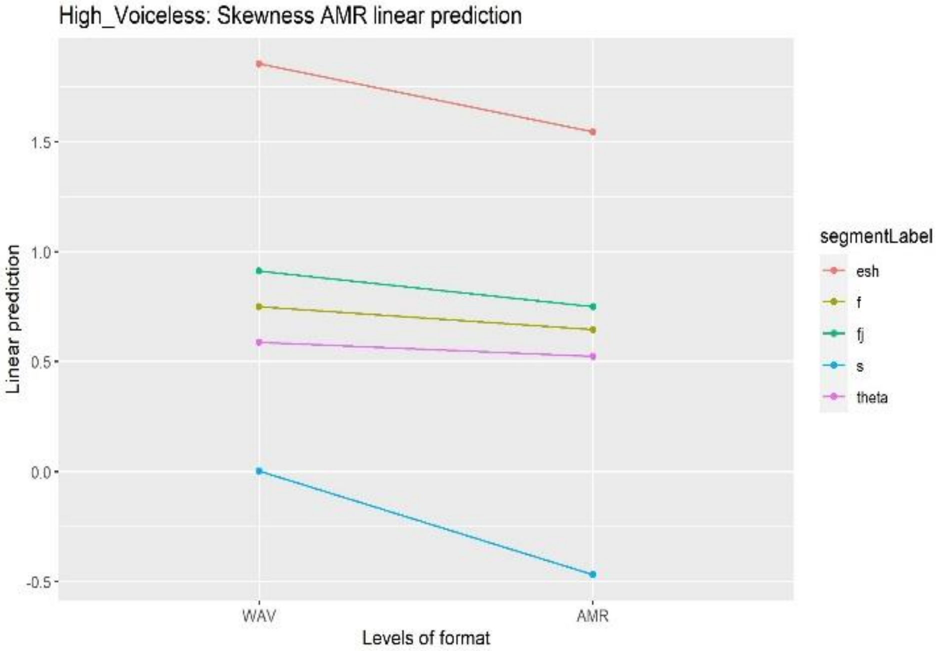


Figure 4.51. Trajectory of the linear predictions in the comparison of WAV and **high** bitrate AMR-WB from the mixed effects models for skewness and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

The distributions are illustrated below in a set of violin plots for each spectral measure. Again, the specific analysis pertaining to each segment will be found in the following sections.

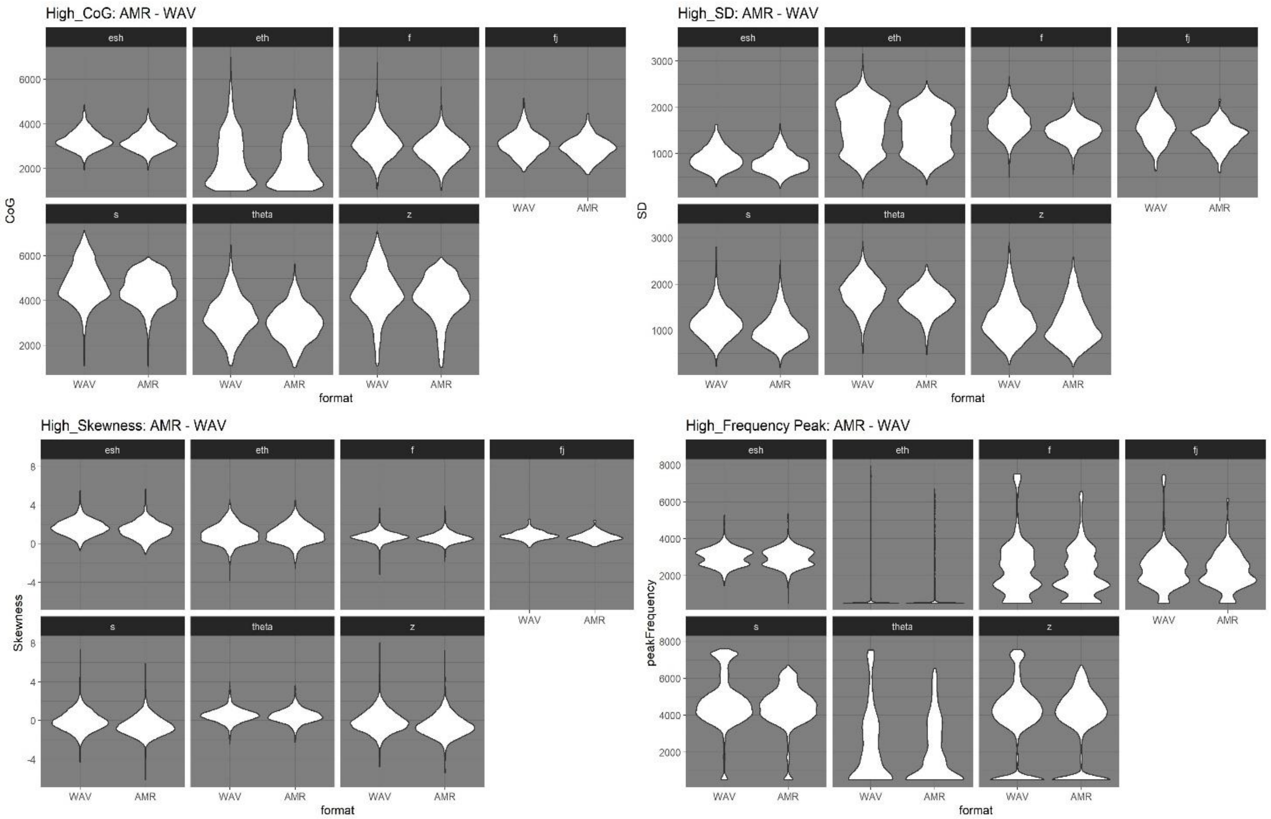


Figure 4.52. Distribution of spectral measure values in WAV baseline and the **High** bitrate AMR-WB codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/*, esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

In brief, the AMR-WB using the high bitrate to different extents decreased all spectral measures for all segments apart from /ð/, which show a very slight increase for skewness. Overall, the distributions maintain their shapes from the WAV files.

4.1.2.1.1 High bitrate: /f/

From the distribution plots, the main observable effect on CoG was a lowering and a slight increase in values around the mean, while SD both presents a lowering as well as more values clearly centred around the mean. For skewness, no notable effect is evident, while the main effect on the frequency peak is due to the upper cut-off together with an increase in values around the mean and lower cut-off. These observations are confirmed by the linear predictions. In addition, it is worth noting that /f/ and /θ/ become more alike following the codec compression.

More specifically, the AMR-WB compression lowered the mean values for all spectral measures by between just over 8 percent for kurtosis to just over 14 percent for skewness (p = <.0001). CoG was lowered by just over 280 Hz or just under 9 percent, while SD was lowered by a 180 Hz or just under 11 percent. Frequency peak showed the biggest change in terms of Hz by 302 Hz, which was a change of just over 12 percent. All p-values and further statistical results can be found in table 4.30 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/f/	280.79	2.82	<.0001
SD	WAV - AMR	/f/	179.65	27.51	<.0001
Skewness	WAV - AMR	/f/	0.10	5.89	<.0001

Table 4.30. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /f/

Apart from skewness, all the maximum values decreased following the codec compression. This was from 6,754 Hz to 5,665 Hz for CoG, from 2,664 Hz to 2,322 Hz for SD and from 7,500 Hz for the frequency peak to 6,563 Hz. Skewness increased from 3.68 to 3.91.

A typical spectrographic representation of /f/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.53. The main noticeable effect here is a general reduction in intensity across the segment as well as an upper frequency limit, which appear to vary slightly across the segment with more frequencies removed towards the final part.

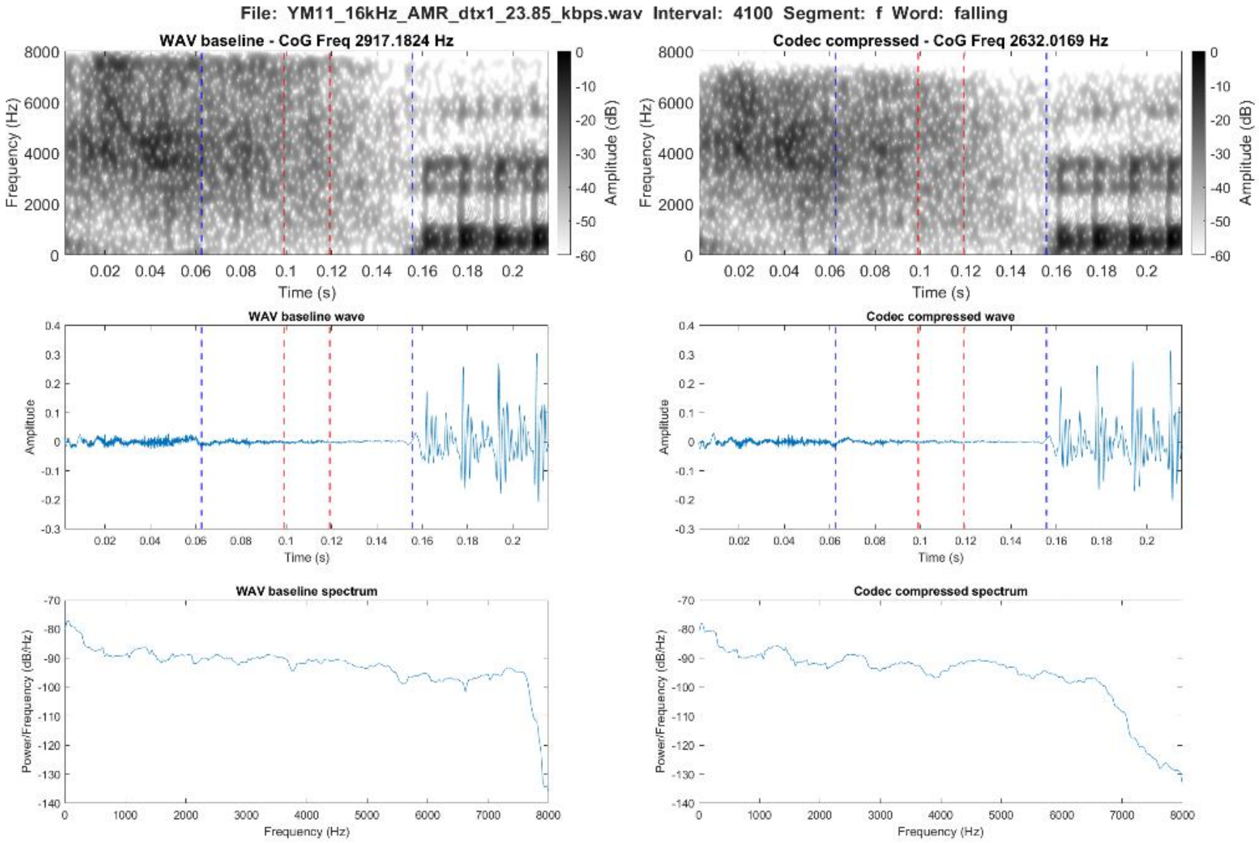


Figure 4.53. Spectrographic comparison of /f/ in the word *falling* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.2 High bitrate: [fʲ]

For [fʲ] the main effects in terms of distribution can be observed for CoG and SD, where both measures lower and particularly SD shows a change in distribution with more values around and below the mean. The frequency peak is mainly affected around the upper-cut off limit, but shows and increase in number of values around the lower frequencies. No notable effect is observable for skewness. Again, the linear predictions also show a downwards trajectory for CoG, SD and skewness.

As with /f/ the mean values for all spectral measures were lowered following the codec compression (p = <.0001 to 0.02). For CoG and frequency peak this was a decrease in mean value by just under 7 percent, or 214 Hz for the former and 168 Hz for the latter. SD showed a lowering by just over 10 percent or 160 Hz. Despite the limited change to the distribution, skewness presented the biggest effect by a change of just under 19 percent, while kurtosis changed by 0.42 or 11 percent. All p-values and further statistical results can be found in table 4.31 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	[fʲ]	214	4.03	<.0001
SD	WAV - AMR	[fʲ]	158.82	6.46	<.0001
Skewness	WAV - AMR	[fʲ]	0.16	2.43	0.02

Table 4.31. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for [fʲ]

Similar to the mean values, all maximum values were lowered. This was by 600 Hz for CoG to 4,450 Hz, and by almost 1,300 Hz for frequency peak to 6,188 Hz. SD was lowered from 2,436 Hz to 2,182 Hz, while skewness lowered by 0.13 to 2.39.

A typical spectrographic representation of [fʲ] in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.54. As previously observed, the main effect is a reduction in intensity across the segment. The upper frequency limit appear slightly lower for the relatively less intense frequencies in the initial part of the segment.

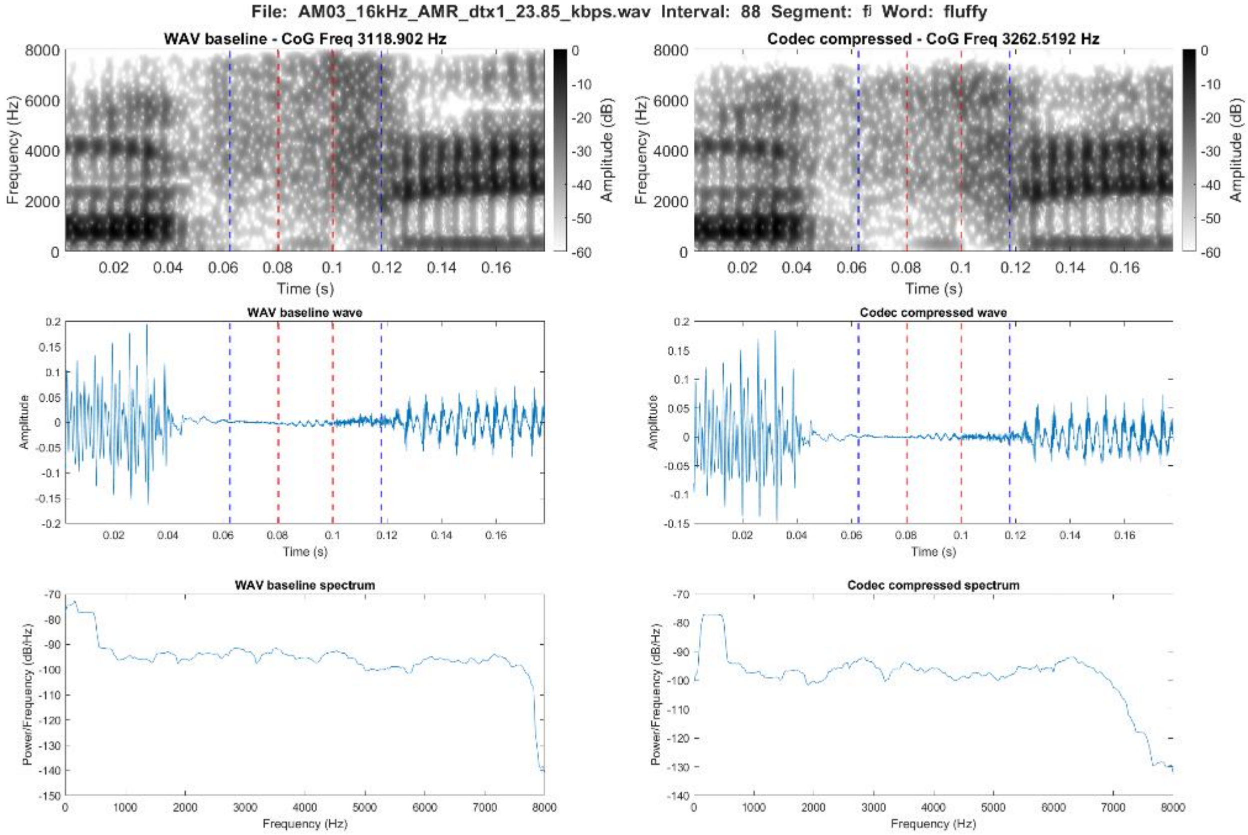


Figure 4.54. Spectrographic comparison of [fʲ] in the word *fluffy* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.3 High bitrate: /θ/

/θ/ shows changes to the distribution in similar ways to /f/. CoG is lowered, while SD is both lowered and have more values centred around the mean. For skewness, only a slight effect is observable, but the distributional shape remains largely unchanged. For frequency peak, the primary effects are again a lowering of the top-most frequencies and an increase in values around the lower cut-off. This is also evident from the linear predictions, where the downwards trajectories are more prominent for CoG and SD.

The mean values followed the same pattern as all the spectral measures were lowered by the AMR-WB. For CoG, SD and skew this was by around 10 to 12 percent. These effects were significant for CoG and SD (p=<.0001). For CoG this was a change of 325 Hz and for SD a change by 181 Hz. The frequency peak lowered similarly to CoG by 334 Hz or just under 15 percent. Kurtosis was lowered by 5.42 percent, which was a decrease of 0.17. All p-values and further statistical results can be found in table 4.32 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/θ/	324.92	29.58	<.0001
SD	WAV - AMR	/θ/	181.28	17	<.0001
Skewness	WAV - AMR	/θ/	0.06	2.22	0.03

Table 4.32. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /θ/

The maximum values lowered across all the spectral measures. For CoG this was from 6,491 Hz to 5,627 Hz, and for SD from 2,436 Hz to 2,182 Hz. The frequency peak lowered by 1 kHz to 6,531 Hz.

A typical spectrographic representation of /θ/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.55. The main effect visible from spectrogram, spectrum and waveform is a general reduction in intensity across the segment.

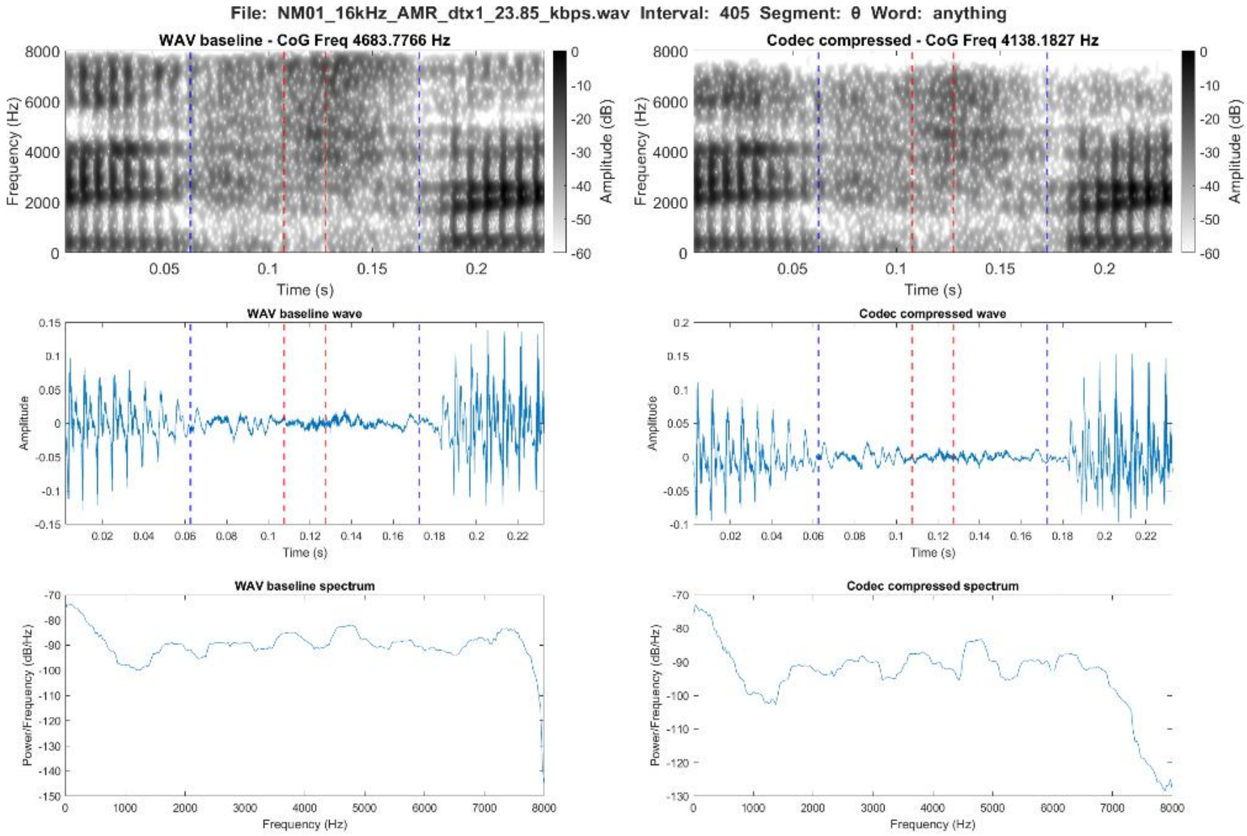


Figure 4.55. Spectrographic comparison of /θ/ in the word *anything* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.4 High bitrate: /s/

From the distribution plots, an overall lowering can be observed for SD and skewness, while CoG and frequency peak are mainly affected in terms of changes to the values at the higher frequencies above the mean. The shape of the distribution changes for CoG, while it apart from the higher values remains almost unchanged for frequency peak. These changes are observable from the linear predictions as downwards trends especially for skewness.

In more detail, for CoG, SD, skewness and frequency peak a decrease in mean value was found following the codec compression (p=<.0001), while an increase was found for kurtosis. The effect on CoG and frequency peak were just under 6 percent, which was a change in Hz of 284 for CoG and 272 for frequency peak. SD was lowered by just under 12 percent or 136 Hz, while skewness presented a change of 0.47. The increase of kurtosis was almost 12 percent or 0.67. All p-values and further statistical results can be found in table 4.33 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/s/	284.24	29.58	<.0001
SD	WAV - AMR	/s/	135.86	30.52	<.0001
Skewness	WAV - AMR	/s/	0.47	39.03	<.0001

Table 4.33. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /s/

Again the maximum values lowered for CoG, SD, skewness, and frequency peak. This was by, more than 1 kHz for CoG from 7,137 Hz to 5,949 Hz, and by almost 1 kHz for frequency peak, which changed from 7,594 Hz to 6,719 Hz. SD lowered from 2,851 Hz to 2,530 Hz, while skewness changed from a maximum value of 7.33 to 5.93.

A typical spectrographic representation of /s/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.56. The reduction in intensity for /s/ is prevalent in the initial and final part of the segment as well as towards the upper frequency limit particularly in the initial part.

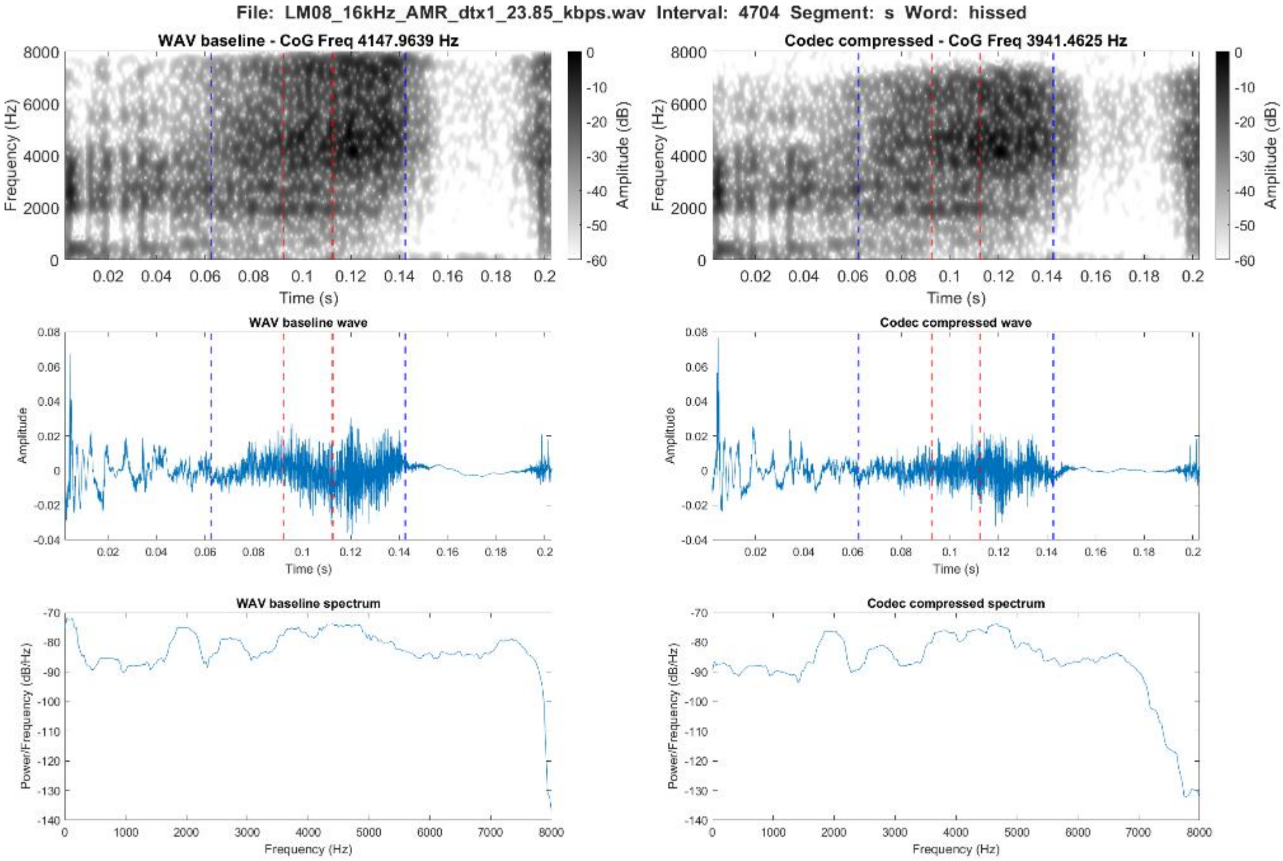


Figure 4.56. Spectrographic comparison of /s/ in the word *hissed* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.5 High bitrate: /ʃ/

For CoG a very slight lowering can be observed from the distribution plots, while the overall shape of the distribution remains the same following the codec compression. SD and skewness also present an overall lowering, while only minor changes are visible to the distribution shape. These observations are confirmed by the linear prediction. The frequency peak shows only an effect of the codec compression by an additional number of values towards the lower cut-off.

All mean values for the spectral measures were lowered following the codec compression for /ʃ/ (p=<.0001 to 0.005). These effects varied between 0.75 percent for frequency peak to just under 18 percent for skewness. This was a change by just 22 Hz for frequency peak. CoG was lowered by just under 2 percent or by 65 Hz, while SD was lowered by 83 Hz or just over 9 percent. Kurtosis lowered by 1.12 or just under 12 percent. All p-values and further statistical results can be found in table 4.34 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/ʃ/	64.94	2.82	0.005
SD	WAV - AMR	/ʃ/	83.04	7.78	<.0001
Skewness	WAV - AMR	/ʃ/	0.31	10.66	<.0001

Table 4.34. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /ʃ/

In contrast to the previous segments, both the maximum value for SD, skewness and peak frequency increased, however slightly, following the AMR-WB compression. This was with 14 Hz for SD to 1,652 Hz, with 63 Hz for the peak frequency to 5,344 Hz, and with 0.18 for skewness to 5.66. CoG followed previous results and decreased from 4,862 Hz to 4,686 Hz.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.57. Again, the main effect is an overall reduction in intensity and a lower upper frequency limit.

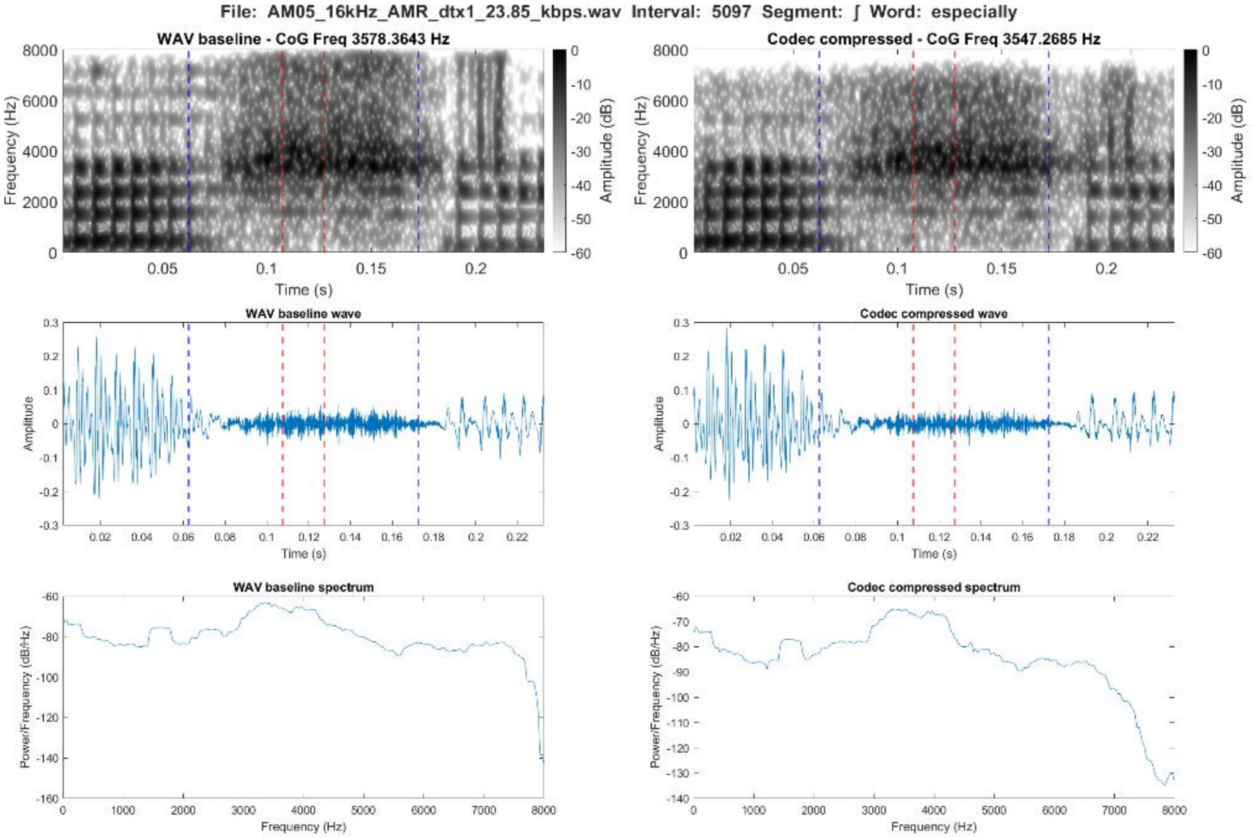


Figure 4.57. Spectrographic comparison of /ʃ/ in the word *especially* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.6 High bitrate: /z/

The effects of on the distribution of /z/ are almost identical to what was observed for /s/. This means main effects on the values at the higher end of the spectrum, and a general lowering of SD and skewness. The linear predictions also show this general decrease and especially so for SD.

The same tendency is seen in the mean value as the effects of the AMR-WB compression on /z/ was again similar to /s/ however smaller (p=<.0001). Apart from kurtosis, all mean values for the spectral measures lowered by just over 5 percent for frequency peak and by just under 8 percent for SD. This was a change of 208 Hz for frequency peak and 99 Hz for SD. CoG lowered slightly more by 242 Hz or just under 6 percent, while skewness was lowered by 0.43. Lastly, kurtosis was increased by just over 5 percent or 0.33. All p-values and further statistical results can be found in table 4.35 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV - AMR	/z/	242.16	8.29	<.0001
SD	WAV - AMR	/z/	99.29	11.45	<.0001
Skewness	WAV - AMR	/z/	0.43	19.90	<.0001

Table 4.35.Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /z/

For /z/ all the maximum values were again lowered. This was from 7,082 Hz to 5,920 Hz for CoG and from 7,563 Hz to 6,719 Hz for frequency peak. SD lowered from 2,901 Hz to 2,584 Hz, while skewness showed a change from 8.05 to 7.29.

A typical spectrographic representation of /z/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.48. In addition to a general reduction in intensity as it has also been observed for the previous segments, for /z/ the frequencies above 4 kHz is particularly reduced.

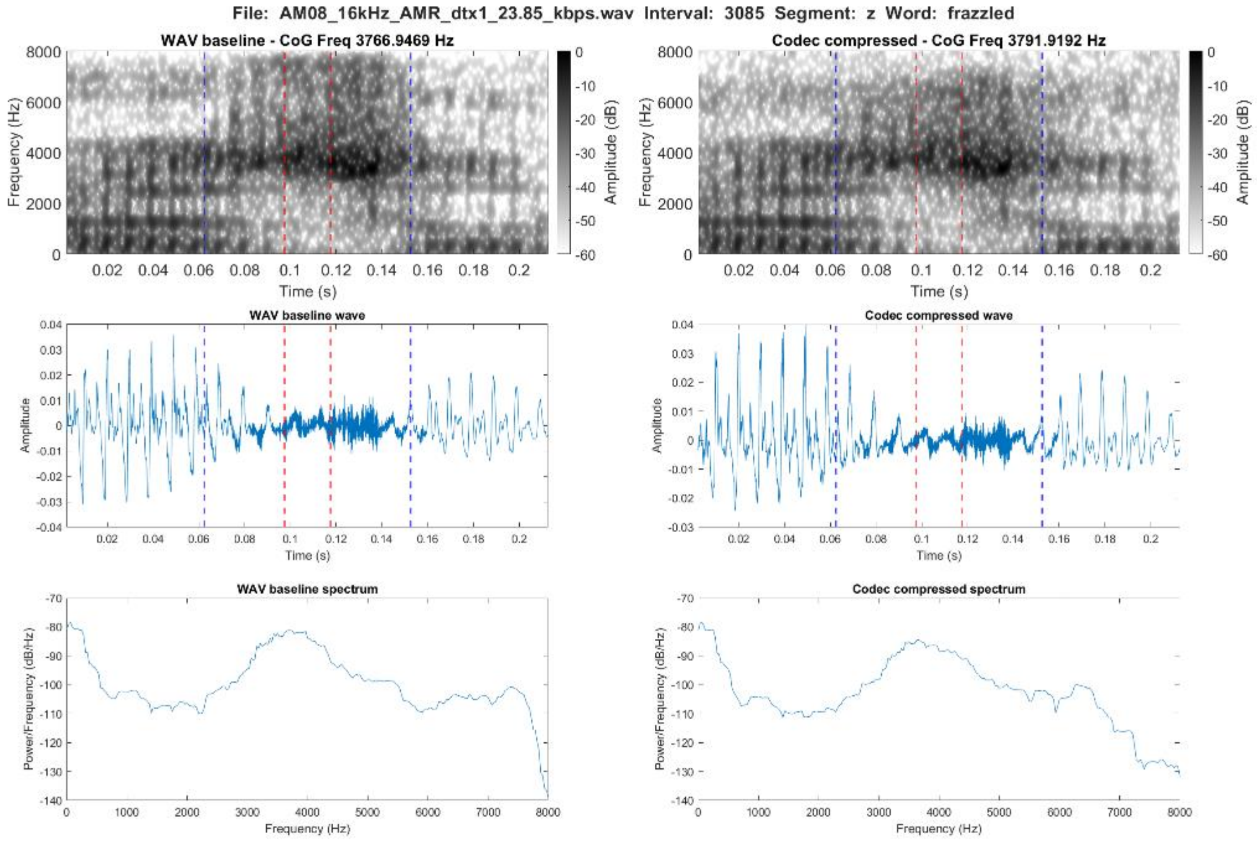


Figure 4.58. Spectrographic comparison of voiced /z/ in the word *frazzled* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.1.7 High bitrate: /ð/

The distribution plots show an increase in number of values towards the lower cut-off for CoG, while for SD the highest and lowest values in the WAV baseline is decreased for the former and increased for the latter. For skewness a slight increase in values above the mean can be observed. For the frequency peak, the only observable effect is caused by the upper cut-off. The linear predictions similarly show the decrease of CoG and SD, though slight for the former, as well as the slight increase of skewness.

These observations are confirmed by the mean values, where apart from skewness, all values for the spectral measures were lowered and significantly so (p=<.0001) for /ð/. This was by just over 2 percent for kurtosis to just under 11 percent for frequency peak, which for the latter meant a change by 134 Hz. CoG and SD both decreased with around 7 percent, which was a decrease by 183 Hz for

CoG and by 104 Hz for SD. The increase in skewness was by just under 3 percent of 0.03. All p-values and further statistical results can be found in table 4.36 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV - AMR	/ð/	183.45	8.29	<.0001
SD	WAV - AMR	/ð/	104.32	10.29	<.0001
Skewness	WAV - AMR	/ð/	-0.03	-1.04	0.3

Table 4.36. Statistical results of the difference between WAV and **high** AMR-WB based on linear prediction models for /ð/

Regardless of the increase in mean value for skewness, all the maximum values were lowered following the codec compression. For CoG and frequency peak this was by over 1 kHz, which meant a change in CoG from 6,963 Hz to 5,554 Hz, and in frequency peak from 7,936 Hz to 6,719 Hz. SD lowered from 3,161 Hz to 2,577 Hz, while SD only changed with 0.05 to 4.51.

A typical spectrographic representation of voiced /ð/ in the WAV baseline and the AMR-WB compression using the high bitrate is presented below in figure 4.59. A general reduction in intensity can again be observed, while the spectrogram show how the formant structure correlating with the transition to the following vowel is clearer following the codec compression.

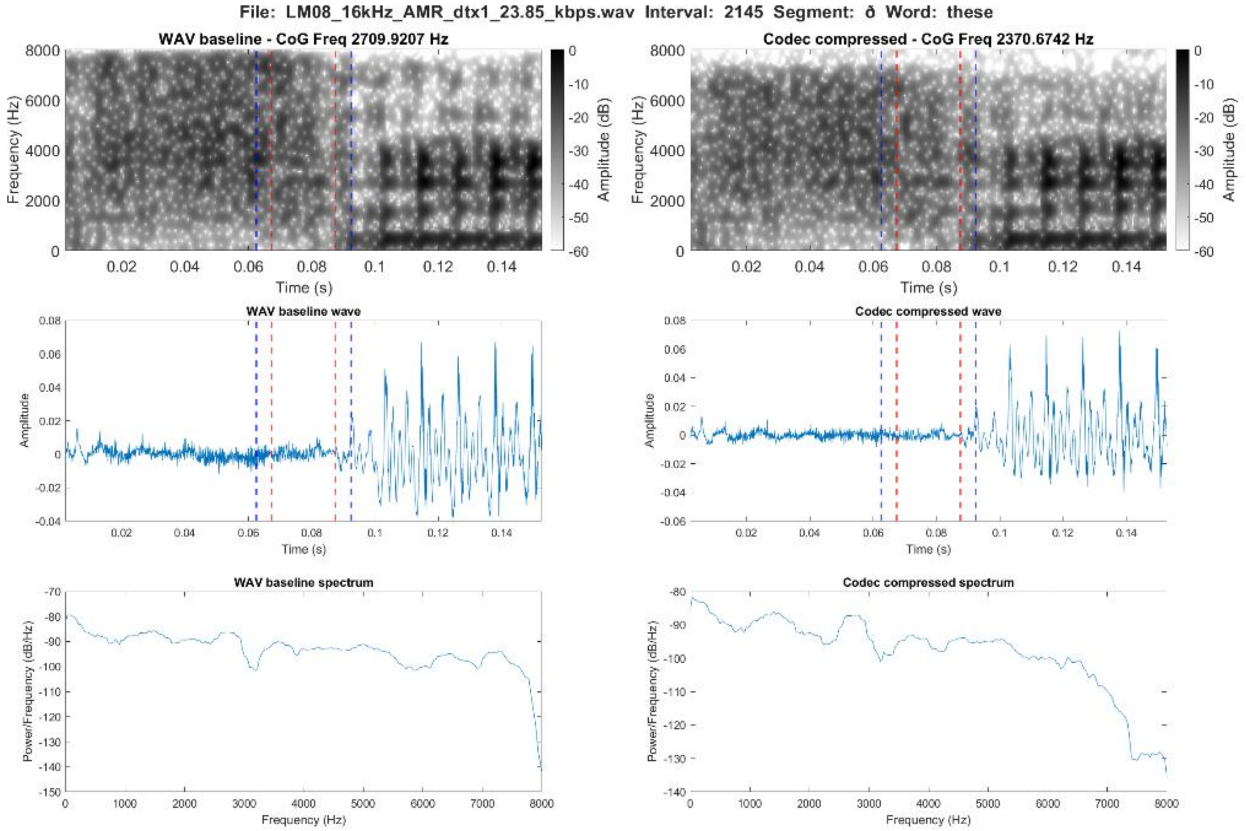


Figure 4.59. Spectrographic comparison of voiced /ð/ in the word *these* in the WAV baseline (left) and the AMR-WB compression at 23.85 kbps (right)

4.1.2.2 High bitrate: MP3

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the MP3 codec in the high quality bitrate dataset.

First, the linear predictions for each spectral measure and the individual segments can be found below (figure 4.60 to 4.65). These indicate the directionality of the changes inferred by the MP3 codec using the high bitrate. The graphs present the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these plots are in the following sections on the individual segments.

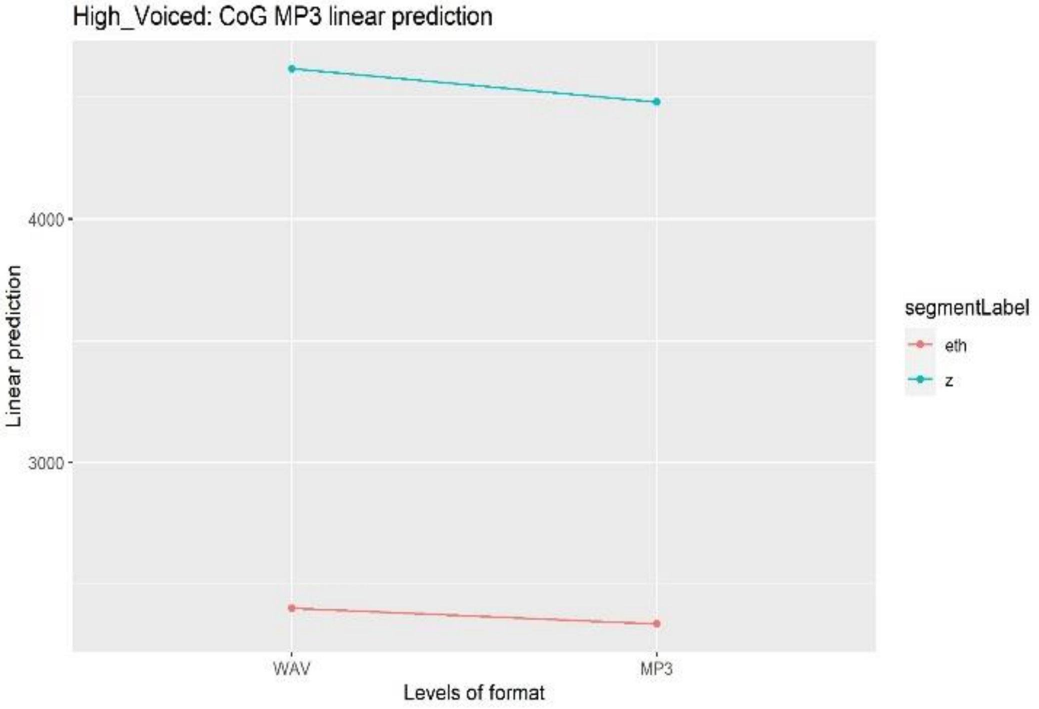


Figure 4.60. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

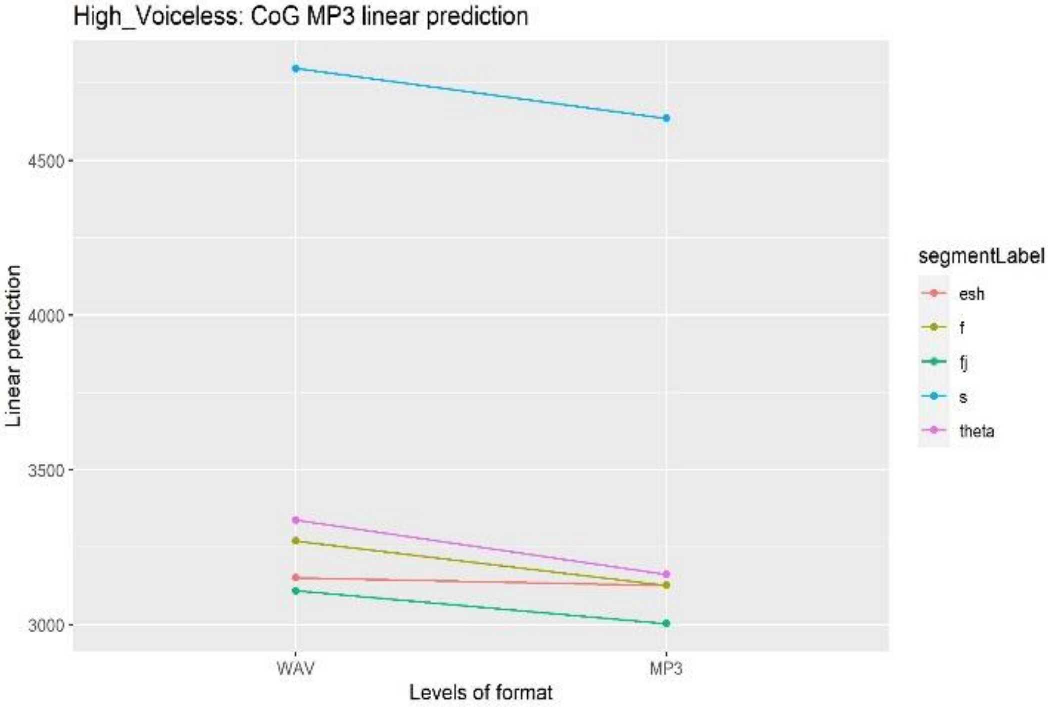


Figure 4.61. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

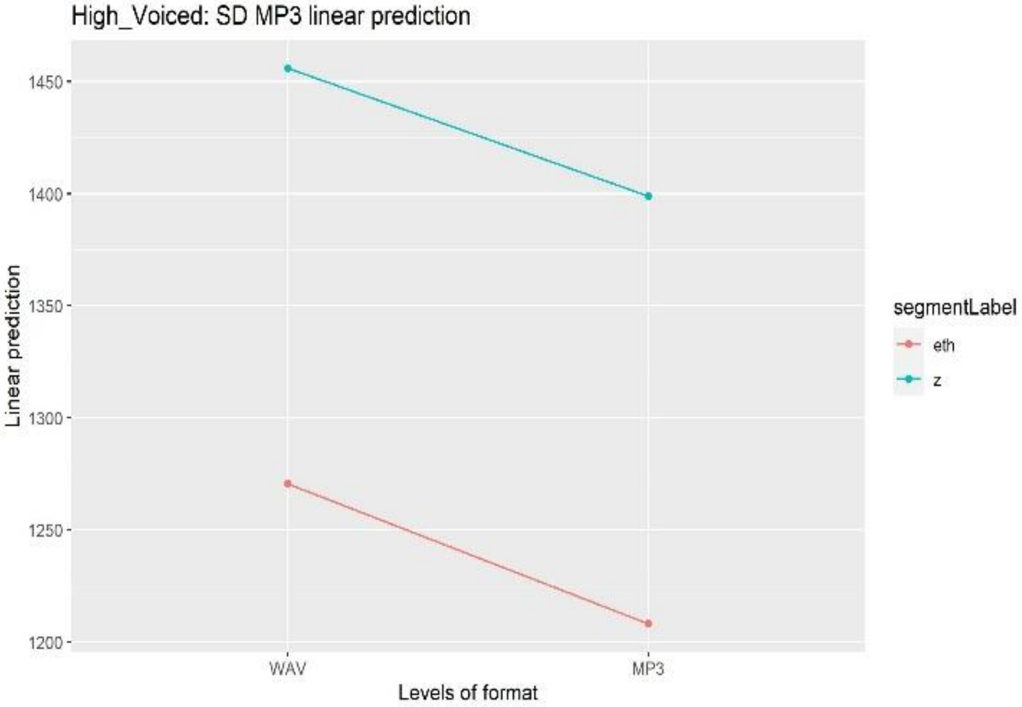


Figure 4.62. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

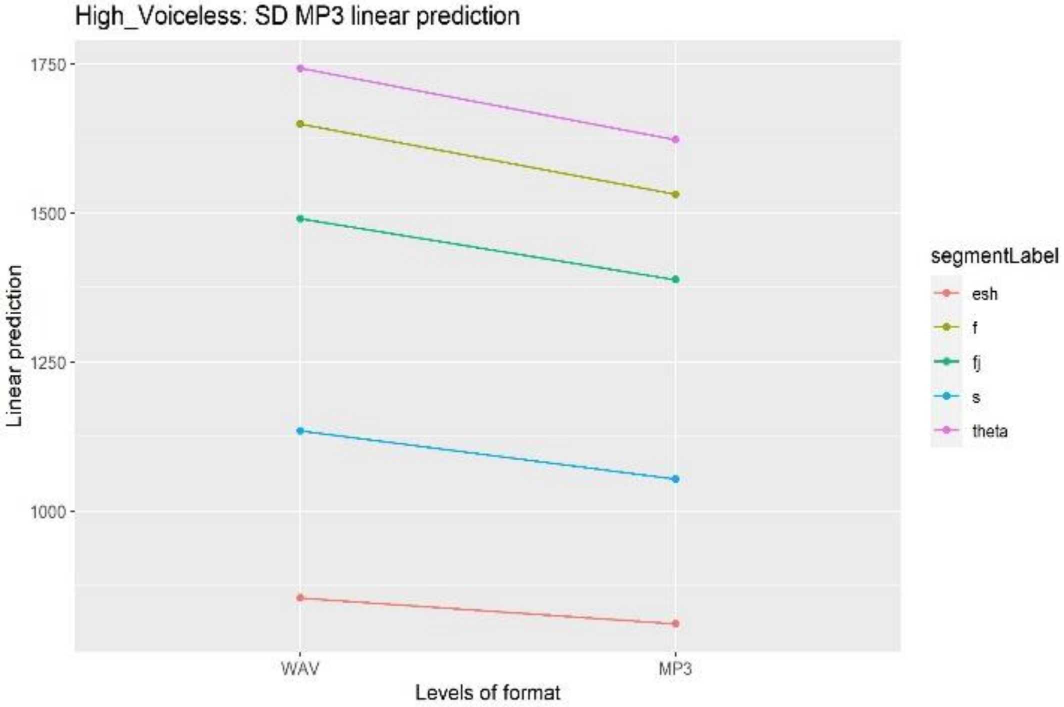


Figure 4.63. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

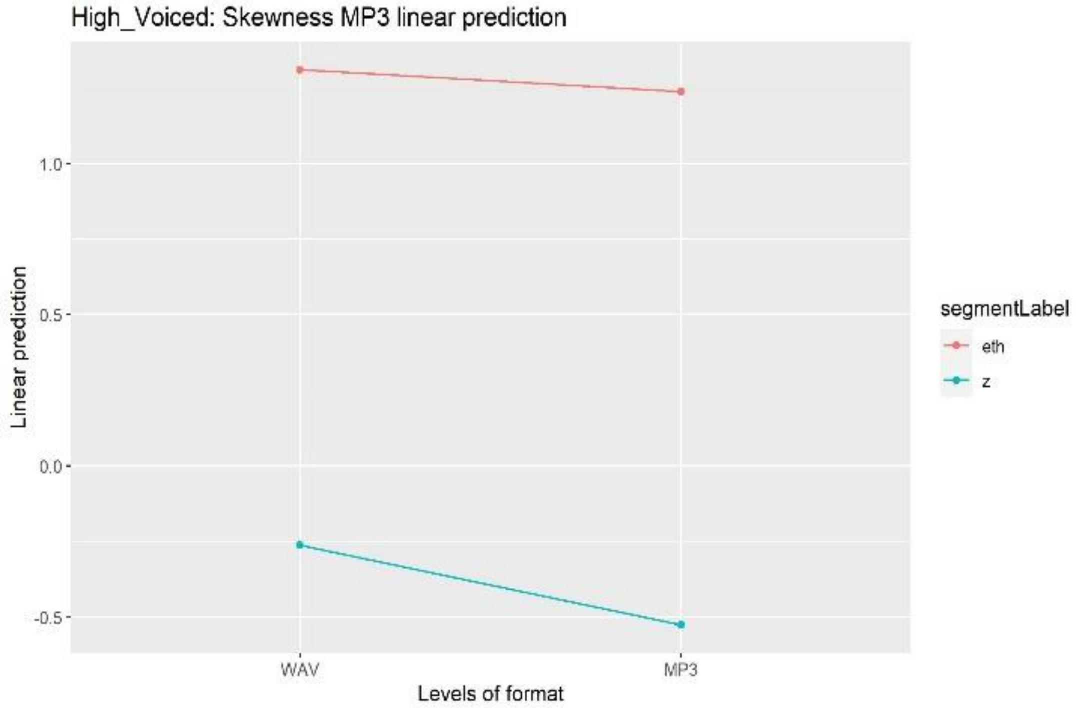


Figure 4.64. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

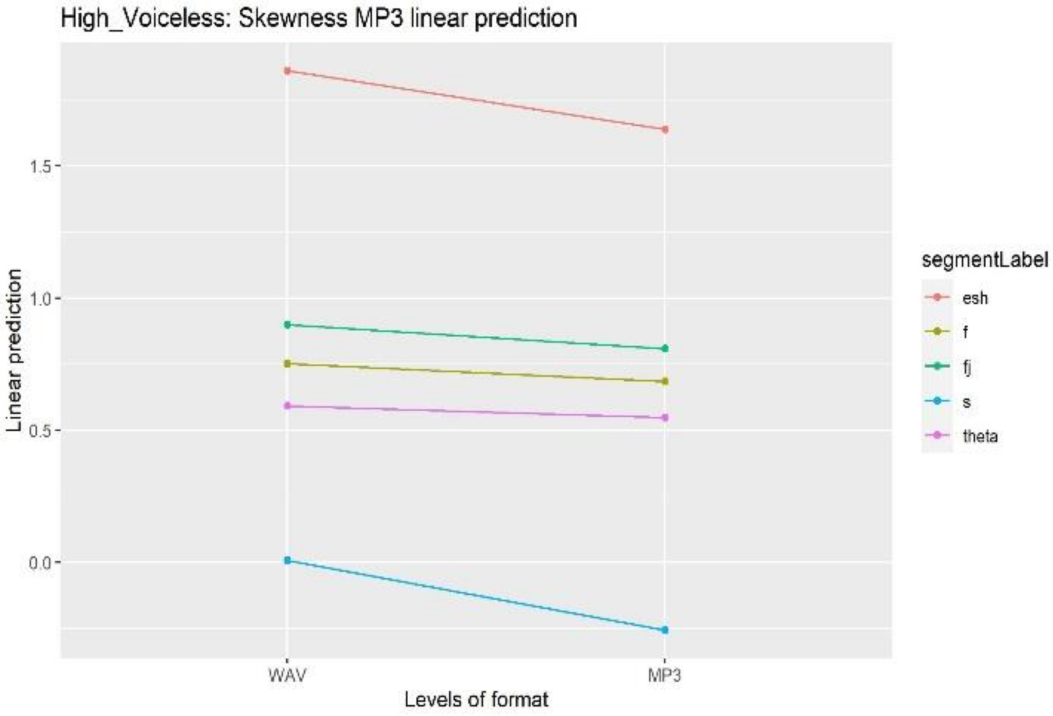


Figure 4.65. Trajectory of the linear predictions in the comparison of WAV and **high** MP3 from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The distributions are illustrated below in a set of violin plots for each spectral measure. Again, the specific analysis pertaining to each segment will be found in the following sections.

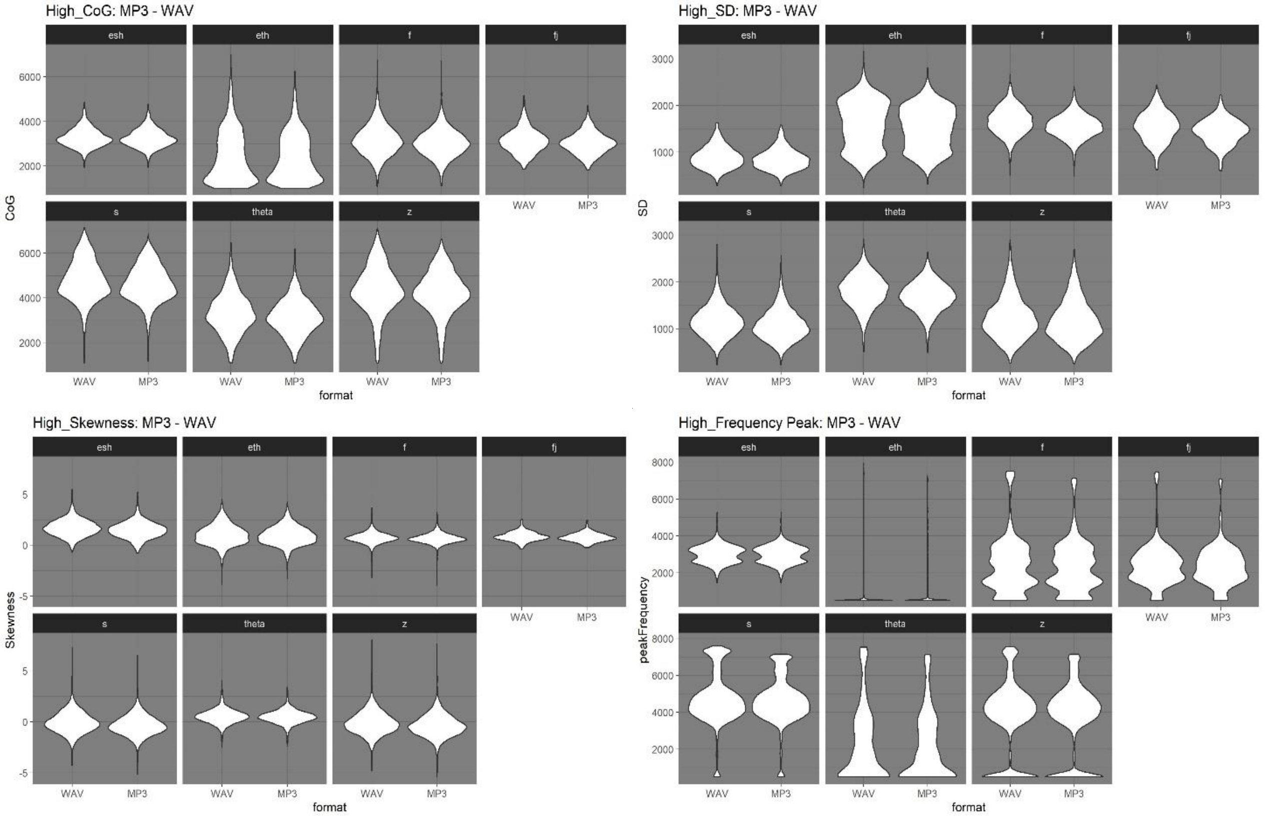


Figure 4.66. Distribution of spectral measure values in WAV baseline and the **High** bitrate MP3 codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

In sum, the linear predictions and distribution plots show how the MP3 compression overall lowered the spectral measures for all segments apart from CoG for /ʃ/, while the distributions show little changes caused by the codec compression.

4.1.2.2.1 High bitrate: /f/

The only notable effect on the distribution of /f/ following the codec compression is to SD, where more values are centred around the mean. The linear predictions reveal a downwards tendency for CoG, which becomes similar to /θ/ and almost identical to /ʃ/. As with the distributions, the clearest downwards trajectory is for SD, while skewness shows little to no effect of the codec compression.

Despite the limited changes to the distribution, the MP3 compression lowered all the mean values of the spectral measures (p<.0001). This was by just over 5 percent for frequency peak to just over 9 percent for skewness. CoG was lowered by 143 Hz or just over 5 percent, while SD lowered by 118 Hz or just over 7 percent. Kurtosis was similarly lowered by just over 4 percent. All p-values and further statistical results can be found in table 4.37 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/f/	142.83	9.92	<.0001
SD	WAV – MP3	/f/	117.92	18.23	<.0001
Skewness	WAV – MP3	/f/	0.07	3.92	<.0001

Table 4.37. Statistical results of the difference between WAV and high MP3 based on linear prediction models for /f/

As with the mean values, all the maximum values for /f/ lowered following the MP3 compression. For CoG, this was a change by 52 Hz to 6,702 and for frequency peak, a change by 375 Hz to 7,125 Hz. SD was lowered from 2,664 Hz to 2,410 Hz, while skewness lowered from 3.68 to 3.25.

A typical spectrographic representation of /f/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.67. Apart from a lower upper frequency limit and a slight reduction in intensity across the segment, no notable effects of the codec compression can be observed.

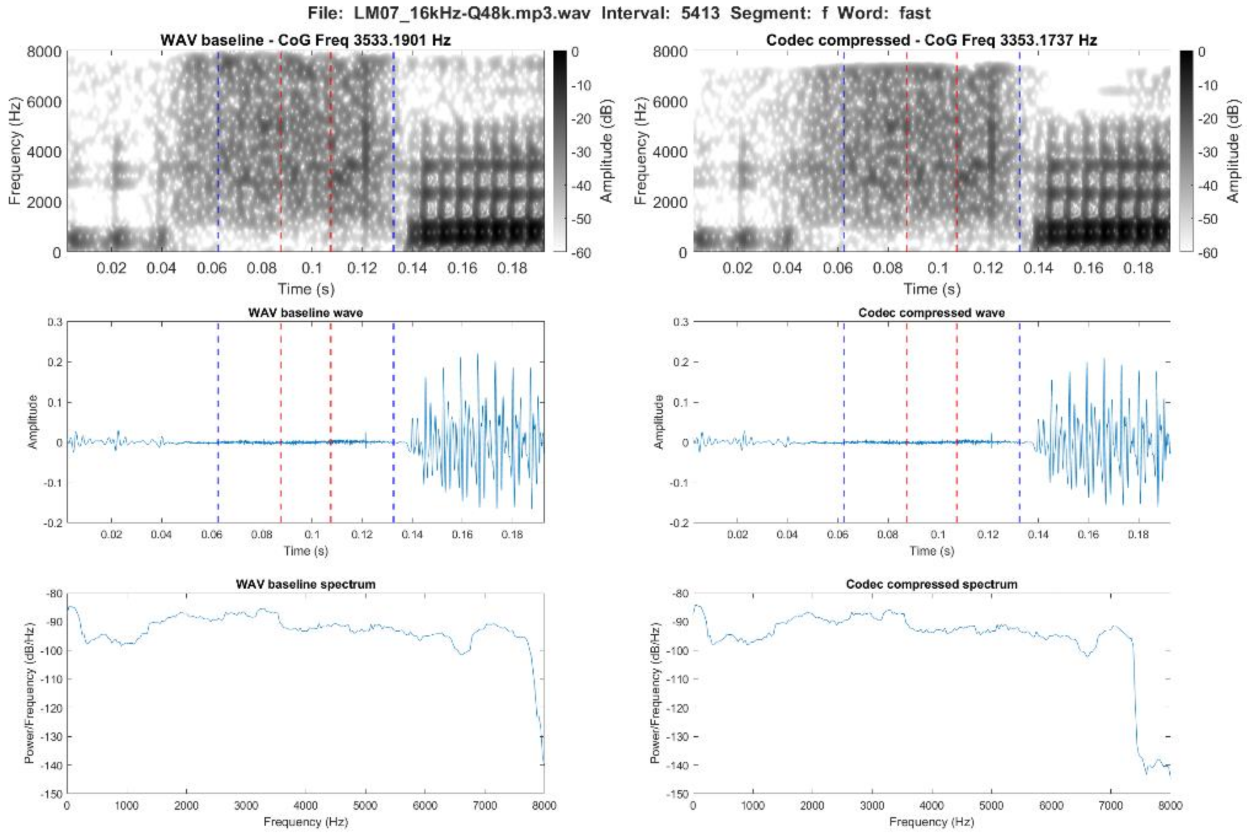


Figure 4.67. Spectrographic comparison of /f/ in the word *fast* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.2 High bitrate: [fʲ]

For CoG and SD a slight lowering can be observed together with a slight change to the distribution again with slightly more values centred around the mean. The linear predictions confirm this pattern and reveal only a very slight downwards trajectory for skewness.

Again, all the mean values were lowered. For CoG and SD this was by around 100 Hz for both or just over 3 percent for CoG (p = 0.05) and just under 7 percent for SD (p<.0001). Frequency peak lowered by 50 Hz or just over 2 percent, while skewness lowered by just over 10 percent. Kurtosis showed a decrease of just over 5 percent. All p-values and further statistical results can be found in table 4.38 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	[fʲ]	106.61	1.97	0.05
SD	WAV – MP3	[fʲ]	102.48	4.21	<.0001
Skewness	WAV – MP3	[fʲ]	0.09	1.39	0.17

Table 4.38. Statistical results of the difference between WAV and high MP3 based on linear prediction models for [fʲ]

All maximum values were also lowered. This was from 5,150 Hz to 4,710 Hz for CoG, and from 7,466 Hz to 7,063 Hz for frequency peak. SD lowered by just over 200 Hz to 2,227 Hz, while skewness decreased by 0.08 to 2.44.

A typical spectrographic representation of [fʲ] in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.68. In contrast to previous segments, an increase in intensity appear following codec compression across the segment, but especially in the final part.

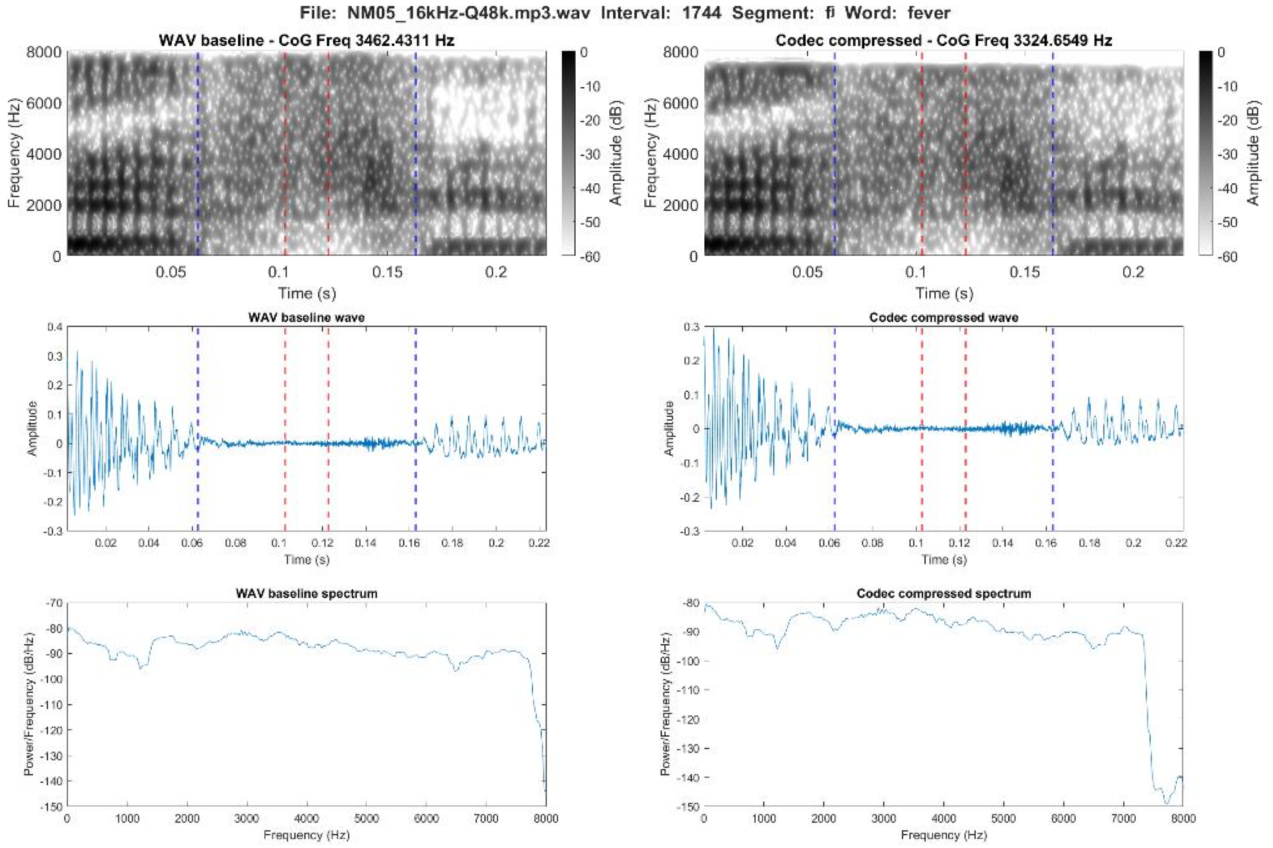


Figure 4.68. Spectrographic comparison of [fʲ] in the word *fever* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.3 High bitrate: /θ/

The changes observed to the distribution of /θ/ are similar to those observed for /f/ and [fʲ]. A lowering can be observed for CoG and more so SD together with a slight change to the distribution and an increased number of values around the mean. For the remaining spectral measures, the effects of the MP3 compression are limited. These tendencies are also true for the linear predictions, where CoG and SD present clear decreases, while skewness remains almost unchanged.

The MP3 compression lowered all the mean values for the spectral measures for /θ/. For CoG this was by just over 5 percent or 176 Hz, while for SD it was a change by just under 7 percent or 121 Hz (p<.0001). Frequency peak lowered by just over 6 percent or 147 Hz. Kurtosis showed a decrease by just under 4 percent, and skewness presented the largest though not significant effect by just over 8 percent or 0.04. All p-values and further statistical results can be found in table 4.39 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/θ/	175.67	7.47	<.0001
SD	WAV – MP3	/θ/	120.44	11.40	<.0001
Skewness	WAV – MP3	/θ/	0.04	1.54	0.12

Table 4.39. Statistical results of the difference between WAV and high MP3 based on linear prediction models for /θ/

As with the previous segments, all maximum values were lowered by the MP3 compression. For CoG this was a decrease by just over 300 Hz to 6,189 Hz, while SD showed a decrease of just over 200 Hz to 2,227 Hz. Frequency peak lowered from 7,531 Hz to 7,125 Hz, and skewness from 4.02 to 3.45.

A typical spectrographic representation of /θ/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.69. As with [fʲ], a general increase in intensity can be seen across the segment following codec compression.

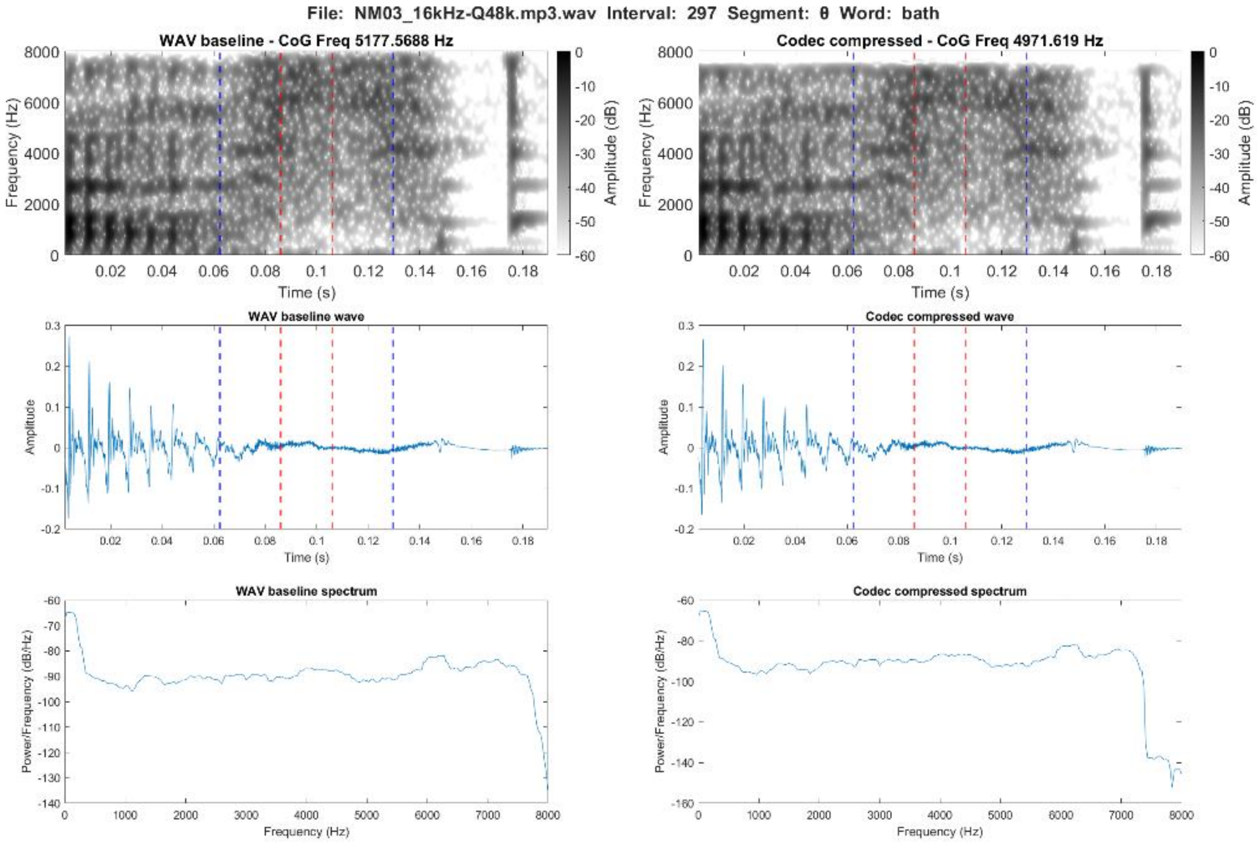


Figure 4.69. Spectrographic comparison of /θ/ in the word *bath* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.4 High bitrate: /s/

From the distribution plots, a lowering is observable for CoG, SD and skewness, which is also evident from the linear predictions. However apart from minor changes to the distribution of values above the mean for CoG and around the mean for SD, the distributions remain largely unchanged. The frequency peak only presents changes to the top-most values.

For /s/ the MP3 compression lowered all mean values (p<.0001) apart from kurtosis, which increased by just over 4 percent or 0.24. The most notable decrease was for skewness, which decreased by just under 309 percent or 0.26. CoG and frequency peak presented changes around 2-3 percent or for CoG 161 Hz and frequency peak 114 Hz. SD lowered by 81 Hz or just under 7 percent. All p-values and further statistical results can be found in table 4.40 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – MP3	/s/	161.03	16.40	<.0001
SD	WAV – MP3	/s/	81.41	18.40	<.0001
Skewness	WAV – MP3	/s/	0.26	22.40	<.0001

Table 4.40. Statistical results of the difference between WAV and high MP3 based on linear prediction models for /s/

The maximum values for the spectral measures all lowered, this was from 7,137 Hz to 6,880 Hz for CoG, and from 2,815 Hz to 2,566 Hz for SD. Frequency peak showed a decrease by just over 400 Hz to 7,156 Hz, while skewness changed from 7,303 to 6.56.

A typical spectrographic representation of /s/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.70. Here again a general increase in intensity can be observed across the segment, while the upper frequency limit is clear just under 8 kHz.

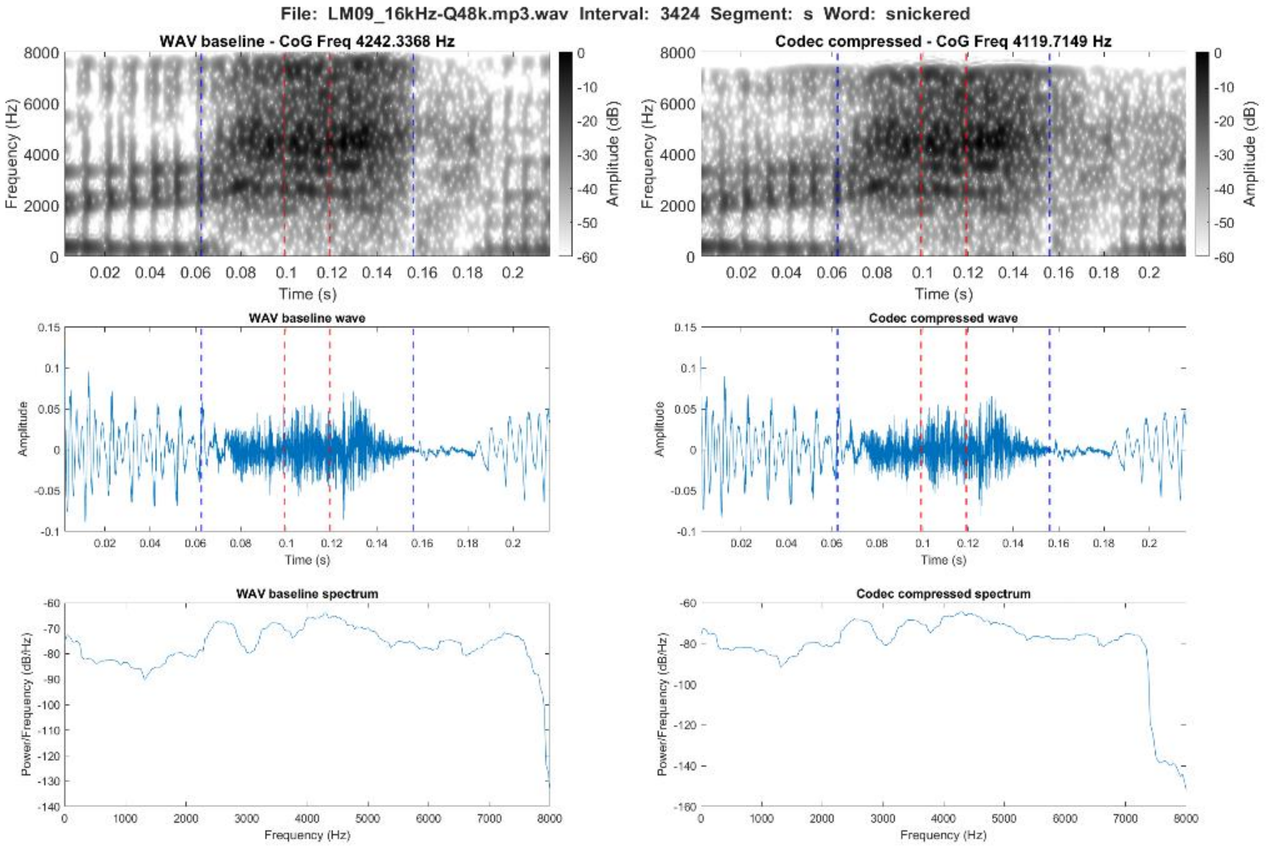


Figure 4.70. Spectrographic comparison of /s/ in the word *snickered* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.5 High bitrate: /ʃ/

No notable changes to the distribution can be observed for any of the spectral measures for /ʃ/. However, from the linear predictions a slight upwards trajectory can be observed for CoG, which makes /ʃ/ similar and almost identical to /f/ and /θ/, but more different from [fʲ] for this measure. SD and skewness both show a decrease, most clearly for skewness.

All mean values for the spectral measures were lowered for /ʃ/. Apart from skewness and kurtosis, which showed changes by just over 10 and just under 13 percent (p<.0001), the effects were limited. For both CoG and frequency peak, the changes were by less than 1 percent and for CoG the 25 Hz change was not significant. Frequency peak lowered by 6 Hz. SD showed a slightly larger change by 43 Hz or just under 5 percent (p<.0001). All p-values and further statistical results can be found in table 4.41 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – MP3	/ʃ/	24.76	1.05	0.29
SD	WAV – MP3	/ʃ/	42.74	4.04	<.0001
Skewness	WAV – MP3	/ʃ/	0.22	7.84	<.0001

Table 4.41. Statistical results of the difference between WAV and high MP3 based on linear prediction models for /ʃ/

For /ʃ/ all the maximum values apart from frequency peak were lowered following the MP3 compression. Frequency peak remained unchanged at 5,281 Hz in both WAV and MP3 compression. CoG lowered from 4,862 Hz to 4,772 Hz, while SD lowered with 45 Hz to 1,583 Hz. Skewness decreased by 0.16 to 5.22.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.71. Only a slight increase in intensity is visible for /ʃ/ following codec compression.

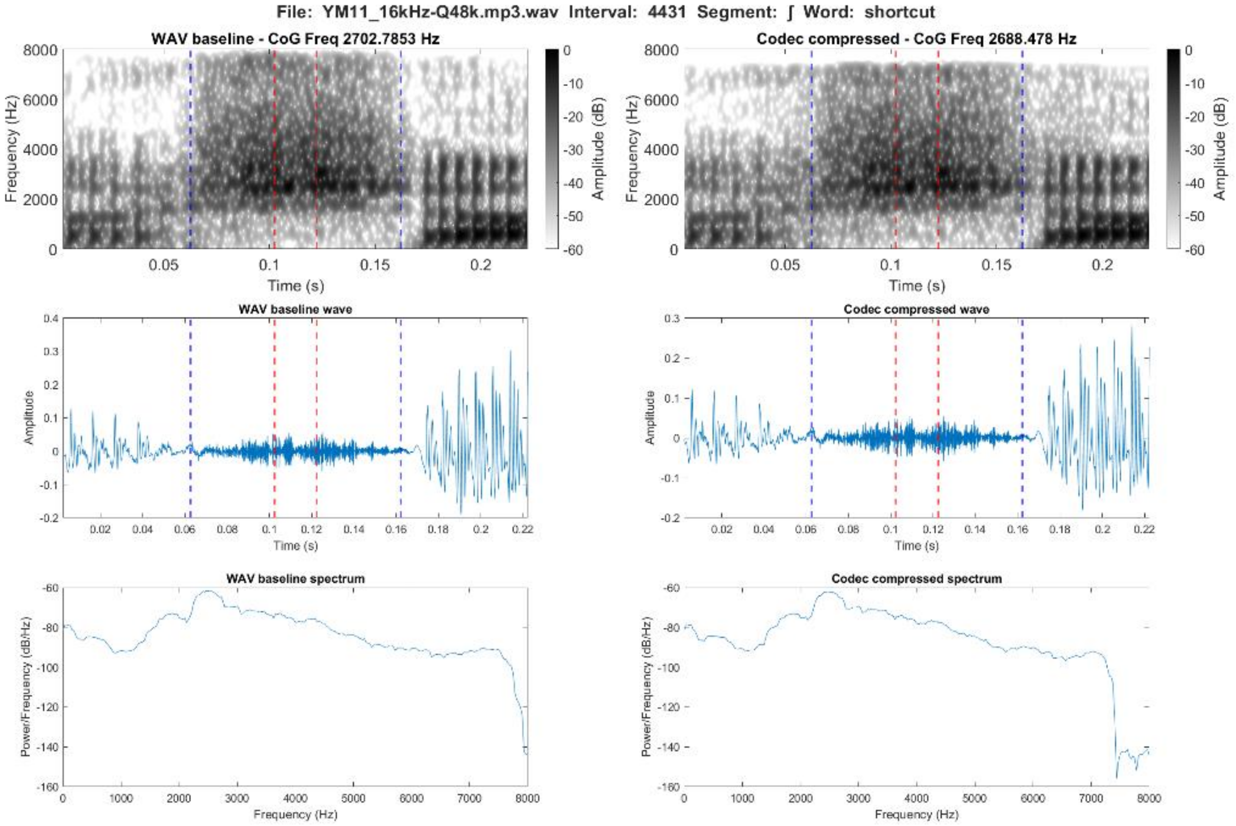


Figure 4.71. Spectrographic comparison of /ʃ/ in the word *shortcut* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.6 High bitrate: /z/

Overall, the changes or lack thereof to the distribution of /z/ following the MP3 compression is almost identical to the effects observed for /s/. This means an overall lowering of SD but little to no notable changes to the rest of the spectral measures. This can also been seen from the linear predictions, where skewness however also shows a more prominent downwards trajectory.

As with most of the previous segments, the MP3 compression lowered all the mean values (p<.0001) apart from kurtosis, which showed a slight increase by just under 1 percent. Again, skewness substantially lowered by just over 226 percent or twice the WAV baseline value, which was the same decrease as for /s/ of 0.26. For CoG the change was just over 3 percent or 137 Hz, while frequency peak lowered by 81 Hz or just over 2 percent. SD decreased by just under 5 percent or 57 Hz. All p-values and further statistical results can be found in table 4.42 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – MP3	/z/	136.78	7.19	<.0001
SD	WAV – MP3	/z/	57.06	6.68	<.0001
Skewness	WAV – MP3	/z/	0.26	12.65	<.0001

Table 4.42. Statistical results of the difference between WAV and high MP3 based on linear prediction models for /z/

All maximum values lowered for /z/ following the MP3 compression. This was by just under 600 Hz for CoG to 6,626 Hz, and by just over 200 Hz for SD to 2,692 Hz. Frequency peak lowered from 7,563 Hz to 7,156 Hz, while skewness lowered from 8.05 to 7.68.

A typical spectrographic representation of voiceless /z/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.72. The main effect of the codec compression again appear as a slight increase in intensity across the segment.

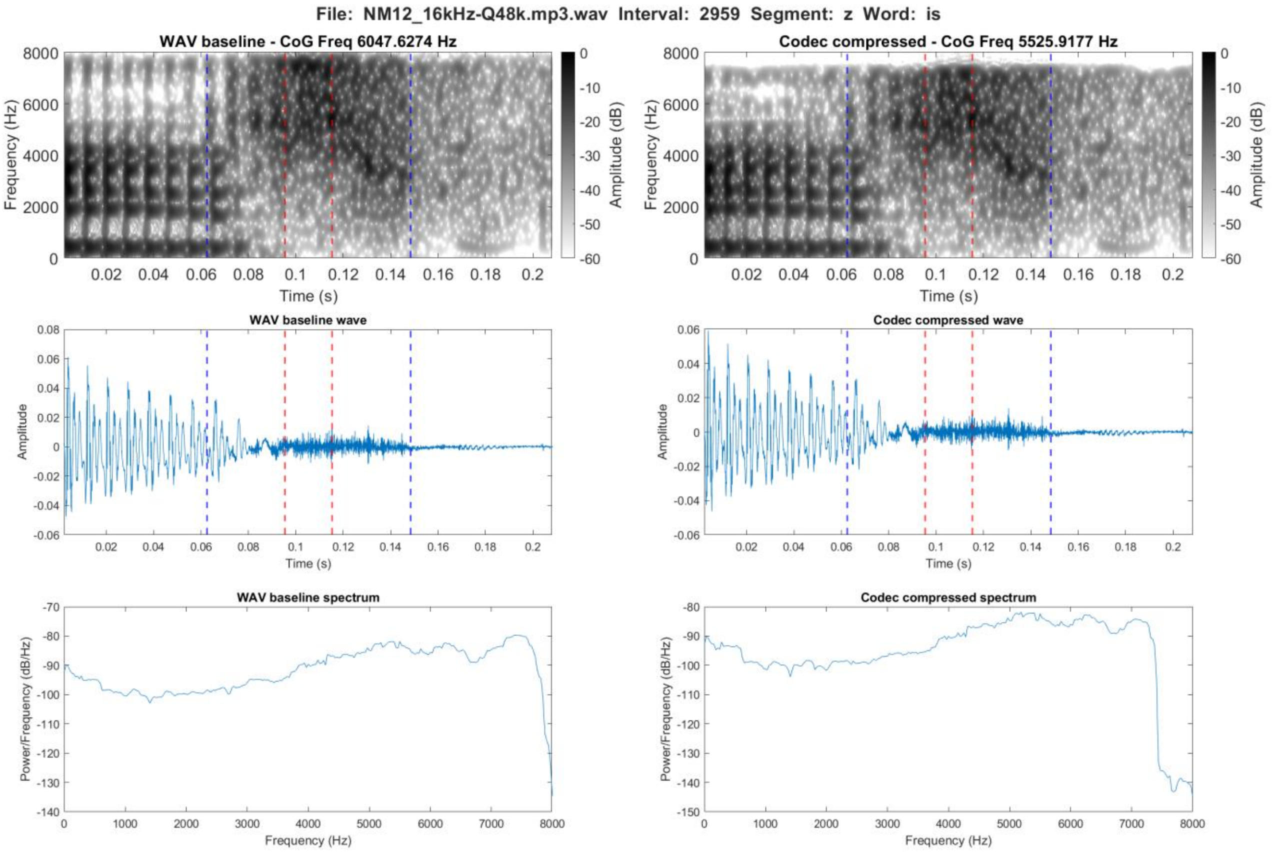


Figure 4.72. Spectrographic comparison of voiceless /z/ in the word *is* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.2.7 High bitrate: /ð/

From the distribution plots a slight increase in number of values towards the lower cut-off appear for CoG following the codec compression, while SD over all maintain the distribution shape, but spread over a smaller range of values. No effect is clearly observable for skewness and frequency peak. The linear predictions show a downwards trends, but only slight decreases for CoG and skewness.

As with the majority of the previous segments, all the mean values for the spectral measures lowered following the codec compression. For CoG and SD, this was with 62 and 63 Hz respectively, or just under 3 percent for CoG (p= .004) and just under 4 percent for SD (p<.0001). Frequency peak lowered by a similar percentage i.e. just over 4 percent, while the decrease in Hz was 53. Similarly, skewness and kurtosis both lowered by around 4 percent (p=0.003). All p-values and further statistical results can be found in table 4.43 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – MP3	/ð/	63.48	2.85	<.004
SD	WAV – MP3	/ð/	62.46	6.25	<.0001
Skewness	WAV – MP3	/ð/	0.07	2.97	0.003

Table 4.43.Statistical results of the difference between WAV and high MP3 based on linear prediction models for /ð/

The maximum values for /ð/ were all lowered by the MP3 compression. For CoG this was from 6,963 Hz to 6,243 Hz, and for SD from 3,161 Hz to 2,803 Hz. Frequency peak lowered by around 700 Hz to 7,281, while skewness lowered by 0.31 to 4.25.

A typical spectrographic representation of plosive /ð/ in the WAV baseline and the MP3 compression using the high bitrate is presented below in figure 4.73. Again, an increase in intensity can be observed, which is here mainly visible from the spectrum.

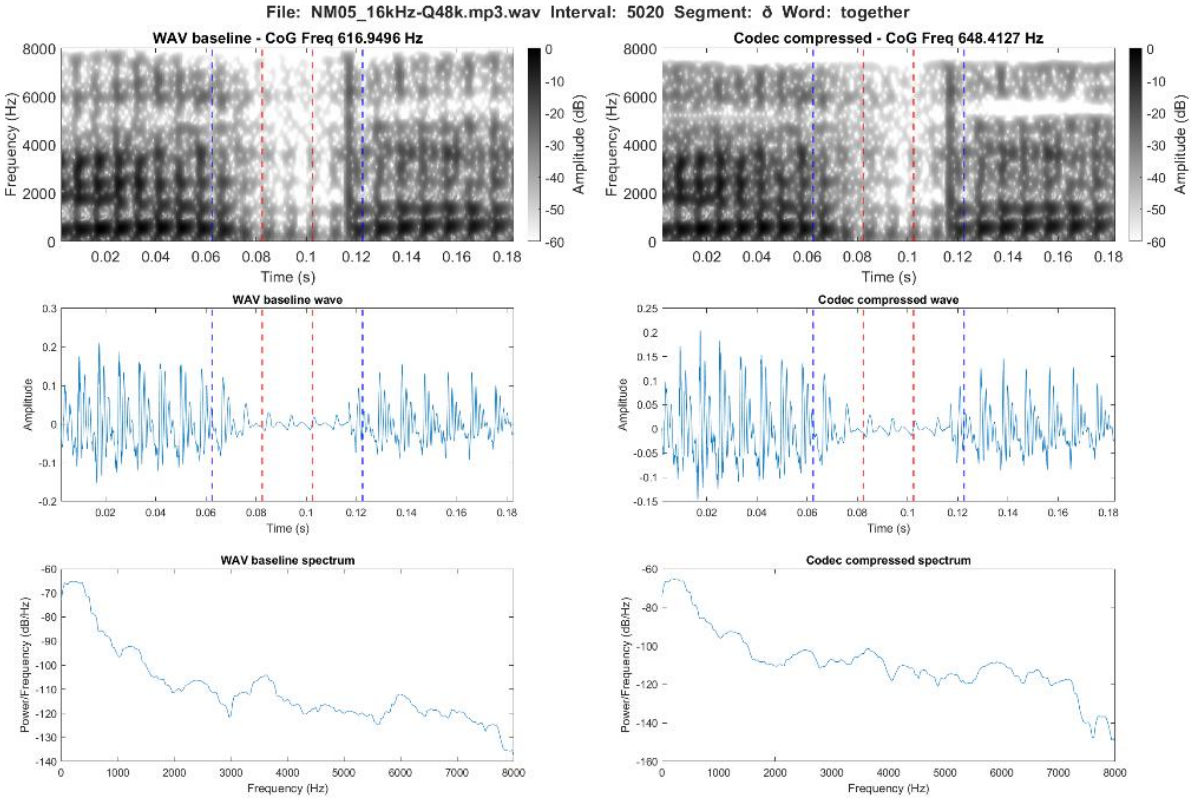


Figure 4.73. Spectrographic comparison of plosive /ð/ in the word *together* in the WAV baseline (left) and the MP3 compression at 48 kbps (right)

4.1.2.3 High bitrate: Opus

This section will present the individual results for each segment and the spectral measures in the comparison between the WAV baseline and the Opus codec in the high quality bitrate dataset.

First, the linear predictions for each spectral measure and the individual segments can be found below (figure 4.74 to 4.79). These indicate the directionality of the changes inferred by the Opus codec using the high bitrate. The graphs present the results as voiced and voiceless segments as this was the grouping made in the mixed effects modelling. The detailed analysis of these plots are in the following sections on the individual segments.

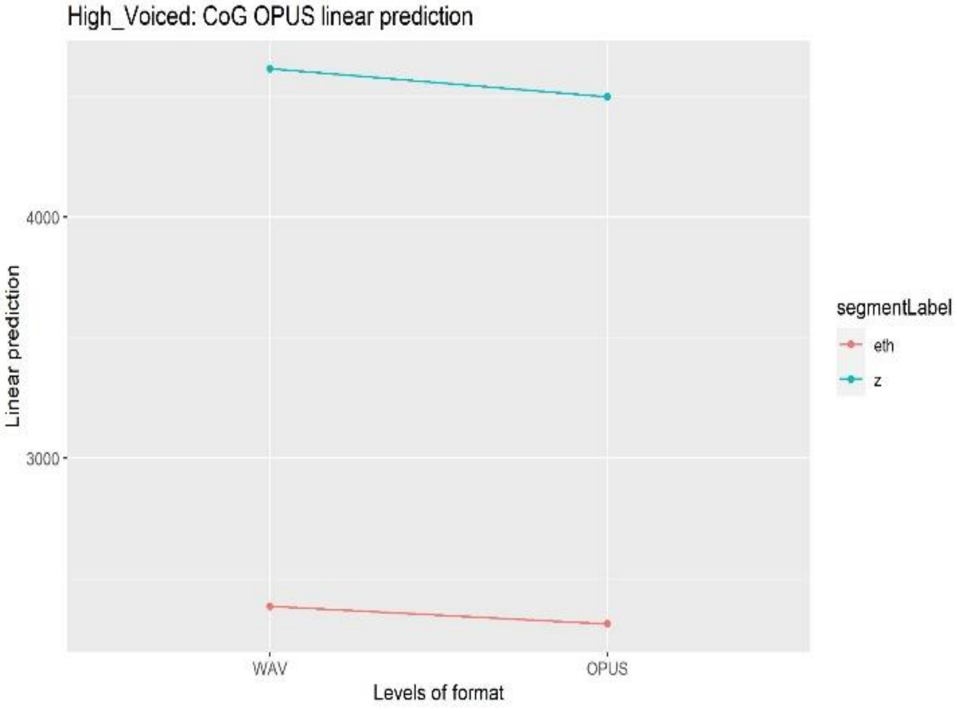


Figure 4.74. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for CoG and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

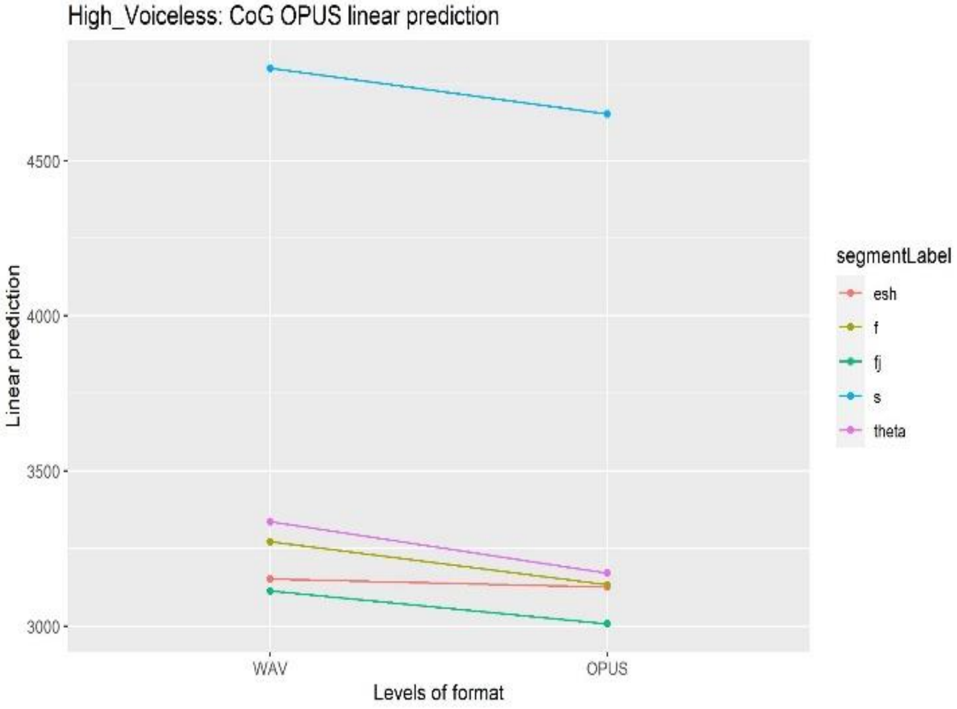


Figure 4.75. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for CoG and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

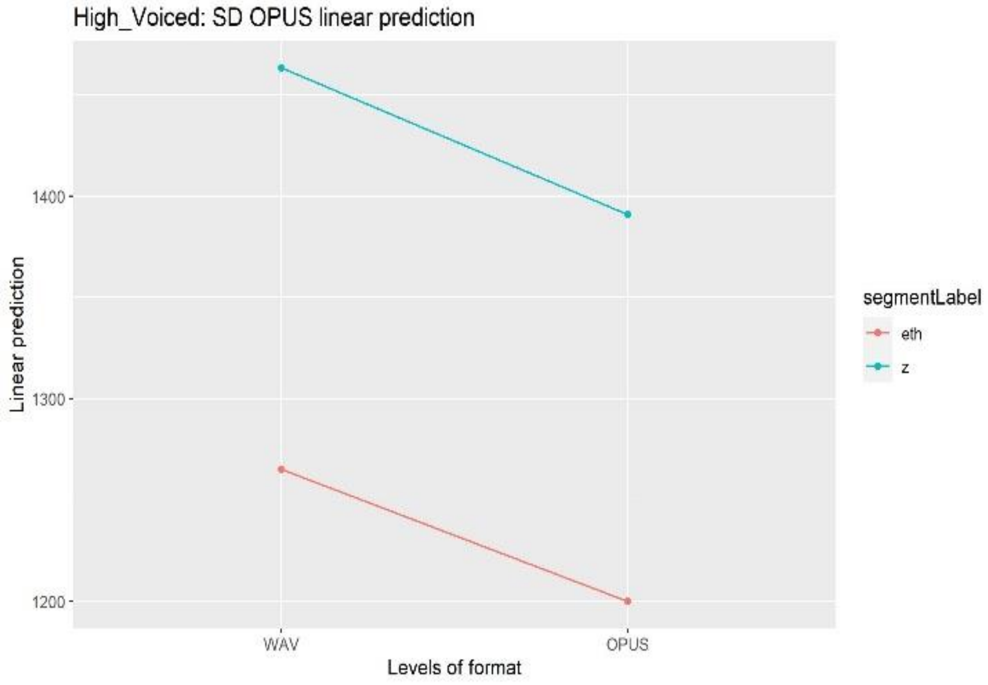


Figure 4.76. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

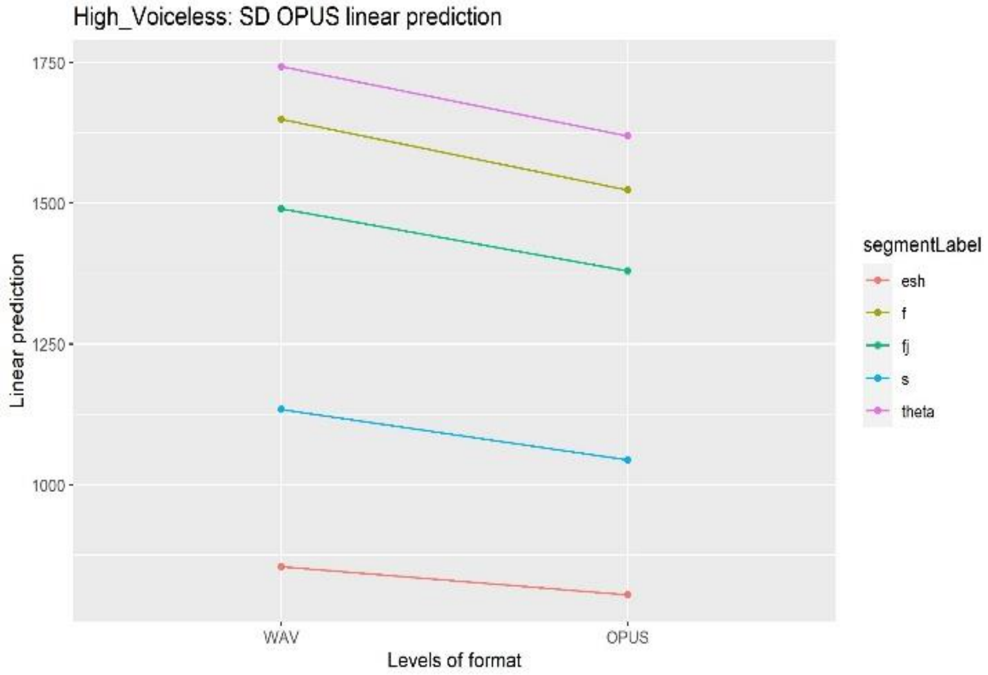


Figure 4.77. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for SD and the voiceless segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

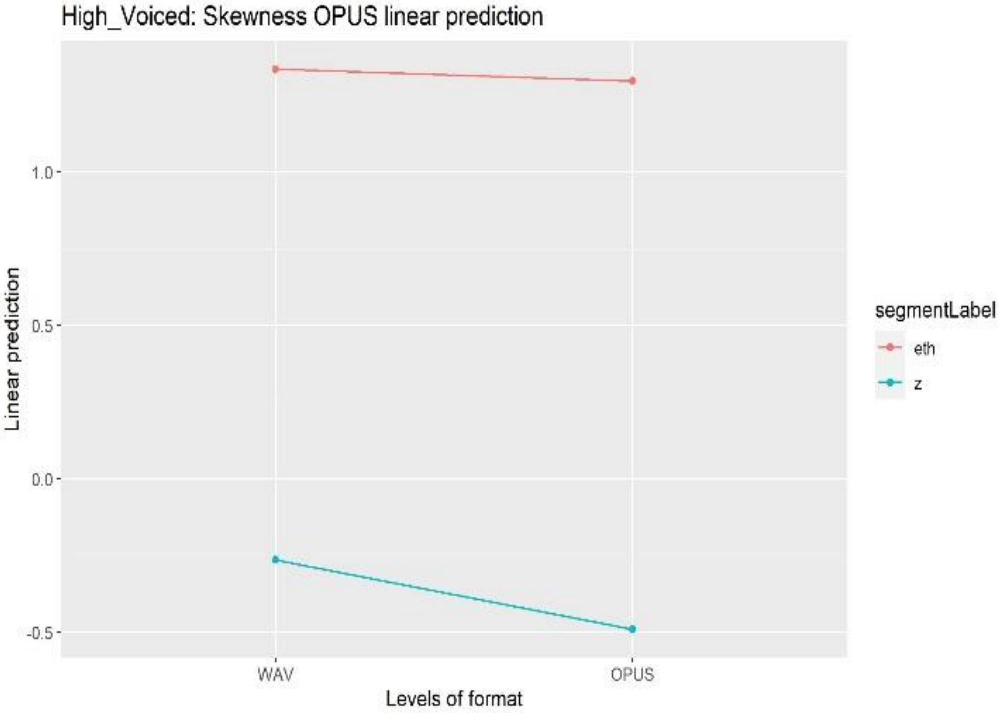


Figure 4.78. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for skewness and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

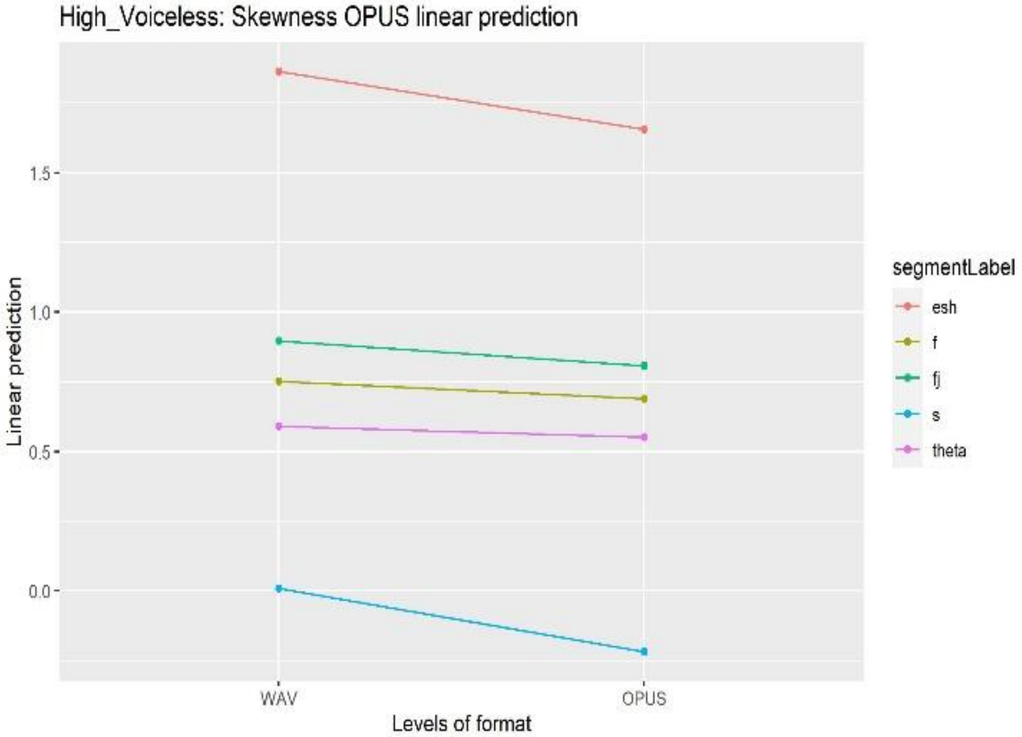


Figure 4.79. Trajectory of the linear predictions in the comparison of WAV and **high** Opus from the mixed effects models for SD and the voiced segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The distributions are illustrated below in a set of violin plots for each spectral measure (see figure 4.80). Apart from some effects on SD seen as some overall lowered distributions and for /f/ and [fʲ] more values centred around the mean, the rest of the spectral measures present very limited changes to the distribution following the Opus compression. Therefore, the individual analysis of each segment in the following sections will not consider the distributional effects further for the Opus codec.

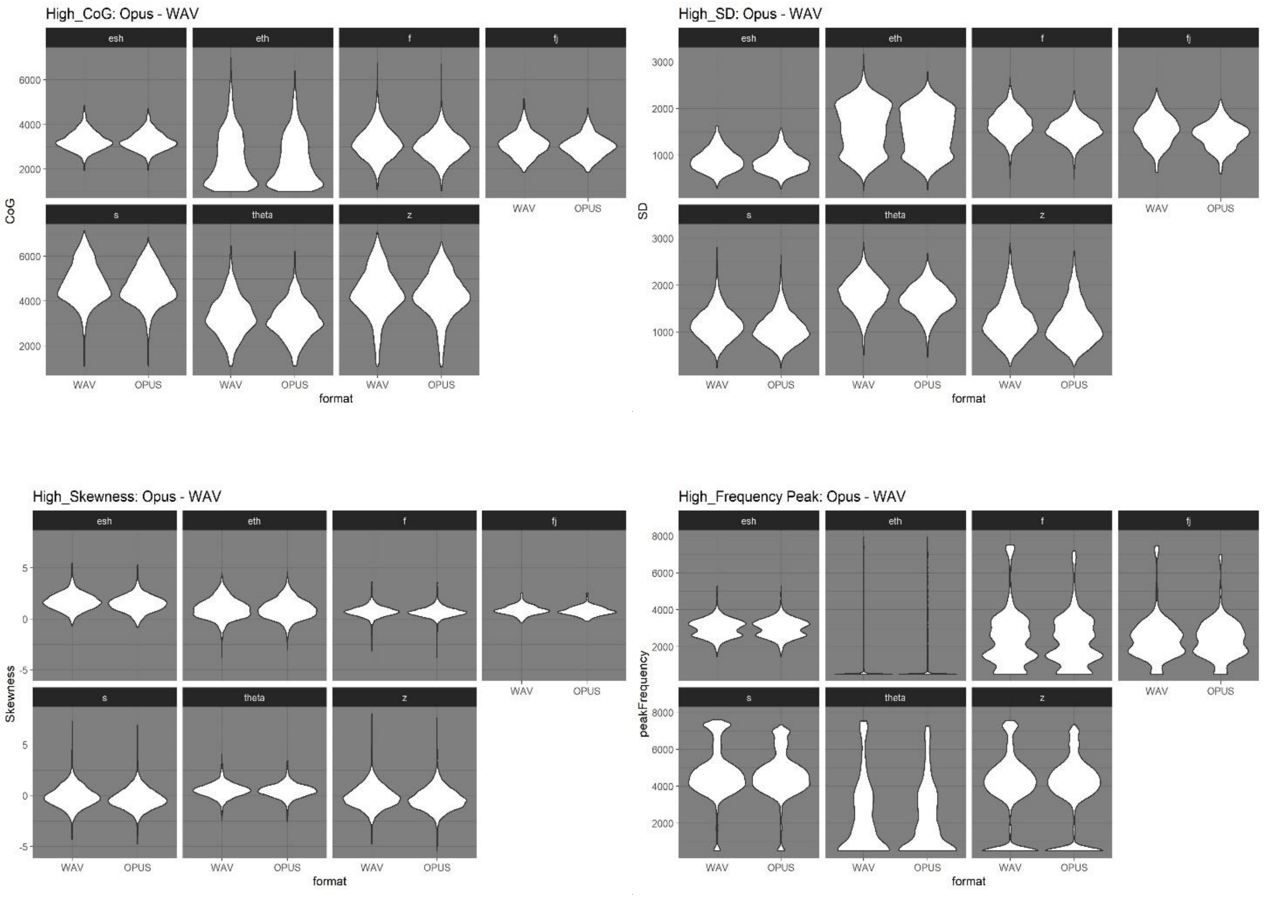


Figure 4.80. Distribution of spectral measure values in WAV baseline and the **High** bitrate Opus codec compression grouped by spectral measure and divided by individual segments. The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* = /ʃ/.

In brief, both linear predictions and distribution plots show limited changes caused by the codec compression. The linear predictions show that the spectral measures are generally lowered or stay similar from WAV to the Opus compression, while the distribution show changes mainly to the higher frequencies.

4.1.2.3.1 High bitrate: /f/

The linear predictions reveal a slight downwards tendency for CoG, while SD more clearly decrease and skewness is overall stable following the codec compression.

In more detail, the Opus compression lowered all mean values of all spectral measures significantly for /f/ (p<.001). This was by just under 3 percent for kurtosis to just over 8 percent for skewness. In between, CoG lowered by just over 4 percent or 139 Hz, while frequency peak lowered by 117 Hz or just under 5 percent. For SD, the change was by 126 Hz or just under 8 percent. All p-values and further statistical results can be found in table 4.44 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/f/	138.93	9.58	<.0001
SD	WAV – Opus	/f/	125.71	19.36	<.0001
Skewness	WAV – Opus	/f/	0.06	3.61	<.001

Table 4.44. Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /f/

As with the mean values, all maximum values lowered following the Opus compression. This was a decrease by 60 Hz for CoG to 6,694 Hz, while SD lowered by around 250 Hz to 2,392 Hz. Frequency peak decreased the most from 7,500 Hz to 7,188 Hz. Skewness lowered from 3.68 to 3.59.

A typical spectrographic representation of /f/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.81. A very slight increase in intensity in mainly the frequencies around 2 kHz in the final part of the segment can be observed. Apart from this, the effect of the codec compression only appear as the effect of the upper frequency limit.

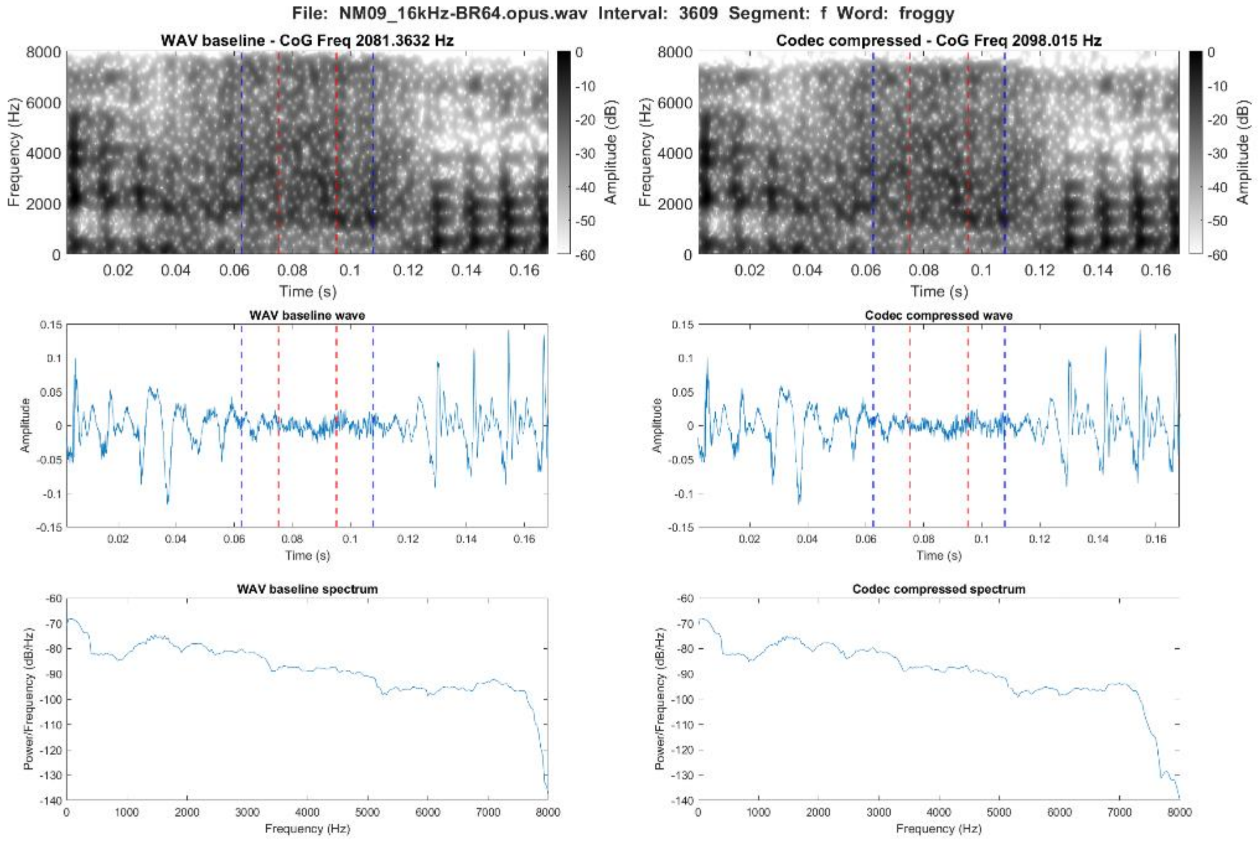


Figure 4.81. Spectrographic comparison of /f/ in the word *froggy* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.2 High bitrate: [fʲ]

The linear predictions show a slight downwards tendency for CoG, a clearer downwards trend for SD and only a very slight decrease for SD.

Again, similar to /f/ all mean values lowered following the Opus compression. For CoG and SD this was by just over 100 Hz for both or by just over 3 percent for CoG (p = 0.05) and by just over 7 percent for SD (p <.0001). Frequency peak lowered by 39 Hz or just under 2 percent, while skewness lowered by just over 10 percent, however, this was not significant. Kurtosis lowered by just under 5 percent. All p-values and further statistical results can be found in table 4.45 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	[fʲ]	106.06	1.94	0.05
SD	WAV – Opus	[fʲ]	110.11	4.50	<.0001
Skewness	WAV – Opus	[fʲ]	0.09	1.37	0.17

Table 4.45. Statistical results of the difference between WAV and **high** Opus based on linear prediction models for [fʲ]

Apart from skewness, which remained almost stable by just a slight increase of 0.02 to 2.54, all maximum values were lowered by the codec compression. This was from 5,150 Hz to 4,724 Hz for CoG, and from 2,436 Hz to 2,191 Hz for SD. Frequency peak lowered by 500 Hz to 6,969 Hz.

A typical spectrographic representation of [fʲ] in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.82. As with the previously observed, an increase in intensity is visible across the segment.

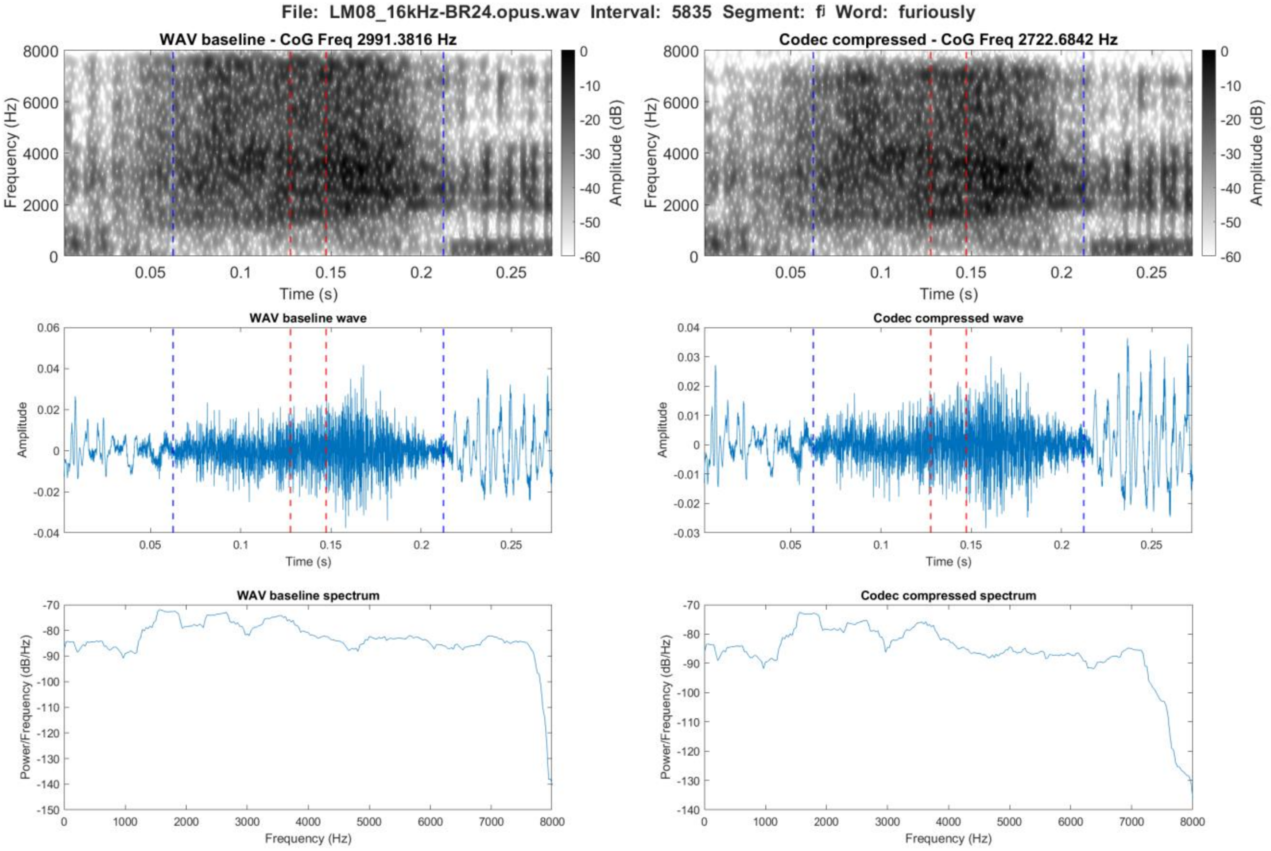


Figure 4.82. Spectrographic comparison of [fʲ] in the word *furiously* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.3 High bitrate: /θ/

As with the previous two voiceless segments, the linear predictions show the clearest decrease for CoG, while skewness remains largely unchanged by the codec compression.

The mean values for the spectral measures all lowered following the opus compression. This was by just under 2 percent for kurtosis to just over 7 percent for skewness. Both SD and frequency peak lowered by around 6 percent, which for SD was a change by 123 Hz (p<.0001) and for frequency peak by 149 Hz. CoG lowered by 5 percent or 166 Hz (<.0001). All p-values and further statistical results can be found in table 4.46 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/θ/	166.11	7.02	<.0001
SD	WAV – Opus	/θ/	123.41	11.64	<.0001
Skewness	WAV – Opus	/θ/	0.04	1.36	0.17

Table 4.46.Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /θ/

As with the mean values, all maximum values lowered following the Opus compression. For CoG this was by just under 270 Hz to 6,224 Hz, while SD lowered by just over 250 Hz to 2,669 Hz. Frequency peak similarly lowered by just over 280 Hz to 7,250 Hz, while skewness lowered from 4.02 to 3.43.

A typical spectrographic representation of /θ/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.83. No notable effect of the codec compression can be observed from this.

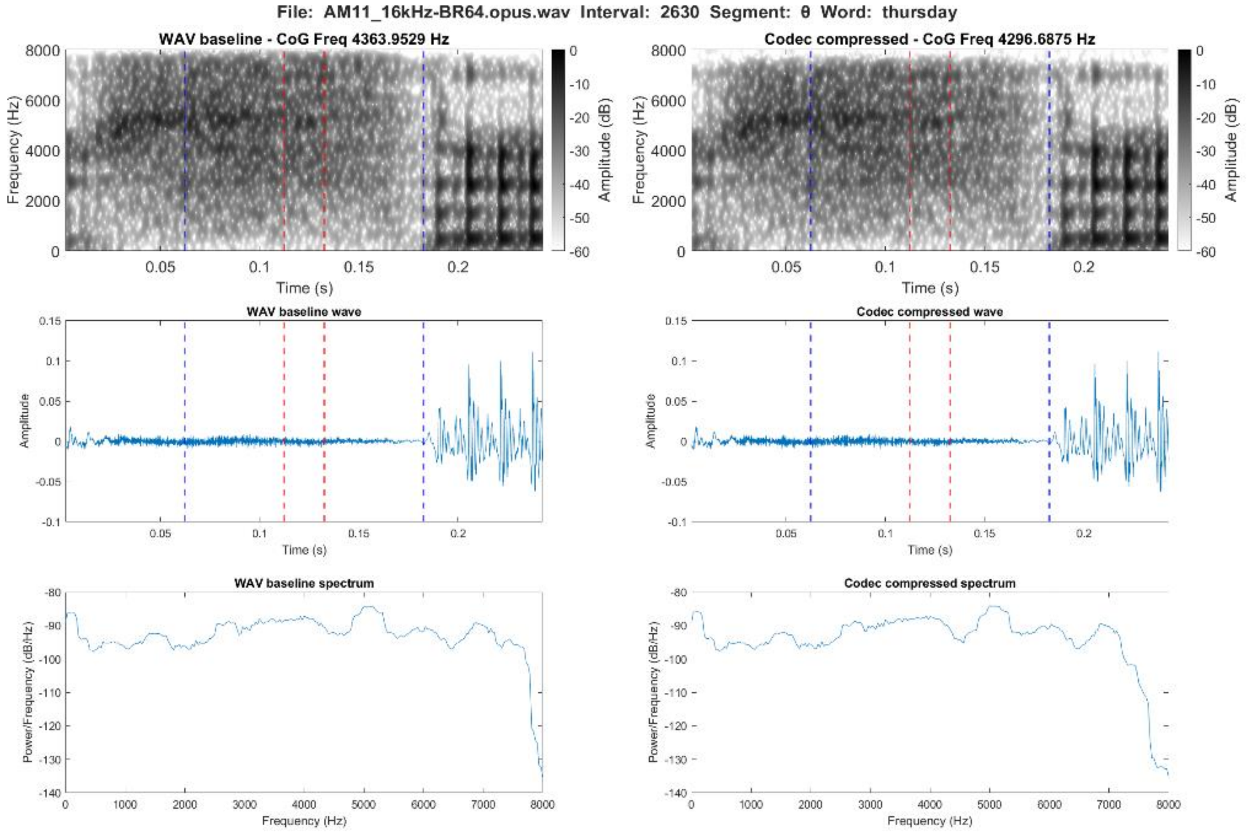


Figure 4.83. Spectrographic comparison of /θ/ in the word *Thursday* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.4 High bitrate: /s/

The linear predictions show that CoG, SD and skewness lowers with similar trajectories following the Opus compression.

For /s/ all mean values for the spectral measures (p<.0001), apart from kurtosis lowered following the codec compression. Kurtosis showed an increase by 5 percent. Skewness presented the biggest change with a decrease by just under 264 percent or 0.23. Both CoG and frequency peak lowered by around 3 percent or for CoG 148 Hz and for frequency peak 122 Hz. SD presented a smaller decrease of 89 Hz, which was just under 8 percent. All p-values and further statistical results can be found in table 4.47 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	/s/	148	14.98	<.0001
SD	WAV – Opus	/s/	89.40	20.20	<.0001
Skewness	WAV – Opus	/s/	023	19.02	<.0001

Table 4.47.Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /s/

The maximum values again all lowered. This was from 7,137 Hz to 6,865 Hz for CoG, and from 2,815 Hz to 2,647 Hz for SD. Frequency peak lowered from 7,594 Hz to 7,313 Hz, while skewness decreased from 7.33 to 6.92.

A typical spectrographic representation of /s/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.84. An increase in intensity can be observed across the segment apart from a slight drop in amplitude around 6 kHz visible from the spectrum.

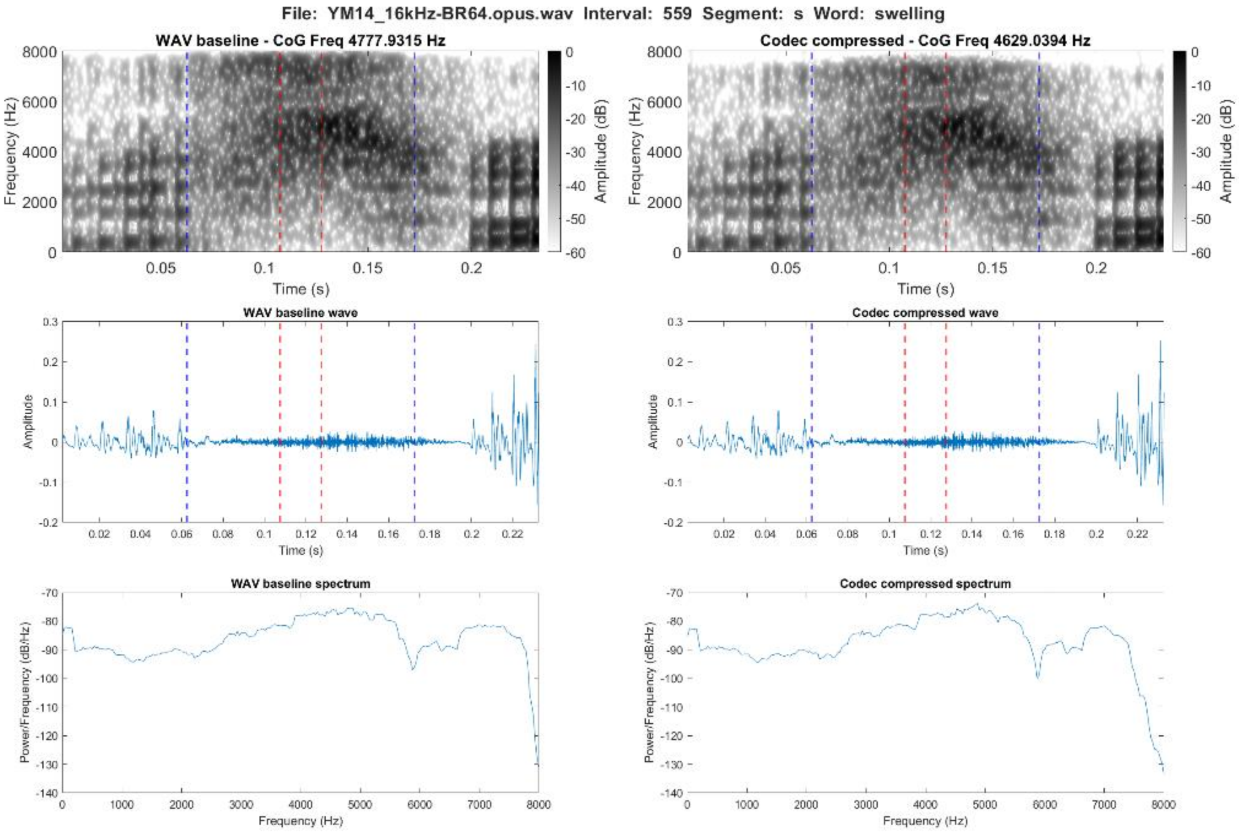


Figure 4.84. Spectrographic comparison of /s/ in the word *swelling* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.5 High bitrate: /ʃ/

For /ʃ/ the linear predictions show a slight increase of CoG, which result in /ʃ/ becoming close to identical to /f/. SD presents a slight downwards tendency, while skewness more clearly lowers.

All mean values for the spectral measures were again lowered. This was by less than 1 percent for CoG and frequency peak, which for CoG meant a change of 25 Hz and for frequency peak a change of just 5 Hz. SD lowered by 50 Hz or just under 6 percent (p<.0001). Skewness presented a change of just over 12 percent (p<.0001), while kurtosis lowered by just over 9 percent. All p-values and further statistical results can be found in table 4.48 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/ʃ/	25.32	1.07	0.29
SD	WAV – Opus	/ʃ/	49.63	4.68	<.0001
Skewness	WAV – Opus	/ʃ/	0.21	7.25	<.0001

Table 4.48. Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /ʃ/

The more subtle changes observed for the mean values were also evident for the maximum values. For CoG this was a change of just under 50 Hz to 4,708 Hz, while SD lowered by 150 Hz to 1,579 Hz. Skewness decreased from 5.48 to 5.34, while frequency peak remained stable with a maximum value of 5,281 Hz in both WAV and codec compression.

A typical spectrographic representation of /ʃ/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.85. No notable effects of the codec compression is visible.

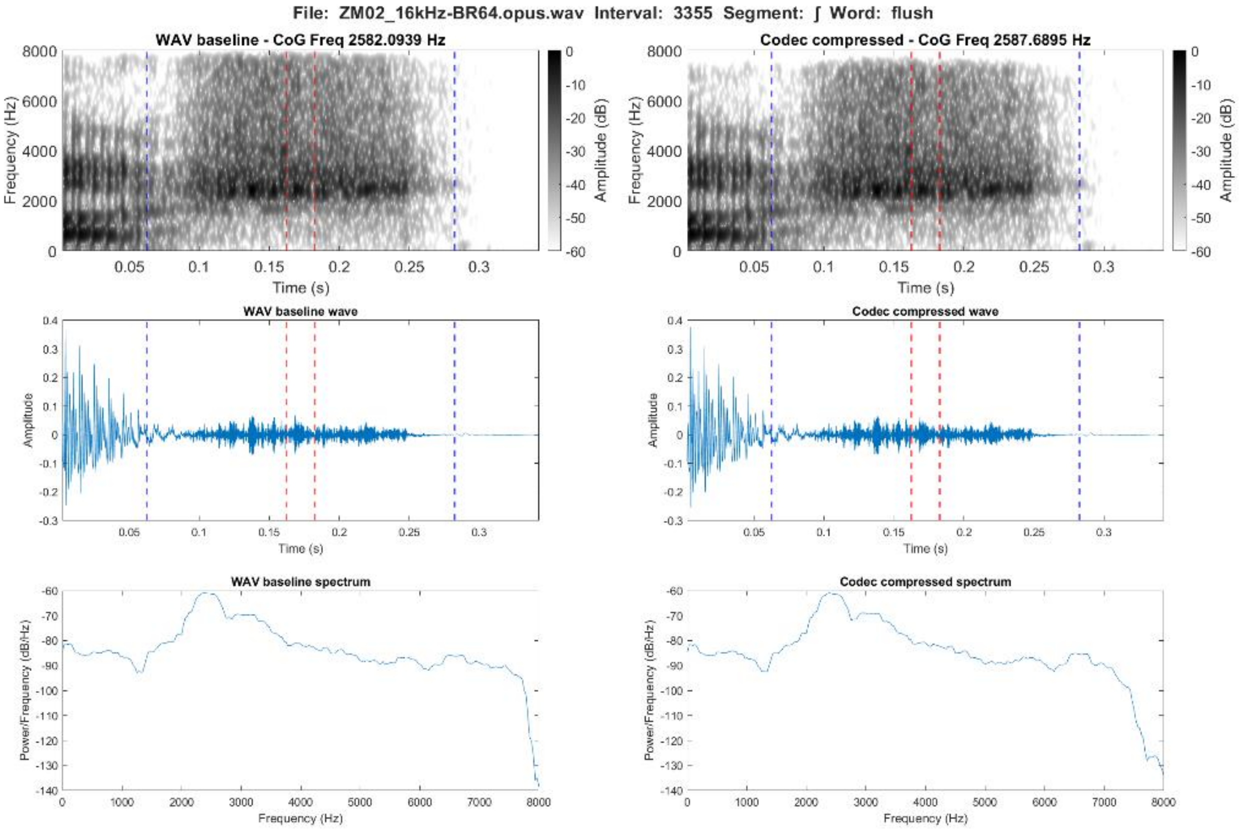


Figure 4.85. Spectrographic comparison of /ʃ/ in the word *flush* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.6 High bitrate: /z/

From the linear predictions, the effect on CoG is limited to only a very slight downwards trend, while SD and skewness more clearly lowers.

As with /s/ the mean value for kurtosis increased following the Opus compression, while the rest of the spectral measures decreased (p<.0001). Kurtosis increased by just over 2 percent. CoG decreased by just under 3 percent or 117 Hz, while frequency peak decreased by just over 2 percent or 81 Hz. SD decreased by 72 Hz or just under 6 percent. Skewness again showed the biggest change by just under 194 percent or, identically to /s/, 0.23. All p-values and further statistical results can be found in table 4.49 below.

Measure	**Contrast**	**Segment**	**Estimate**	**t.ratio**	**p.value**
CoG	WAV – Opus	/z/	116.74	6.05	<.0001
SD	WAV – Opus	/z/	72.34	8.35	<.0001
Skewness	WAV – Opus	/z/	0.23	10.65	<.0001

Table 4.49. Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /z/

All the spectral measures, which lowered in mean values, also lowered in maximum value following the Opus compression. For CoG this was from 7,082 Hz to 6,634 Hz, and for SD from 2,901 Hz to 2,732 Hz. Similar to SD, frequency peak lowered by around 200 Hz to 7,344 Hz. Skewness changed from 8.05 to 7.67.

A typical spectrographic representation of voiced /z/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.86. From the spectrogram /z/ the comparison between WAV and Opus mainly reveals a minor decrease in intensity, which is confirmed by the spectrum.

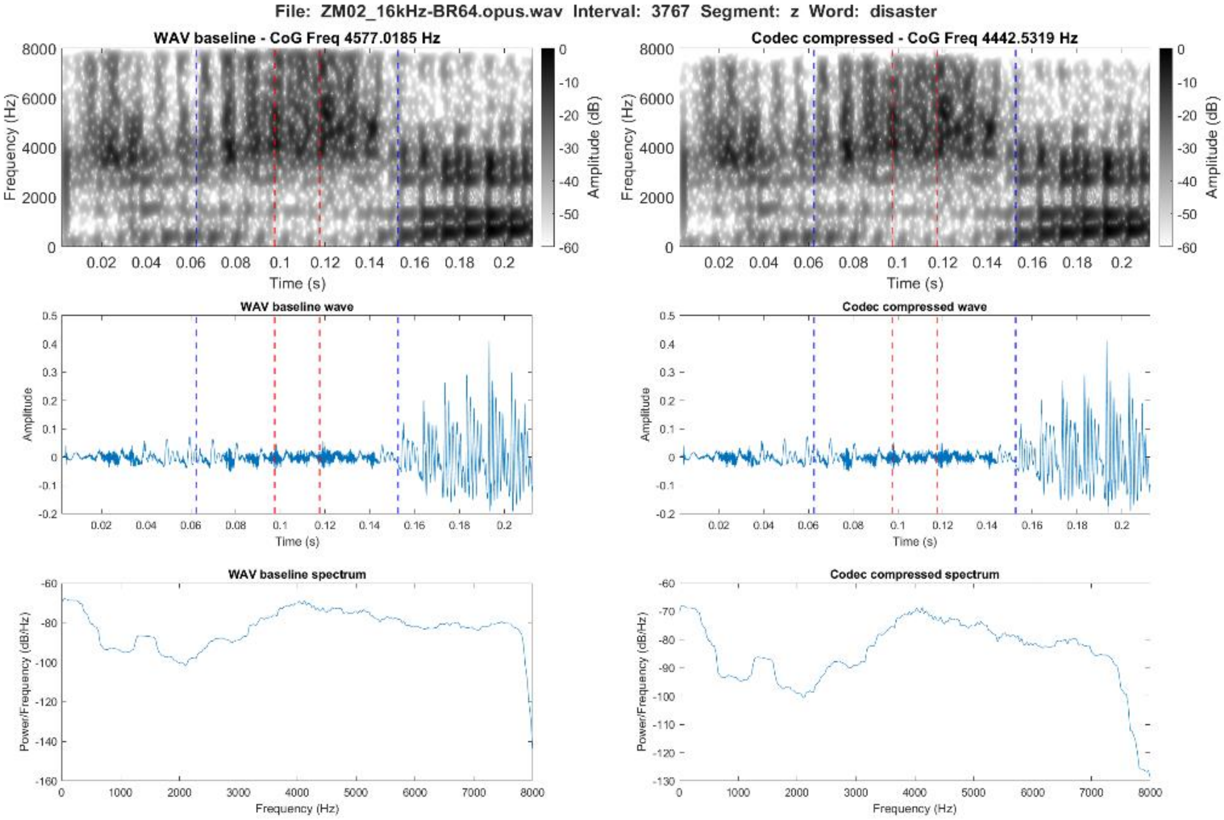


Figure 4.86. Spectrographic comparison of voiced /z/ in the word *disaster* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.2.3.7 High bitrate: /ð/

From the linear predictions, the effect of the codec compression is limited for CoG and skewness, which show very slight downwards to stable trajectories, while SD more clearly lowers.

All the mean values for the spectral measures were again lowered. For all measures apart from frequency peak, this was by 3 to 4 percent. For CoG, this was a decrease of 73 Hz (p = 0.001) and for SD 65 Hz (p<.0001). Skewness lowered by 4.23 and kurtosis by 0.21. Lastly, frequency peak lowered similarly to CoG with 72 Hz, which is a change by just under 6 percent. All p-values and further statistical results can be found in table 4.50 below.

Measure	Contrast	Segment	Estimate	t.ratio	p.value
CoG	WAV – Opus	/ð/	72.76	3.22	0.001
SD	WAV – Opus	/ð/	65.11	6.43	<.0001
Skewness	WAV – Opus	/ð/	0.04	1.55	0.12

Table 4.50. Statistical results of the difference between WAV and **high** Opus based on linear prediction models for /ð/

Apart from frequency peak, which remained stable at 7,938 Hz in both WAV baseline and Opus compression, the three other spectral measures lowered in maximum values. This was for CoG from 6,963 Hz to 6,402 Hz, and for SD from 3,161 Hz to 2,789 Hz. Skewness decreased by 0.10 to 4.66.

A typical spectrographic representation of voiced /ð/ in the WAV baseline and the Opus compression using the high bitrate is presented below in figure 4.87. Apart from a slight reduction in intensity, no notable effects are visible in the comparison between the WAV and codec compressed files.

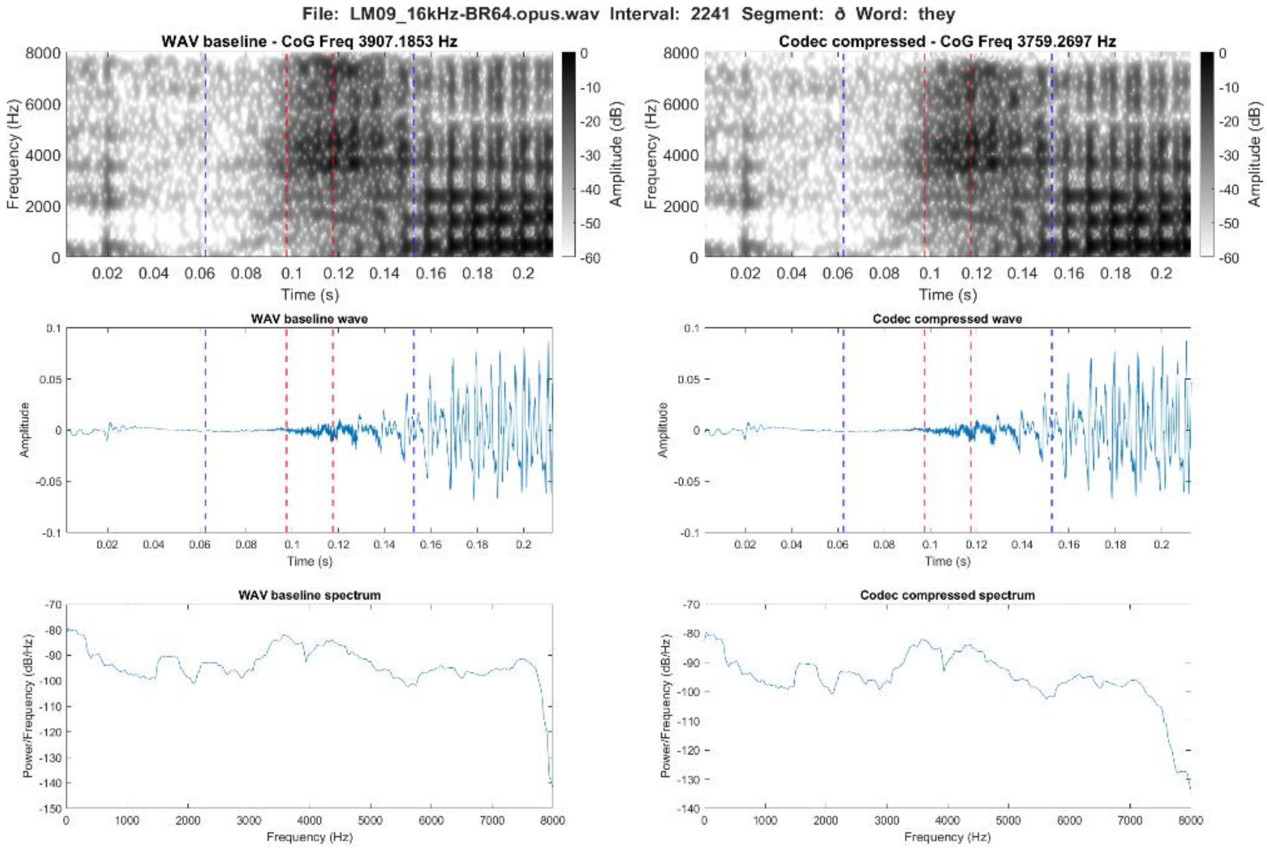


Figure 4.87. Spectrographic comparison of voiced /ð/ in the word *they* in the WAV baseline (left) and the Opus compression at 64 kbps (right)

4.1.3 1 kHz tokens

This section will present a more broad analysis of the 1 kHz tokens divided into two sections, one for each quality level i.e. low bitrates and high bitrates.

To recap, the 1 kHz tokens are the ones, where one or more tokens in either codec compression or WAV baseline have a CoG below 1 kHz. These are of interest because any movement from above 1 kHz to below 1 kHz due to any of the three codec-compressions potentially suggest substantial effects on the original WAV baseline. All the 1 kHz tokens were paired with their counterparts in the other codecs and baseline and all together removed from the main dataset in order not to skew the dataset (see section 3.3.5 for further details on this).

The following analysis is based upon the dataset with only the actual pairs (one below and one above 1 kHz, both below 1 kHz or both under 1 kHz (type A, B and C), see table 4.51 below) in comparison to the balanced dataset of 1 kHz tokens. For both the low and the high bitrates, none of these pairs were of /ʃ/ or [fʲ].

Type of pair	Description
A	*Both tokens with the same segment number have CoG values below 1 kHz*
B	*of two tokens with the same segment number, only the one in the codec compression has got a CoG value below 1 kHz*
C	*of two tokens with the same segment number, only the one in the baseline has got a CoG value below 1 kHz*

Table 4.51. Description of types of pairs in the datasets including only the 1 kHz tokens when in subsets of WAV baseline and one codec compression.

Due to the limited amount of tokens in this dataset mixed effects modelling was not used for the analysis. Hence, the analysis is based on descriptive statistics for the individual segments and spectrographic analysis of selection of the individual pairs.

The spectrographic representations were for comparative reasons chosen to be of the same segments as what those presented in the baseline study. This enables a more clear comparison of the effects of the codec compression under the different bitrates.

4.1.3.1 Low bitrates

The number of pairs for the 1 kHz tokens in the low bitrate condition including all tokens in all codecs and WAV and pairs in group A, B, and C, amounted to 3,893 pairs or 4.56 % of the entire dataset. For the different codecs, this is 1,484 pairs in the AMR-WB compression, 1,168 pairs in the MP3 compression, and 1,241 pairs in the Opus compression.

The number of 1 kHz tokens varies for each segment and codec compression, thus the exact number of rejected pairs for each segment in each codec compression are presented in table 4.52 below. This shows how most rejected pairs across the three codec compressions are of /ð/ and more specifically, these mainly occur in the word *the*.

Segment	**WAV-AMR-WB** **6.6 kbits** Group A, B & C	**WAV-MP3** **16 kbits** Group A, B & C	**WAV-Opus** **12 kbits** Group A, B & C
/f/	18	13	12
/θ/	12	14	8
/s/	9	4	5
/z/	109	47	58
/ð/	1336	1090	1158

Table 4.52. Number of rejected pairs in groups B and C for each segment in each of the codec compression with the **low** bitrates

The trajectory of the changes from the WAV baseline to each codec compression including group A, B, and C expressed through the CoG mean values are presented below (see figure 4.88 to 4.90). These show how, apart from /ð/ in the MP3 compression, all CoG mean values were lower following codec compression. However, it was only for the AMR-WB compression (excluding /θ/ and /ð/), and for /θ/ in the MP3 compression that these changes were from above 1 kHz in the WAV baseline to below 1 kHz in the codec compression. This was because the mean values express a combination of both decreases, increases and pairs with both tokens below the 1 kHz baseline, which means that these values illustrate a broad tendency, but hide individual token behaviour.

Now, for /ð/ the effect of the codec compression varies as the AMR-WB and Opus lowered the CoG mean, while MP3 increased it. Regardless of codec, the trajectory remained below 1 kHz. It is notable that following the MP3 compression /z/ and /f/ became almost identical in CoG mainly as a consequence of the lowering of /z/, while /z/ and /s/ became almost identical and very similar to /f/ in the Opus compression.

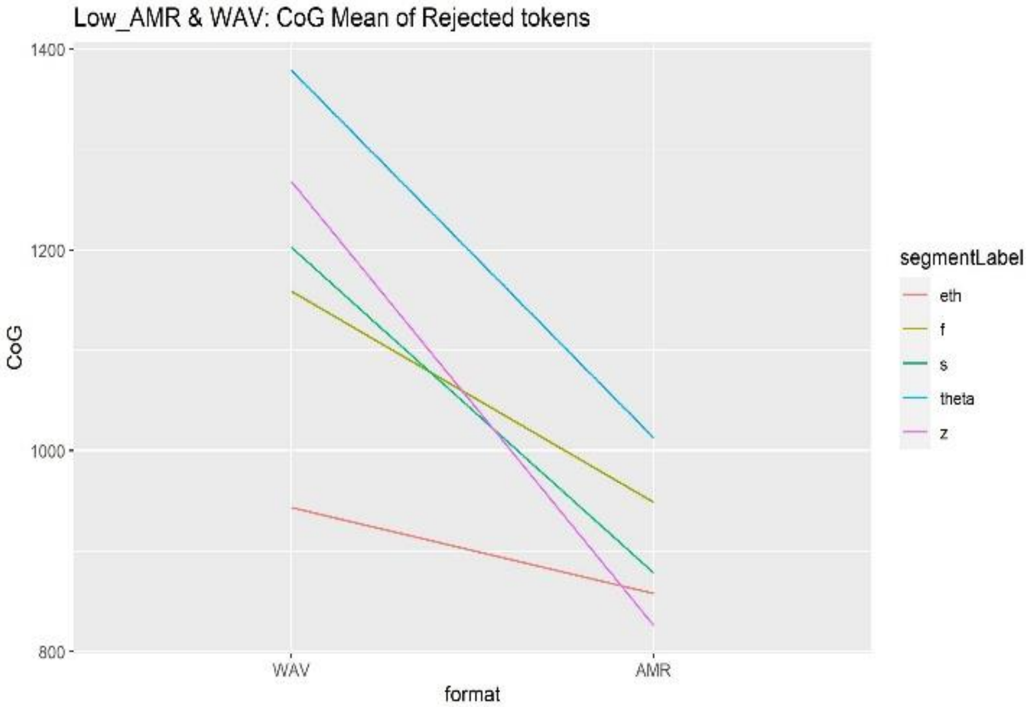


Figure 4.88. Trajectory of mean CoG values in 1 kHz tokens in the **Low** bitrates from the baseline WAV to AMR-WB incl. group A, B and C.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/. *Rejected tokens* = 1 kHz tokens.

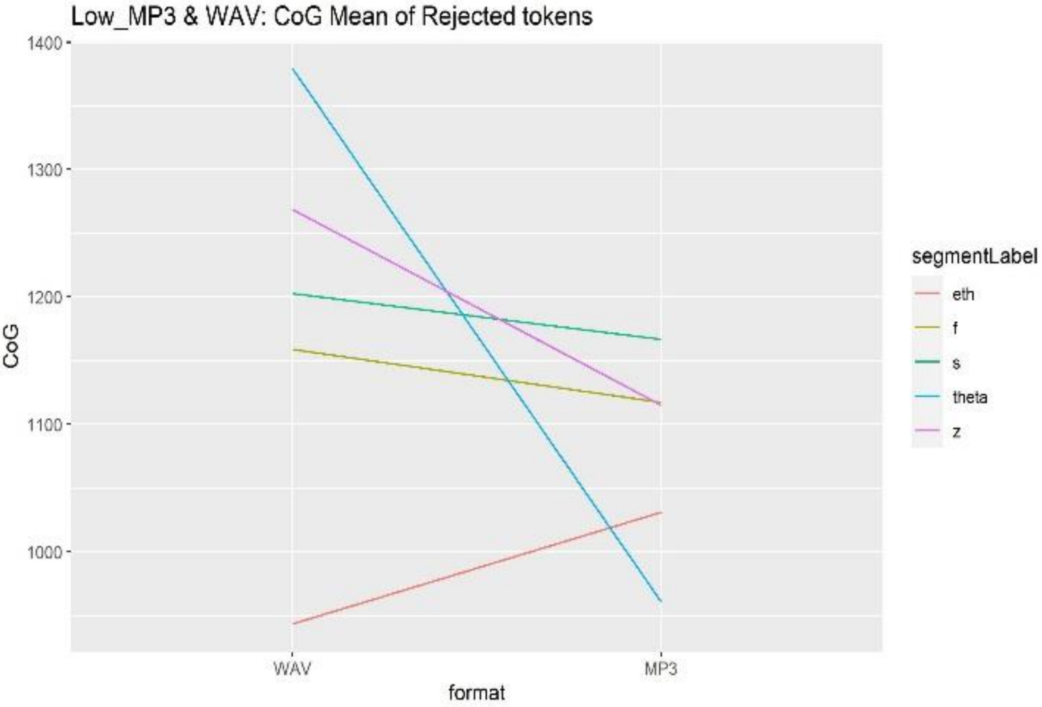


Figure 4.89. Trajectory of mean CoG values in 1 kHz tokens in the **Low** bitrates from the baseline WAV to MP3 incl. group A, B and C.
The symbols are interpreted as follows *Theta* = /θ/, *esh* = /ʃ/, *fj* = [fʲ], and *esh* /ʃ/. *Rejected tokens* = 1 kHz tokens.

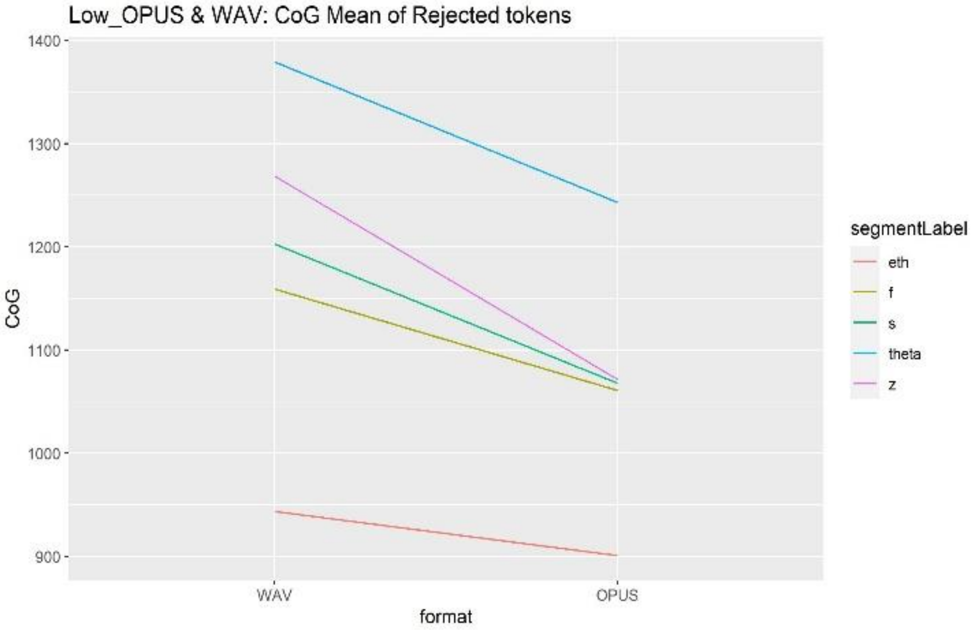


Figure 4.90. Trajectory of mean CoG values in 1 kHz tokens in the **Low** bitrates from the baseline WAV to Opus incl. group A, B and C.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/. *Rejected tokens* = 1 kHz tokens.

Overall, the mean values for the spectral measures including all groups showed decreases for CoG and SD for all segments and codec compressions, while a more mixed pattern appeared for skewness and kurtosis. The changes to frequency peak were often not substantial and varied between increases, decreases and one unchanged value for /z/ in the Opus compression. All mean values for the baseline and each codec compression is presented in table 4.53 below.

Seg	Codec	Bitrate (kbps)	CoG (Hz)	SD (Hz)	Skew (Hz)	Kurt (Hz)	Freq. Peak (Hz)
/f/	AMR	6.6	949	637	3.52	22.17	594
/f/	MP3	16	1117	660	2.32	11.76	676
/f/	Opus	12	1061	779	3.21	19.53	612
/f/	**WAV**	**NA**	**1159**	**904**	**3.12**	**19.18**	**592**
/s/	AMR	6.6	878	672	3.05	19.05	514
/s/	MP3	16	1167	823	1.62	7.98	517
/s/	Opus	12	1068	876	2.63	15.48	514
/s/	**WAV**	**NA**	**1203**	**1002**	**2.37**	**13.76**	**531**
/z/	AMR	6.6	826	715	4.23	29.81	512
/z/	MP3	16	1115	981	2.35	10.12	510
/z/	Opus	12	1072	1046	3.05	16.74	511
/z/	**WAV**	**NA**	**1269**	**1269**	**2.52**	**12.64**	**511**
/θ/	AMR	6.6	1013	962	3.16	14.54	524
/θ/	MP3	16	961	757	2.99	15.97	528
/θ/	Opus	12	1243	1284	2.85	13.34	517
/θ/	**WAV**	**NA**	**1379**	**1461**	**2.92**	**14.79**	**516**
/ð/	AMR	6.6	858	641	4.00	30.62	512
/ð/	MP3	16	1031	733	2.38	12.05	520
/ð/	Opus	12	901	716	3.81	27.53	511
/ð/	**WAV**	**NA**	**944**	**784**	**3.70**	**25.73**	**513**

Table 4.53. Mean values for all spectral measures for the **low** bitrate 1 kHz tokens for each segment, WAV baseline, and codec compression (increase from WAV marked in yellow; decrease from WAV marked in blue).

Lastly, spectrographic representations were made for all rejected segments, and visually inspected by the author. These revealed how effects observed for the average bitrate including the reduction of especially high frequency content was also present in the low bitrates, but tended to include more general reduction across the spectrum. Again, non-encoding was observed, and the MP3 codec under this bitrate quality exhibits even more prominent removal of frequency content from around 4.5 kHz. Examples of these observations are given in figure 4.91 to 4.93 below with spectrographic representations of a selected set of segments and codecs illustrative of the general tendencies. There are no examples from Opus as the despite showing similar tendencies, the effects were generally less prominent in comparison to AMR-WB and MP3.

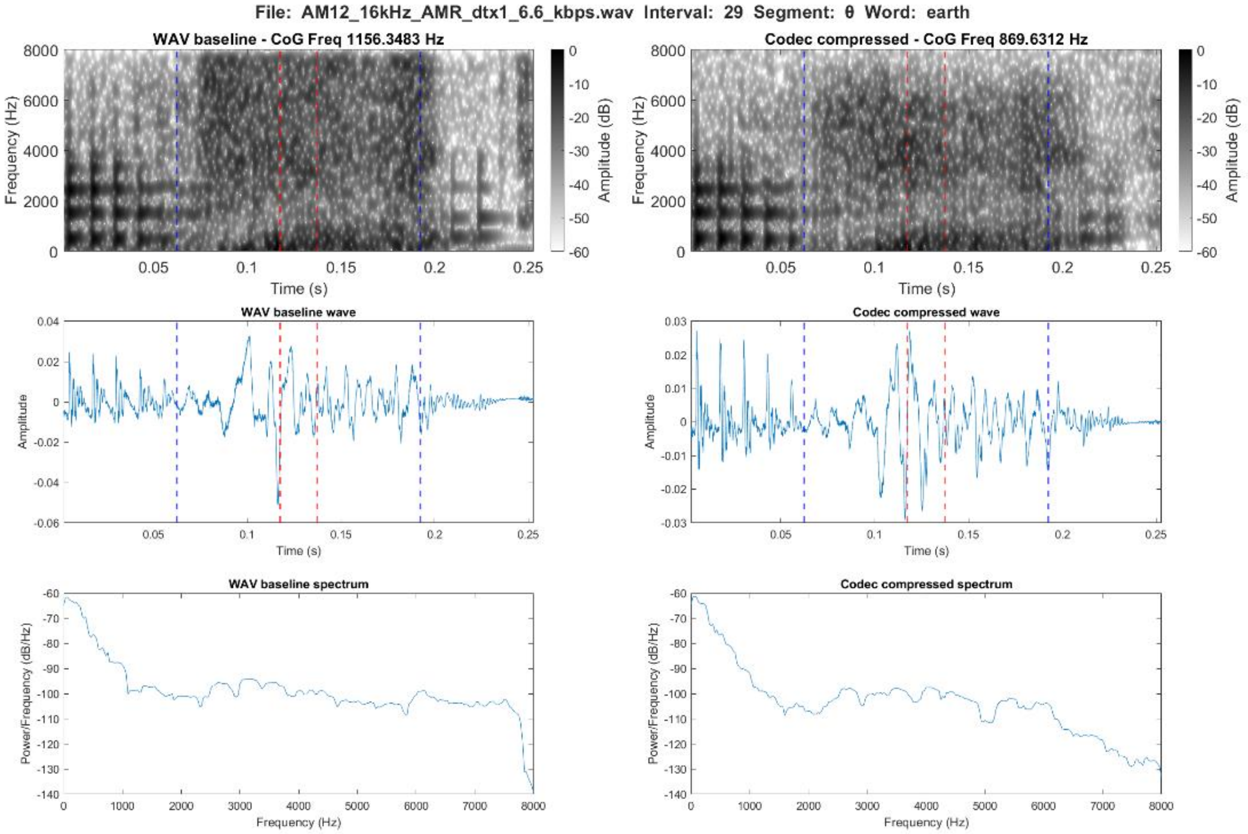


Figure 4.91. Main reduction of higher frequency content illustrated as /θ/ in the word *earth* in the AMR compression

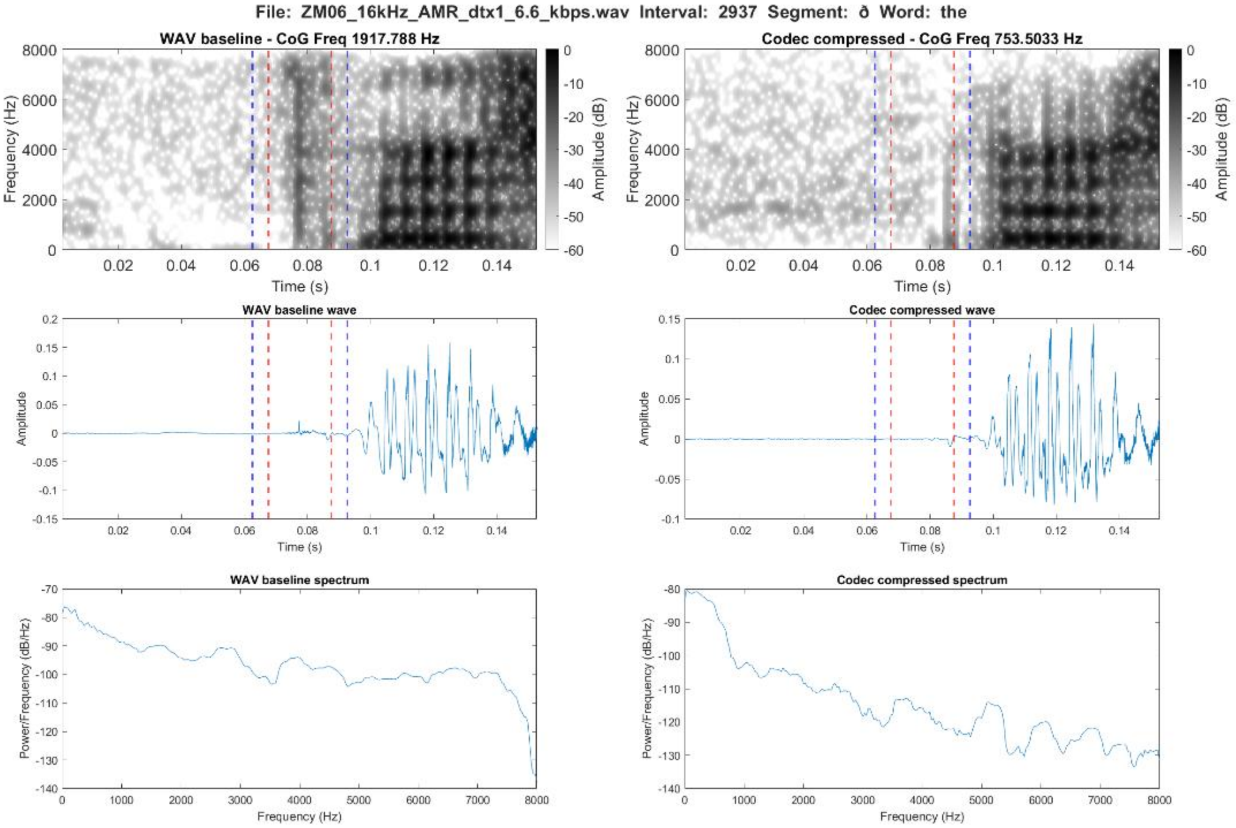


Figure 4.92. Non-transmission with insertion of comfort noise illustrated as /ð/ in the word *the* in the AMR-WB compression

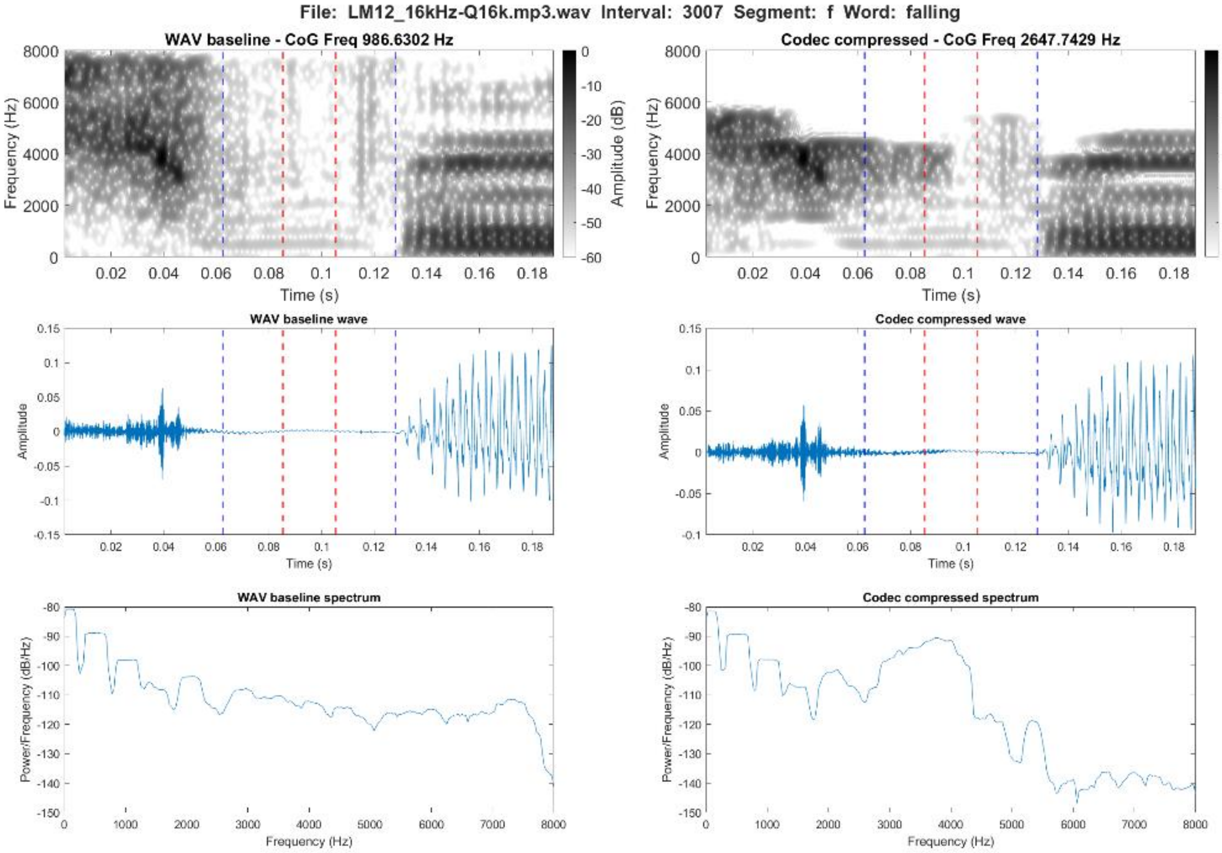


Figure 4.93. Increase caused by insertion of new frequency content by the MP3 compression illustrated as /f/ in the word *falling* in the MP3 compression

4.1.3.2 High bitrates

The number of pairs for the 1 kHz tokens in the high bitrate condition amounted to 3,306 tokens or 3.88% of the entire dataset including all tokens in all codecs and WAV and pairs in group A, B and C. This equals a total of 1,153 rejected pairs in the AMR-WB compression, 1,064 pairs in the MP3 compression, and 1,089 pairs in the Opus compression.

The number of 1 kHz tokens varies for each segment and codec compression, thus the exact number of rejected pairs for each segment in each codec compression are presented in table 4.54 below. As with the low bitrate data, most of the rejected pairs across the three codec compressions are of /ð/ and in the word *the*.

Segment	**WAV-AMR-WB** **23.85 kbits** Group A, B & C	**WAV-MP3** **48 kbits** Group A, B & C	**WAV-Opus** **64 kbits** Group A, B & C
/f/	12	7	7
/θ/	11	5	7
/s/	4	3	3
/z/	48	34	35
/ð/	1078	1015	1073

Table 4.54. Number of rejected pairs in groups B and C for each segment in each of the codec compression with the **High** bitrates

The trajectory of the changes from the WAV baseline to each codec compression including group A, B, and C expressed through the CoG mean values are presented below in figure 4.94 to 4.96. These show how the only two segments moving from across the 1 kHz boundary in terms of mean values were /θ/ in the AMR-WB compression and /f/ in the MP3 compression. For /θ/ in the AMR-WB compression, this was a decrease from above 1 kHz in the baseline WAV to below 1 kHz in the AMR-WB compression, while the reverse was the case for /f/ in the MP3 compression.

In addition, apart from /θ/ all segments regardless of codec compression had their mean values below 1 kHz. The CoG mean for /ð/ increased in both AMR-WB and MP3, while it remained unchanged by Opus. This was generally true for Opus, which showed limited effects for all segments below 1 kHz.

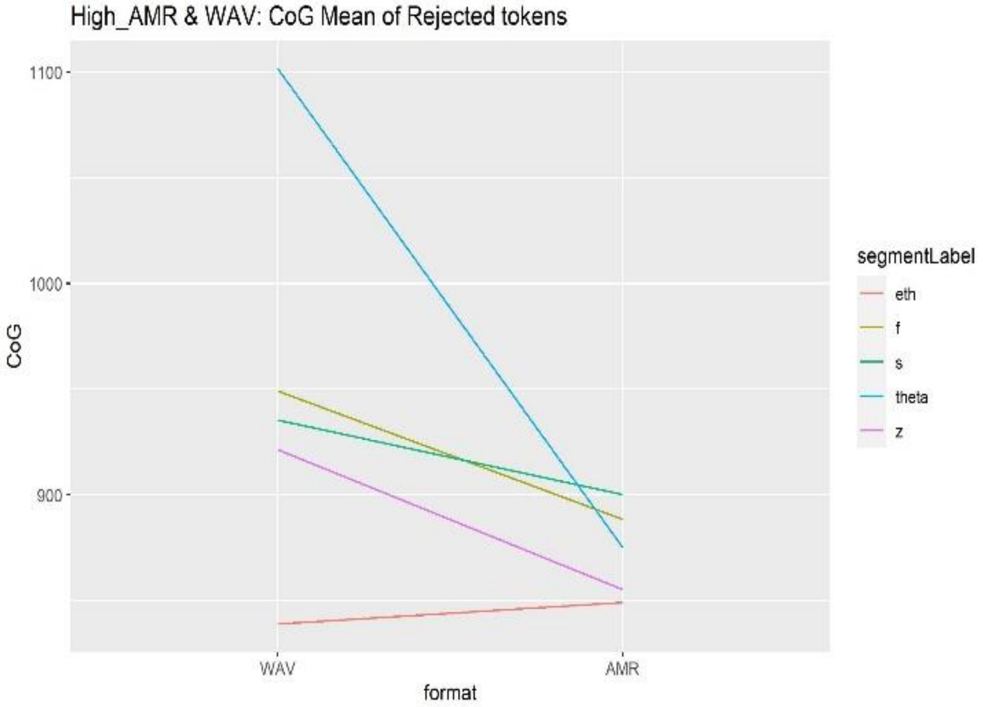


Figure 4.94. Trajectory of mean CoG values in 1 kHz tokens in the **High** bitrates from the baseline WAV to AMR-WB incl. group A, B and C.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/. *Rejected tokens* = 1 kHz tokens.

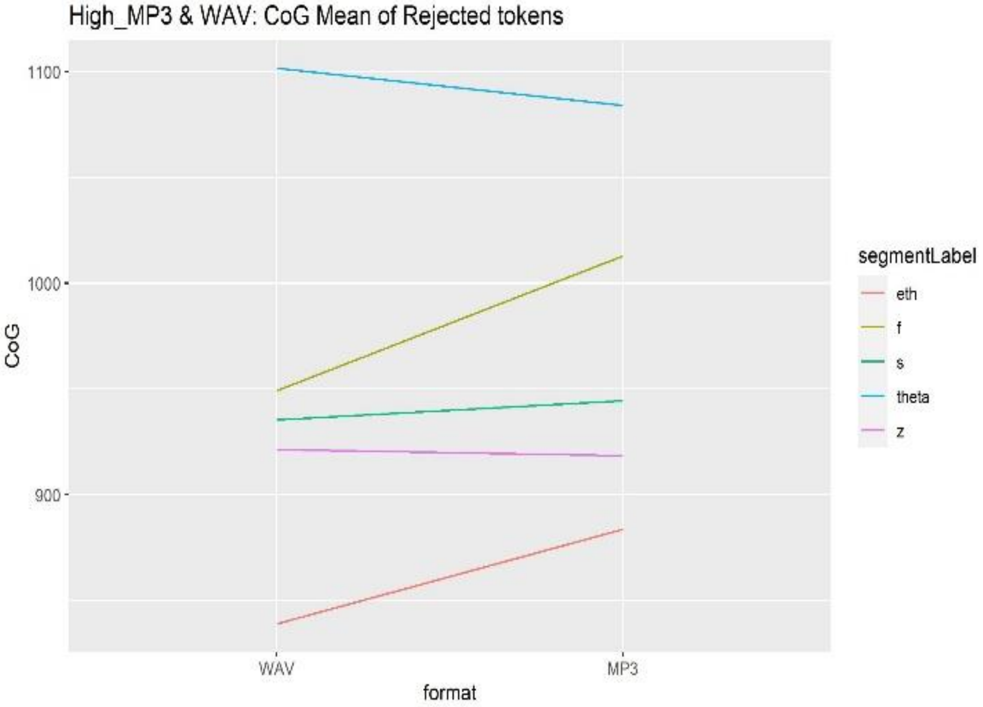


Figure 4.95. Trajectory of mean CoG values in 1 kHz tokens in the **High** bitrates from the baseline WAV to MP3 incl. group A, B and C.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/. *Rejected tokens* = 1 kHz tokens.

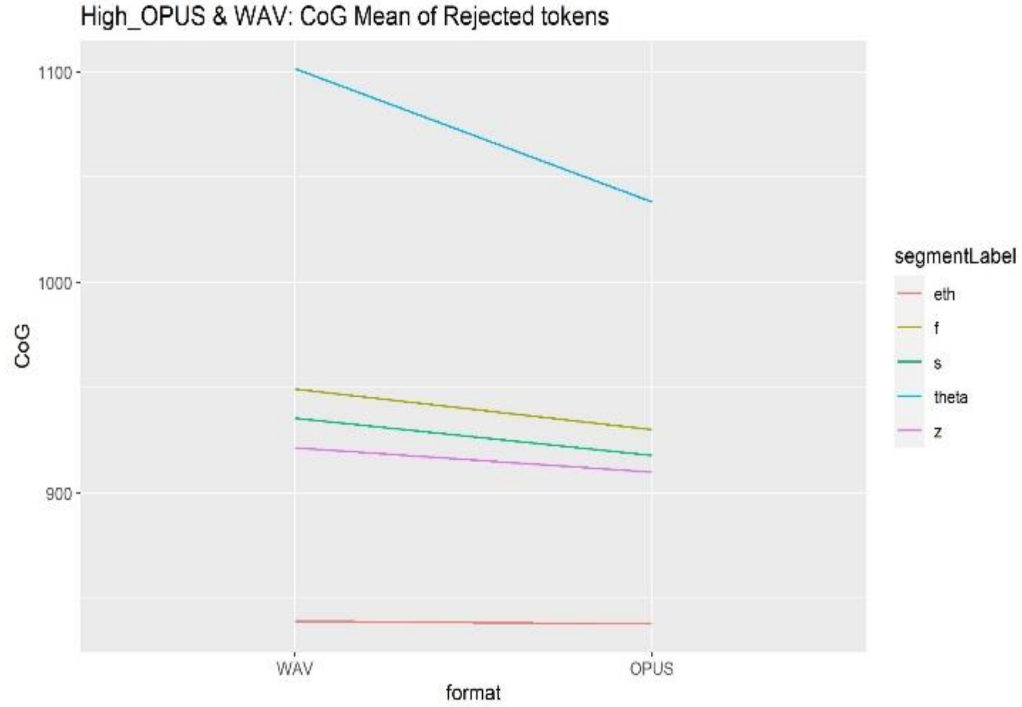


Figure 4.96. Trajectory of mean CoG values in 1 kHz tokens in the **High** bitrates from the baseline WAV to Opus incl. group A, B and C.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/. *Rejected tokens* = 1 kHz tokens.

The mean values for the spectral measures including all groups show how almost all segments in one or more codec compressions both decreased and increased for each spectral measure, while frequency peak almost consistently stayed unchanged for all of the codec compressions. CoG and SD were still generally lowered, while for skewness and kurtosis, the pattern appeared both segment and codec dependent with /ð/ as the only token decreasing across all codec compressions. All mean values for the baseline and each codec compression are presented in table 4.55 below.

Seg	Codec	Bitrate (kbps)	CoG (Hz)	SD (Hz)	Skew (Hz)	Kurt (Hz)	Freq. Peak (Hz)
/f/	AMR	23.85	889	615	3.77	27.33	583
/f/	MP3	48	1013	721	3.43	23.94	596
/f/	Opus	64	930	677	3.54	23.48	596
/f/	**WAV**	**NA**	**949**	**719**	**3.66**	**24.94**	**596**
/s/	AMR	23.85	900	658	3.61	27.61	547
/s/	MP3	48	945	712	3.40	23.39	547
/s/	Opus	64	918	690	3.44	24.25	547
/s/	**WAV**	**NA**	**935**	**732**	**3.47**	**24.52**	**547**
/z/	AMR	23.85	855	755	3.94	24.60	515
/z/	MP3	48	919	853	3.58	21.29	515
/z/	Opus	64	910	841	3.64	21.92	515
/z/	**WAV**	**NA**	**921**	**873**	**3.70**	**22.82**	**516**
/θ/	AMR	23.85	875	880	4.08	22.91	520
/θ/	MP3	48	1084	1170	3.33	16.91	509
/θ/	Opus	64	1038	1104	3.66	20.52	509
/θ/	**WAV**	**NA**	**1102**	**1199**	**3.62**	**20.49**	**509**
/ð/	AMR	23.85	849	684	4.10	29.44	510
/ð/	MP3	48	884	728	3.78	24.99	510
/ð/	Opus	64	838	670	4.09	29.03	510
/ð/	**WAV**	**NA**	**839**	**680**	**4.21**	**31.05**	**510**

Table 4.55. Mean values for all spectral measures for the 1 kHz tokens in **High** bitrates divided by segment, WAV baseline, and codec compression (increase from WAV marked in yellow; decrease from WAV marked in blue).

Lastly, spectrographic representations were made for all rejected segments, and visually inspected by the author of which examples are given below (see figure 4.97 to 4.98). The most prominent observation here is the fact that the non-encoding observed in both the average and low bitrate is not present in the high bitrate. Apart from this, the spectrograms show similar tendencies as seen previously, but less pronounced and they are mainly related to the upper cut off. Opus presented limited changes using the high bitrate, and thus non of the examples below are from this codec compression.

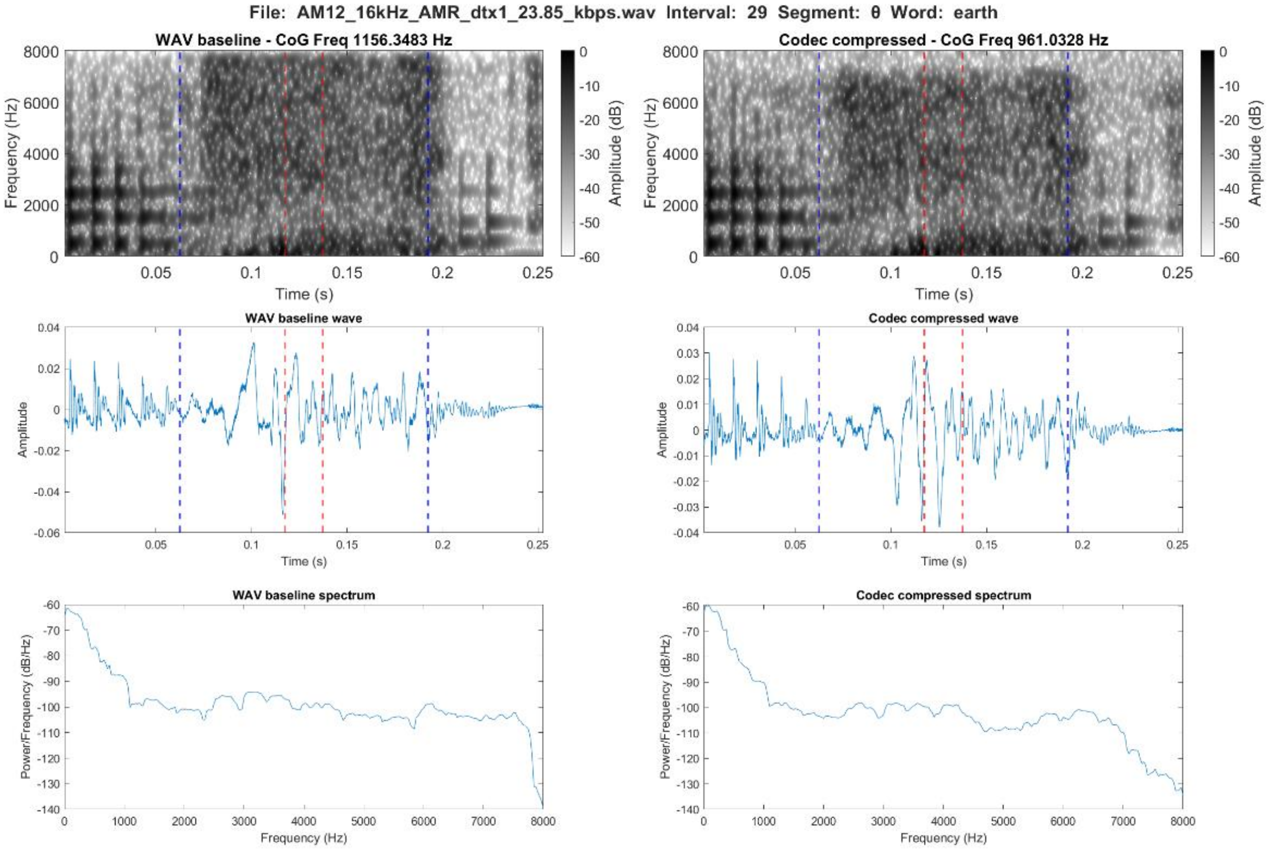


Figure 4.97. Main reduction of higher frequency content illustrated as /θ/ in the word *earth* in the AMR compression

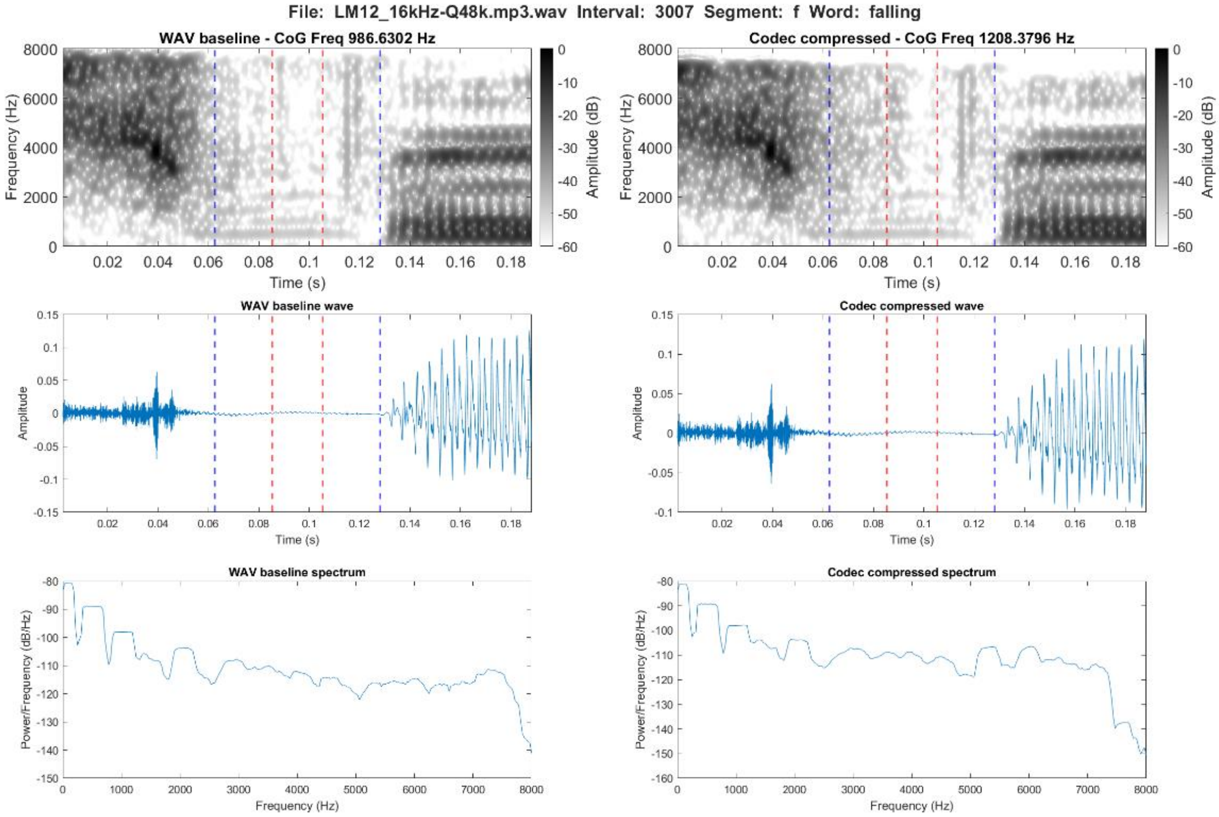


Figure 4.98. Increase caused by insertion of new frequency content by the MP3 compression illustrated as /f/ in the word *falling* in the MP3 compression

4.1.4 Bitrate comparison

This section does not present any new results, but rather a summary of the results obtained in this and the previous baseline study. This is done in order to more easily compare the broader differences in effects of the codec compressions using different bitrates.

The comparison is made via two main tables, consisting of the percentage change in mean values for each of the spectral measures, for each of the segments in the three codecs across the three bitrate qualities i.e. low, average and high. The tables are divided into two for aesthetic reasons, and table 4.56 contains the results for CoG, SD, and frequency peak, while table 4.57, below it, presents the results from skewness and kurtosis.

Seg	Codec	CoG (%) Low*	CoG (%) Average*	CoG (%) High*	SD (%) Low	SD (%) Average	SD (%) High	Freq. Peak (%) Low	Freq. Peak (%) Average	Freq. Peak (%) High
/f/	AMR	-8.34	-8.73	**-8.87**	-8.34	**-11.04**	-10.74	-9.32	**-12.96**	-12.35
/f/	MP3	**-21.75**	-4.08	-4.51	**-21.75**	-7.16	-7.05	**-11.84**	-4.47	-5.04
/f/	Opus	-4.04	**-13.09**	-4.39	-4.04	**-12.33**	-7.52	-5.93	**-19.41**	-4.78
[fʲ]	AMR	-5.91	-6.18	**-6.84**	-5.91	-8.97	**-10.29**	-4.25	**-10.28**	-6.97
[fʲ]	MP3	**-18.02**	-3.34	-3.41	**-18.02**	-7.27	-6.64	-2.96	**-3.69**	-2.07
[fʲ]	Opus	-2.00	**-10.93**	-3.39	-2.00	**-12.63**	-7.13	-0.76	**-11.64**	-1.61
/s/	AMR	**-6.88**	-5.94	-5.94	-10.65	-6.35	**-11.67**	-7.02	**-7.58**	-5.73
/s/	MP3	**-14.88**	-3.52	-3.36	**-25.75**	-5.98	-6.97	**-9.90**	-2.71	-2.41
/s/	Opus	-3.24	**-6.27**	-3.09	-5.79	-5.62	**-7.68**	-3.12	**-5.40**	-2.57
/z/	AMR	**-10.09**	-6.97	-5.58	-3.71	-2.04	**-7.99**	**-14.52**	-9.89	-5.40
/z/	MP3	**-15.14**	-3.82	-3.15	**-17.56**	-2.83	-4.60	**-10.42**	-2.38	-2.11
/z/	Opus	-4.72	**-7.22**	-2.69	-2.88	-3.46	**-5.83**	-5.26	**-7.39**	-2.11
/ʃ/	AMR	0.09	-0.03	**-1.99**	-2.53	1.38	**-9.26**	0.30	**-1.06**	-0.75
/ʃ/	MP3	**-4.76**	-0.55	-0.76	**-30.26**	-4.11	-4.77	-0.23	**-0.27**	-0.20
/ʃ/	Opus	0.10	**-3.69**	-0.77	-2.40	**-13.37**	-5.54	0.32	**-1.14**	-0.16
/θ/	AMR	-10.60	**-11.71**	-9.79	**-13.91**	-12.31	-10.03	-13.29	**-18.79**	-14.54
/θ/	MP3	**-26.14**	-5.43	-5.30	**-33.20**	-7.33	-6.67	**-21.22**	-8.41	-6.40
/θ/	Opus	-6.20	**-14.99**	-5.01	-6.77	**-10.99**	-6.83	-10.52	**-25.57**	-6.47
/ð/	AMR	-12.57	**-11.83**	-7.48	-13.33	**-18.65**	-6.62	-15.91	**-17.00**	-10.80
/ð/	MP3	**-16.53**	-2.66	-2.59	**-24.66**	4.19	-3.96	**-13.29**	-3.45	-4.26
/ð/	Opus	-6.64	**-11.06**	-2.97	-7.14	**-17.42**	-4.13	-8.37	**-18.48**	-5.83

* Low (bitrate): AMR- WB 6.6 kbps, MP3 16 kbps, and Opus 12 kbps
* Average (bitrate): AMR-WB 12.65 kbps, MP3 32 kbps, and Opus 24 kbps
* High (bitrate): AMR-WB 23.85 kbps, MP3 48 kbps, and Opus 64 kbps

Table 4.56. Percentage change in mean values of CoG, SD, and frequency peak from the WAV baseline to the three codec compressions for each segment across the three bitrate qualities i.e. low, average, and high. Colours are indicating the direction of the change i.e. blue = decrease and yellow = increase, while the result from the bitrate with the greatest effect for each segment in each codec is marked in bold.

Seg	Codec	Skew	Skew	Skew	Kurt	Kurt	Kurt
		(%)	(%)	(%)	(%)	(%)	(%)
		Low*	Average*	High*	Low	Average	High
/f/	AMR	**-19.15**	-12.21	-14.05	-6.99	-6.56	**-8.14**
/f/	MP3	**-59.25**	-12.28	-9.14	**-22.37**	-5.73	-4.27
/f/	Opus	**-13.48**	13.09	-8.45	-5.88	**10.62**	-2.91
[fʲ]	AMR	**-20.56**	-12.75	-18.62	-9.35	-7.33	**-11.00**
[fʲ]	MP3	**-64.14**	-13.42	-10.39	**-23.06**	-6.97	-5.44
[fʲ]	Opus	**-18.23**	2.16	-10.33	**-8.18**	8.13	-4.69
/s/	AMR	-287.60	-144.44	**550.27**	3.25	1.03	**11.81**
/s/	MP3	**-1010.3**	-288.89	308.71	**48.73**	3.30	4.17
/s/	Opus	-192.11	**-366.67**	263.50	-0.19	**12.70**	5.00
/z/	AMR	-142.49	14.29	**367.91**	0.20	**-6.27**	5.37
/z/	MP3	**-665.54**	-207.14	226.37	**20.46**	-1.59	0.56
/z/	Opus	-116.98	**-207.14**	193.75	-3.42	**6.05**	2.17
/ʃ/	AMR	-16.66	-1.04	**-17.98**	**-17.53**	-7.18	-11.89
/ʃ/	MP3	**-73.98**	-14.81	-12.94	**-42.99**	-12.42	-10.37
/ʃ/	Opus	**-16.35**	-13.85	-12.04	**-14.33**	3.26	-9.03
/θ/	AMR	**-12.09**	0.77	-12.08	**-5.46**	-2.71	-5.42
/θ/	MP3	**-38.66**	-10.39	-8.19	**-13.83**	-4.30	-3.60
/θ/	Opus	-5.26	**33.69**	-7.27	-3.91	**11.89**	-1.70
/ð/	AMR	**10.02**	9.90	2.54	**15.03**	-0.14	-2.38
/ð/	MP3	**-33.67**	-11.49	-7.09	**-46.60**	-13.43	-9.65
/ð/	Opus	3.44	**13.25**	-3.77	5.72	**7.14**	-4.35

* Low (bitrate): AMR- WB 6.6 kbps, MP3 16 kbps, and Opus 12 kbps
* Average (bitrate): AMR-WB 12.65 kbps, MP3 32 kbps, and Opus 24 kbps
* High (bitrate): AMR-WB 23.85 kbps, MP3 48 kbps, and Opus 64 kbps

Table 4.57. Percentage change in mean values of skewness and kurtosis from the WAV baseline to the three codec compressions for each segment across the three bitrate qualities i.e. low, average, and high. Colours are indicating the direction of the change i.e. blue = decrease and yellow = increase, while the result from the bitrate with the greatest effect for each segment in each codec is marked in bold

The tables show how, apart from /ʃ/, which was generally not substantially affected by the codec compressions, CoG, SD, and frequency peak are all lowered by the codec compressions. However, the effects on these three measures are both codec and segment dependent. MP3 is the only codec where the low bitrate is almost consistently the bitrate with the greatest effect. For AMR-WB and Opus, the bitrate with the greatest effect vary by segment and spectral measure. For Opus, the average bitrate generally produced the greatest effects of the codec compressions, while for AMR-WB the pattern is more varied. /θ/ was generally the segment most affected by the codec compressions.

For skewness and kurtosis, the low bitrates more consistently produced the greatest effects across segments and codecs. Again, the average bitrate for Opus tends to cause more notable effects than the low and high bitrates, while /s/ and /z/ are the segments most affected by the codec compression for these two measures. Despite being the most affected, it is worth noting that they were not affected equally, which indicates a phonetic distinction between the two regardless of the phonological distinction made by the MFA.

An additional comment in the comparison between the bitrates relates back to figures 4.1 to 4.3 in section 4.4., which illustrates the different upper frequency limits for the three codecs and each bitrate quality. This showed how the average and high bitrate for MP3 behaved similarly, while the low bitrate resulted in a markedly lower frequency limit. For AMR-WB, the average and low bitrate behaved similarly, while for Opus the changes in bitrate had little effect on the upper frequency limit. Thus, the results are again codec dependent in terms of the effect of bitrate.

Finally, the TH-fronting and /s/-retraction have previously been mentioned. /s/-retraction correlates with a lower CoG for /s/ relative to /ʃ/, however in the current data the difference in mean values for CoG between the two in all conditions is more than 1 kHz. Therefore, these results will not be compared further, but will be discussed in the discussion in section 4.5 and the main discussion and conclusion in Chapter 6. On the other hand, is has been observed how /f/ and /θ/ do become more alike based on the linear predictions. In table 4.58, the differences in mean values between the two sounds across the three codecs and three bitrate qualities are presented.

Diff. between	Codec	Bitrate (kbps)	CoG (Hz)	SD (Hz)	Skew	Kurt	Freq. Peak (Hz)
/f/ & /θ/	AMR	6.6	54	112	0.12	0.07	231
/f/ & /θ/	MP3	16	35	93	0.03	0.19	351
/f/ & /θ/	Opus	12	64	116	0.12	0.06	251
/f/ & /θ/	**WAV**	**NA**	**140**	**133**	**0.19**	**0.13**	**155**
/f/ & /θ/	AMR	12.65	41	98	0.12	0.06	260
/f/ & /θ/	MP3	32	103	123	0.18	0.14	228
/f/ & /θ/	Opus	24	71	143	0.13	0.18	258
/f/ & /θ/	**WAV**	**NA**	**153**	**135**	**0.21**	**0.2**	**145**
/f/ & /θ/	AMR	23.85	108	133	0.17	0.1	178
/f/ & /θ/	MP3	48	119	133	0.18	0.17	170
/f/ & /θ/	Opus	64	125	137	0.19	0.16	178
/f/ & /θ/	**WAV**	**NA**	**152**	**136**	**0.21**	**0.19**	**145**

Table 4.58. The difference in mean values between /f/ and /θ/ in the WAV baseline and the three codec compressions using the three different bitrate qualities

The differences in mean values between the two sounds show how generally, the lower the bitrate, the more similar the two segments become in terms of these spectral measures. This is apart from frequency peak, where the two segments become more distinct in comparison to WAV files. The smallest effect is again found in the high bitrates. For none of the codecs or bitrates all the measures become identical or close to identical.

4.5 Discussion and Conclusion

This study aimed to answer the following three research questions:

> **BitrateRQ1:** What do the included measures indicate about the effect of different bitrates on fricatives for each codec compression?
>
> **BitrateRQ2:** To what extent are the observed effects codec dependent? And, what is the interaction between the codecs and bitrates based on these effects?
>
> **BitrateRQ3:** In what way does this inform potential phonetic implications of codec compression and digital transmission on fricatives in view of varying bitrates in view of linguistic research?

For BitrateRQ1 and BitrateRQ2, the initial predictions for this study were primarily based on the findings from the baseline study with the main hypothesis that the effects would be enhanced in the lower bitrates and limited in the higher bitrates. This prediction was based on the fact that the higher the bitrate, the more data is available to represent the speech input, and thus, the more accurate the acoustic representation of the original input in the codec compressed signal.

This prediction was found to be partially correct. This is in the way that the effects were indeed relatively greater in the low bitrate in comparison to the high bitrates overall, but only the MP3 consistently presented greater effects in the low bitrate in comparison to the average and high bitrates. For AMR-WB and Opus, the most substantial effects were at times found for both the average and the high bitrate. In support of this prediction is the fact that the number of 1 kHz tokens were for most segments twice as many in the low bitrate than in the high bitrate, while the difference in bitrate clearly affected the spectral measures. This can be seen e.g. from the effects of the MP3 compression, which in the low bitrate stands out especially from Opus. MP3, which was the codec with the smallest effect in the average bitrate, had substantial effects on the spectral measures as well as the spectrographic representations, while in the high bitrates it again behaved at times identical to Opus. It was further observed that a drop in spectral amplitude typically occurred for MP3 in the low bitrate around 4-6 kHz. This finding cannot be explained from the present data, but appear across segments.

This leads back to the prediction that, based on the results from the baseline study, the AMR-WB and Opus compression would have the greatest effects on the spectral measures. This prediction turned out to be bitrate dependent and essentially not confirmed by the low and high bitrate results. This is in the way that in both these conditions AMR-WB was amongst the two codecs with the greatest effect on the spectral measures but in the low bitrate condition, this was together with MP3, while in the high bitrates MP3 and Opus behaved almost identical and with overall limited effects.

A greater effect was expected on the high frequency content across segments and codec compressions, which indeed turned out to be the case as a correlate of the limited bandwidth. Hence, the effects on e.g. maximum values and distribution patterns were more evident in the low and average bitrates than the high bitrates. The lowest cut-off value imposed by the codec compressions was found for MP3 in the low quality bitrate.

More specifically, not considering the 1 kHz tokens, the codec compressions overall significantly lowered the spectral measures in comparison to the WAV baseline. This is apart from a number of instances of skewness and kurtosis, which was increased in both bitrate qualities. For /ʃ/ a very slight change under 1 percent in the Opus and AMR-WB low bitrate compression also appeared as an increase in CoG and frequency peak. As expected /ʃ/ generally showed little to no effect of the codec compression, but did in certain cases become more similar or identical to /f/ and /θ/ i.e. in high bitrate MP3 and in Opus identical CoG to /f/ based on the mixed effects models.

It is interesting to note that as expected both CoG and SD were lowered for /s/ in both bitrates, but that in the low bitrates /s/ and the voiced counterpart /z/ are affected differently by the codec compressions, while in the high bitrates the effect on the two segments are very similar especially looking at the MP3 and Opus codec. Taking into consideration that MP3 and Opus overall had very similar effects in the high bitrates across segments, this suggests a clear role of bitrate as well as codec type on the observed effects. Furthermore, the fact that MP3 is not a speech codec appears mainly to play into spectral accuracy in the lower and average bitrates.

/f/ and /θ/ were expected to show the biggest differences between the WAV baseline and the codec compressions. This also included these two segments to have more tokens were moved to the

1 kHz dataset due to their low intensity and high frequency content. This prediction was not borne out consistently. More tokens of /ð/ and /z/ were moved to the 1 kHz dataset from the main dataset. The 1 kHz tokens with the clearest downwards trajectory of mean CoG was of /θ/ e.g. only in high bitrate Opus and MP3 no movement from above to below 1 kHz.

Specifically for /f/, despite the apparent similarity to [fʲ], was, based on the baseline study, expected to be less affected by the codec compressions regardless of bitrate in the comparison with /f/. This prediction was correct particularly in the low bitrates, and less so in the higher bitrates, where the effects as mentioned previously becomes less prevalent for all segments. From the 1 kHz tokens, a greater effect on /f/ is clearly observable as no tokens of [fʲ] was rejected from the main dataset. This suggests that the allophonic distinction does have influence on the effect of the codec compression, which in turn is related to the preceding and following sounds. This will be discussed further in the final discussion and conclusion in Chapter 6.

From the baseline study, it was clear that /ð/ was often not a true fricative and a large number of tokens were rejected from the main dataset because WAV and codec compression had CoG values below 1 kHz. /ð/ was expected to show both increases and decreases and particularly an increase in the MP3 compression with the low bitrate. This was not found as all spectral measures were lowered for /ð/ apart from certain cases of skewness, which increased. However, skewness showed increases more generally in the 1 kHz tokens. This however suggest that there is a difference in the way the codec compressions handle frequency information in the lower part of the spectrum i.e. below the set limit of 1 kHz to the information above it.

Another observation relates to the effect on skewness. A substantially bigger effect was found on skewness e.g. up to a change of up to 1000 times the WAV value in the codec compression, when the skewness value was already negative in the WAV condition. This particularly substantial change was found for MP3 in the low bitrate. In other words, the codec compression affected the segments with more energy centred below the mean and moved this further below the mean relative to the segments with a positive skewness and more energy centred above the mean.

In terms of the 1 kHz tokens, the general expectation that more tokens would be rejected as a function of a lower bitrate was confirmed. However, the overall number of pairs rejected were never more than a few percent regardless of codec compression and bitrate. In addition, most of the tokens rejected were of /ð/ and /z/ with a mean CoG below 1 kHz in both WAV and codec compression. The tokens, which were in fact presenting a change from above to below 1 kHz or vice versa following the codec compression, showed different tendencies based on the codec type and bitrate. Overall, the main effects were on the high frequency content, but for the average and low bitrates the reductions in frequency content and intensity extended more broadly across the spectrum. The non-encodings were only observed for the average and low bitrates, while the high bitrates generally represented the acoustic content more accurately.

Thus, despite the number of 1 kHz tokens being limited, even a few tokens being non-encoded might be important for acoustic analysis based on shorter speech samples e.g. in forensic phonetic analysis of incriminating recordings. Moreover, the fact that this inconsistency and non-encoding are observed in controlled conditions again suggest that these tendencies can be expected to be more prevalent in live transmitted speech and when background noise is present, which the codecs are designed to remove.

For BitrateRQ3, as previously, the main perspectives here are on forensic phonetics and sociolinguistics. In the baseline study, it was mentioned how the results had implications for speaker comparisons, profiling as well as content determination. This is because changes to or lack of spectral information mean that sounds might become less distinct and harder to interpret without relying on context and increasing the potential influence from priming and bias. The results here and the comparison between the bitrates for the codecs show that these effects and implications of the codec compressions are more likely to occur in the lower bitrates especially for MP3, while the speech signal in the Opus compression using the high bitrate is almost identical to the WAV baseline. As the effects are codec and bitrate dependent, the advice to be aware of the basic workings of digital transmission, the codec used and the equipment and hardware are repeated here.

In a sociolinguistic perspective, TH-fronting and /s/-retraction have been the focus, and will also be so here.

In terms of TH-fronting, /f/ and /θ/ are alike for most spectral measures prior to any codec compression, however become close to identical in terms of CoG, and in low bitrate AMR-WB also skewness. TH-fronting is generally researched as a perceptual factor (e.g. E. Wood 2003; Bailey 2016; Stuart-Smith et al. 2013), and the fact that some of the spectral measures become more alike cannot be seen as directly correlated with TH-fronting. Nevertheless, the fact that codec compressions limit the distance between the two segments in terms of CoG and at times skewness between the two sounds for these measures makes it more likely to be present perceptually. The fact that the frequency peak is lowered in the codec compressions also has the potential to affect this feature, as the fronting would result in a higher frequency peak. To what extent this means that more or less TH-fronting might be perceived by listeners following codec compression requires further perceptual studies on the data.

These effects are highly codec and bitrate dependent, but do illustrate the potential of TH-fronting to happen as a consequence of the codec compressions at different bitrates.

For /s/-retraction, it is as previously mentioned, not possible to give an exact number or percentages of Hz required to constitute this feature without doing relational calculation between /s/and /ʃ/ (e.g. Baker, Archangeli, and Mielke 2011). It is clear that the CoG for /s/lowered in the codec compressions, while /ʃ/ again was one of the least affected segments. However, in none of the bitrates, does the CoG mean value for /s/ lower to or below the value of /ʃ/ The changes are nevertheless significant across bitrates, which still makes this a relevant avenue of research. This is particularly in the MP3 low bitrates, where the frequencies above 4 kHz are in most cases non-encoded. MP3 is used for download of sound files and this thus, underlines the importance of awareness of the bitrate with which the files are encoded and downloaded.

This study confirm the importance of bitrates in the evaluation of the effect of codec compression on fricatives, and again caution its use without taking these effects into account. In an applied perspective, this means that the low frequency energy is more prevalent in codec compressed speech and thus, will overall favour male speakers. This will be discussed further in the main discussion in Chapter 6.

Finally, the results reported here suggest that the fricatives more often than not significantly change in terms of spectral measures following any codec compression at any of the bitrates investigated here. This means codec compressed speech should be handled with caution in any linguistic study, and not taken as an alternative to direct high quality recordings even at high quality bitrates. The fact that these results appear in controlled conditions, makes it even further relevant to investigate in what way live transmission and non-optimal conditions e.g. in background noise influence the performance of the codecs. This will be the topic of Chapter 5.

In summary, the effects on the spectral measures are both codec, segment and bitrate dependent, but a general pattern appears in the sense that for most measures and segments apart from skewness and kurtosis, the codec compressions have the energy lower in the spectrum. This is potentially due to the filtering effect, which means the higher frequencies are not present in the codec compressed files. The extent to which this happens is however, dependent on the segment, bitrate and the codec. These effects are relevant to acoustic phonetics including sociophonetics and forensic phonetics as most of these effects are statistically significant. Furthermore, the effects at times lead certain sounds to become more alike or even identical for certain measures, which have previously been found to distinguish e.g. sibilants from non-sibilants (e.g. Blacklock 2004; Shadle and Mair 1996; Jongman, Wayland, and Wong 2000). In addition, the spectral changes are potentially, especially in the average and low bitrates, a channel for diffusion of sociolinguistic variation based on acoustic variation e.g. TH-fronting. These hypotheses will require further research into sociolinguistic as well as perceptual aspects, which is highly encouraged, but beyond the scope of the current project.

Thus, this study underlines the conclusion of the previous study to be cautious when using digitally transmitted speech for data-collection especially for acoustic and segmental purposes (e.g. Sanker et al. 2021, Leemann et al. 2020, Siegert and Niebuhr 2021). In addition, this study has shown how bitrate must be taken into account on a codec dependent basis.

Chapter 5 : Spectral implications of live transmission

5.1 Introduction

The previous two studies established how, even in controlled conditions with no external factors, the codec compression significantly affected the spectral characteristics of fricatives. This was shown to happen at both low, high and average bitrates, and generally with the greatest effects in the low bitrates and the smallest effects in the high bitrates.

These controlled conditions with no external factors will never occur in an everyday scenario of digital transmission of speech, where the signal is sent live over the network. Therefore, it is essential to get a better understanding of how speech, and here particularly fricatives are actually affected in scenarios closer to real life. The current study will present results from live transmission using mobile phones and the AMR-WB network without a data connection and thus, using the 3G network to ensure that the network access did not vary between VoLTE (4G or 5G) and the AMR-WB 3G connection.

More specifically, three main factors are relevant to consider in live transmission: a) bitrate varying dynamically over time depending on the available network capacity and connection; b) hardware i.e. devices used and their components; and c) automatic removal of background noise by the mobile phones.

Firstly, the dynamically varying bitrate means that a digital transmission like this cannot be guaranteed to maintain one set bitrate throughout the entirety of a call and may vary between the available bitrates depending on network traffic and location (Whitrow 2019). The present study partly controls this variable by using only one static location for all transmissions. However, to what extent the network traffic varies over the span of the recording cannot be controlled, but will potentially be evident from the recording quality.

Secondly, the hardware is essential as different phone models implement different loudspeakers and microphones as well as Active Noise Cancellation (ANC) technology. These variables can be controlled by using the same hardware in this case the same phone models. However, the exact technology behind each of these is often not disclosed by the manufacturers due to market competition. This is especially true for the ANC technology. This technology is meant to filter background noise additionally to what the network, and encoding process will do during transmission (Kottayi et al. 2016).

Thirdly, removal and limitation of background noise from the signal is one of the key features both in digital transmission, and for the present study due to the acoustic similarity between fricatives and potential background noise. The codecs remove background noise based primarily on periodicity of the signal, as noise is aperiodic. This is done primarily using VAD, which is installed as part of the signal chain of the codec (See section 2.6 for further details, 3GPP 2020b). However, most fricatives are also aperiodic and in that way, they are likely, to different extents, to be mistaken for noise and only partially transmitted or at times not transmitted at all. A few instances of the latter were already attested in the average bitrates under controlled conditions. The exact impact and workings of this noise reduction cannot be controlled, but by controlling the type and level of the background noise across recordings, the methodology remains replicable.

In summary, these variables can to varying extents be controlled, but will inevitably regardless of methodological setup, vary in ways that cannot be predetermined. However, not doing this type of research for that specific reason leaves a gap in our knowledge about how digital transmission in fact affects speech and how data sampled from live transmission might be different from what is seen under controlled conditions. Overall, this can potentially lead to false judgements on the gravity of the influence of digital transmission on speech, and the use and/or misinterpretation of this type of speech in linguistic research. In that way, this study is a pilot study, which aims to give a preliminary view of how the speech perceived in everyday phone calls are acoustically different from high quality studio recordings.

The present study thus aims to give an indication of real life effects on fricatives of digital transmission and more specifically mobile phone transmission with and without background noise. It will do so by recording the live transmission between two mobile phones of the same read speech used in the previous studies i.e. the *Chicken Little story* to maintain comparability. To ensure a range of real life scenarios are considered, the recordings are done in three conditions, one where the speech is directly replayed into the sending phone via a cable removing any effects of the handset microphone or background noise, and two where the speech is replayed via a KEMAR dummy-head with the sending phone held to the ear with and without background noise replayed via an array of six loudspeakers (see details on setup and equipment in section 5.3.2 to 5.3.4).

From the perspective of ecological validity, the experiment will not completely replicate a real life scenario for a number of reasons, which will also be addressed in the methodology in section 5.3.4

on procedure and technical setup. The methodological choices made for this study i.e. using a laboratory setup rather than a real time recording of a mobile phone call from a busy street, was made because replicability as well as comparativeness were prioritised over a completely ecologically valid setup. The latter would have introduced further factors including varying levels and types of background noise, speaker variability e.g. dynamic changes in voice quality to accommodate the surrounding environment. These factors would not have been possible to assess with the amount of previous research available and within the timeframe of this thesis. It would also have limited the comparability with the two previous studies. Taken together, this study prioritised singling out the effect of the AMR-WB in live transmission with and without background noise, which despite the number of additional variables and factors in a true real life scenario must be the first step in assessing the effects of digital transmission in real life scenarios.

In sum, the study will work with the following three research questions:

LiveRQ1 What do the included measures indicate about the effect of digital live transmission on speech?

LiveRQ2: How does introducing background noise to the speech signal to be transmitted, affect the spectral measures and overall output signal?

LiveRQ3: In what way do these findings help understand the consequences of digital transmission on fricatives in real life scenarios and inform further research?

5.2 Predictions

The effects observed for the AMR-WB codec in the two previous studies are expected to be more pronounced in the live transmission due to the introduction of hardware, limited network access, packaging, package loss, and particularly noise reduction. Overall, by using a dummy head and adding background noise in a lab environment, the study is expected to mimic a closer to real life scenario of a phone call than the previous studies.

With the introduction of background noise the voiceless fricatives /f/, and /θ/ are expected to be affected more relative to their voiced counterparts. This is because they are now subject to not only the codecs general speech identification, but also the potential misidentification of aperiodic speech sounds as background noise.

In addition to background noise, the previous studies have shown that in certain cases the codec compressions insert what is most likely comfort noise, where no actual fricative was present. This sound increases CoG and in that way imitates the fricatives. When background noise is added to the recordings this effect is expected to be more prominent in the live transmission if the fricatives are mistaken for noise and not encoded but compensated for by the comfort noise. In terms of packaging, package loss and limited network access, these are expected to result in more tokens in the 1 kHz tokens dataset.

Focusing on specific segments, the predictions are similar to and based on the results of the previous studies. /f/ and /θ/ are expected to be affected by clearly lowered CoG as well as decreases in intensity observable from the spectrographic representations. Following the previous results, these two sounds are expected to become more alike especially in terms of CoG following the live transmission. It is likely that these two fricatives will also become more similar to some of the other fricatives e.g. /ʃ/ as previously observed due to the lowered values and more limited frequency span of the transmission channel. However, the extent to which this will happen is unclear from the previous results.

[fʲ] is predicted to again be less affected than /f/, but follow the general pattern of lowered CoG, SD and frequency peak, while skewness and kurtosis are also expected to decrease.

For /ʃ/ the effects are expected to be limited based on the results of the previous studies, where even in the low bitrates, the effects on the spectral measures of this sound was not substantial. The same is true of /ð/, which however is expected to have a number of tokens in the 1 kHz tokens dataset due to changes in CoG from above to below 1 kHz or vice versa. This is based on the previous results, where /ð/ was the only sound observed not to be encoded by the AMR-WB as well as the fact that it mainly occurs in initial position and here often consists of one short burst of energy e.g. produced as a plosive. Overall, this makes /ð/ prone to be mistaken for noise and not transmitted due to either the ANC or the AMR-WB codec.

For /s/ and /z/, the primary effects are expected to be seen on the higher frequency content due to the upper cut-off. This is based on the intensity of /s/, which from the previous results appear to render it more robust to the codec compression, while the voicing and formant structure of /z/ makes the lower frequencies more likely to be preserved in the transmission.

As with the previous studies, the results of this study are expected to have implications across linguistic fields as well as informing further research into the impact on digital transmission of speech on sociolinguistic variation as well as forensic phonetic casework.

5.3 Methodology

This experiment was conducted on the 7th of March 2023 in the Listening Room of the Audio Lab at the School of Physics, Engineering and Technology, University of York. Philip Harrison and Andrew Chadwick assisted with the setup of the experiment as well as the exporting. In addition, Philip Harrison, as in the studies from the previous Chapters, assisted with the spectral analysis and generation of spectrograms.

The purpose of the experiment was to obtain recordings of speech live transmitted via real mobile phone calls in three different conditions. These were:

1. **Direct condition**: The speech signal was replayed directly from a computer via a cable to the transmitting mobile phone. The transmitted signal output was transferred via the headphone socket on the receiving phone and recorded to a computer. In that way, this condition illustrates the codec compression and transmission effects alone with no background noise or other acoustic effects or influences.

2. **No noise condition**: The speech signal was replayed via a KEMAR dummy head (see section 5.3.4 for photo illustration) and picked up by the mobile phone's microphone with the signal transmitted via a live phone call to the receiving mobile phone. The transmitted signal output was transferred via the headphone socket on the receiving phone to a computer. In that way, this condition illustrates the codec and transmission effects with low ambient background noise with the acoustic effects of mount, head and interactions with the acoustic environment (see section 5.3.8. for details on setup).

3. **With noise condition:** The speech signal was replayed via a KEMAR dummy head whilst traffic noise was replayed via a 6 loudspeaker ambisonic array (i.e. an audio system capable of reproducing the directional and acoustic properties of the recorded sound from two or more speakers). This gives a more natural sound field in comparison to the sounds originating in a single loudspeaker. For current study, this means a setup more reflective of a real life scenario. The combined speech and traffic noise was picked up by the mobile phone's microphone and transmitted via a live phone call to the receiving mobile phone. The transmitted signal output

was transferred via the headphone socket of the receiving phone to a computer. In that way, this condition illustrate codec and transmission effects in replayed traffic noise including the acoustic effects of mouth, head and interactions with the acoustic environment (see section 5.3.8. for details on setup).

The methodology for this study is, for some aspects e.g. in terms of segmentation and data extraction, identical to the baseline study. Therefore, this section will at times only contain brief overviews and /or refer the reader to more detailed information from the previous sections in the baseline study in Chapter 3.

5.3.1 Corpus and Participants

The corpus used for this study is the same as for the two previous studies, but in this case only one speaker per accent group was used. This is again from the *You Came to Die?! corpus* (Best et al., 2012-2015), which consists of thirty male and thirty female speakers, all native speakers of English and aged between 18 and 41. The participants spoke five different accents of English with six speakers of each accent. These were Australian (AUS), New Zealand (NZL), London (LON), Newcastle (NCL), and York (YRK) English. Every speaker was recorded reading a set of nonsense words in /zVbə/context; reading real keywords, and reading a phonologically balanced version of the *Chicken Little story* (approximately 10 minutes) (see transcript in appendix 1).

The current study uses the 10 minute reading of the *Chicken Little story* produced by the male speakers however; in contrast to the previous study, this is only from one speaker for each accent. This was: AM03, LM02, NM01, YM01, and ZM02. These speakers were chosen solely based on the fact that they were the first for each accent. The decision to limit the amount of data is based on the procedural setup and the amount of time it would take to play back the recordings of all 30 speakers (See section 5.3.7 for procedure details). Thus, the study investigates the speech from 5 male speakers. For additional details on the corpus and participants, the reader is referred to in Chapter 3, section 3.3.1.

5.3.2 Materials

Again, the dataset elicited /s, z, f, θ, ð, ʃ/and [fʲ] in read speech from an approximately 10 minute long reading of the *Chicken Little* story in varying phonetic contexts and in word-initial, word-medial, and word-final position (e.g. <fear>, <painful>, <safe>; see full transcript in appendix 1). For this experiment, only the speech audio files from 5 speakers i.e. one for each accent were used.

Including the speech from all 5 speakers resulted in a dataset with a total of 14,280 fricatives with one quarter in the Original WAV format and the rest in the live transmitted conditions across the three previously described conditions i.e. direct, and with and without background noise. This is 3,570 tokens per condition including the WAV baseline. As in the previous studies, the number of tokens of each fricative were not equal as it was determined by the phonetic content of the Chicken Little story. In sum, 5.1 below presents the total number of tokens per segment as well as their distribution across word position i.e. initial, medial and final.

Segment	**Total** (all conditions)	**Initial position**	**Medial position**	**Final position**
/f/	2264	1476	420	368
[fʲ]	160	100	60	0
/s/	4832	2576	1092	1164
/z/	2612	52	400	2160
/ʃ/	856	560	196	100
/θ/	864	204	324	336
/ð/	2692	2284	216	192

Table 5.1. Number of tokens per segment in full dataset with different criteria

5.3.3 Equipment

The equipment used in the experiment are listed in table 5.2 below including a brief description of its use. For further details see section 5.3.8 on procedure and technical setup.

Equipment	Description
6 loudspeaker ambisonic array	Located in the Listening Room of the Audio Lab, School of Physics, Engineering and Technology, University of York. The 6 loudspeakers are Genelec 8040A Studio Monitors arranged in a front, back, left, right, up, down array.
2 x Samsung Galaxy S8 owned by the Audio Lab	The two smartphones used as the transmitting and receiving mobile phones.
Windows PC (Windows 10) with Ferrofish A32 audio interface	Computer used for all replay and recording
Tascam iXZ (Teac Europe GmbH 2022)	Mobile device audio interface connected to the input of the transmitting phone in the Direct condition.
Radial X-Amp (Radial Engineering 2023)	Amplifier providing impedance matching. Only used for the recordings in the Direct condition to connect the output of the computer replaying the speech material to the Tascam iXZ.
AMbiX	Ambisonic decoder. This is a convolution plugin with custom made filters for the specific loudspeaker array, which equalises the speakers for level and frequency response.
GRAS 45BC KEMAR Head and Torso with Mouth Simulator (GRAS Sound & Vibration 2023)	Dummy head used for the replay in the conditions with and without noise.

2 x Microphone stands	One used to hold the transmitting phone next to the dummy head, and the second used to hold the sound level meter during measurements of sound pressure level.
2 x Vodafone sim cards	One in each Samsung Galaxy S8 to enable the call to be established between the two phone using the same network. Vodafone was chosen as it is one of the most widespread networks in England (Vodafone 2023)
Reaper software (version 6.74) (Cockos Incorporated 2023)	Software used for all replay and recording.
NTi XL2 sound level meter	Sound level meter used to measure the sound pressure level of the speech produced by the speaker in the KEMAR dummy head as well as the traffic noise.
Various cables	Used to connect the receiving mobile phone to the computer as well as the Radial X-amp amplifier and Tascam iXZ in the Direct condition.
Tape measure	Used to measure the distance between the phone and the KEMAR dummy head's mouth as well as the distance between the KEMAR dummy head's mouth and microphone stand with the sound level meter.

Table 5.2. List of equipment with brief descriptions of its use.

5.3.4 Procedure and Technical setup

Firstly, the general setup of the experiment included three main components; a: the 6 loudspeaker array, b: the KEMAR dummy head including the transmitting mobile phone and c: the computer and related setup used for replay and recording including the receiving mobile phone. The 6 loudspeaker array was actively used for the playback of the traffic noise in the With noise condition, while the loudspeaker rig formed the main part of the acoustic environment in the No noise and With noise conditions. Throughout the recordings using the KEMAR dummy head, a curtain was drawn around the entirety of the array to minimise reverberation and acoustic disturbance from the rest of the room and surroundings.

The KEMAR dummy head was placed in the centre of the loudspeaker array and positioned using lasers installed in the array to ensure equal distribution of the background noise around the KEMAR dummy head in the With noise condition. As KEMAR was positioned, the transmitting phone was placed in a manner that replicated a regular phone conversation with the loudspeaker of the mobile phone placed by the ear and the microphone of the phoned angled towards the mouth. As the dummy head is primarily made of plastic apart from the ears, which are made of rubber, the transmitting mobile phone was placed only touching the top of the ear to avoid any vibrations in the plastic caused by the playback of the audio being picked up in the transmission. The distance between the mouth of the dummy head and the phone was measured to 5 cm. This setup was chosen to make the setup as close to an everyday mobile phone call as possible. The exact setup is depicted below in figure 5.1.

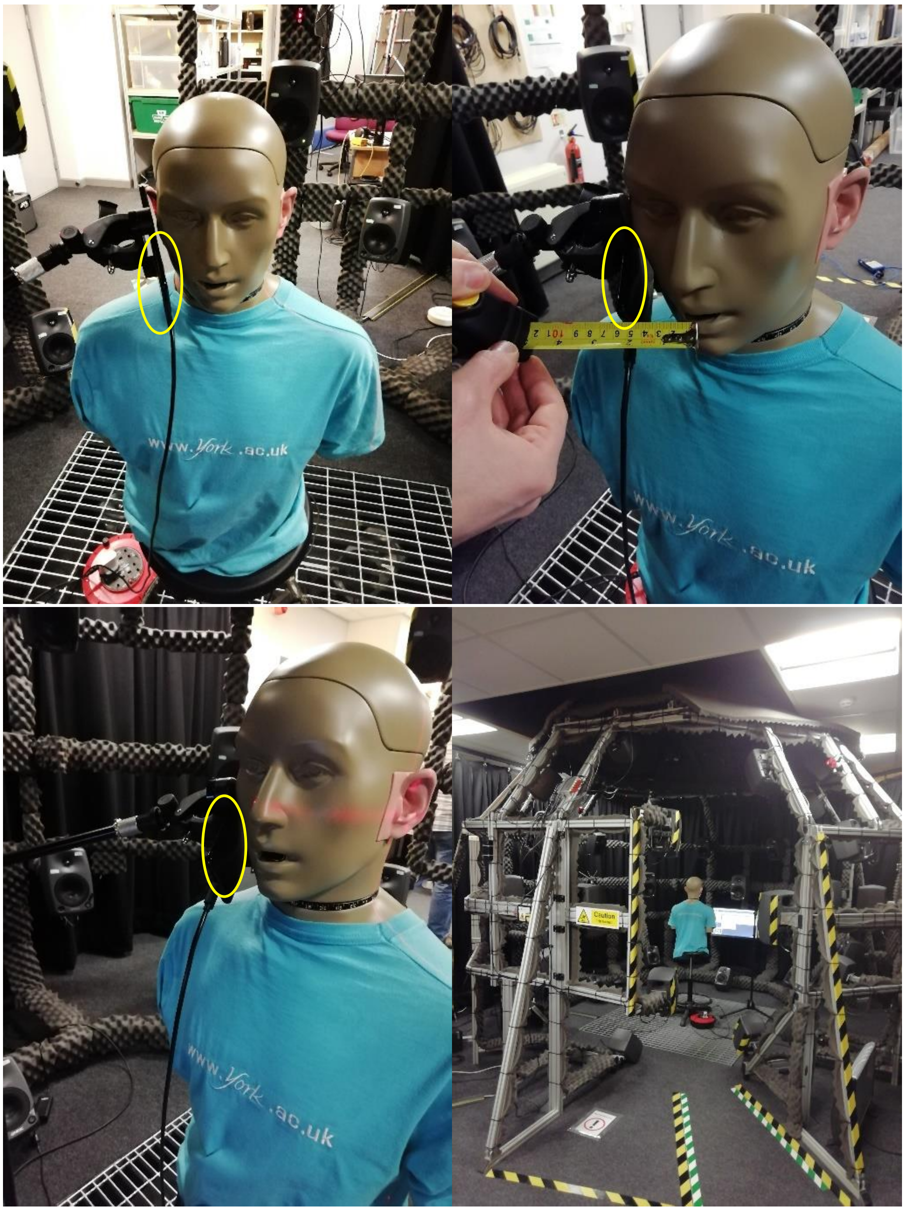


Figure 5.1. Setup of dummy head and transmitting phone
including measure of distance between the mouth of the KEMAR dummy head and the microphone receiving the speech input on the phone as well as the laser based placement in the 6 loudspeaker array. Transmitting mobile phone marked with yellow circle for clarity.

As it is also evident from figure 5.1, the acoustic environment in the loudspeaker array also included a screen placed in front of the KEMAR dummy head as well as the microphone stand used to hold the transmitting mobile phone.

The third part of the setup outside the loudspeaker array consisted of the Windows 10 computer with the Ferrofish audio interface, the receiving phone, the Radial X-Amp and the Tascam iXZ as well as the various cables used to connect the equipment together.

For all conditions, a call was established between the two Samsung Galaxy S8 with the receiving phone initiating the call and the transmitting phone receiving said call. The receiving phone had its microphone turned off to avoid any unwanted feedback due to the proximity of the transmitting and receiving phone. The scenario where the receiving phone would be in this close proximity to the transmitting phone in a real life phone call is highly unlikely, which means a scenario where the receiving phone would pick up feedback from the transmitted signal is highly unlikely. Thus, this decision does not affect the ecological validity of the study, but simply illustrate the effect of the transmission alone.

The mobile phones had the data and internet connection switched off to ensure stable transmission using the 3G network as the 4G and 5G (VoLTE) network access was not considered to be guaranteed across a phone call of this length i.e. just over 50 minutes per condition. As it was shown in the main introduction in Chapter 1, 3G is still the most widespread network and available to almost 95 percent of the world's population (International Telecommunication Union (ITU) 2022b), which founds this decision. However, it must be acknowledge that the scenario is just one possible scenario as most mobile phone calls today must be assumed to be done with an enabled data connection. Instead of risking results, where it would not be possible to track and report which network generation was used i.e. 3G, 4G or 5G (VoLTE), replicability and control of this variable was prioritised.

It is worth mentioning that the current setup does not allow for the influence of the Lombard effect i.e. the tendency for speakers to increase vocal effort, when speaking in loud environments to enhance audibility (Lombard 1911), which would be expected in a real life scenario of a live mobile phone call especially in variable background noise. The inclusion of this factor would require live recordings, which for similar reasons as mentioned above and in the introduction to this study is beyond the scope of this study.

The following sections will consider the specific procedure details for each of the three conditions i.e. Direct, No noise, and With noise.

5.3.4.1 Direct condition

The original speech files were replayed from the computer via the Radial X-amp connected to the Tascam iXZ. The Radial x-amp was used to match the impedance of the computer's audio interface output to the instrument input of the Tascam iXZ. The Tascam iXZ input was set to instrument and its output was connected to the headphone socket of the transmitting Samsung Galaxy S8, which also functions as a microphone input. The input level dial on the Tascam iXZ, does not have any influence on the output level of the device.

For the receiving Samsung Galaxy S8, a cable was plugged into the headphone socket, which then fed to the computer's audio interface. This signal sent to the computer was recorded on a new track within the same project file as the original source audio. The output level of the receiving phone was set to full, and was connected to the line input of the computer's audio interface with no gain change i.e. gain at 0 dB.

5.3.4.2 No noise condition

Firstly, the track in Reaper containing the speech files had a convolution filter /correction filter on its output, which corrected for the non-flat frequency response of the loudspeaker within the KEMAR dummy head. Without this compensating filter, the frequency response of the loudspeaker would boost the higher frequencies relative to the lower frequencies. The compensated output from the computer was then fed to the KEMAR dummy head.

The replay level of the speaker in KEMAR was set to what was perceptually judged prior to experiment by the author and supervisor to be a reasonable level for normal speech. This level was subsequently measured as 58.4 dBA RMS with an 84.6 dBA peak at 1 meter from the mouth of KEMAR for the first minute of the first speaker i.e. AM03. At 5 cm from the mouth i.e. by the microphone of the transmitting Samsung Galaxy S8, the level was measured as 80.2 dBA RMS with a peak at 106.3 dBA. These measurements were made using a NTi XL2 sound level meter.

The original sound files were then replayed from KEMAR. The microphone of the transmitting Samsung Galaxy S8 then picked up this acoustic signal, which was transmitted to the receiving phone. The output from the receiving Samsung Galaxy S8 was then fed to the computer's audio interface via the phone's headphone socket and recorded on a new track within the same project as the original audio as in the Direct condition.

5.3.4.3 With noise condition

The physical setup and recording of the signal in Reaper in this condition is identical to that described for the No noise condition above. This is except the replay of the traffic noise via the 6 loudspeaker array. Since the traffic noise recording was shorter than the speech samples the noise recording was repeated, and there was an overlap with the start of the next traffic noise sample of approximately 7.8 seconds. To ensure a smooth transition and no increase in overall amplitude, a crossfade was applied across repetition. The traffic noise audio files were added to the Reaper project as a separate track. Within Reaper, the output of this track was routed via the ambisonic decoder i.e. AmbiX, in order for the correct decoded signal to be fed to each of the 6 speakers in the loudspeaker array.

Again, the level of the traffic noise was to set a perceptually realistic level judge by the author and supervisor. This was then measured over the first minute using the NTi XL2 sound level meter. This showed that the traffic noise had a RMS sound pressure level of 70.5 dBA with a peak of 92.6 dBA in the centre of the loudspeaker array at the position of the KEMAR dummy head. The playback of the traffic noise recording began before the speech samples and continued past the end of each speech sample. This was done to ensure uninterrupted background noise across all speech samples.

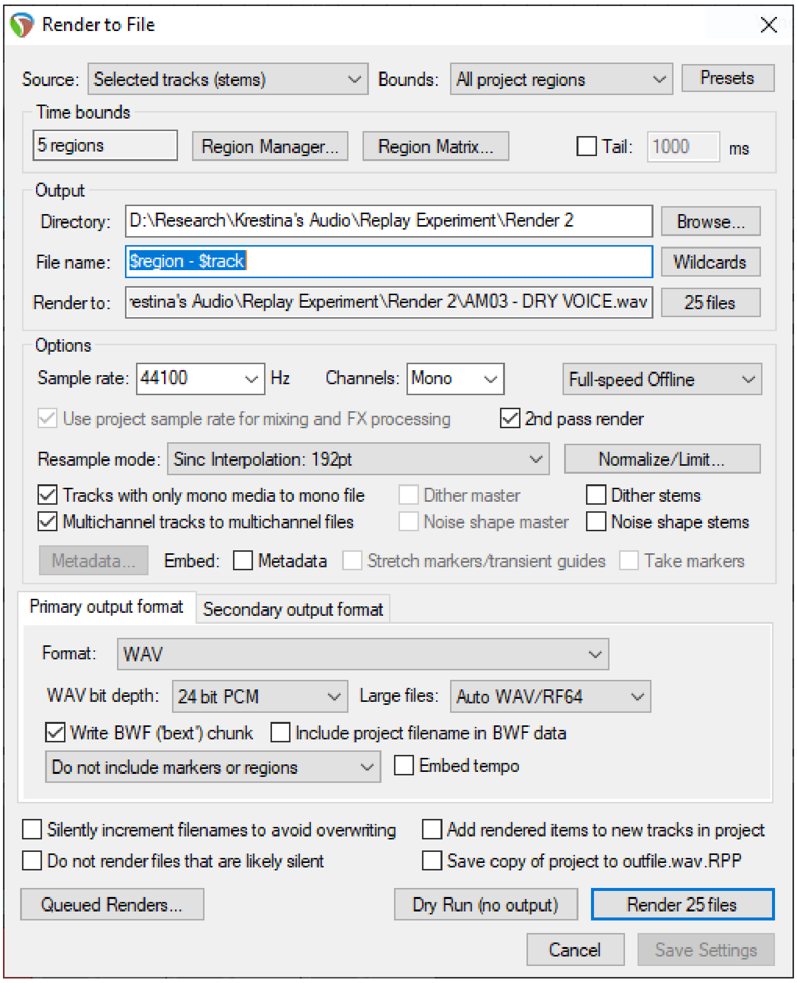
5.3.5 Sound files

The sound files to be transmitted were the original 44.1 kHz WAV files of the Chicken Little reading. The 44.1 kHz files were not down-sampled as the input signal was intended to replicate real speech and thus, the full frequency range was needed. All MATLAB and Reaper related analysis described here was done by Philip Harrison.

The speech files used for the three conditions were added to a single project with Reaper (Cockos Incorporated 2023). They were placed on a timeline with a gap of between 16 and 20 seconds to avoid the files overlapping and two tracks were used with identical audio in each. One track was used for the Direct condition, while the other was used for the No noise and With noise conditions as a correction filter was applied for the latter.

The sound file used in the With noise condition to replicate background noise was an open access online available ambisonic B-Format (AmbiX) file from SoundField by the microphone manufacturer RØDE (Soundfield by RØDE 2023). Specifically, this was track 61 *St Kilda Road Traffic* (Schutze, n.d.). The traffic noise had a duration of 00:05:31.7, which is shorter than the speech samples. Therefore, multiple instances of the file (2 to 3) were used to ensure continuous traffic sounds for each recording. In addition, within the Reaper project, the amplitude of all the speech files was adjusted via level normalisation using a target RMS level of -30 dB. This was done to ensure an equivalent replay level across speakers and to maximise the replay level without clipping the signal. The transmitted signal was being recorded in real time to the same project file.

Regions were added to the continuous recordings of the output from the receiving phone to demarcate the speech from each of the 5 speakers. The recordings of the transmitted speech for each speaker and each condition were then exported as separate files. The settings illustrated in figure 5.2 below were used to render the recording for each speaker for each of the three conditions.



Render to File

Source: Selected tracks (stems) Bounds: All project regions Presets

Time bounds

5 regions Region Manager... Region Matrix... □ Tail: 1000 ms

Output

Directory: D:\Research\Krestina's Audio\Replay Experiment\Render 2 Browse...

File name: $region - $track Wildcards

Render to: restina's Audio\Replay Experiment\Render 2\AM03 - DRY VOICE.wav 25 files

Options

Sample rate: 44100 Hz Channels: Mono Full-speed Offline

☑ Use project sample rate for mixing and FX processing ☑ 2nd pass render

Resample mode: Sinc Interpolation: 192pt Normalize/Limit...

☑ Tracks with only mono media to mono file □ Dither master □ Dither stems

☑ Multichannel tracks to multichannel files □ Noise shape master □ Noise shape stems

Metadata... Embed: □ Metadata □ Stretch markers/transient guides □ Take markers

Primary output format | Secondary output format

Format: WAV

WAV bit depth: 24 bit PCM Large files: Auto WAV/RF64

☑ Write BWF ('bext') chunk □ Include project filename in BWF data

Do not include markers or regions □ Embed tempo

□ Silently increment filenames to avoid overwriting □ Add rendered items to new tracks in project

□ Do not render files that are likely silent □ Save copy of project to outfile.wav.RPP

Queued Renders... Dry Run (no output) Render 25 files

Cancel Save Settings

Figure 5.2. The settings in Reaper used to render the recording for each speaker for each of the three conditions

Lastly, the inspection of the mobile phone transmitted files revealed an offset caused by the transmission (i.e. a delay) between the original file and their live transmitted counterparts. This delay was variable across the duration of the recordings. Consequently, the alignment of each segment (including surrounding speech) was calculated independently in a modified version of the main

MATLAB script described below using correlation to ensure that the same frame of the segment was analysed both in the original and the transmitted versions.

Since the offset of the original WAV files and the transmitted version had different offsets, it was decided to as part of the main spectral analysis script to calculate the offset separately for each token prior to the spectral analysis. This script was a MATLAB script written by Philip Harrison (Harrison 2022; MathWorks Inc. 2010). The spectra was then extracted as before using multitaper analysis (see details in section 3.3.5 on data extraction).

5.3.6 Segmentation

This study used the same forced aligned and corrected TextGrids as the baseline study. Thus, the reader is referred to section 3.3.4 for further details.

5.3.7 Data extraction, measurements, and statistical analysis

For the spectral analysis, a lower frequency limit was again set to 500 Hz in order to avoid the influence from mains hum and voicing, while the upper analysis frequency was set to 8 kHz. This was done consistently across all recordings both the original WAV as well as the experimental conditions. The upper analysis frequency was set based on the same criteria as previous studies i.e. to ensure comparability with codec compression as well as previous results.

CoG, SD, skewness, kurtosis and frequency peak were measured and extracted using a modified version of the MATLAB script written by Philip Harrison for the two previous studies (Harrison 2022; MathWorks Inc. 2010). The measures were taken from the 20 ms central frame (see section 3.3.5 for further details on this). In addition to analysing the 15 files generated in the experiment i.e. 1 for each of the 5 speakers in each condition, the spectral analysis was also performed on the original audio files to allow direct comparison.

As mentioned the measurements were again obtained using multitaper analysis instead of the more traditional LPC, periodogram, and FFT (see section 3.3.5 for further details on this). For this reason, no additional windowing or pre-emphasis was applied.

Next, the tokens to be rejected were identified i.e. any token with a CoG below 1 kHz in one or both conditions. The number of 1 kHz tokens were very limited and therefore, it was decided to not do any separate analysis on these and keep them as part of the primary dataset. The number of tokens with a CoG below 1 kHz per segment is presented below in table 5.3.

Segment	**Total** (CoG below 1 kHz)
/f/	2
[fʲ]	0
/s/	0
/z/	18
/ʃ/	0
/θ/	2
/ð/	401

Table 5.3. Total number of tokens for each segment with CoG below 1 kHz

Divided by conditions 130 of the tokens below 1 kHz occurred in the Original condition, 98 in the Direct condition, 116 in the No noise condition, and lastly 79 tokens in the with condition.

Spectrograms, waveforms and spectra for all segments were generated and extracted and a representative selection of tokens across segments, speakers and conditions were inspected and will be presented in section 5.4. on results. This was done to qualitatively inspect the potential effects of the live transmission.

Lastly, all statistical analysis was done in R (RStudio Team 2019; RCore Team 2020). For this study, the analysis consisted of descriptive anlysis as well as distribution plots rather than inferential analysis. The decision not to include mixed effects models, was based on time constraints as well as the reduced number of data points in the individual datasets, which were substantially less than for the previous studies. Consequently, no p-values will be presented in this study. However, the changes in e.g. mean and maximum values will still give clear indications of the gravity of the observed effects of the transmission.

5.4 Results

All results presented in this section are obtained from one main dataset of spectral measures extracted from the recordings obtained from the experiment as described above in section 5.3. The results will be presented individually for each condition and segment in the coming sections. However, before these more detailed results are presented, some more general observations and summary statistics will be included.

Firstly, the limitations in bandwidth for the present study is not assessed via a white noise signal as in the previous Chapters. This is because the white noise signal would need to be part of the experimental setup, which was not the case, and thus a live transmission of the white noise signal is not available at present. From visual inspection of the representative selection of spectrograms, which will be presented in the following sections, the bandwidth appear to be around 8 kHz as expected based on the sample rate of the codecs. This indicate, that a good connection has been available.

This observation is supported by the peak frequency maximum values, which are presented below for each segment and condition in table 5.4. These values are not representative of the true frequency peaks due to the analysis band being between 500 Hz and 8 kHz. Because, the files have an actual upper frequency limit at 22 kHz, the values and the similarities between them are an artefact of the multitaper analysis e.g. the fact that the values are 10 Hz above the upper frequency limit of the analysis. However, they still provide an indication of condition and segment dependent tendencies as well as the available bandwidth i.e. any value at 8,010 Hz show that the actual peak is above this limit, but not with how much.

From these values, it is evident how the effect of the live transmission is again dependent on condition, and to an even greater extent segment-dependent with over 3,000 Hz changes for /f/, and no changes for /s/ despite the both having a maximum value of 8,010 in the original WAV baseline.

Seg.	**Condition**	**Freq. Peak Maximum values (Hz)**
/f/	Direct	4393
/f/	No_noise	6202
/f/	With_noise	4522
/f/	**Original**	8010
[fʲ]	Direct	3876
[fʲ]	No_noise	3919
[fʲ]	With_noise	3919
[fʲ]	**Original**	7924
/s/	Direct	8010
/s/	No_noise	8010
/s/	With_noise	8010
/s/	**Original**	8010
/z/	Direct	7881
/z/	No_noise	7881
/z/	With_noise	8010
/z/	**Original**	8010
/ʃ/	Direct	3919
/ʃ/	No_noise	3704
/ʃ/	With_noise	4134
/ʃ/	**Original**	3962
/θ/	Direct	6202
/θ/	No_noise	7795
/θ/	With_noise	6718
/θ/	**Original**	8010
/ð/	Direct	7752
/ð/	No_noise	7795
/ð/	With_noise	7924
/ð/	**Original**	8010

Table 5.4 Maximum values frequency peak from the central frame for each fricative in each condition.

Secondly, the mean values for each spectral measure, which are presented below in table 5.5, reveal how all spectral measures across segments apart from skewness for [fʲ] are to varying degrees affected by the live transmission. Moreover, it is evident how this effect varies depending on segment and condition e.g. there is almost no effect on /ʃ/. The more specific analysis for each segment will be done in the following sections, where the condition dependent results are presented.

Seg.	Condition	CoG	SD	Skew	Kurt	Freq. Peak
		(Hz)	(Hz)	(Hz)	(Hz)	(Hz)
/f/	Direct	2462	1113	0.80	4.15	2044
/f/	No_noise	2610	1180	1.24	5.70	2214
/f/	With_noise	2316	1199	1.41	6.90	1798
/f/	**Original**	**3474**	**1935**	**0.70**	**2.95**	**2765**
[fʲ]	Direct	2506	1015	0.90	4.74	2010
[fʲ]	No_noise	2635	1074	1.39	6.41	2182
[fʲ]	With_noise	2506	1076	1.50	8.71	2020
[fʲ]	**Original**	**3378**	**1804**	**0.90**	**3.36**	**2613**
/s/	Direct	4061	950	-0.17	7.52	4088
/s/	No_noise	4255	1123	0.28	6.03	4211
/s/	With_noise	4268	1181	0.05	6.21	4225
/s/	**Original**	**5088**	**1343**	**0.17**	**4.48**	**4975**
/z/	Direct	3571	1026	-0.46	7.55	3232
/z/	No_noise	3739	1131	0.09	7.00	3273
/z/	With_noise	3615	1227	-0.09	6.78	3244
/z/	**Original**	**4431**	**1365**	**0.31**	**5.77**	**3919**
/ʃ/	Direct	2940	572	1.05	8.06	2872
/ʃ/	No_noise	2918	578	1.68	12.89	2873
/ʃ/	With_noise	2915	600	1.59	12.65	2857
/ʃ/	**Original**	**3200**	**939**	**2.15**	**10.63**	**2930**
/θ/	Direct	2477	1240	0.76	4.12	1831
/θ/	No_noise	2727	1356	1.02	5.02	2083
/θ/	With_noise	2076	1172	1.69	9.02	1577
/θ/	**Original**	**3539**	**1986**	**0.51**	**3.19**	**2435**
/ð/	Direct	1834	1073	1.40	8.01	1028
/ð/	No_noise	1897	1095	1.47	9.29	1065
/ð/	With_noise	1534	826	1.60	10.04	1026
/ð/	**Original**	**2132**	**1404**	**1.86**	**11.92**	**1177**

Table 5.5. Mean values for all spectral measures for each fricative in each individual condition and the original WAV files.

Table 5.6 below depicts the directionality and magnitude of the changes observed to the mean values in Table 5.5. Overall, it is clear from these values that apart from skewness and kurtosis all spectral measures regardless of segment and condition are lowered in comparison to the original recording. The magnitude of these changes are again segment and condition dependent and will be presented more in depth in the following sections.

Seg	Codec	CoG (Hz)	CoG (%)	SD (Hz)	SD (%)	Skew (Hz)	Skew (%)	Kurt (Hz)	Kurt (%)	Freq. Peak (Hz)	Freq. Peak (%)
/f/	Direct	1012	-29	822	-42	-0.10	15	-1.20	41	721	-26
/f/	No_noise	864	-25	755	-39	-0.54	78	-2.75	93	551	-20
/f/	With_noise	1158	-33	736	-38	-0.71	102	-3.95	134	967	-35
[fʲ]	Direct	872	-26	789	-44	0	0	-1.38	41	603	-23
[fʲ]	No_noise	742	-22	731	-40	-0.49	54	-3.04	90	431	-16
[fʲ]	With_noise	872	-26	728	-40	-0.60	67	-5.34	159	593	-23
/s/	Direct	1027	-20	393	-29	0.34	-201	-3.04	68	887	-18
/s/	No_noise	833	-16	220	-16	-0.11	68	-1.55	35	765	-15
/s/	With_noise	820	-16	162	-12	0.11	-68	-1.74	39	751	-15
/z/	Direct	860	-19	340	-25	0.77	-246	-1.78	31	687	-18
/z/	No_noise	691	-16	235	-17	0.22	-71	-1.24	21	647	-16
/z/	With_noise	816	-18	138	-10	0.40	-127	-1.01	18	675	-17
/ʃ/	Direct	260	-8	367	-39	1.10	-51	2.57	-24	58	-2
/ʃ/	No_noise	282	-9	361	-38	0.47	-22	-2.25	21	57	-2
/ʃ/	With_noise	285	-9	338	-36	0.56	-26	-2.01	19	73	-2
/θ/	Direct	1062	-30	746	-38	-0.25	48	-0.93	29	604	-25
/θ/	No_noise	812	-23	630	-32	-0.51	100	-1.83	57	352	-14
/θ/	With_noise	1464	-41	814	-41	-1.18	232	-5.83	183	858	-35
/ð/	Direct	298	-14	331	-24	0.46	-25	3.91	-33	149	-13
/ð/	No_noise	235	-11	309	-22	0.39	-21	2.63	-22	112	-10
/ð/	With_noise	598	-28	578	-41	0.26	-14	1.88	-16	152	-13

Table 5.6. Differences in mean values between baseline (Original) and the different recording conditions in Hz and percentage. Colours indicate the direction of the change. (i.e. blue = decrease; yellow = increase, no colour = no effect)

Lastly, the difference in mean values between /f/ and /θ/ are presented below in Table 5.7. This is done to illustrate how different or similar these two segments are as a consequence of the codec compressions, which is again related to TH-fronting. /s/-retraction has also been mentioned in previous sections and Chapters, but again the difference between the two segments is regardless more than 1 kHz, and the results will therefore not be presented in more detail here.

Diff. between	Condition	CoG (Hz)	SD (Hz)	Skew	Kurt	Freq. Peak (Hz)
/f/ & /θ/	Direct	15	127	0.04	0.03	213
/f/ & /θ/	No_noise	117	176	0.22	0.68	131
/f/ & /θ/	With_noise	240	27	0.28	2.12	221
/f/ & /θ/	**Original**	**65**	**51**	**0.19**	**0.24**	**330**

Table 5.7. The difference in mean values between /f/ and /θ/ in the WAV baseline (i.e. original) and the three recording conditions

Table 5.7 shows that the change in spectral values for /f/ and /θ/ is not consistent across the recording conditions, and none of the measures became identical. For CoG in the conditions with and without background noise, the two segments became more distinct, while in the Direct condition are only different by 15 Hz. For SD, a similar pattern appeared, but here they were less distinct in the condition with noise, while they became more distinct in the two other conditions. This can also be seen for skewness and kurtosis, while the frequency peak mean values were consistently more alike in the live transmission. However, for the frequency peak, the distance between the segments never became less than 100 Hz.

5.4.1 Direct condition

This section will present the individual results for each segment and the spectral measures in the comparison between the original WAV and the live transmission in the Direct condition.

First, the trajectory of the mean values for each spectral measure and the individual segments can be found below (figure 5.3 to 5.6). These indicate the directionality of the changes imposed by the transmission.

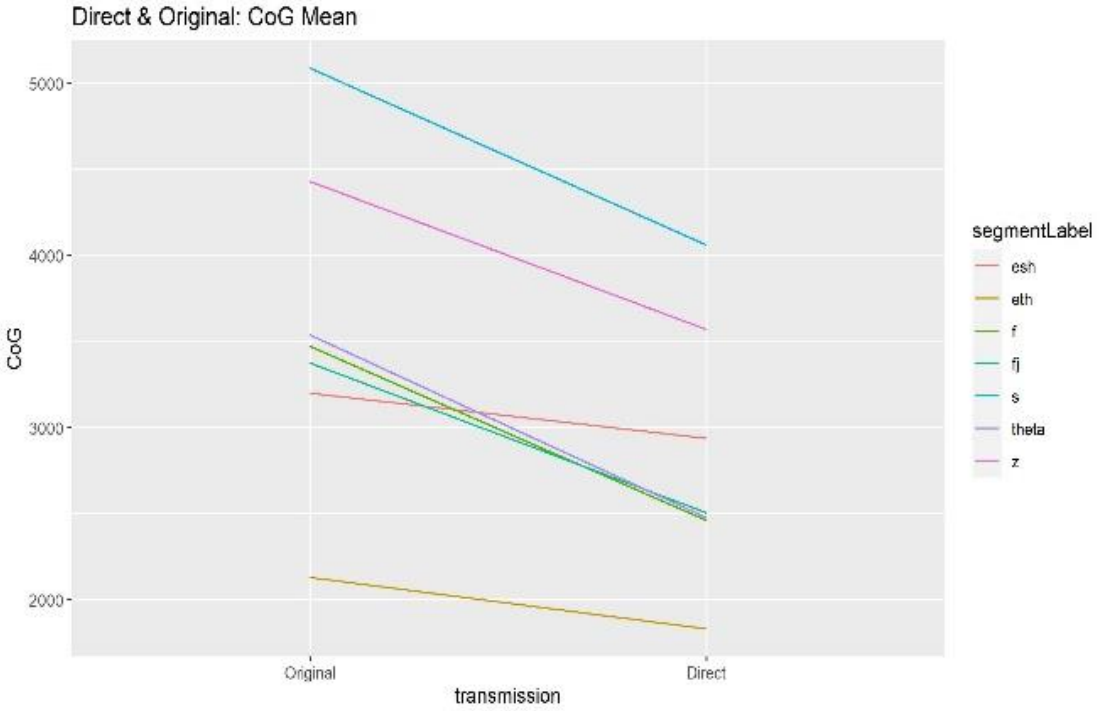


Figure 5.3. Trajectory of mean values for CoG and individual segment in the comparison between the original WAV files and the Direct condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

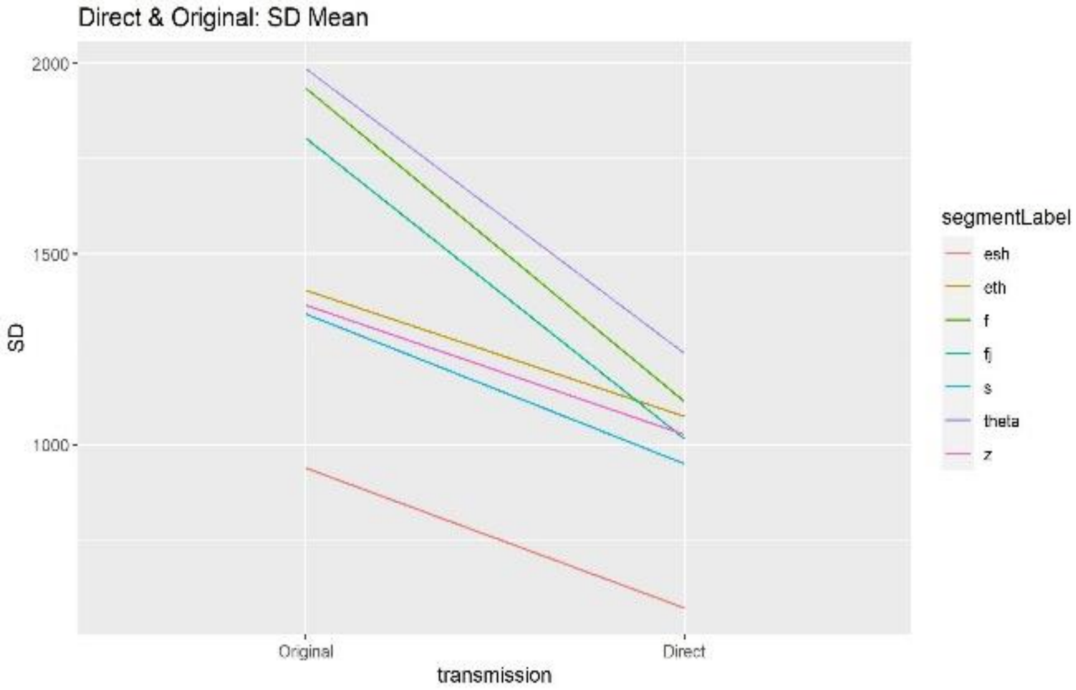


Figure 5.4. Trajectory of mean values for SD and individual segment in the comparison between the original WAV files and the Direct condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

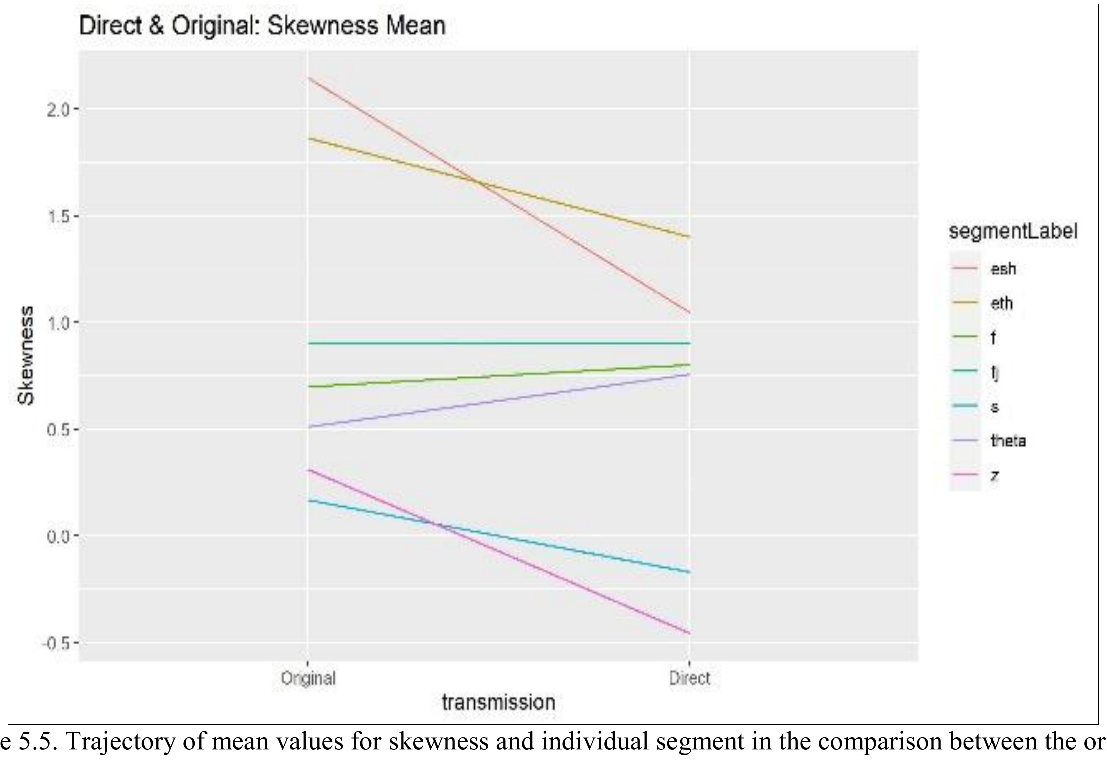


Figure 5.5. Trajectory of mean values for skewness and individual segment in the comparison between the original WAV files and the Direct condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

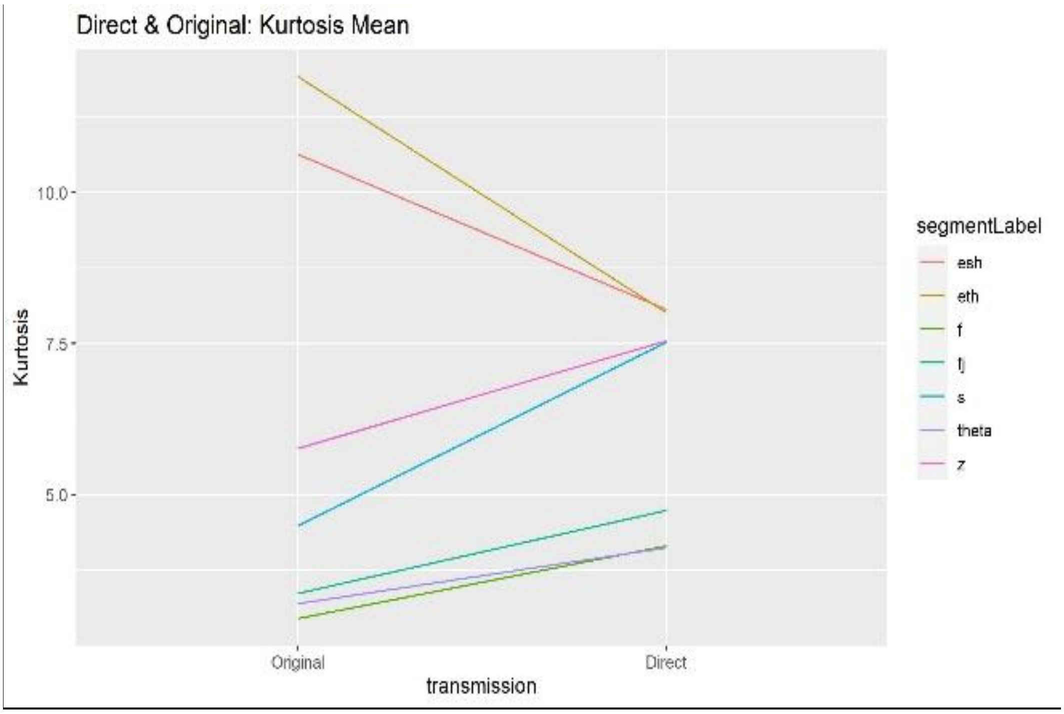


Figure 5.6. Trajectory of mean values for kurtosis and individual segment in the comparison between the original WAV files and the Direct condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The trajectory plots indicate that based on the mean values, /f/ and /θ/ are generally the two sounds, which tend to become the most alike for all four spectral measures following the direct transmission. However, for e.g. CoG these two segments are already similar in the original WAV files. For SD, [fʲ] and /z/ become almost identical, while /θ/ and especially /f/ become more like /ð/. For skewness, again /θ/ becomes more like /f/, while little change can be observed in the relation between the other segments for this measure. Finally, kurtosis shows some varying patterns. For this measure a number of segments become identical, this includes /ð/ and /ʃ/, /z/ and /s/, /θ/ and /f/, while [fʲ] and /θ/ become more distinct.

The distributions are illustrated below in a set of violin plots for each spectral measure (figure 5.7 to 5.9). The specific analysis pertaining to each segment will be found in the following sections.

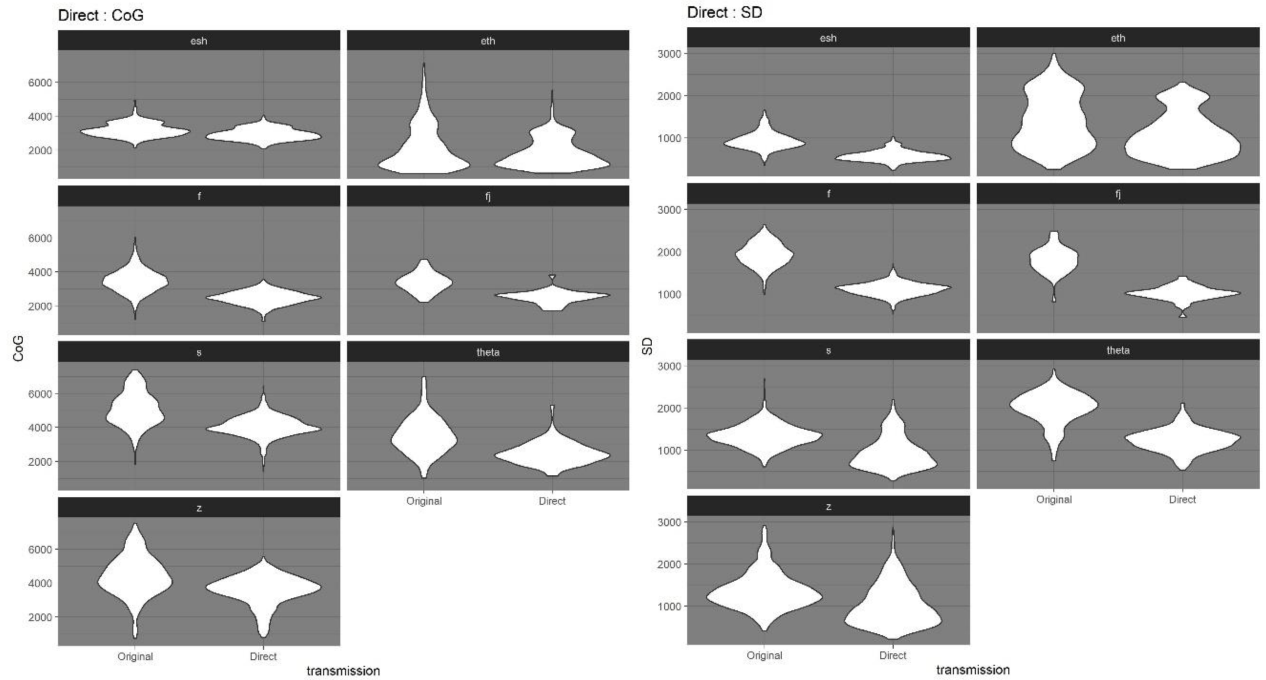


Figure 5.7. Distribution of CoG and SD values in Original baseline and the Direct condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

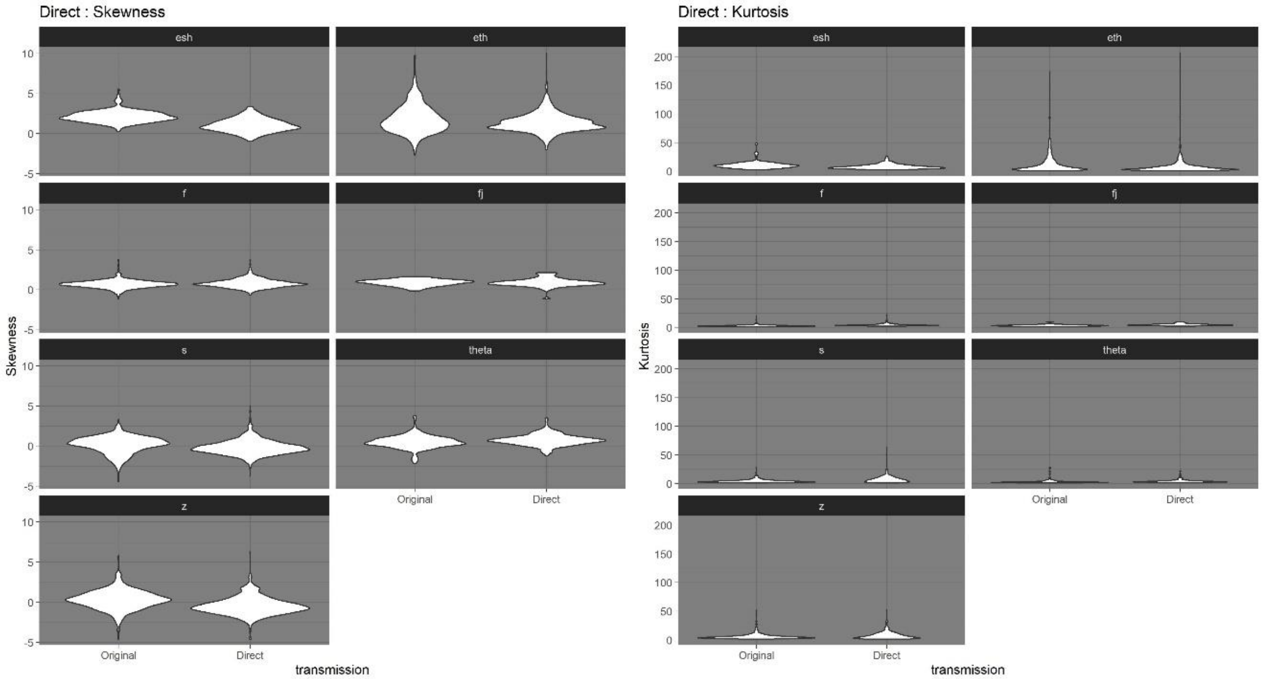


Figure 5.8. Distribution of skewness and kurtosis values in Original baseline and the Direct condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^j], and eth = /ð/.

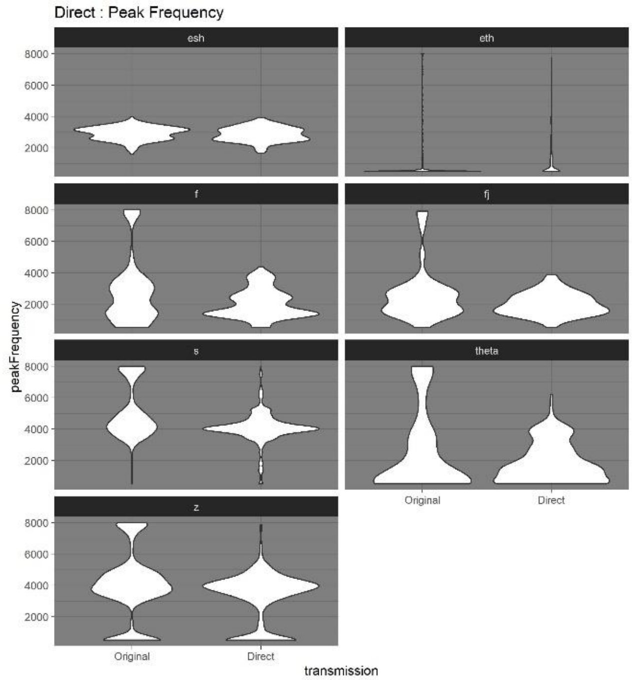


Figure 5.9. Distribution of frequency peak values in Original baseline and the Direct condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

In brief, the live transmission in the Direct condition consistently lowered CoG and SD, while more mixed patterns with both decreases and increases can be observed for skewness and kurtosis. This is also evident from the distribution plots, where it is clear that some of the greatest effects are found for SD, while frequency peak also clearly lowers apart from /ʃ/ and /ð/.

5.4.1.1 /f/

The distribution plots illustrate how the most prominent effects in terms of changes to distribution shape can be observed for CoG, SD and frequency peak. Both CoG and SD show a lowered overall distribution, this is especially true for SD, while both measures also present a more condensed distribution following the direct transmission. This is in the sense that the values are spread over a smaller range of frequencies and more prominently around the mean. The violin plot indicate a clear lowering of frequency peak with a cut-off just over 4 kHz, and a centring of values lower in the spectrum. The effect on skewness and kurtosis are limited when looking at these plots.

In more detail, the mean values revealed how for /f/ CoG, SD, and frequency peak all lowered in the Direct condition, while skewness and kurtosis both increased following transmission. For CoG and frequency peak this was by just below 30 percent or 1,012 Hz for CoG and 721 Hz for frequency peak. SD was relatively more affected by the transmission by a change of 42 percent or 822 Hz. The effect on skewness was 15 percent, which was an increase of 0.10. For kurtosis, a similar effect to SD was observed by a change of 41 percent of 1.20.

Finally, the comparison of the spectrogram for the baseline and the Direct condition (see example in figure 5.10), revealed how especially the high frequency content was reduced in the Direct condition. In addition, the waveform and spectrum reveal clearly lowered intensity across the segment together with a more even distribution of amplitude across the spectrum especially above 5 kHz.

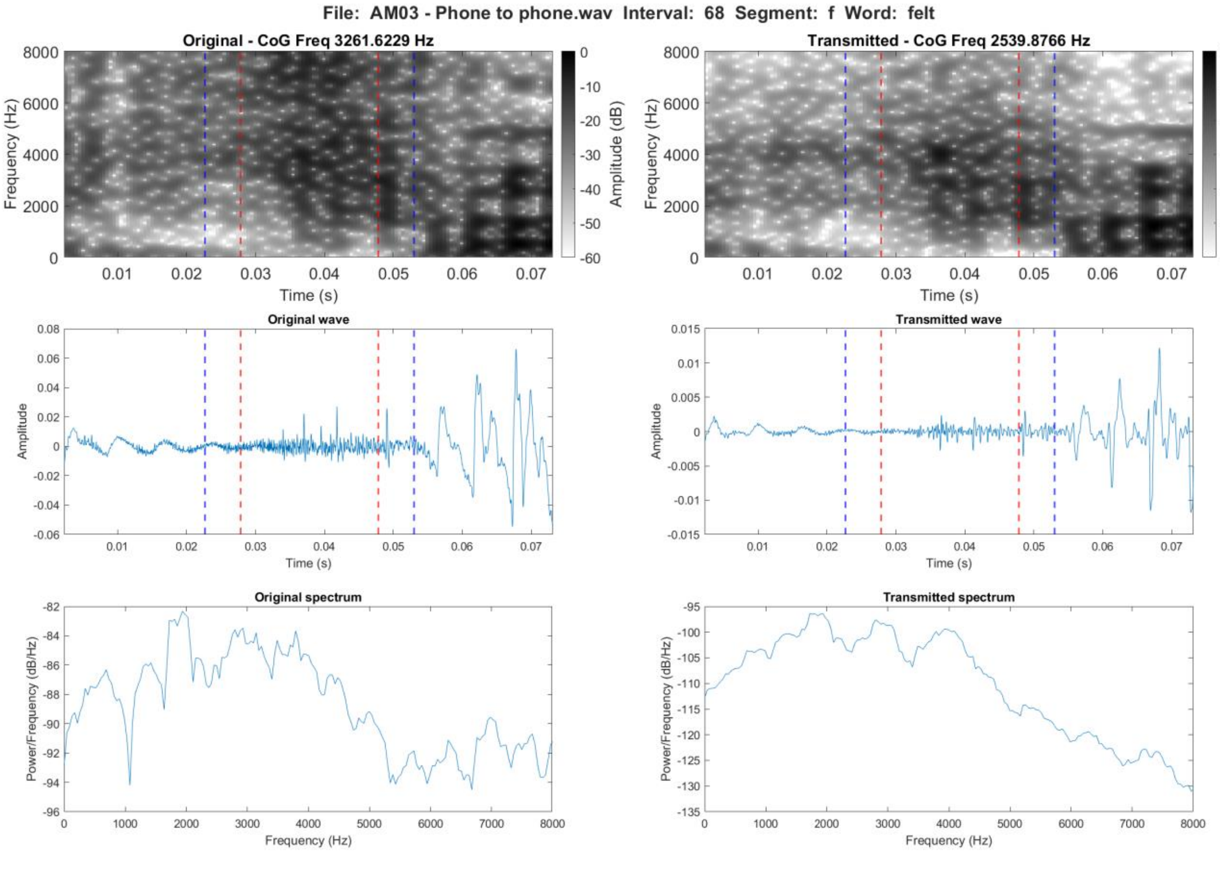


Figure 5.10. Spectrographic comparison of /f/ in the word *felt* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.2 [fʲ]

Similar to /f/ the most visible effects in terms of distribution are found for CoG, SD and frequency peak. In addition, the observed effects for particularly CoG and SD appear similar for the two segments with a lowering and centring of values around the mean. For frequency peak, the changes to the distribution follow a similar pattern to /f/, but with a more uniform distribution following transmission similar to the shape observed for CoG and SD. For skewness and kurtosis, no clearly visible effects are observable from the plots apart from a slight increase in outliers for skewness.

This is expressed in the mean values in the following way. [fʲ] showed similar patterns to /f/, however with no observed effect of transmission on skewness. CoG lowered by 26 percent or 872 Hz and frequency peak by 23 percent or 603 Hz. For SD and kurtosis, the effect was again similar to /f/ with a change of 44 percent or 789 Hz for SD and 41 percent or 1.38 for kurtosis.

As with /f/, the spectrographic representation reveals a primary effect of the transmission on the high frequency content above 4 to 5 kHz as well as an overall lower intensity and a smoothed spectrum (see figure 5.11). However, in contrast to /f/, the formant structure of [fʲ] becomes clearer following transmission.

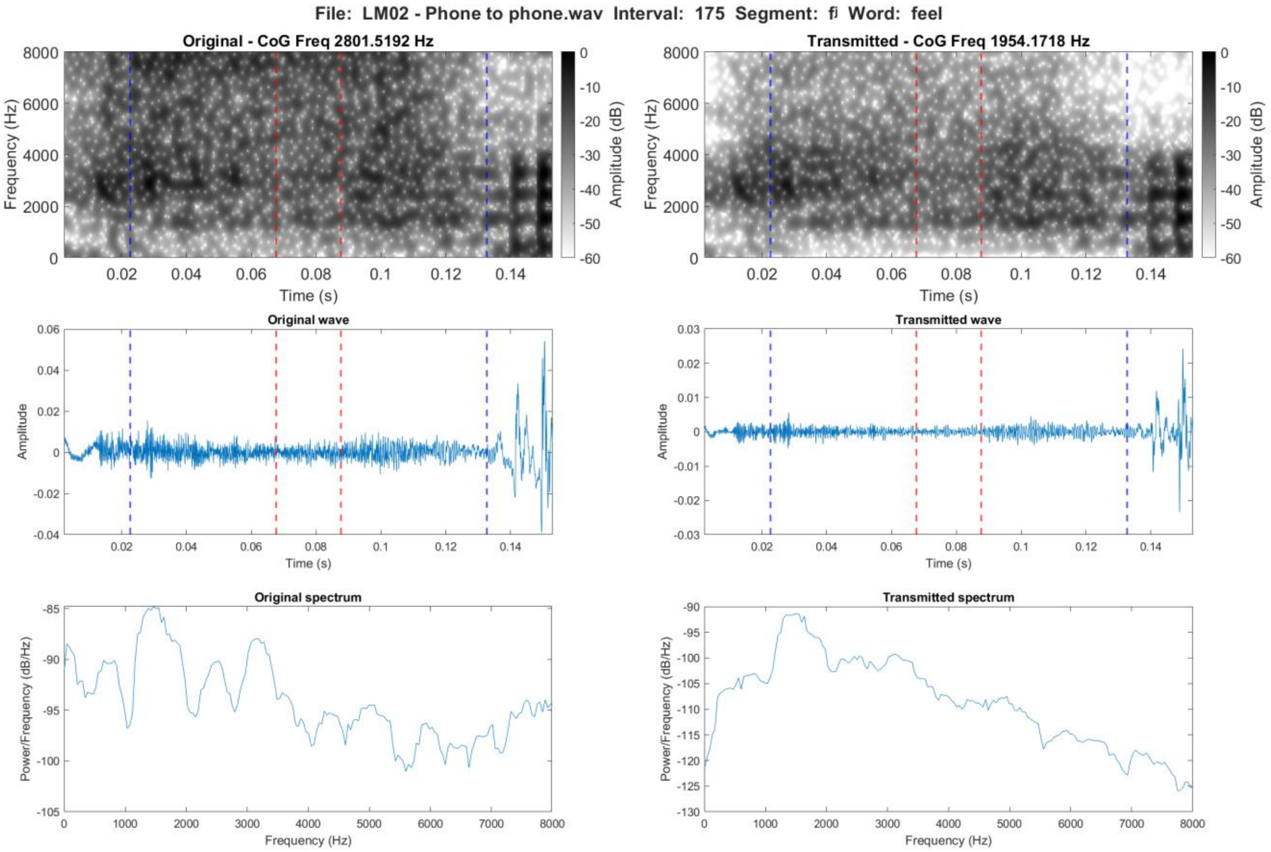


Figure 5.11. Spectrographic comparison of [fʲ] in the word *feel* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.3 /θ/

A clear decrease is observable for CoG, SD and frequency peak based on the violin plots. For CoG and SD this is in a similar manner to what was observed for /f/ and [fʲ], where the distribution is narrower across a smaller span of frequencies and more centred around the mean. For frequency peak, the lowering results in a comparatively smaller change in the shape of the distribution, but a clear increase of values in the lower part of the spectrum. Again, little effect is observable from the plots for skewness and kurtosis.

Despite the lack of observable effect on the distribution, as with /f/ and [fʲ], skewness and kurtosis increased in the Direct condition, while the rest of the spectral measures lowered in mean value. The observed effects were for all measures apart from kurtosis, which increased by 29 percent, larger than what was observed for /f/ and [fʲ]. For CoG, this meant a decrease of 30 percent or 1,062 Hz and for frequency peak a decrease by 25 percent or 604 Hz. SD decreased by 38 percent or 746 Hz, while skewness increased by 0.25 or 48 percent.

The spectrogram in figure 5.12 shows how the baseline present a high level of variation in both intensity across the frequency span, while its live transmitted counterpart presents a much more regular pattern. Similar to the previous segments, the transmission lowered the amplitude, while the spectrum reveal additional variation in particularly above 4 kHz, which was not present in the WAV baseline.

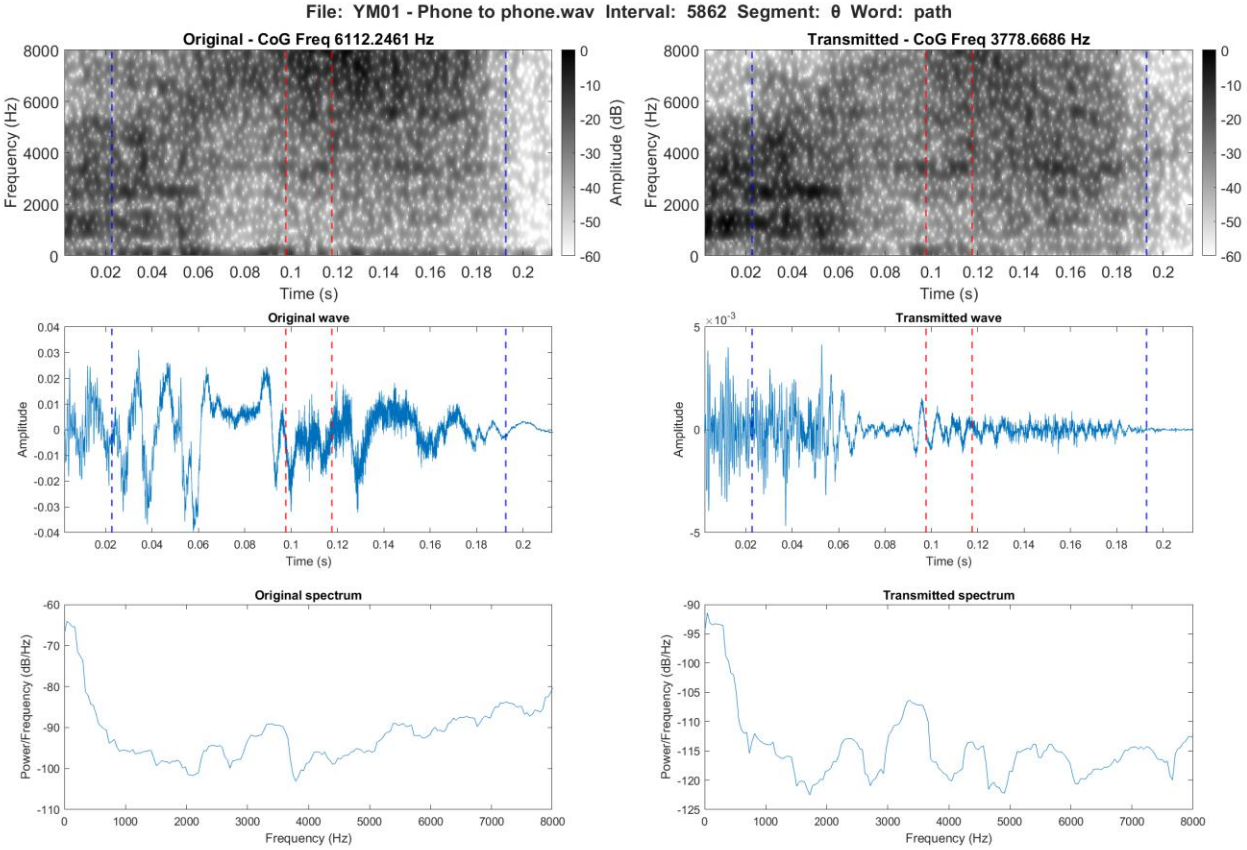


Figure 5.12. Spectrographic comparison of /θ/ in the word *path* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.4 /s/

For /s/, the distribution plots, reveal tendencies similar to /f/ and [fʲ] for CoG. This means that CoG is lowered and more clearly centred around the mean with a slightly narrower distribution in terms of frequencies. For SD a lowering is also clearly observable, but here the values are more widespread and have a more substantial set of values placed above the mean. A lowered and more narrow distribution is observable for skewness, while a slight increase appear to be visible for kurtosis. For frequency peak the distribution is more centred around the mean, while the frequency range appears unchanged.

All mean values, apart from kurtosis, were lowered in the Direct condition for /s/. This was by 20 and 18 percent for CoG and frequency peak respectively, or 1,027 Hz for CoG and 887 Hz for frequency peak. SD was lowered by 29 percent, which was slightly more than CoG. For SD this was a change of 393 Hz. Skewness showed the third biggest effect observed across all segments and conditions. This was a lowering of 201 percent or 0.34. For kurtosis, an increase of 68 percent or 3.04 was observed.

The spectrograms and spectra (see figure 5.13) show how the high frequency content from around 5 kHz for /s/ is reduced following the transmission, while the main concentration of energy around 4 kHz for the segment are intensified. Both the waveforms reveal that the transmission has minimised irregularities and frication, while the spectrum is largely identical to the WAV baseline up until around 5 kHz, where a clear drop appear.

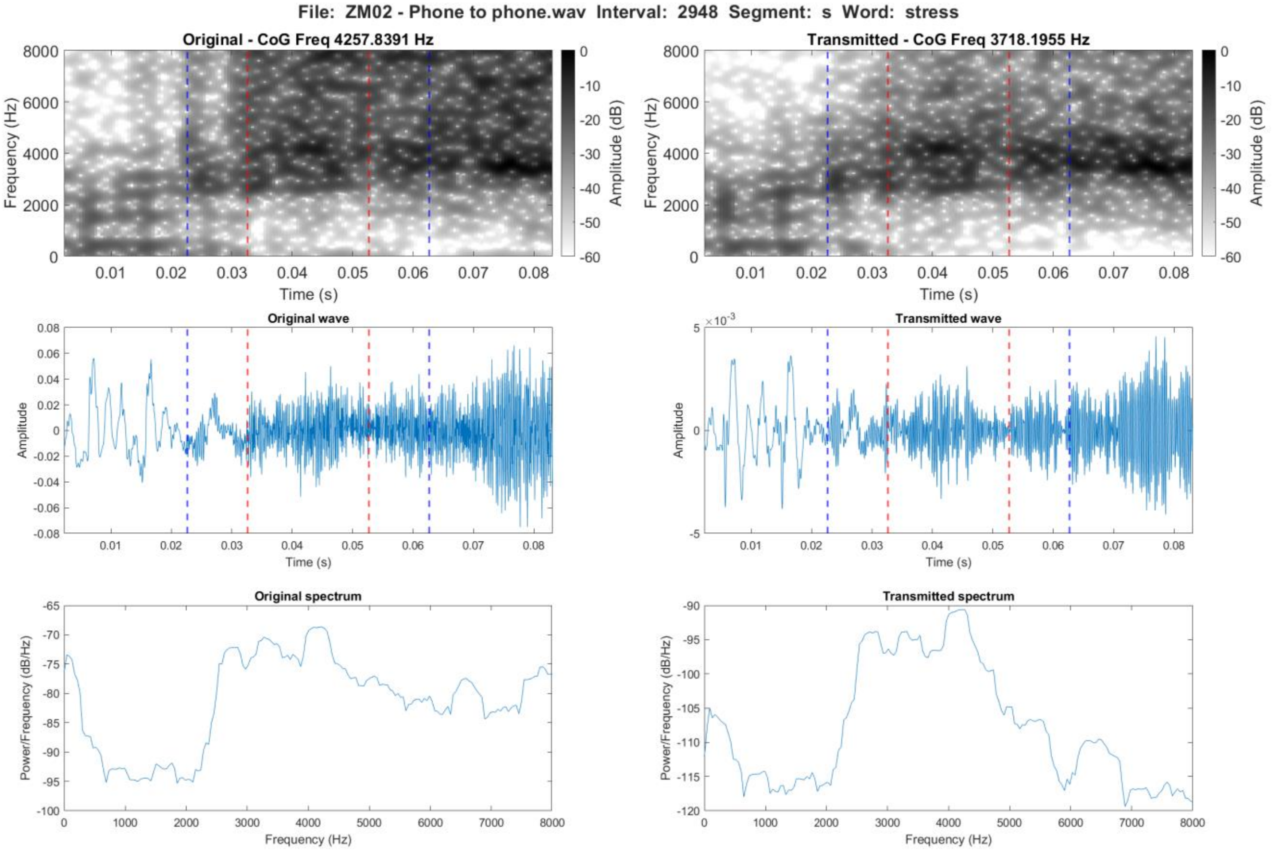


Figure 5.13. Spectrographic comparison of /s/ in the word stress in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.5 /ʃ/

From the distribution plots, slight lowering is observable for CoG, while more clear lowering and centring of values around the mean is visible for SD. For skewness, the effect on the distribution of /ð/ following the transmission is similar to what has previously been observed for CoG and SD. The values are lowered and distributed over a narrower range of frequencies around the mean. For kurtosis, a slightly narrower and lowered distribution can be observed, while only showing minor changes and it maintains the overall shape and place of distribution.

CoG and especially frequency peak showed comparatively smaller effects from what has previously been observed. For CoG, this was a decrease of 8 percent or 260 Hz, while for frequency peak this was a decrease of 2 percent or 58 Hz. The remaining measures all decreased similarly to previous segments. This was by 39 percent or 367 Hz for SD, 51 percent or 1.10 for skewness, and 24 percent or 2.57 for kurtosis.

Similar to previous segments, /ʃ/ is reduced in the frequencies at the higher end of its range i.e. above around 5 kHz, while the most intense frequencies towards its lower range between 3 and 4 kHz are intensified (see figure 5.14). This is also evident from the waveform, where a noticeable increase in amplitude is visible from the baseline to the transmission, while the shape of the spectrum remains unchanged from the WAV baseline.

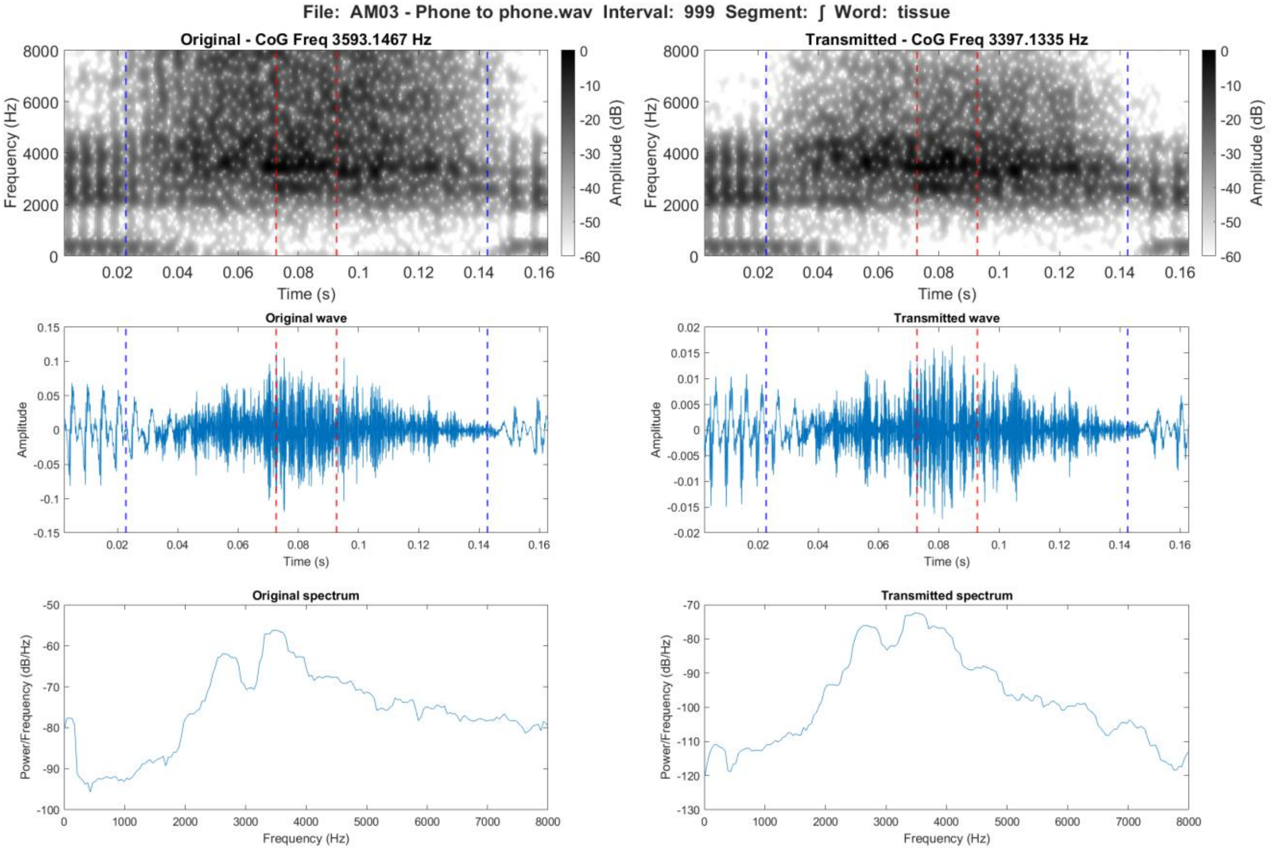


Figure 5.14. Spectrographic comparison of /ʃ/ in the word *tissue* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.6 /ð/

From the distribution plots, CoG and SD show a lowering of the top-most values and an increase in the number of values towards the lower end of the spectrum and around the mean. For skewness, the higher values appear to be centred more clearly around the mean rendering a more narrow distribution, while kurtosis appears to generally be lowered following the transmission. As the frequency peak is already relatively low in comparison to the other fricatives, little to no effect is observable from the violin plots.

More specifically, all mean values for the spectral measures decreased for /ð/ in the Direct condition. This was by between 13 percent and 33 percent, the former being a 149 Hz change for frequency peak and the latter a 3.91 change for kurtosis. CoG decreased almost identically to frequency peak by 14 percent or 298 Hz, while SD decreased by 24 percent or 331 Hz. Skewness presented an effect similar to SD with a decrease of 25 percent or 0.46.

The spectrographic representation shows how in this example /ð/ is fully voiced (figure 5.15). Again, the higher frequency content above 4 kHz is reduced, while the formant structures are intensified in terms of amplitude and present a comparatively more regular pattern across the segment.

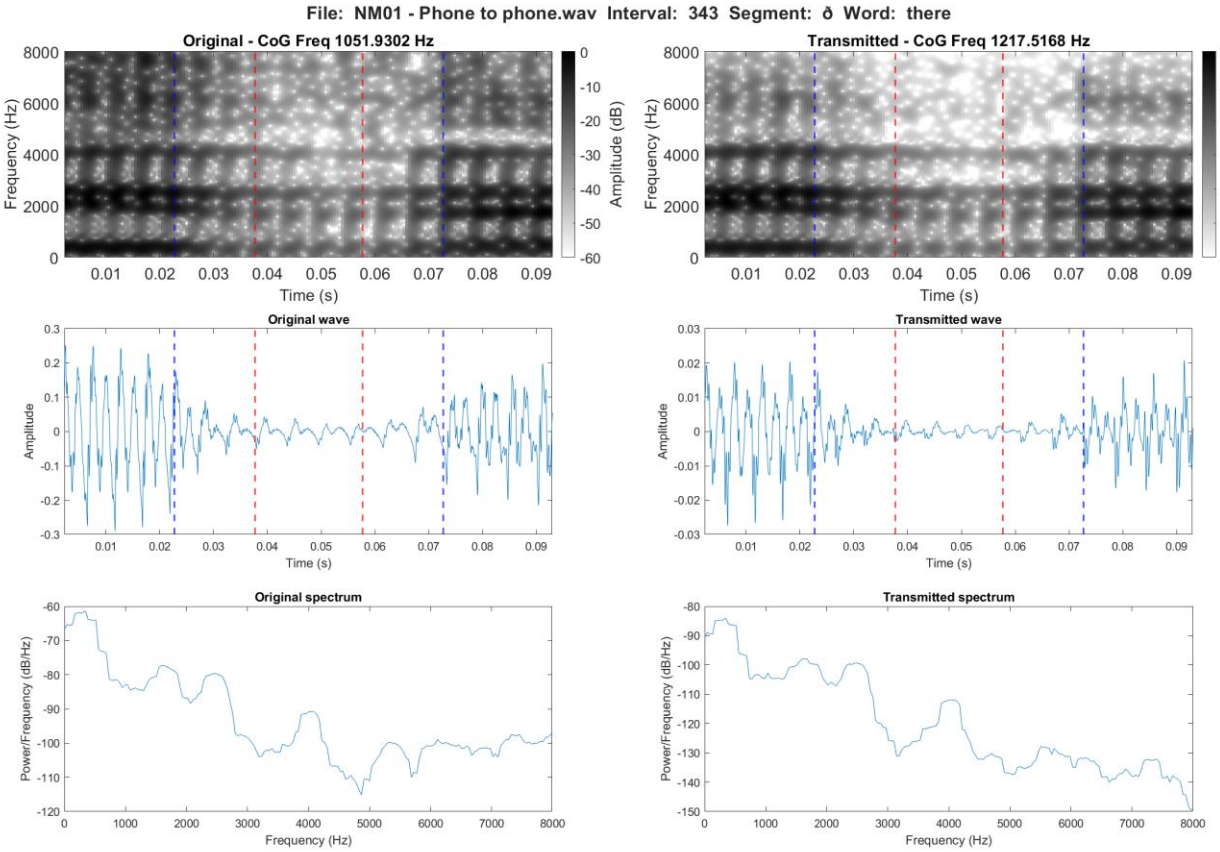


Figure 5.15. Spectrographic comparison of /ð/ in the word *there* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.1.7 /z/

The distribution of /z/, shows lowered values for CoG and especially SD, while frequency peak is also lowered, but the primary effect is observable in the high frequency content. The shape of the distribution values following the transmission largely follows the previously observed patterns, however with more values clearly present at below the mean. For SD, the distribution is affected by the transmission similarly to /s/ rendering a more triangular shape with most values towards the lower half of the spectrum. Skewness likewise appear to lower, while the distribution of the kurtosis values indicates an increase. Lastly, the main effect on the distribution of frequency peak is as with /s/, a centring of values around the mean.

/z/ behaved identically to /s/ in terms of which measures presented decreases and increases, while the magnitude of the changes also behaved similarly apart from kurtosis, where the increase was a little more than half of what was observed for /s/ i.e. 31 percent. Frequency peak presented an identical change in percentages, but a smaller change in absolute terms at 687 Hz. CoG lowered by 19 percent, which is 860 Hz, while SD lowered by 25 percent or 340 Hz. Lastly, skewness showed the largest observed change for any segment or measure. This was a decrease of 246 percent or 0.77. It should be noted that both /z/ and /s/ have negative mean values for skewness already in the Original recording.

Apart from a reduction in the higher frequency content above 4 kHz, little change is observable from the spectrogram in comparison between the baseline WAV and the Direct condition (see figure 5.16). However, from the waveform and spectrum, it is evident that for /z/ the transmission is smoothing both irregularities, while lowering the amplitude.

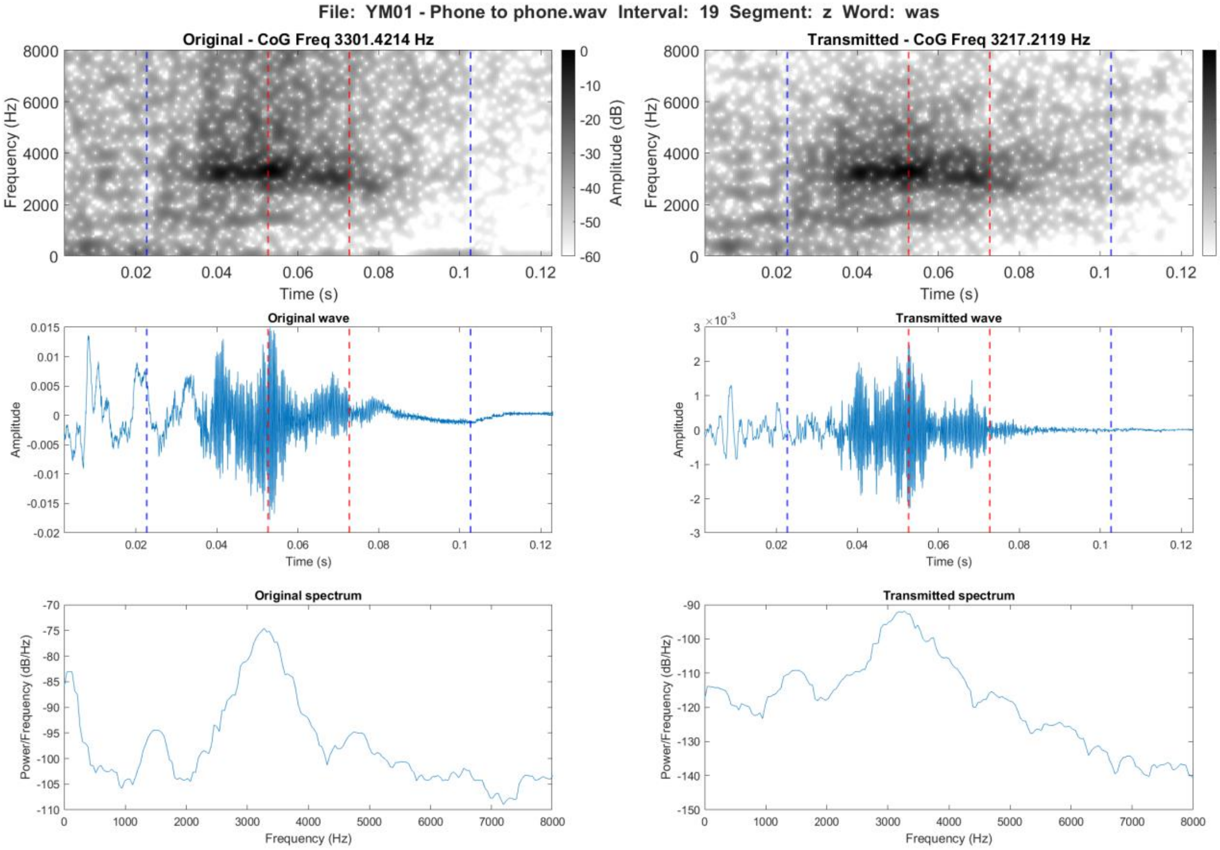


Figure 5.16. Spectrographic comparison of /z/ in the word *was* in the WAV baseline (left) and the live transmission in the Direct condition (right)

5.4.2 No noise condition

This section will present the individual results for each segment and the spectral measures in the comparison between the original WAV and the live transmission in the No noise condition.
First, the trajectory of the mean values for each spectral measure and the individual segments can be found below (figure 5.10 to 5.13). These indicate the directionality of the changes resulting from the transmission.

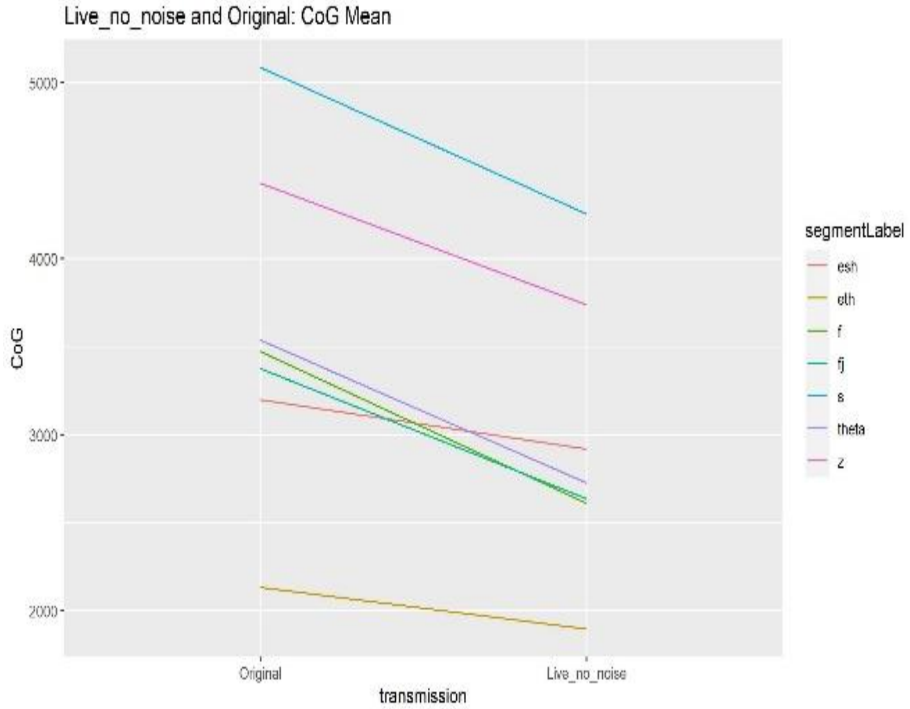


Figure 5.17. Trajectory of mean values for CoG and individual segment in the comparison between the original WAV files and the No noise condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

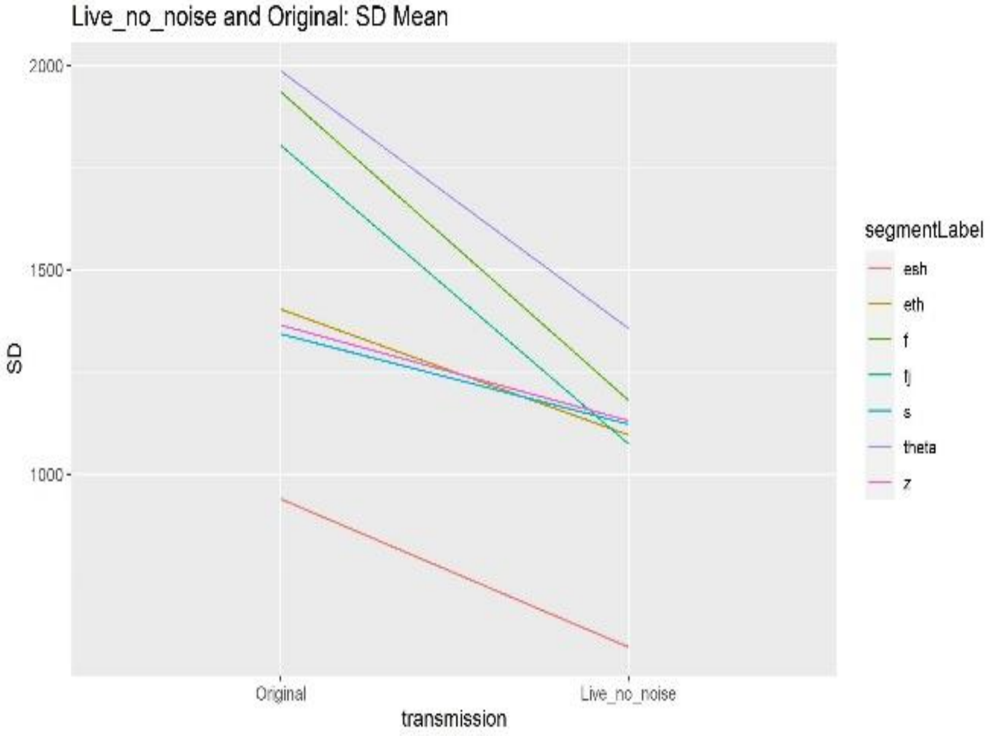


Figure 5.18. Trajectory of mean values for SD and individual segment in the comparison between the original WAV files and the No noise condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

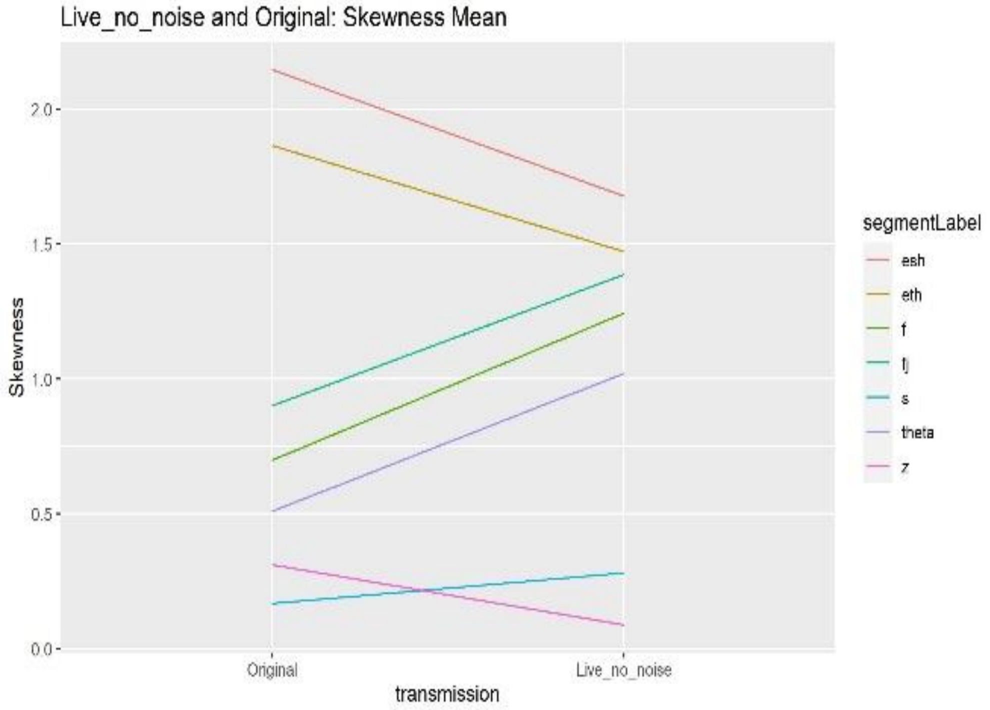


Figure 5.19. Trajectory of mean values for skewness and individual segment in the comparison between the original WAV files and the No noise condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

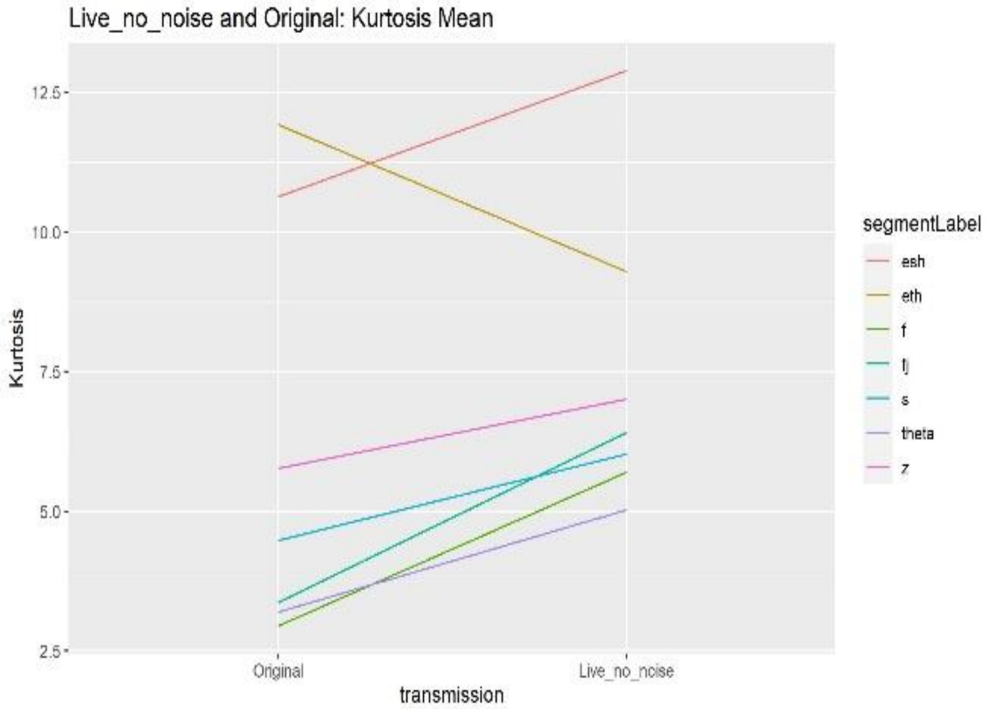


Figure 5.20. Trajectory of mean values for kurtosis and individual segment in the comparison between the original WAV files and the No noise condition.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The trajectory plots indicate that based on the mean values, the No noise condition results in varying tendencies for each segment and spectral measure. For CoG, the mean values consistently decrease. /θ/ and /f/ again become more alike, but otherwise little change in the relation between the segments for this measure can be observed. In terms of SD, /s/ and /z/ become slightly more alike following the transmission, but were already similar in the original WAV files. In addition, [fʲ] becomes almost identical to /ð/ for SD and more similar to /z/ and /s/, while /f/ become more similar to /z/, /s/, and /ð/. [fʲ] and /ð/ become very similar in the No noise condition looking at skewness, while [fʲ], /f/, and /θ/ generally have values closer to /ð/ and /ʃ/ following transmission. Lastly, /ð/ and /ʃ/ become more distinct in terms of kurtosis. A similar, but less pronounced, tendency is observed for /θ/ and /f/, while [fʲ] becomes more similar to /s/.

The distributions are illustrated below in a set of violin plots for each spectral measure (figure 5.14 to 5.18). Again, the specific analysis pertaining to each segment will be found in the following sections.

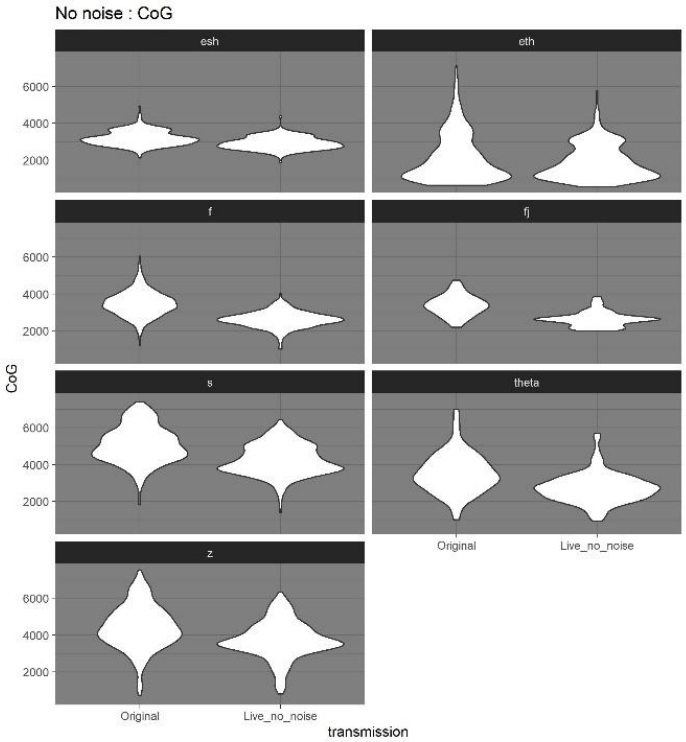


Figure 5.21. Distribution of CoG in Original WAV baseline and the No noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

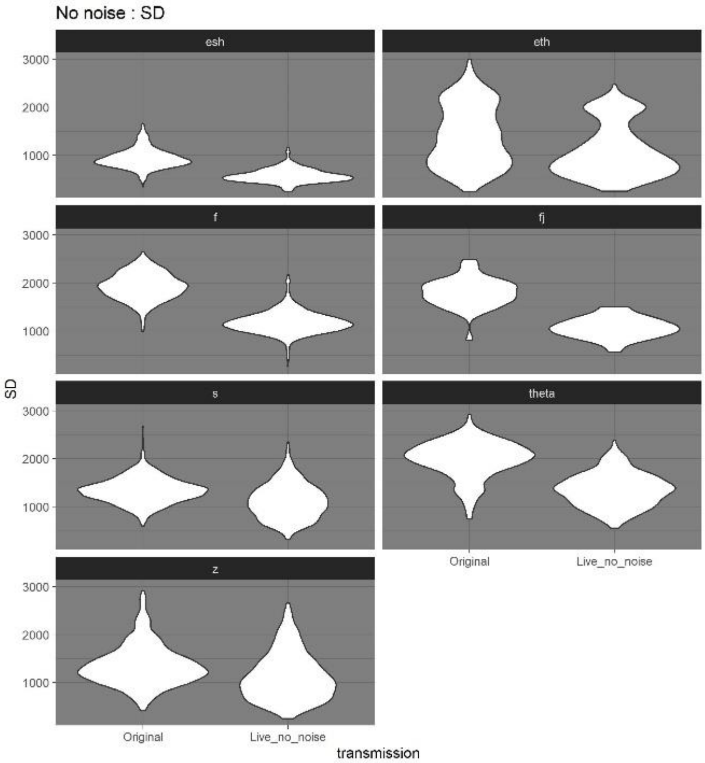


Figure 5.22. Distribution of SD values in Original WAV baseline and the No noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

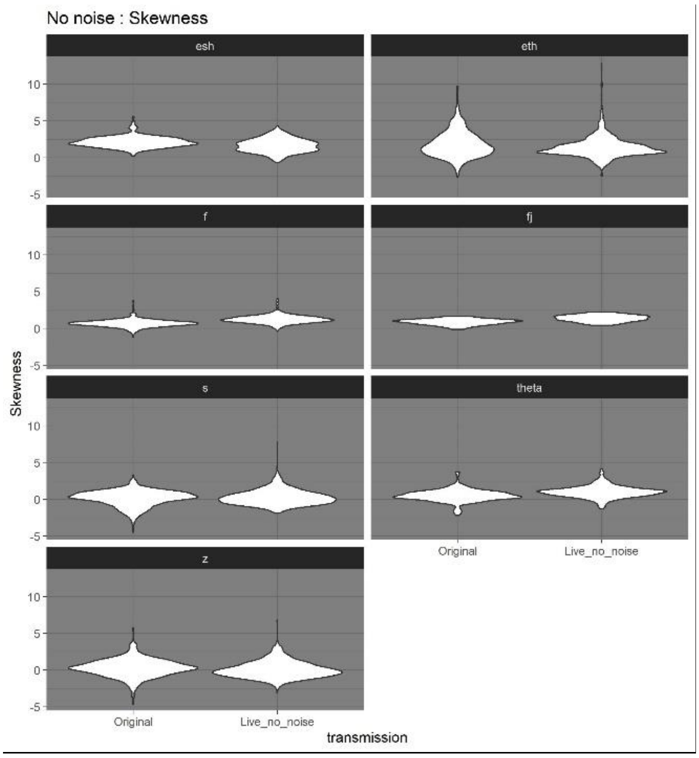


Figure 5.23. Distribution of skewness in Original WAV baseline and the No noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

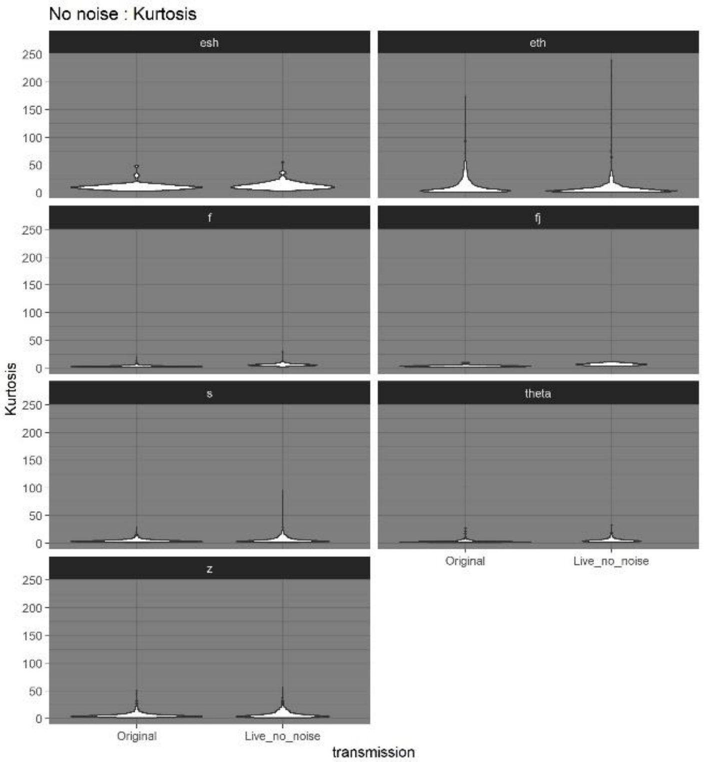


Figure 5.24. Distribution of kurtosis in Original WAV baseline and the No noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

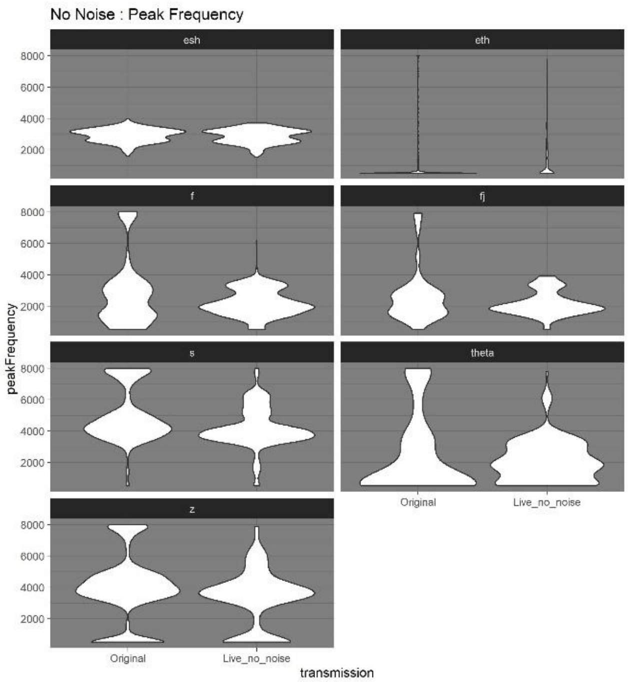


Figure 5.25. Distribution of peak frequency in Original baseline and the No noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [f^{j}], and eth = /ð/.

In sum, the condition without noise lowered all mean values for CoG and SD, while skewness and particularly kurtosis also increased. The distribution plots show how particularly SD was affected by the live transmission in this condition.

5.4.2.1 /f/

From the distribution plots, the overall effect on /f/ for CoG and SD appear similar. Both of these measures appear to lower and have more values centred around the mean following the transmission. For CoG, this means a more narrow distribution, while SD appears to maintain its frequency range. An increase can be observed for both skewness and kurtosis with little change to the overall shape of the distribution for skewness, but a slightly wider range of values for kurtosis. Lastly, a lowered and more narrow distribution can be observed for frequency peak following the transmission.

For /f/ the mean values confirmed the tendencies observed based on the distributions. CoG, SD, and frequency peak again decreased following transmission, while skewness and kurtosis both increased. The effect of the transmission was similar for CoG and frequency peak in terms of percentages with a change of 25 percent or 864 Hz for CoG and a change of 20 percent or 551 Hz for frequency peak.

SD lowered by 39 percent or 755 Hz. The largest effect was found for kurtosis, which increased by 93 percent or 2.75, while skewness increased by 78 percent or 0.54.

The spectrographic representation as well as the waveform reveal a clear reduction across the frequency span of /f/ following the transmission in the No noise condition (see figure 5.26). It is interesting to note how the slight formant structure, which is part of the transition into the following vowel, is enhanced following transmission. The spectrum confirm the decrease in amplitude and reveal clear variation not present in the WAV baseline.

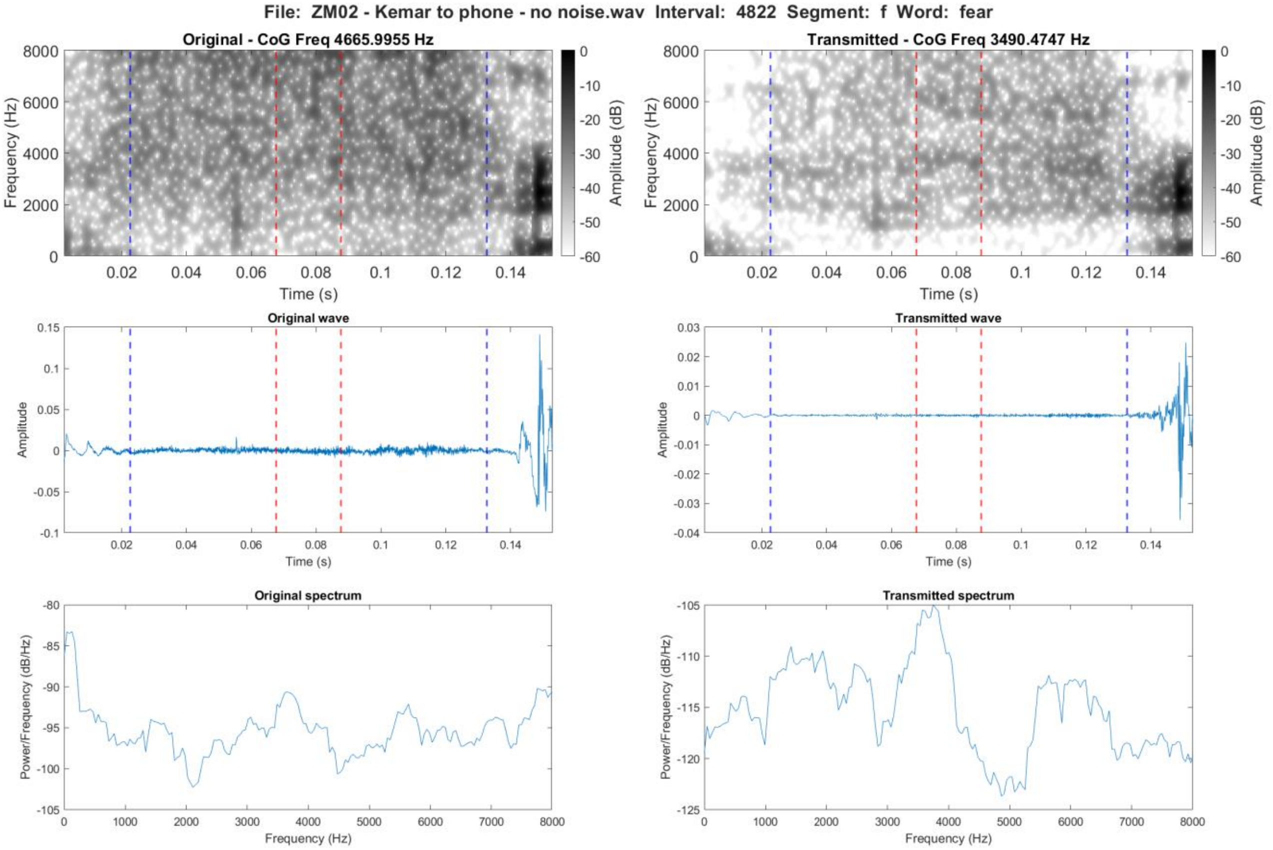


Figure 5.26. Spectrographic comparison of /f/ in the word *fear* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.2 [fʲ]

The distribution plots show how the effect on CoG and SD for [fʲ] are similar to what was observed for /f/. This means a general lowering and a distribution with more values centred around the mean. For both skewness and kurtosis an increase can again be observed, while frequency peak clearly lowers and has more values centred around the mean in a relatively narrower distribution.

As with /f/, skewness and kurtosis increased, while the rest of the spectral measures decreased following the transmission with no noise. CoG decreased by 22 percent or 742 Hz, while frequency peak decreased less by 16 percent or 765 Hz. SD presented a change similar to the Direct condition with a change of 40 percent or 731 Hz. The increase to skewness was 54 percent or 0.49, while it was 90 percent or 3.04 for kurtosis.

From the spectrographic representation, it can be seen how the main effect of the transmission on [fʲ] is to the higher and lower frequency content above and below 3-4 kHz, while the formant structure is again made more prominent by the transmission (see figure 5.27). This is in contrast to /f/, where the reduction was more general. It is worth noting here that [fʲ] is in medial position, while the example of /f/ above was in initial position. As previously observed, the irregularities typical for fricatives and evident as variations in amplitude is evened out by the transmission and the waveform in turn appear more regular. From the spectrum a smoothing is observable particularly from around 4.5 kHz.

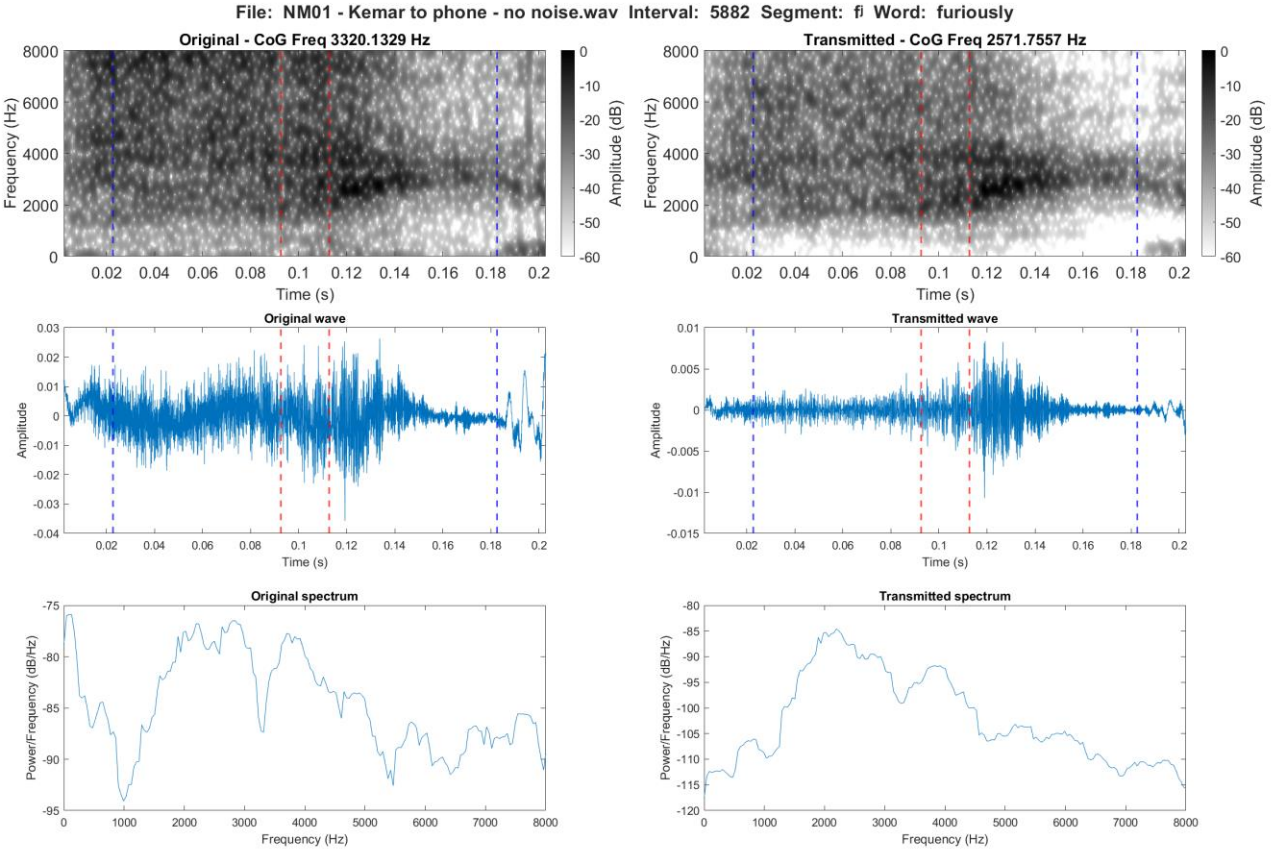


Figure 5.27. Spectrographic comparison of [fʲ] in the word *furiously* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.3 /θ/

The distributions of the spectral measures for /θ/ shows visible changes following the No noise transmission. For CoG and SD this can be seen as clear lowering and for CoG specifically, centring of values around the mean. SD shows a slightly different pattern with more values present below the mean. For skewness an increase can be observed with little change to the overall shape of the distribution, while kurtosis also appear to slightly increase and has its values distributed over a slightly larger span of values following transmission. For frequency peak, both the values above and below the mean are affected by the transmission, which appear as an overall lowering, but with fewer values at the lowest frequencies present in the original recording.

The increases observed for skewness and kurtosis were also visible from the changes to the mean values, while the remaining spectral measures all decreased. For CoG this was by 23 percent or 812 Hz, while it was a little less, namely 14 percent or 352 Hz for frequency peak. SD lowered by 32 percent or 630 Hz. The largest effect was found for skewness, which increased by 100 percent or 0.51, while kurtosis increased by 57 percent or 1.83.

Similar to /f/, a clear reduction across frequencies can be observed from both the spectrographic representation and waveform (see figure 5.28). These reductions are substantial and almost eliminate the frication at the point of measurement. From the waveform, it can be seen how the initial part is clearly irregular in the original waveform, but appears as a regular soundwave. Thus, what in the original waveform appear as part of the fricative is given more voiced characteristics following transmission and made more similar to the preceding vowel. From the spectrum some variation is introduced following the codec transmission, particularly evident as a drop around 4 kHz.

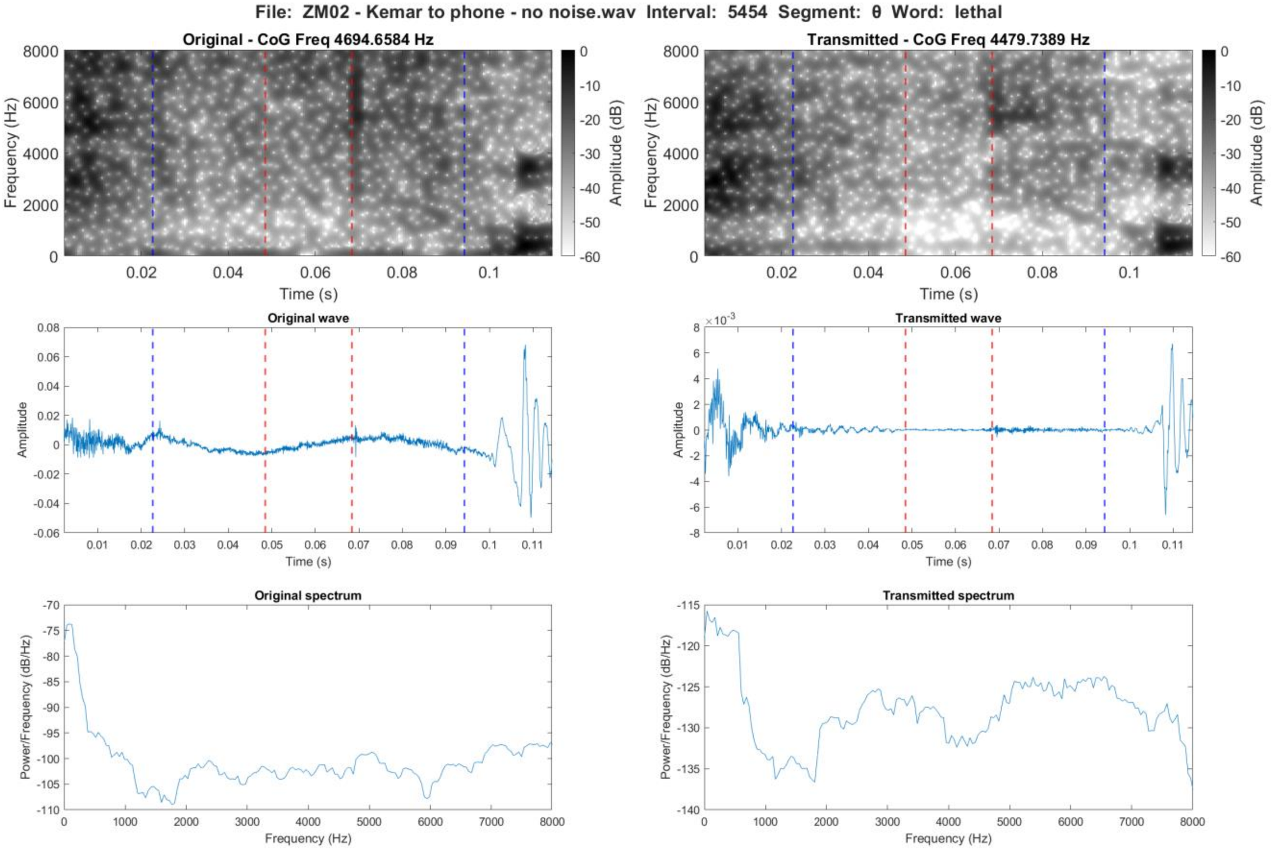


Figure 5.28. Spectrographic comparison of /θ/ in the word *lethal* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.4 /s/

The distribution plots illustrate how CoG is lowered for /s/ with more values centred around the mean. For SD the effect of the transmission is similar to what was observed for /θ/ with an overall lowering and a more even distribution of values i.e. fewer values clearly centred around the mean. For skewness, the lower values appear to increase for /s/ following transmission, while the overall distribution is slightly lowered. An increase can also be observed for kurtosis. For frequency peak, a lowering as well as the tendency for more values to be centred around the mean can again be observed. Despite this, more values can also be seen to occur below the mean.

For the mean values, while one of the largest effects and decreases was found for skewness in the Direct condition, the condition without noise presented an increase of skewness of 68 percent or 0.11 for /s/. Kurtosis also increased, this was by 35 percent or 1.55. For both CoG and SD a decrease of 16 percent could be observed, which was a change of 833 Hz for CoG and 220 Hz for SD. Frequency peak followed this trend with a decrease of 15 percent or 756 Hz.

For /s/, the spectrographic representation (Figure 5.29) as well as waveform present an overall increase in intensity and frequency information especially in the lower frequencies following the no noise transmission. The spectrum is relatively more smooth following the transmission.

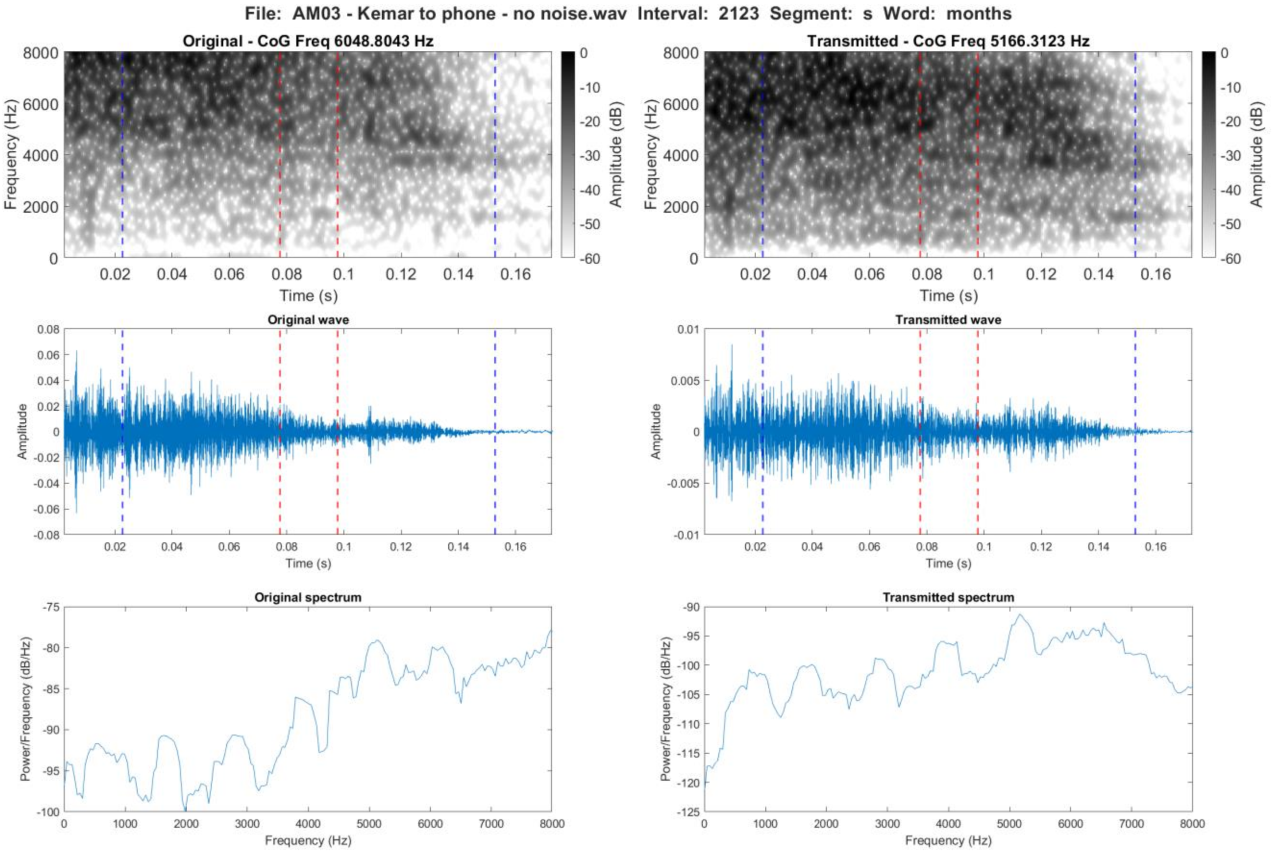


Figure 5.29. Spectrographic comparison of /s/ in the word *months* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.5 /ʃ/

For /ʃ/ the distribution of CoG lowers, but only with minor changes to the overall shape of the distribution. SD on the other hand also lowers, but here the distribution is as previously observed more centred around the mean and appears to have a range over a narrower span of frequencies. The opposite effect can be observed for skewness, where the values are less centred around the mean, but appears more equally distributed across a similar value span. Nevertheless, skewness still appear to be lowered. Kurtosis presents an increase, which is again with the values more evenly distributed. Lastly, no substantial changes are observable for frequency peak.

As previously observed, /ʃ/ presented minor effects to CoG and frequency peak in the transmission without noise. This was a decrease of 9 percent or 282 Hz for CoG and 2 percent or 57 Hz for frequency peak. An almost identical effect to the Direct condition was observed for SD, which lowered by 38 percent or 361 Hz. Skewness also decreased, which was by 22 percent or 0.47, while kurtosis was the only measure increasing in this condition. This was by 21 percent or 2.25.

The effects on /ʃ/ based on the spectrographic representation (see figure 5.30) is more limited in comparison with the previous segments. The main effect appears on the high and low frequency content, where the information below the first formant like structure has not been transmitted. The waveform presents a smaller reduction in amplitude, while the spectrum appear similar between the WAV baseline and the transmission.

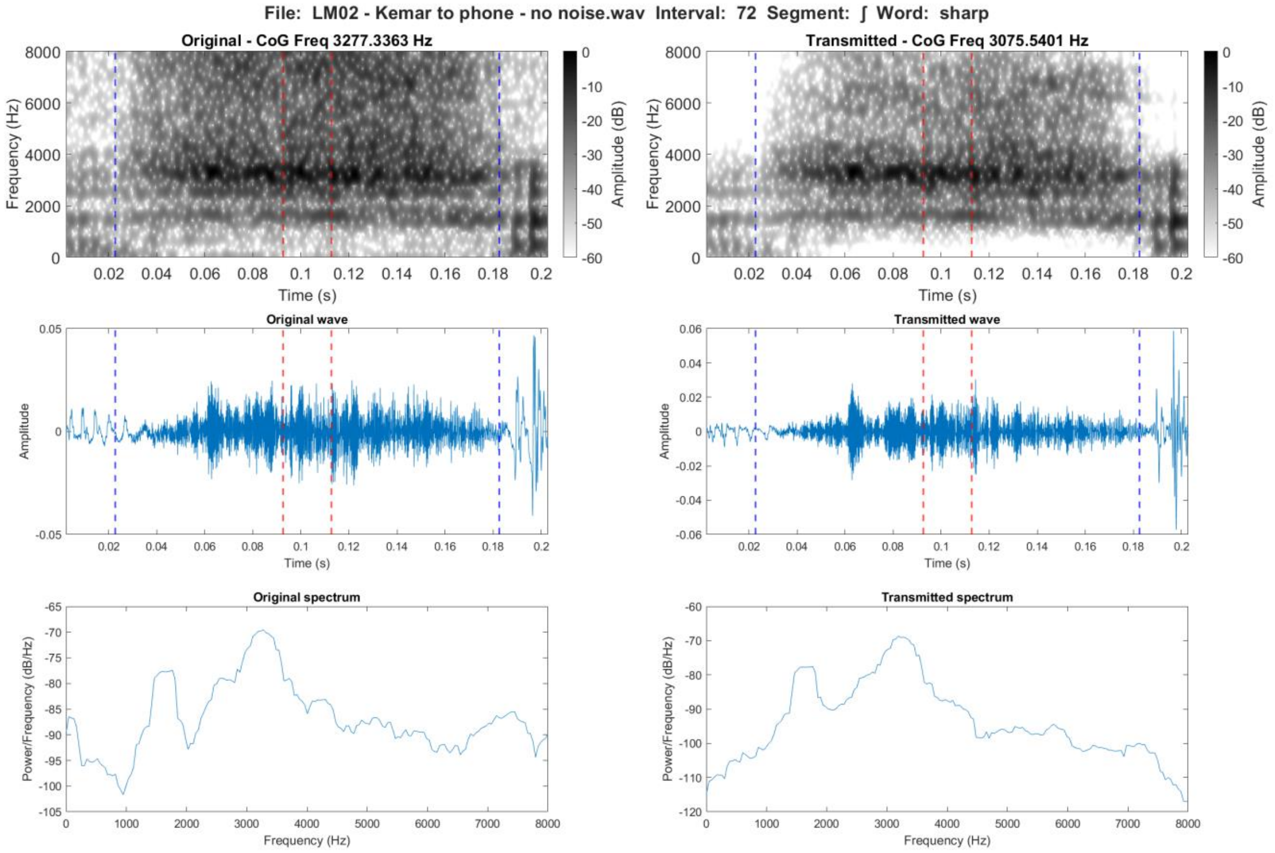


Figure 5.30. Spectrographic comparison of /ʃ/ in the word *sharp* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.6 /ð/

From the distribution plots, a lowering and increase in values towards the lower half of the spectrum is evident for CoG and especially SD. For both skewness and kurtosis an increase in outliers in the higher frequencies are visible. In addition, skewness and kurtosis present an increase in values around the mean, which also means a slightly narrower distribution of the main body of values. Finally, frequency peak show little to no visible effect of the transmission in the violin plots.

For the mean values, all spectral measures decreased for /ð/ following transmission. For SD, skewness, and kurtosis this was by 21 to 22 percent, which specifically meant a change of 309 Hz for SD, 0.39 for skewness, and 2.63 for kurtosis. CoG and frequency peak decreased by 11 and 10 percent respectively, which was a change of 235 Hz for CoG and 112 Hz for frequency peak.

For /ð/ the spectrographic representation shows a substantial reduction, which for the central frame is close to a non-transmission (figure 5.31). The main part of the frequency information maintained from the WAV baseline is in the final frame, where the formant structure in the transition to the following vowel is enhanced. Moreover, the fricative lost some of its acoustically plosive like structure in the no noise transmission. The waveform confirms this observation, where the variation and frication observed in the WAV baseline is clearly reduced following transmission. The spectrum appear with more variation following the transmission.

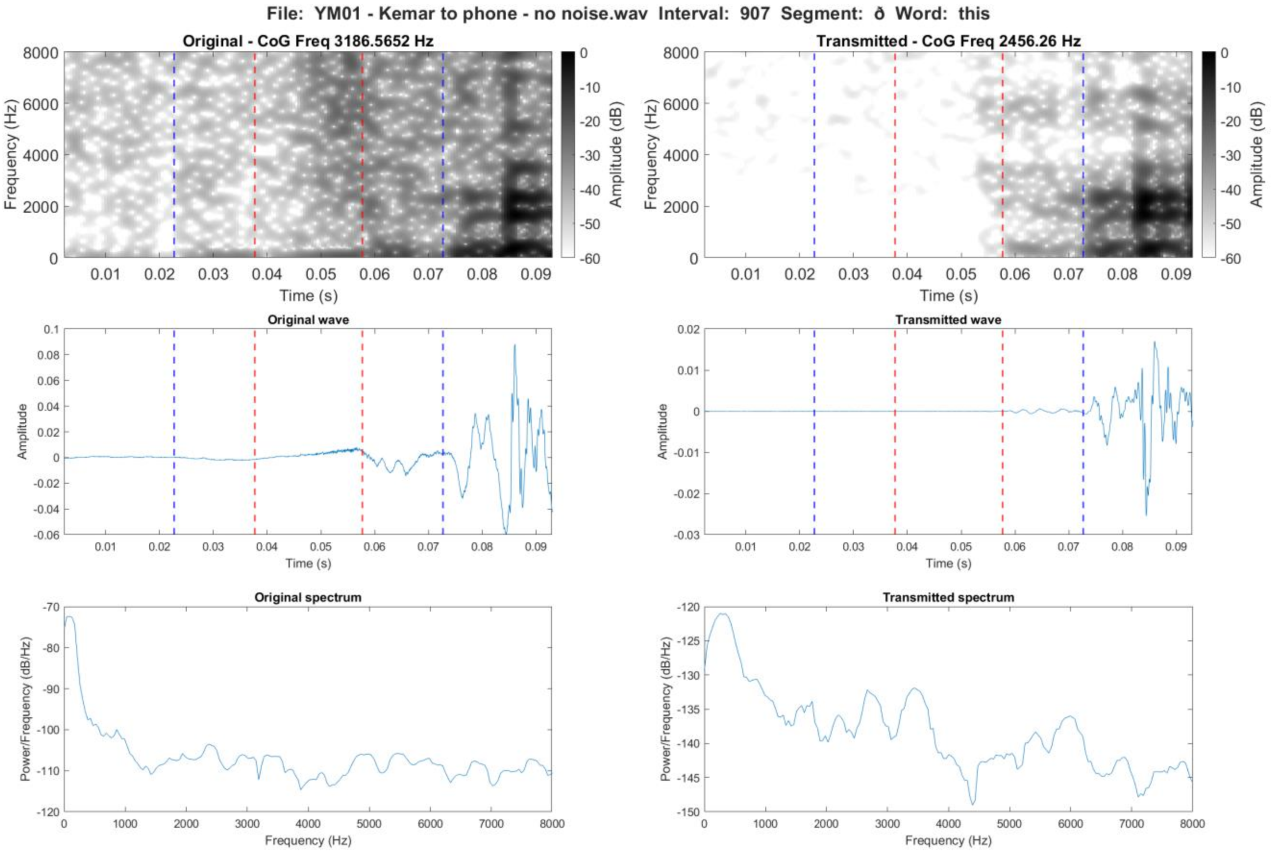


Figure 5.31. Spectrographic comparison of /ð/ in the word *this* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.2.7 /z/

The effects on the distribution of /z/ are generally very similar to what was observed for /s/. This means that CoG and SD appear to be lowered following the transmission, while CoG has more values centred around the mean and for SD the values are more evenly distributed. For skewness, a slight lowering can be observed, while kurtosis as with /s/, can be observed to increase. The effect on frequency peak is also very similar to /s/ with a lowering and centring of values.

This can also be observed for the mean values, where /z/ again behaved similarly to /s/ in this condition. This meant a decrease of CoG, SD and frequency peak by 16 and 17 percent, which was a change of 691 Hz for CoG, 235 Hz for SD, and 647 Hz for frequency peak. The main difference between the two was in terms of skewness, where /z/ presented a decrease of 71 percent or 0.22. For kurtosis, /z/ increased by 21 percent or 1.24.

In contrast to /s/, the spectrographic representation shows an overall decrease in intensity mainly for the high frequency content (See figure 5.32). The waveform shows a slightly reduced frication and variation in amplitude following the transmission, while the shape of the spectrum is overall similar to the WAV baseline.

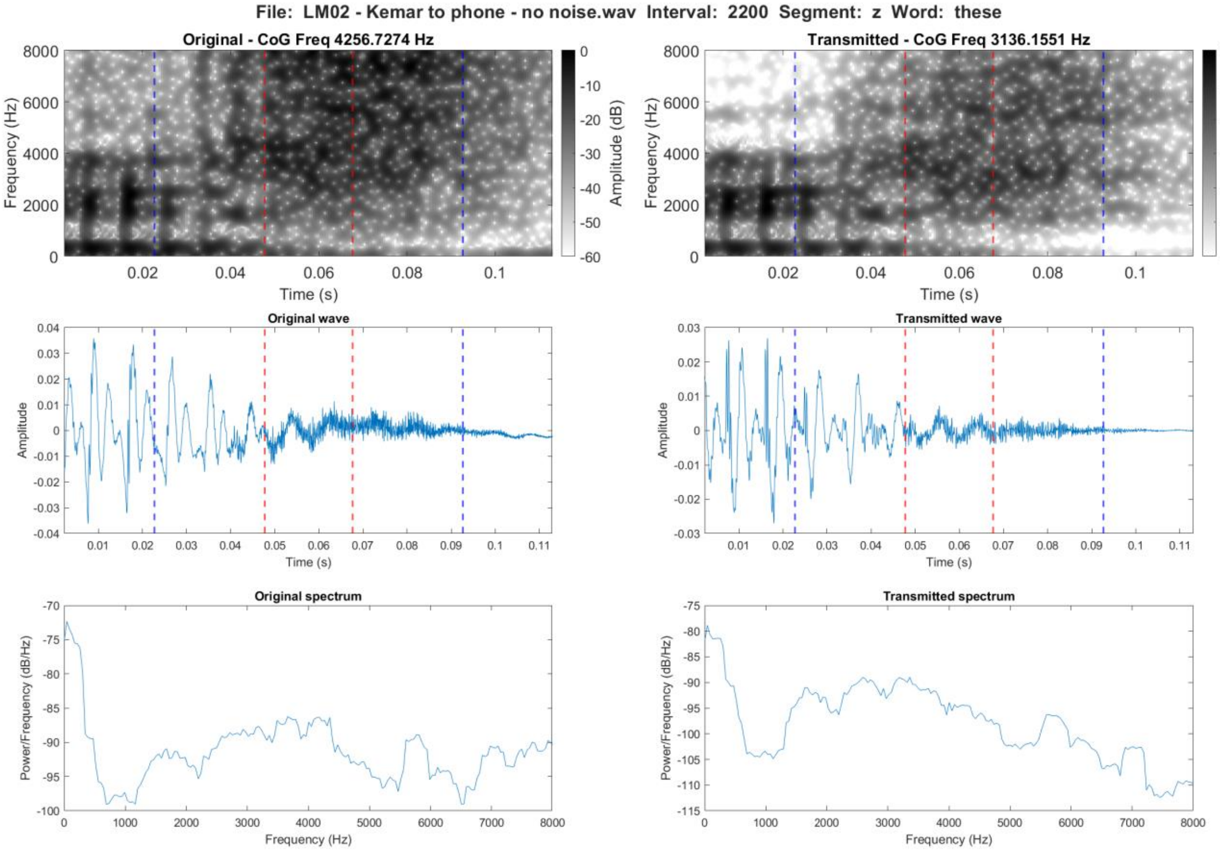


Figure 5.32. Spectrographic comparison of voiced /z/ in the word *these* in the WAV baseline (left) and the live transmission in the No noise condition (right)

5.4.3 With noise condition

This section will present the individual results for each segment and the spectral measures in the comparison between the original WAV and the live transmission in the With noise condition.

First, the trajectory of the mean values for each spectral measure and the individual segments can be found below (figure 5.19 to 5.22). These indicate the directionality of the changes caused by the transmission.

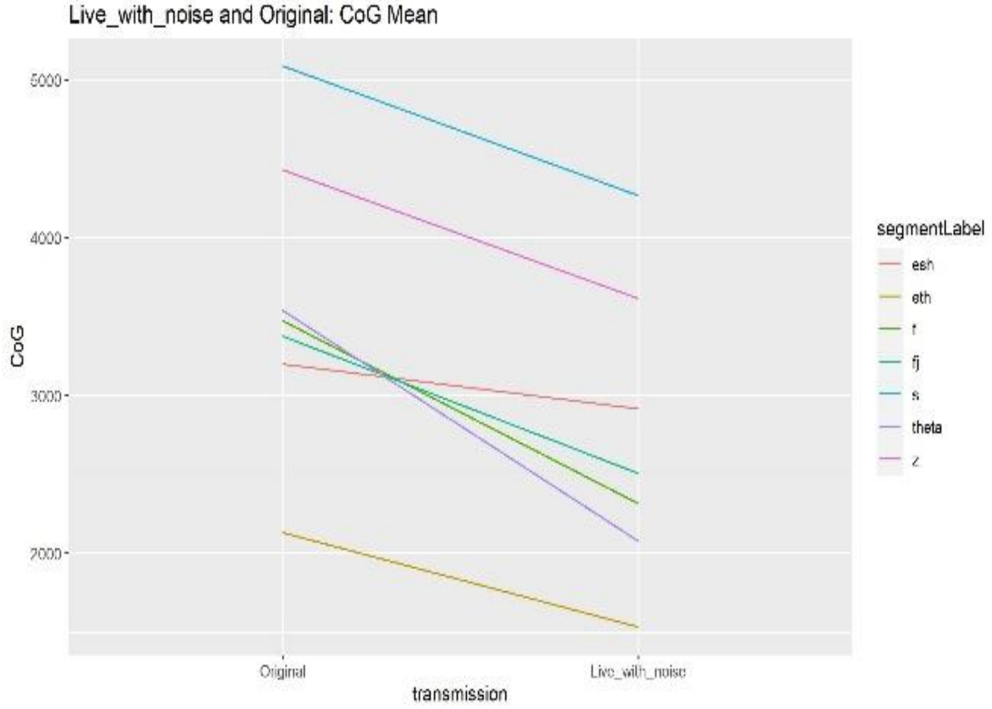


Figure 5.33. Trajectory of mean values for CoG and individual segment in the comparison between the original WAV files and the With noise condition.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

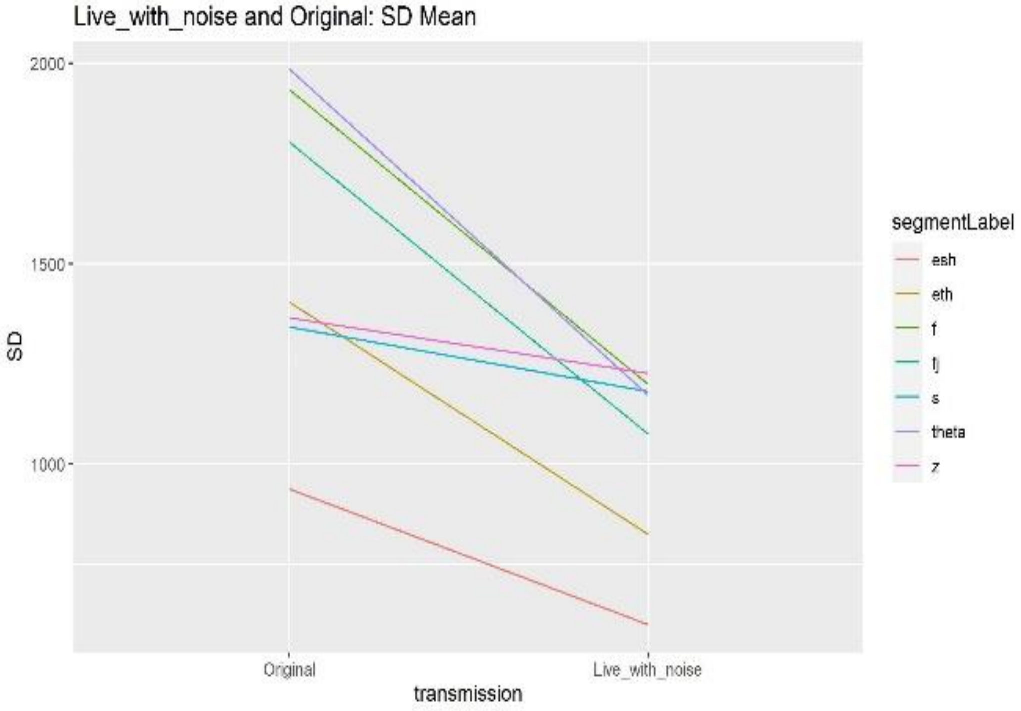


Figure 5.34. Trajectory of mean values for SD and individual segment in the comparison between the original WAV files and the With noise condition.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

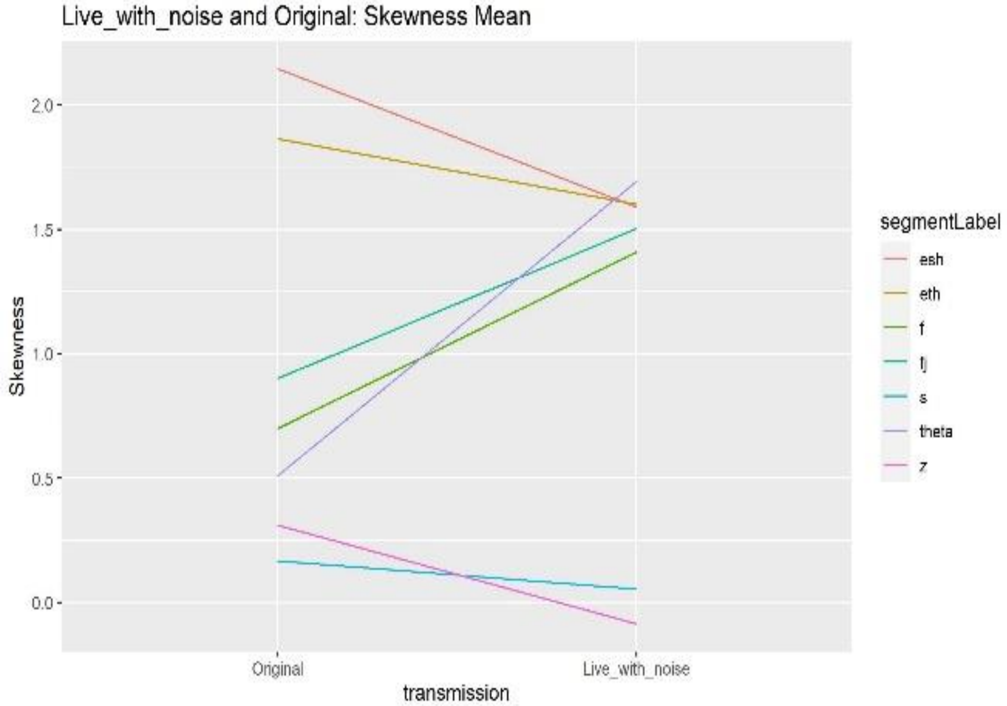


Figure 5.35. Trajectory of mean values for skewness and individual segment in the comparison between the original WAV files and the With noise condition.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

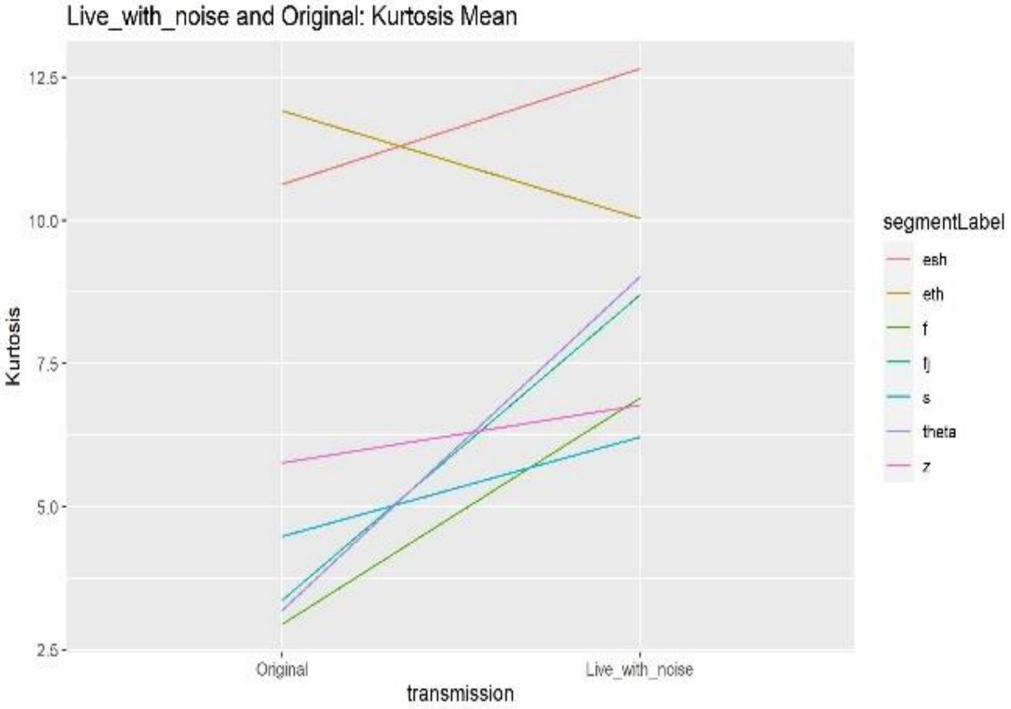


Figure 5.36. Trajectory of mean values for kurtosis and individual segment in the comparison between the original WAV files and the With noise condition.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

The trajectory plots for the mean values indicate how none of the segments become more similar following the transmission with background noise. On the other hand, /f/, [fʲ], and /θ/ appear to become slightly more distinct. For SD, /f/ and /θ/ behave in a similar manner and both become almost identical to /s/ and closer to /z/. In addition, a very slight tendency for /f/ and /θ/ to become more similar can be observed for SD. Skewness again shows a varied pattern, where /ʃ/ and /ð/ become almost identical following transmission, while /f/, [fʲ], and /θ/ become more like /ð/ and /ʃ/. Lastly, for kurtosis /ʃ/ increases, while for /ð/ it decreases and the two become clearly more distinct for this measure. Similarly, /θ/ and [fʲ] become clearly more distinct from /f/, while /f/ become almost identical to /z/.

The distributions are illustrated below in a set of violin plots for each spectral measure (figure 5.23 to 5.27). Again, the specific analysis pertaining to each segment will be found in the following sections.

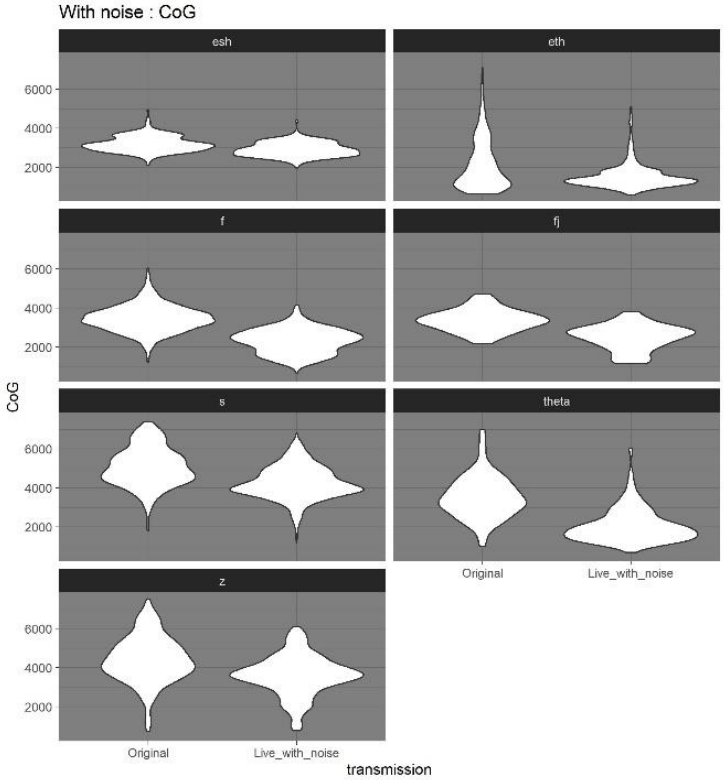


Figure 5.37. Distribution of CoG in Original WAV baseline and the With noise condition grouped by spectral measure and divided by individual segments.
The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

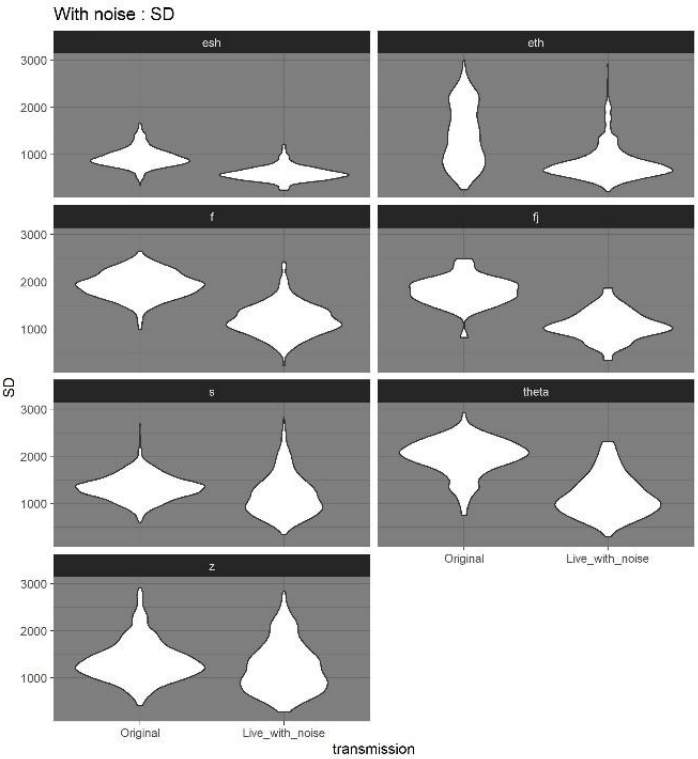


Figure 5.38. Distribution of SD in Original WAV baseline and the With noise condition grouped by spectral measure and divided by individual segments.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

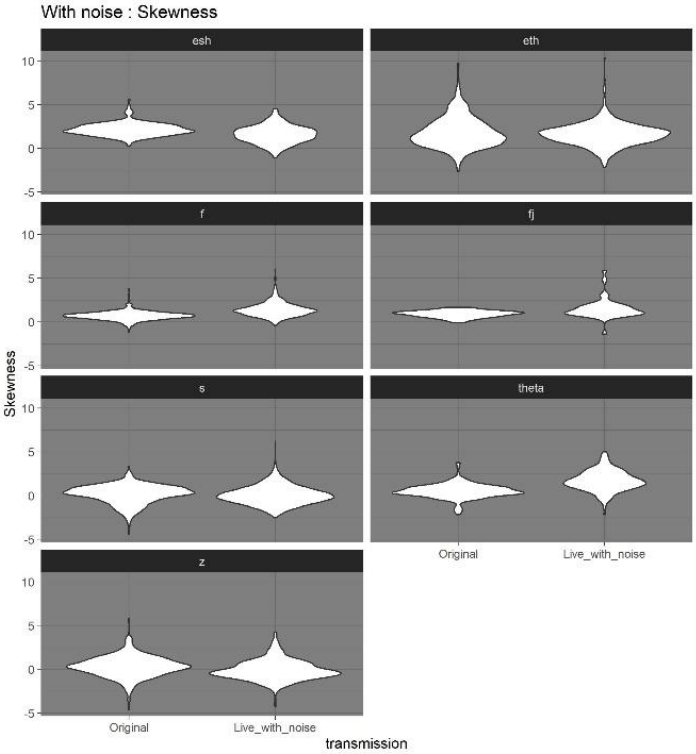


Figure 5.39. Distribution of skewness in Original WAV baseline and the With noise condition grouped by spectral measure and divided by individual segments.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

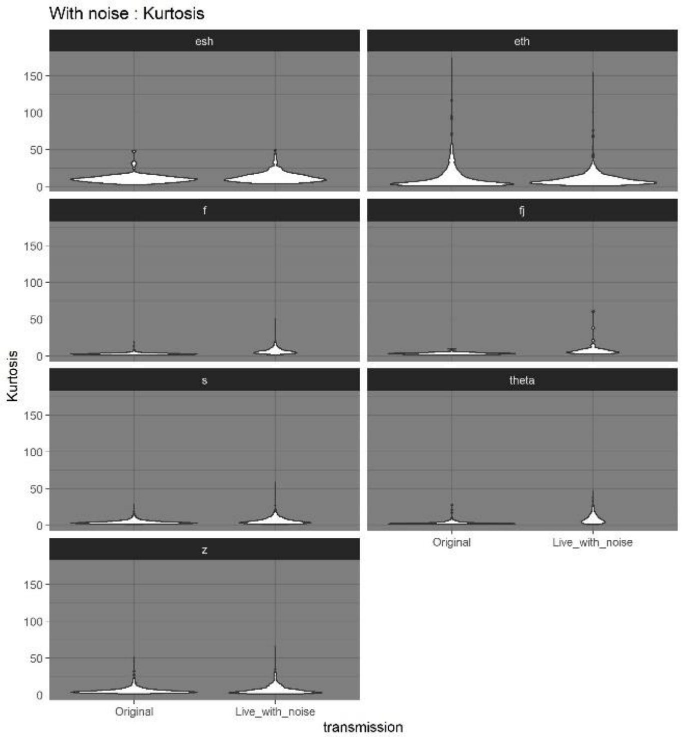


Figure 5.40. Distribution of kurtosis in Original WAV baseline and the With noise condition grouped by spectral measure and divided by individual segments.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

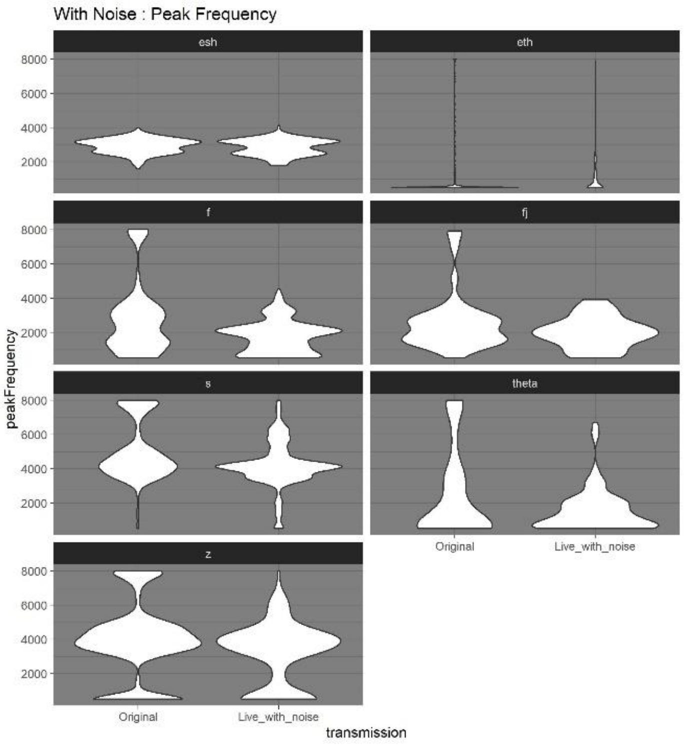


Figure 5.41. Distribution of peak frequency in Original WAV baseline and the With noise condition grouped by spectral measure and divided by individual segments.

The symbols are interpreted as follows Theta = /θ/, esh = /ʃ/, fj = [fʲ], and eth = /ð/.

5.4.3.1 /f/

The distribution plots reveal a lowering of CoG and SD for /f/ with more values below the mean. In contrast, skewness and kurtosis increase and present a distribution following transmission with more evenly distributed values. For frequency peak, a clear lowering is observable with more values centred around the mean as well as towards the lower extreme of the present values.

For /f/, the effects on mean values observed for CoG, SD and frequency peak were similar with decreases between 33 percent and 38 percent. For CoG this was a change of 1,158 Hz, for SD a change of 736 Hz, and for frequency peak a change of 967 Hz. Skewness and kurtosis both increased by over 100 percent following transmission. For skewness, this was a change of 102 percent or 0.72 and for kurtosis, a change of 134 percent or 3.95.

The spectrographic representation as well as the waveform present clear changes to /f/ following the live transmission with noise (see figure 4.42). The acoustic variation and frication identifying /f/ in the WAV baseline are for the most part not transmitted, but replaced with what appear as more general background noise. In addition, the spectrogram shows some low frequency information, not present in the original WAV file, is inserted in the transmission as an extension of the previous vowel. The spectrum is generally smoothed and presents less of the variation found in the WAV baseline.

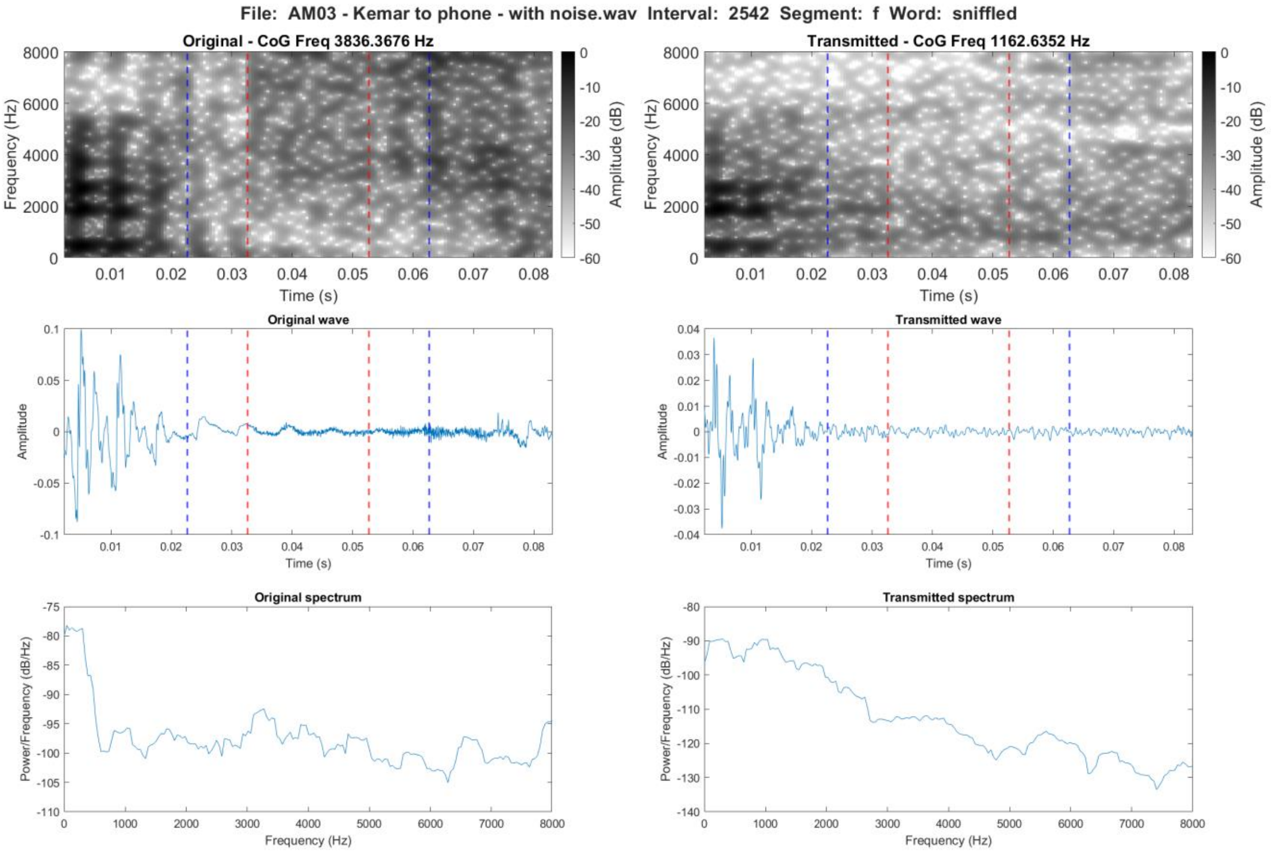


Figure 5.42. Spectrographic comparison of /f/ in the word *sniffled* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.2 [fʲ]

Again [fʲ] behaves similarly to /f/ in terms of the distribution of spectral measures in the With noise condition. Thus, CoG is lowered with more values below the mean, while SD is also lowered, but has more values centred around the mean in comparison to the original recording. Both skewness and kurtosis are increased and distributed over a wider range of values. Lastly, the frequency peak values are lowered with a narrower distribution with more values centred around the mean.

As in the No noise condition, the mean values for skewness and kurtosis increased, while the rest of the spectral measures decreased following transmission. Almost identically to the No noise condition, SD decreased by 40 percent or 728 Hz. CoG decreased by 26 percent or 872 Hz, while frequency peak decreased by 23 percent or 593 Hz. The increase of skewness was by 67 percent or 0.60, while kurtosis increased by 159 percent or 5.34.

The spectrographic representation reveals how the main reduction in energy for [fʲ] following transmission is to the initial part of the segment as well as the frequency content above and below the most intense mid frequencies (see figure 5.43). These relatively more intense frequencies have a slightly clearer formant structure in the transmitted file. This can also be seen from the waveform, where this part of the segment is spread over a wider amplitude range.

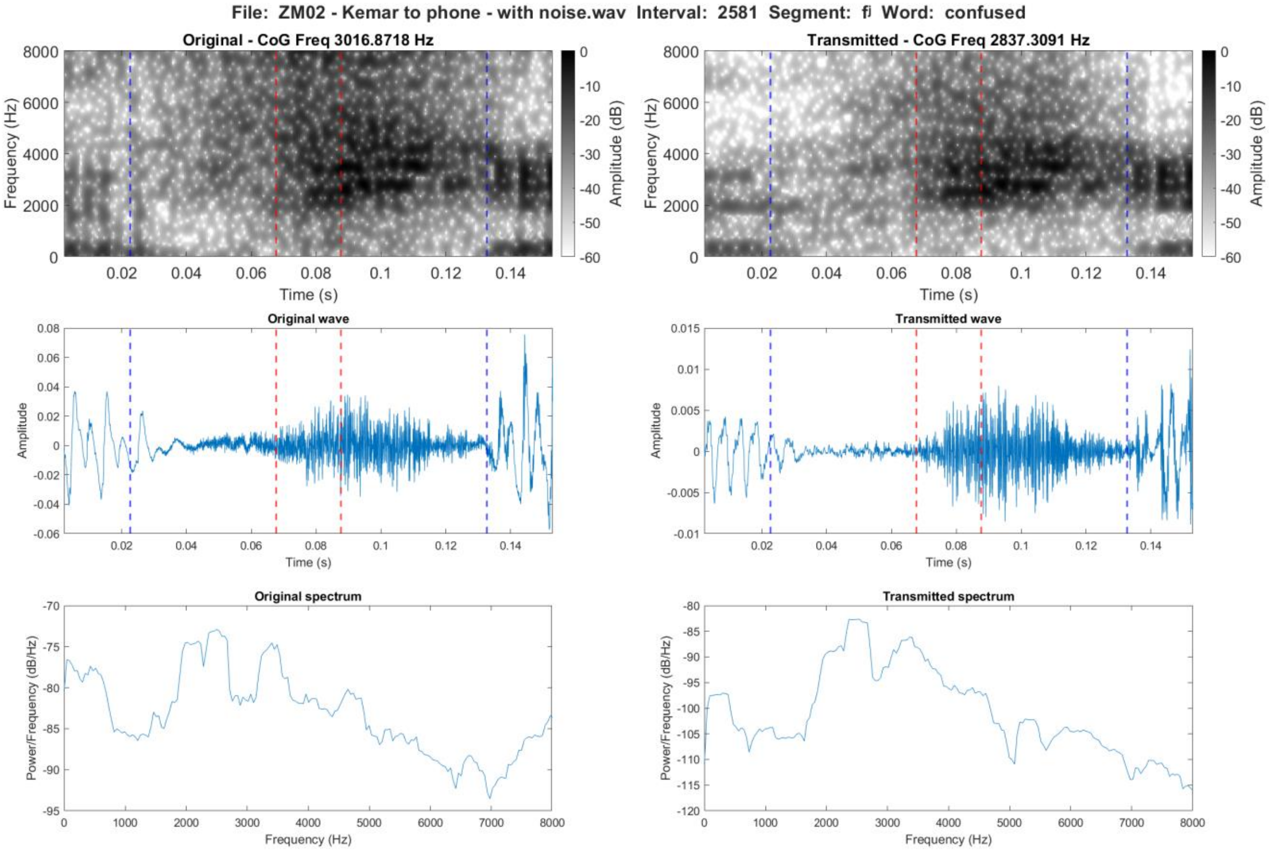


Figure 5.43. Spectrographic comparison of [fʲ] in the word *confused* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.3 /θ/

From the distribution plots, /θ/ shows clear lowering of both CoG and SD in similar ways to what was observed in the previous conditions including more values centred around the mean for CoG, and values more evenly distributed for SD. Both skewness and kurtosis present increases, and again a more even distribution of values less centred around the mean. Finally, frequency peak is clearly lowered and has a greater amount of values present towards the lower extreme of its frequency range.

The mean values showed how for both CoG and SD, the transmission resulted in decreases of 41 percent. This was for CoG a decrease of 1,464 Hz and for SD a decrease of 814 Hz. Frequency peak also decreased following transmission. This was similar to /f/, as /θ/ decreased by 35 percent or 858 Hz. Skewness and kurtosis both increased. For skewness, this was the second largest effect observed across all measures, conditions and segments. Specifically, skewness increased by 232 percent or 1.18, while kurtosis increased by 183 percent or 5.83. In comparison to the segments previously presenting these relatively large effects, both skewness and kurtosis had positive values in both the original recording and With noise condition.

Similar to /f/, /θ/ is generally reduced in intensity with the primary frequency information maintained following the transmission, which is related to the transition from the preceding and following vowel (see figure 5.44). The waveform revealed that the amplitude was reduced in the comparison between the WAV baseline and the transmitted file. The same can be observed from the spectrum, where clear variation also appear, which cannot be observed in the WAV baseline.

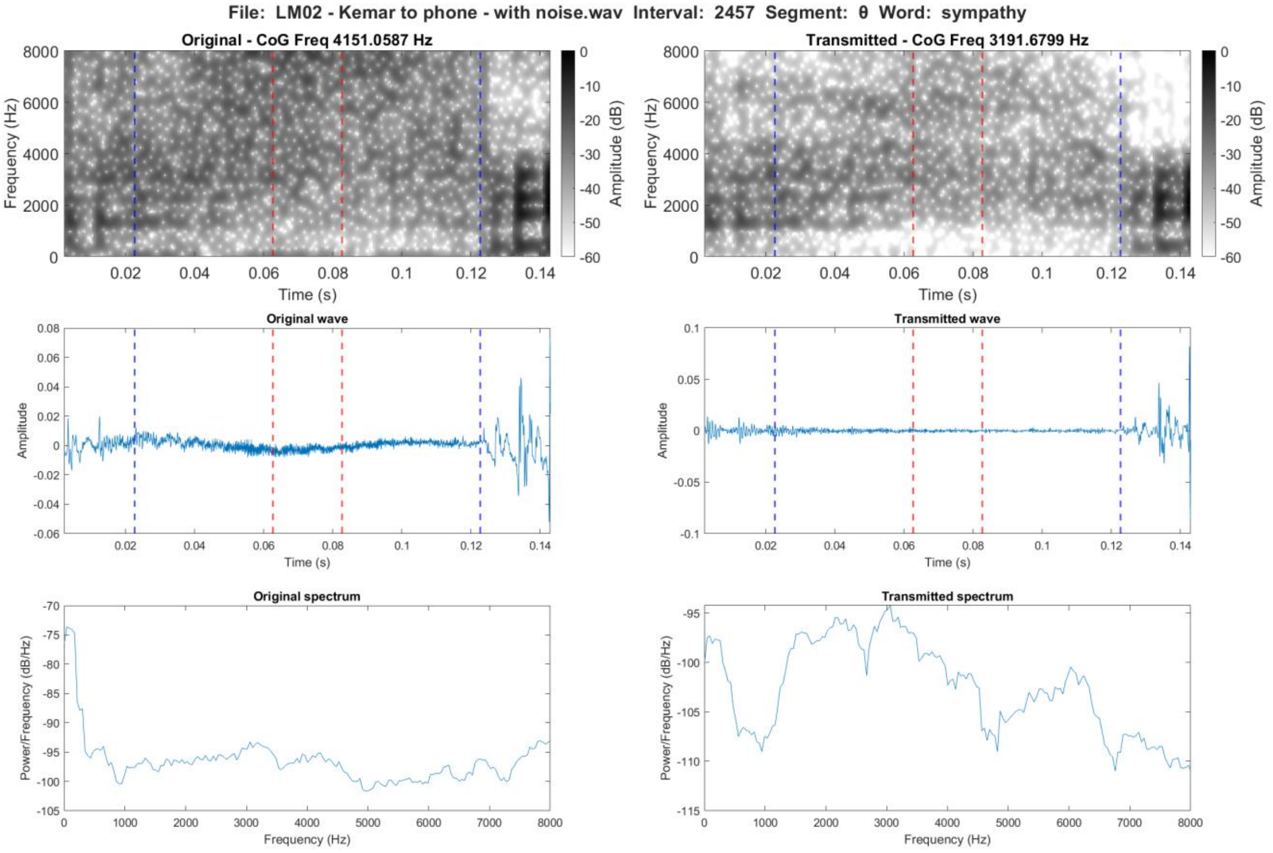


Figure 5.44. Spectrographic comparison of /θ/ in the word *sympathy* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.4 /s/

The distributions of the CoG and SD for /s/ are both lowered following the transmission with noise. This is in a similar way to what has previously been observed. Thus, CoG is lowered and has more values centred around the mean, while SD is lowered and have distribution with more even distributed values. A slight overall lowering can be observed for skewness with the lowest values in the original recording increasing. In comparison, kurtosis appears to increase following transmission, while frequency peak again lowers and has more values centred more narrowly around the mean.

For the mean values, the effects for /s/ in the With noise condition were again similar for CoG, SD and frequency peak. This was a decrease to the No noise condition for CoG by 16 percent or 820 Hz. For frequency peak, the percentage change was also identical to the previous condition at 15 percent, while this was a decrease of 751 Hz. SD decreased by 12 percent or 162 Hz, while skewness showed the biggest effect with a decrease of 68 percent or 0.11. Kurtosis was the only measure, which increased. This was by 39 percent or 1.74.

The spectrographic representation shows how the most intense frequencies in the WAV baseline, which align with the formant structure of the following vowel at the end of the segment, are enhanced in the with noise transmission (see figure 5.45). As in the No noise condition, acoustic information here likely related to the background noise is inserted across the lower frequencies below 4 kHz. Overall, the waveform again presents a reduction in amplitude with less frication. The spectrum overall maintain its shape from the WAV baseline.

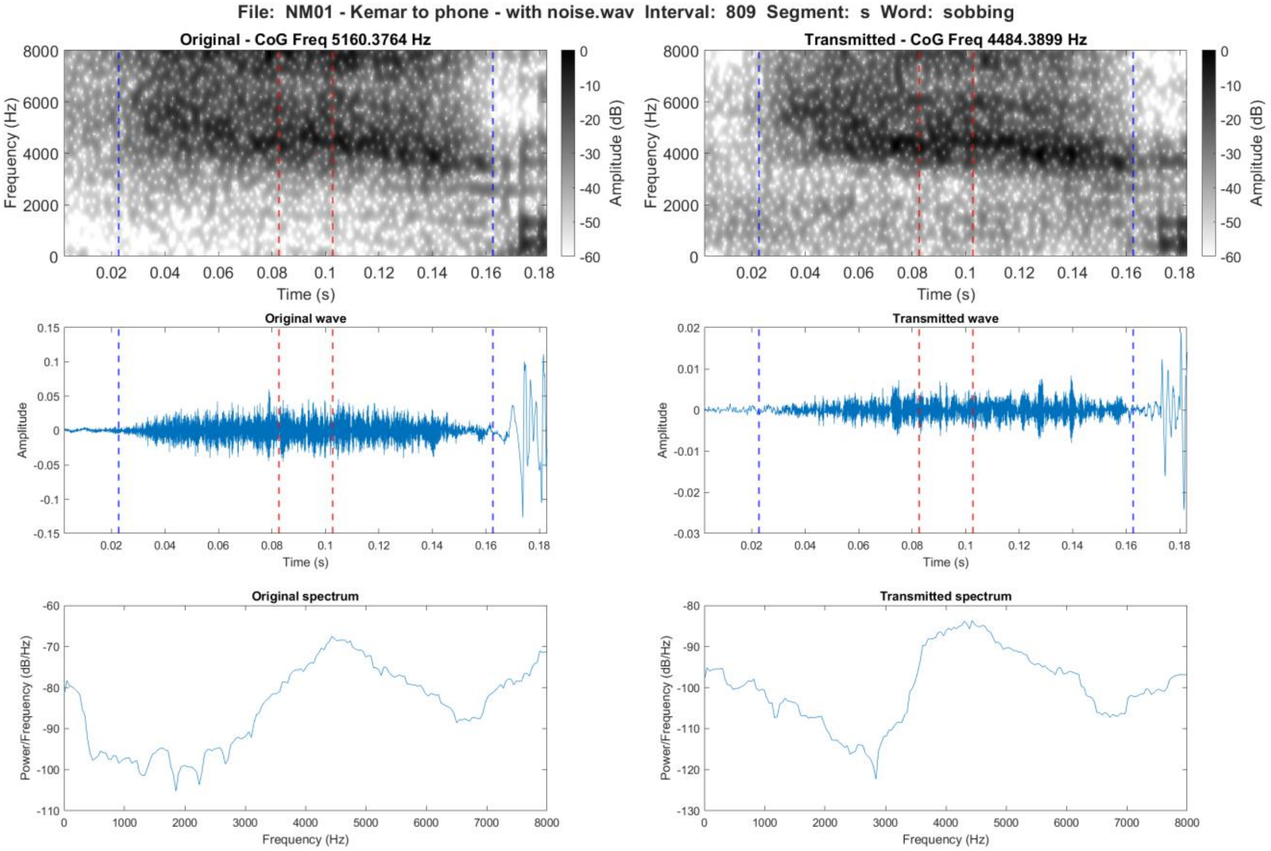


Figure 5.45. Spectrographic comparison of /s/ in the word *sobbing* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.5 /ʃ/

For /ʃ/ the distribution of CoG and SD are both lowered, which for CoG entails only a minor change to the overall shape of the distribution, while SD has more values centred around the mean. A decrease can be observed for skewness, while kurtosis shows an increase. Both skewness and kurtosis present a more even distribution of values following the with noise transmission. As in previous condition, little effect is observable for the distribution of frequency peak.

For the mean values, relatively minor effects were observed for CoG and frequency peak in this condition. For CoG this was a decrease of 9 percent or 285, which is almost identical to what was observed in the condition without noise. The decrease of frequency peak for /ʃ/ was identical across conditions at 2 percent, which in this condition was a decrease of 73 Hz. SD decreased by 36 percent or 338 Hz, while skewness decreased by 26 percent or 0.56. The only measure increasing was kurtosis, which increased by 19 percent or 2.01.

Apart from a slight reduction in the higher frequency content above 4 kHz and a tendency to more clear formant structure, the effect on /ʃ/ is limited based on the spectrographic representation (see figure 5.46). As previously observed, the transmission results in a more clearly centred amplitude around the 0 amplitude line. The spectrum remain similar in shape to the WAV baseline until around 5 kHz, where a drop appears.

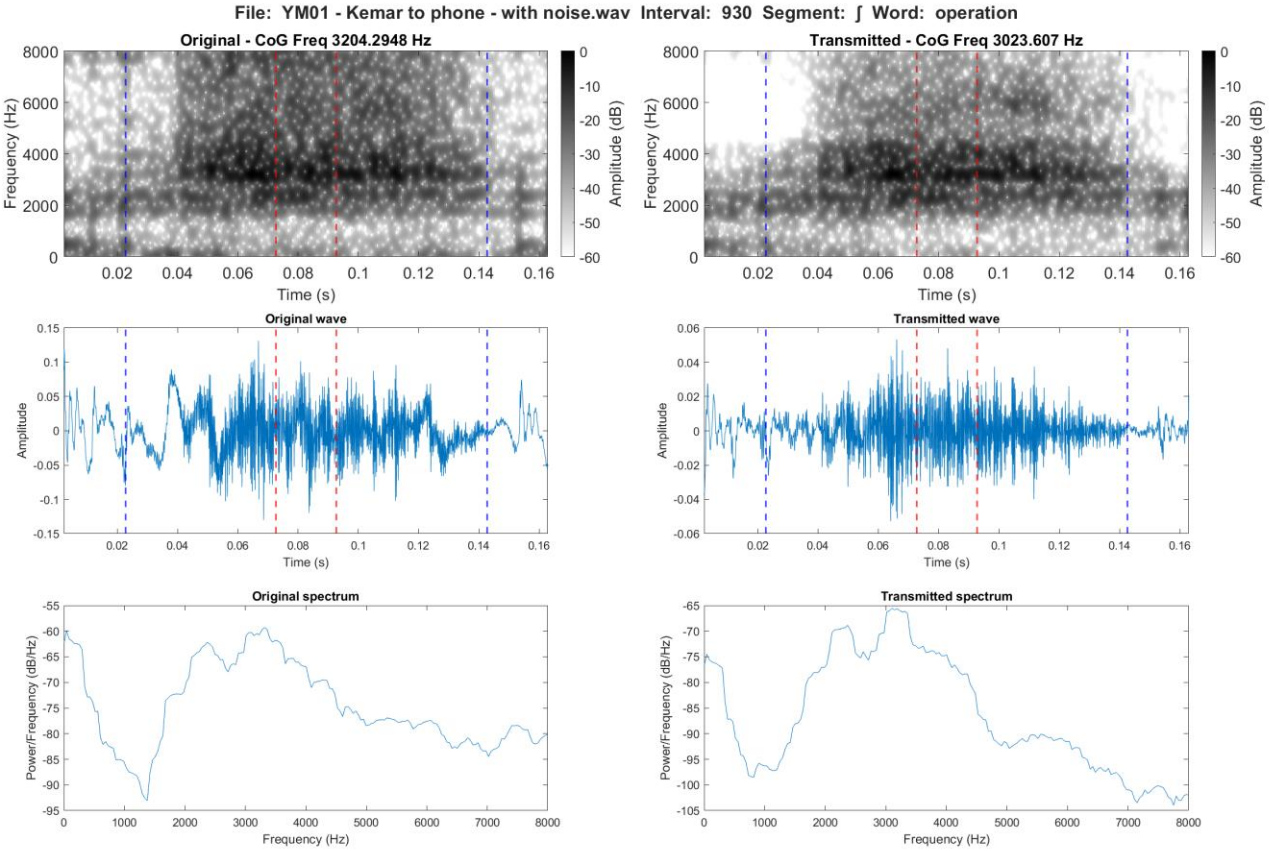


Figure 5.46. Spectrographic comparison of /ʃ/ in the word *operation* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.6 /ð/

For /ð/ the distribution plots show similar tendencies to what have previously been observed for CoG and SD. This means that both these values are lowered and have more values centred around the mean. In addition, CoG also presents a more narrow distribution. Skewness and kurtosis both show increases, which for SD again entails more values centred around the mean. The effect on frequency peak is less clear as the frequency peak is less clearly observable from the violin plots, but appears to show a slight increase.

All the mean values for each of the spectral measures decreased following transmission. This was for CoG by 28 percent or 598 Hz, which was a larger change than what was observed in any of the previous conditions. The same can be said for SD, which lowered by 41 percent or 578 Hz. The frequency peak was less affected, and despite the observation from the distribution plot, decreased by 13 percent or 152 Hz. Similarly, skewness lowered by 14 percent or 0.26, while kurtosis lowered by 16 percent or 1.88.

The spectrographic representation and waveform reveal how the transmission has reduced /ð/ to a point where the segment appears at so low an intensity that it might be interpreted as background noise or the continuation of the preceding sound (see figure 5.45). The comparison of the spectra, reveal patterns of variation in amplitude in the transmitted frame, not present in the WAV baseline.

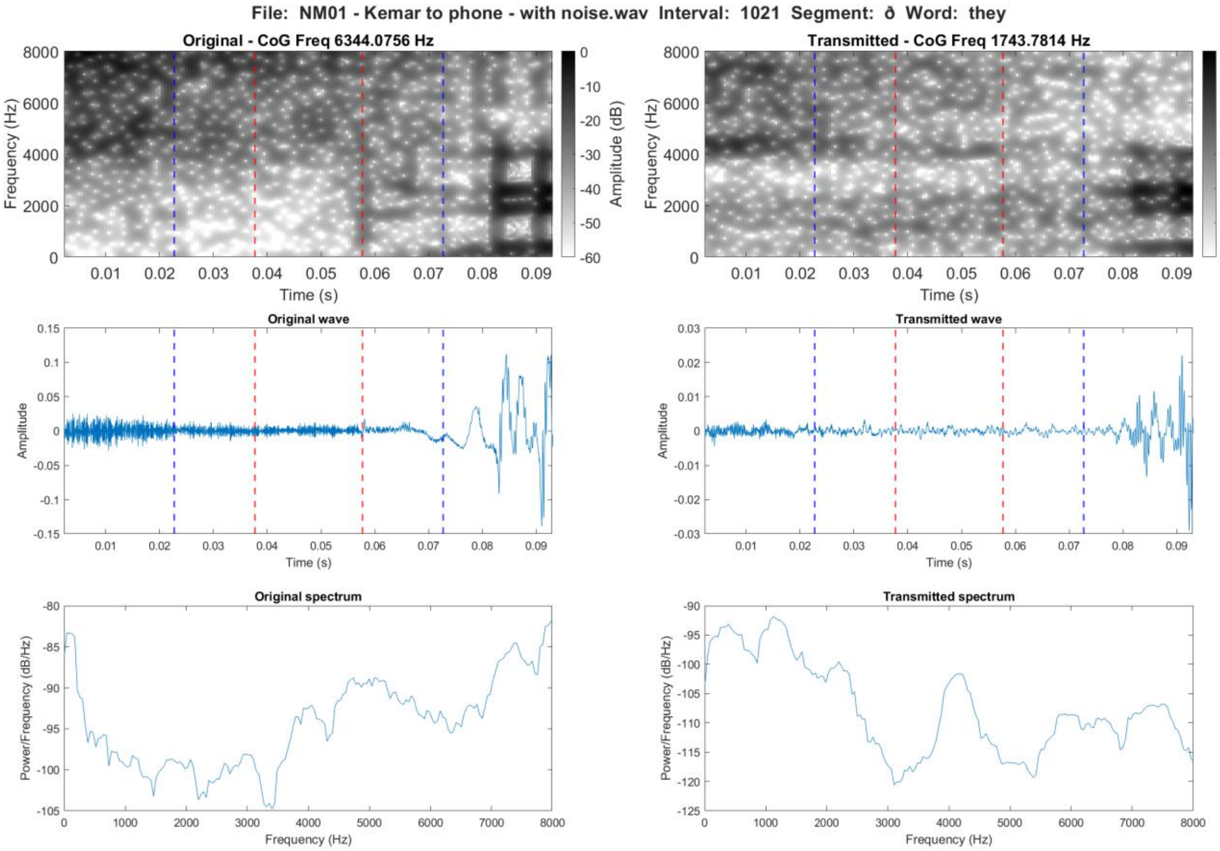


Figure 5.47. Spectrographic comparison of /ð/ in the word *they* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.4.3.7 /z/

As in previous conditions, the effects observed for the distribution of /z/ following the transmission is similar to what has been observed for /s/. This means a lowered and more centred distribution of CoG, and a lowered and a comparatively more even distribution of the SD values. For skewness a decrease can be observed with a only little change to the overall shape of the distribution. Kurtosis can also be observed to slightly increase, while the effect on frequency peak is less clear, but appears as an overall lowering with the changes mainly occurring around the values above and below the mean.

In terms of mean values, all spectral measures apart from kurtosis again decreased following transmission. For CoG and frequency peak, this was by 17-18 percent or 816 Hz for CoG and 675 Hz for frequency peak. SD presented a decrease of 10 percent or 138 Hz, while skewness decreased by 127 percent or 0.40. Kurtosis increased by an identical amount to CoG in percentage terms i.e. 18 percent or 1.01.

The spectrographic representation shows how the transmission has inserted frequency information aligned with the formant structure of the preceding vowel around 2 kHz, which were not present in the original WAV file (see figure 5.48). Moreover, the higher frequency content above 4 kHz is reduced in the background noise condition. The insertion of frequency information in the lower frequencies are also clearly visible from the spectrum, which also shows the gab in amplitude around 3 kHz.

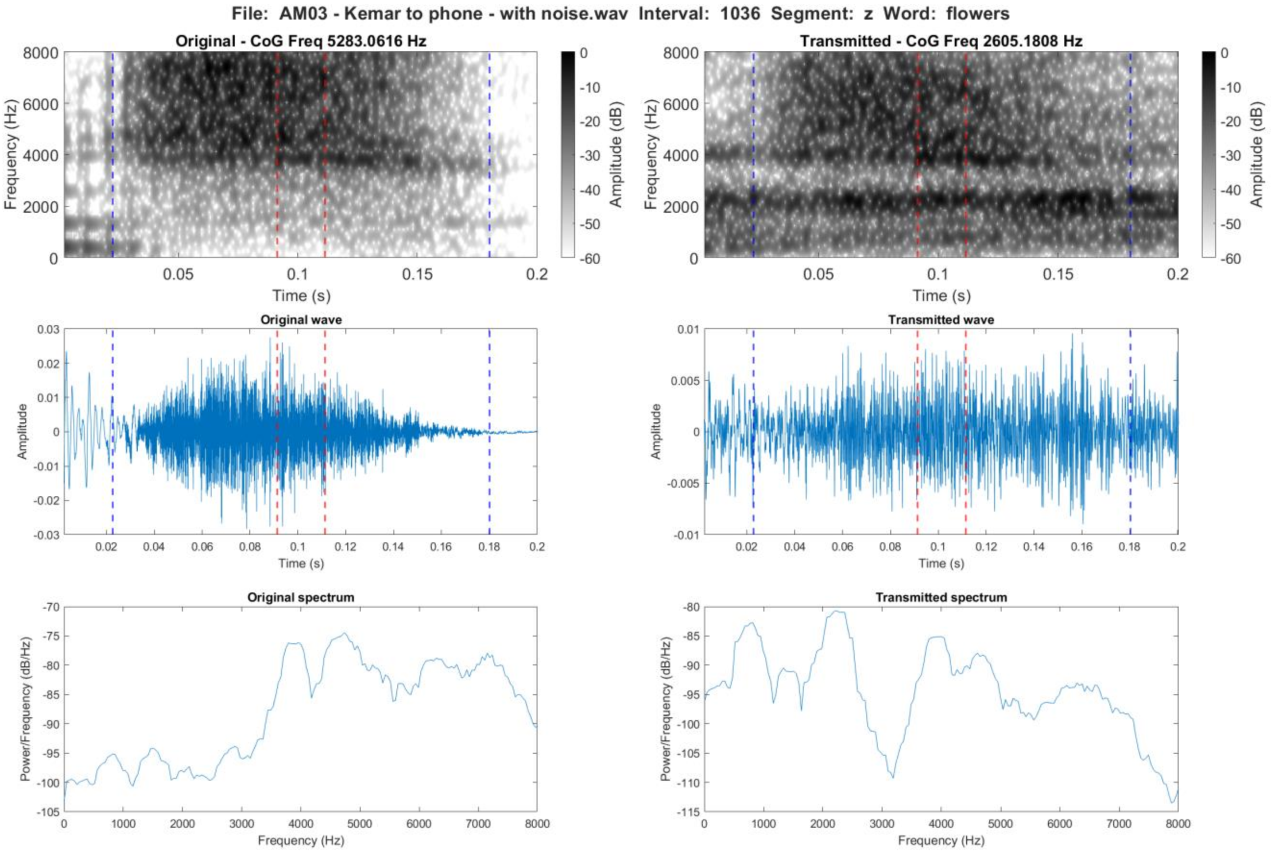


Figure 5.48. Spectrographic comparison of /z/ in the word *flowers* in the WAV baseline (left) and the live transmission in the With noise condition (right)

5.5 Discussion and Conclusion

This Chapter was concerned with three research questions:

LiveRQ1 What do the included measures indicate about the effect of digital live transmission on speech?

LiveRQ2: How does introducing background noise to the speech signal to be transmitted, affect the spectral measures and overall output signal?

LiveRQ3: In what way do these findings help understand the consequences of digital transmission on fricatives in real life scenarios and inform further research?

Broadly, the live transmission consistently lowered both CoG and SD for all segments and across all three conditions, while skewness and kurtosis presented both decreases and increases. This relates back to the first main prediction, which was the expectation that the observed effects would be greater than what had been observed in the previous studies with the AMR-WB transmission. The results do overall confirm this prediction. The results further confirm the prediction that /f/ and /θ/ would be two of the most affected segments due to their relatively lower intensity. However, it should be kept in mind that the WAV baselines are different for the controlled studies and the present study i.e. the two controlled studies had a down-sampled 16kHz baseline, while the this study used a 44.1 kHz baseline. The frequency analysis was done over the same 500 Hz to 8 kHz band as the research in the previous Chapters, which as mentioned had an influence on all the spectral measures, but particularly peak frequency.

Subsequently, the effects were predicted to be more prominent for the voiceless fricatives. This was true for /f/, [fʲ], /θ/, where especially the effect on the frequency peak maximum was more pronounced for these three segments. This effect was condition dependent, and was generally observed to be more pronounced in the Direct condition than the two conditions using the KEMAR dummy head both with and without background noise. Despite being voiceless, /s/ showed no effects on the maximum value for the frequency peak in the Direct condition, and overall was not affected, based on this measure, to the same extent as the other voiceless fricatives. One of the main distinctions between these fricatives is their intensity level, which indicates that this feature plays a bigger role in the effect of the transmission in comparison to the broader distinction between voiced or voiceless. It can also be mentioned here that /s/, in a large part, showed similar effects to /z/ despite the voicing distinction between the two. As with the previous studies, the smallest effects are generally observed on /ʃ/.

For SD, the size of the effects is similar for all segments, including /ʃ/. The live transmission’s tendency to affect SD is particularly evident from comparison of the distributions prior to and following the transmission. These effects including the lowered CoG, indicating that the live transmission overall lowered the frequency content of the fricatives, while making it less dispersed and thus, it occurs across a narrower range of frequencies.

For skewness and kurtosis, the effects were as mentioned more mixed. The effects on skewness indicate that the transmission has a relatively greater effect on the acoustic content of segments with the energy centred below the mean in comparison to the segments with positive skewness values or segments increasing following transmission, where the effects are smaller. This is unexpected and establishing the exact reason for this is beyond the scope of the present project. The only segment showing similar high percentage changes, but with an increase was /θ/ in the With noise condition, where an increase of over 200 percent was also observed.

For kurtosis, values mainly increased in the live transmissions, which indicate that the peakedness of the fricatives, especially in the With noise condition increase in comparison to the original WAV files.

In addition, the biggest relational changes were observed for these two measure, which indicates that the live transmission affects the overall energy distribution for the fricatives. This is key as skewness has previously been shown to be one of the distinguishing features for these fricatives (Shadle and Mair 1996). Thus, when e.g. /z/ goes from a higher skewness value to a clearly lower skewness value than /s/, or /θ/ goes from being higher to being lower in skewness than /f/ and [fʲ], this likely renders the measure unreliable in terms of distinguishing these fricatives.

Whether these findings make the segments more or less distinct is dependent on segment, condition and spectral measure. The distinction between /f/ and /θ/ and the distance between the two based on the spectral measures are also dependent on the condition. For none of the conditions do the two segments become identical or almost so for all measures. For CoG both the condition with and without noise in fact make the two segments more distinct. Whether this is perceptually perceived as less or more TH-fronting needs to be established through further research. However, the fact that the relation between the /f/ and /θ/ is affected by the live transmission makes it a relevant discussion and topic for further research in sociolinguistic.

Overall, /f/, [fʲ], and /θ/ group together apart from a few instances e.g. kurtosis in the With noise condition, where /f/ becomes clearly more distinct from /θ/ and /f/. It was predicted, based on the previous studies, that [fʲ] would lower, but less so than /f/. This turned out not be the case based on the relative effects expressed through the percentage changes in the mean values, which were overall similar for the two.

It was expected that /f/ and [fʲ] would potentially become more similar to /ʃ/. This was generally not borne out despite a few instances where the low intensity fricatives' mean values become more like /ʃ/ and /ð/ e.g. for skewness in the With noise condition.

A larger number of tokens were predicted to be moved to the 1 kHz tokens dataset. However, the number of tokens qualifying to be moved were very limited and thus, this prediction was not borne out and all tokens remained in the main dataset. This dataset is comparatively smaller to the two previous studies, and part of the reason for the lower number of potential 1 kHz tokens is in that way an artefact of the relatively smaller dataset. The predictions that most of these tokens would be of /ð/ was however, true. The reason for this could potentially be related to both voicing as well as a number of tokens being produced as plosives.

The qualitative analysis generally confirmed the patterns observed based on the spectral measures. Both more smoothed and more varied spectra were observed. It should be noted here that part of this effect is down to the multitaper analysis, but that this analysis was used for both the WAV files and the live transmitted files. In that way, this cannot be the only reason. The effects are varying depending on codec and segment, which suggests that either the codec compressions have an effect on the spectra or that the multitaper analysis is sensitive to the effects of the type of codec compression. This will be discussed further in the main discussion and conclusion in Chapter 6.

This part of the analysis showed that across the three conditions, the transmission had the most substantial effect on the amplitude relative to the surrounding speech, which were both lowered. This is particularly for a main concentration of energy above and below around 4 kHz. In addition, the spectra revealed changes to the amplitude pattern both seen as more smoothed as well as more varied intensities across the segments. This was typically from around 1 kHz. Another general observation across all the live transmission conditions related to this was irregularity of the waveform, which often showed limited frication. Thus, it is clear that especially the intensity of the signal is affected by the live transmission.

More specifically, the transmission caused the higher frequencies usually above 4 kHz to be reduced, while the formant structures were intensified. These effects were observed in all conditions, but more distinctively in the conditions without and with noise. This is particularly evident from the heavy reduction, which tends towards a non-transmission of /ð/ in the No noise condition. For /z/ formant structure appears in the With noise condition, which was not originally present. Overall, the way the fricatives are affected appear dependent on the preceding and following vowel as the formant structures are often enhanced based on the segmental context. The exact effect of vowel context require further research. In that way, this these results illustrate how live transmission both remove and insert frequency information in the comparison to the original WAV input.

In relation to the insertion of frequency information by the transmission, it was predicted that more comfort noise would be added to the signal because the particularly the voiceless fricatives would not be encoded and transmitted. In the previous Chapter, this was often observed as an increase in CoG for the fricatives from below to above 1 kHz. As mentioned, no tokens were in fact moved to the 1 kHz tokens dataset, and from the spectrographic analysis, this feature does not appear more prevalent in the live transmission. However, this is based on a limited set of examples, and it can therefore not be ruled out that this tendency is not actually present.

Specifically, in relation to the insertion of background noise, it was expected that more tokens would be moved to the 1 kHz tokens dataset, due to being heavily reduced and potentially not transmitted, when background noise was introduced to the transmission. No 1 kHz dataset was made, but an essential part of the study was the qualitative analysis of a selection of spectrograms, waveforms and spectra for the individual segments in each condition. This analysis amongst other things showed both large reductions for a number of segments, and potential non-encoding of /ð/ in initial position in both the condition without and with background noise. It was predicted based on the two previous studies that the effect of the transmission with noise transmission would result in most non-encodings of /ð/ particularly in this condition. This prediction could not be confirmed based on this study.

It should be kept in mind that both the sound pressure level of the background noise and the playback of the recordings through the KEMAR dummy head did not vary as much as might be expected in a real-life mobile phone conversation. Further research is needed to establish to what extent e.g. the Lombard effect, less controlled background noise, or variation in distance between the mouth and the phone would affect the quality of the output signal. In terms of quality, it is also worth noting here that the observed frequency range in the live transmitted recordings indicate that good network access has been available during the experiment. Again, further research in e.g. more rural areas is needed to establish the effects in these less optimal conditions.

This observation is related to the point mentioned both in the introduction as well as methodology for this Chapter about ecological validity of this study. The overall and long-term goal of this type of research is to understand the acoustic implications of digital transmission in any live transmission and in that way in any everyday scenario including this type of speech. However, to do this a better understanding of the individual factors and variables e.g. the transmission on its own and background noise at a controlled sound pressure level is required.

Taken together, this study has provided a first insight into the effects of live digital transmission on fricatives, and in that way, given a glimpse of what is included acoustically in every day real life mobile phone conversations. One of the key findings is how the less intense voiceless fricatives are more affected by the live transmission, which means that these in real life mobile phone conversations are likely to be of most interest for e.g. perceptual research as well as sociolinguistics.

In conclusion, it is clear that despite the background noise not always presenting substantially different changes to the signal in comparison to the other conditions outside the qualitative analysis, the live transmission alters the relation between the fricatives based on the spectral measures. Moreover, the qualitative analysis of spectrograms, waveform and spectra revealed clear effects of all live transmission conditions and even tendencies to non-transmission. The effects were especially evident for level of frication and amplitude. Thus, it is important to do these measurements with caution and not directly compare live transmitted speech, regardless of condition, and high quality studio recordings unless this comparison is the purpose of the research. In addition, the fact that the direct recording, which would be expected to be the best quality with the least effect, turns out to at times be the condition with the greatest effects, calls for further research and caution when judging quality and using datasets of this type of digitally transmitted speech.

Chapter 6 : Main Discussion and Conclusion

This final chapter of the thesis will discuss and sum up the research and results presented in the previous chapters and their implications.

Overall, one of the main motivations for this research was to create an interdisciplinary relevant project, which could bridge between e.g. sound engineering including the more technical aspects of speech communication, and fields of linguistics e.g. sociolinguistics and forensic phonetics. The project aimed to fulfil this ambition as previously mentioned by e.g. choice of measurements and the technical components i.e. different popular codecs, bitrates, and live transmission. It was also a main concern to explain some of these technical aspects of digital conversion and transmission in a way that was accessible for linguists with little to no prior knowledge on the topic.

More specifically, the thesis investigated the fricatives /f/, /θ/, /s/, /ʃ/, /z/, /ð/ and [fʲ] and measuring the first four spectral moments i.e. CoG, SD, skewness, and kurtosis as well as frequency peak compressed with different codecs, using different bitrates and in live transmission. The decision to investigate consonants was based on the fact that consonants have generally received less attention particularly in research on digital conversion and transmission of speech. Fricatives were chosen specifically because of their aperiodic and noise-like quality as this was expected to make them more vulnerable to the noise reduction features in the codec compressions. The spectral measures were chosen to make the results relevant and applicable in linguistics as well as other fields working with acoustics. In sum, the intention was to create a baseline for future research to progress from, which encourages researchers across fields to engage with the results from their specific viewpoints.

It aimed to answer the following three research questions:

mainRQ1: How are fricatives affected acoustically by digital transmission by different codecs at different qualities (i.e. different bitrates) in comparison to direct high quality microphone recordings?

mainRQ2: In what way does live transmission affect the acoustic implications of digital transmission on fricatives in comparison to high quality recordings and codec compression under controlled conditions?

mainRQ3: How can the understanding of the acoustic implications of digital transmission on consonants be applied in linguistic research (e.g. in forensic phonetics and sociolinguistics) and beyond?

The following sections will provide answers to these research questions via different perspectives on the results. The first two subsections will mainly be related to the first two mainRQs and provide summaries of the main findings (sections 6.1 and 6.2), while the following four sections will be related to mainRQ3 and address some of the implications for linguistic research including for sociolinguistics and forensic phonetics (see sections 6.3 to 6.5). These sections will be followed by a section on some methodological considerations and their implications (section 6.6). Finally, section 6.7 will also be based on mainRQ3 and provide some future and interdisciplinary perspectives beyond linguistics, and will be followed by the conclusion in section 6.8.

Before engaging with these sections, one question has arisen during this project, namely, *how big does the effects need to be to have an impact?* This is not a straightforward question to answer as it depends on the perspective from which it is asked, and the research in question.

Nevertheless, a simple answer can in fact be given: *the effects have an impact regardless of their magnitude*. An understanding of any underlying acoustic effects of codec compression and digital transmission are needed in order to improve and better explain other aspects such as perception, and from a more practical perspective to avoid reporting acoustic characteristics of speech, which is in fact an artefact of codec compression, digital transmission, or measurement method.

6.1 Summary of the main findings

The first research question was focused on the concrete acoustic implications of codec compression using different codecs and bitrates in controlled conditions. This question was answered via the two first studies, where both used software to codec compress the WAV files. The first study in Chapter 3 created a baseline investigating the AMR-WB, Opus, and MP3 codec using only one average bitrate for each, while the second study in Chapter 4 applied different bitrates i.e. an additional low and high bitrate for each codec compression.

All three codecs presented significant changes to the original WAV files, but before considering these effects in more detail, the potential limitations in bandwidth in digital transmission must be acknowledged. The two studies in Chapter 3 and 4 made an estimation of the effective encoding bandwidth based on the encoding of a white noise signal. All three codecs presented slightly lowered upper frequency limits, but the main effect was seen for MP3 in the low bitrate. Here the signal was cut-off at around 4 to 5 kHz, which affected the output signal substantially. The bandwidth is as expected dependent on the bitrate as well as the type of codec, which means that these two factors must be considered in unison in any linguistic research using digitally transmitted speech. The fact that the quality of the output signal is a combination of bitrate and codec type leads back to the point that newer technologies aim to provide better quality at lower bitrates. In other words, they aim to provide better quality, but using less data to do so. This potentially explains why the newer Opus technology performs better regardless of bitrate in comparison to the other two codecs and why the upper frequency limit is largely unchanged from the WAV baseline. The take home point here is that limitations in bandwidth are inevitable in codec compression, but when using higher bitrates the limitation is less and often between 7 and 8 kHz. From the perspective of most fricatives, this means that lowered frequency peaks and CoG must be expected in any codec compression. Lastly, extra care should be taken when analysing MP3 files as the effects in the low bitrates made the fricatives spectrographically unrecognisable in comparison to the WAV files.

The codec and segment dependent tendencies can be illustrated via the largest and smallest effects caused to the mean values of the spectral measures by the codec compressions. To allow comparison and more general observations across the codecs and bitrates, the approach in the following paragraphs concerned with mainRQ1 will discuss the results and their implications based on the individual spectral measures.

In short, most of the effects of the codec compressions were found to be significant especially in the average and low bitrates.

Another related aspect, is how the speakers produced different phonetic variations of /ð/ apart from as a voiced fricative. These were as an approximant and secondly as a plosive, which is not uncommon for voiced fricatives and /ð/ specifically. For the results, this again means that they are not strictly representative of the effect on phonetically voiced fricatives, but on the other hand, it allows the analysis to expand and give an indication of the effects on other consonants i.e. plosives. The observed examples of this plosive production were primarily voiceless.

For CoG, the codec compressions almost consistently lowered the mean values in comparison to the original WAV files. The only exception was found for /ʃ/ in the low bitrate, where an increase by under 1 percent was observed. In the low bitrate across codecs the smallest percent change was found for /ʃ/ in the AMR-WB, which lowered by 0.09 percent, while the largest effect was found for /θ/ in the MP3 compression by a lowering of 26.14 percent. In the average bitrate, the smallest effect was again found for /ʃ/ in the AMR-WB. Here the CoG mean was lowered by 0.03 percent, while the largest change was found in the Opus codec with a 14.99 percent lowering of /θ/. For the high bitrate, it was again /ʃ/, which presented the smallest change with a lowering of less than 1 percent and /θ/, which showed the largest change by a lowering of 9.79 percent. However, for the high bitrate, the smallest effect appeared in MP3, while the largest effect appeared in the AMR-WB. These effects illustrate how the codec compression generally results in more energy relatively lower in the spectrum. As will become evident from the rest of the spectral measures, the effects on CoG reveal a general pattern of segment dependent tendencies.

Similar to CoG, SD lowered in all codec compressions and bitrates following all of the three codec compressions apart from /ʃ/ in the AMR-WB compression using the average bitrate, where the mean value increased by 1.38 percent. This was also the smallest effect found in the average bitrate, while the largest effect was found for /ð/ in AMR-WB by a lowering of 18.65 percent.

In the low bitrate the smallest effect was again found for /ʃ/, which lowered by 2.53 percent, while the largest effect was found for /θ/ in the MP3 compression, where the mean value lowered by 33.20 percent. In contrast, the smallest effect in the high bitrate was found for /ð/ in the MP3 compression, where the mean value lowered by 3.96 percent, while the largest effect was an 11.65 percent lowering of /s/ in the AMR-WB. In that way, across the codec compressions, the energy is less dispersed across the spectrum in comparison to the original WAV files.

For skewness, both decreases and increases were observed. The low bitrate had the greatest effect by a lowering of 1,010.3 percent for /s/ in the MP3 compression. The smallest change using this bitrate was an increase by 3.44 percent for /ð/ in the Opus compression. In the average and high bitrate it was again /s/ which was most affected by the codec compression. This was a decrease by 366.67

percent in the average bitrate and by just over 550 percent in the high bitrate, both in the AMR-WB compression. The smallest effect in the average bitrate was a decrease by 0.77 percent in the AMR-WB for /ʃ/, while the smallest effect in the high bitrate was a decrease of 3.77 percent for /ð/ in the Opus compression. Similar to /s/, /z/ showed high percentage changes in the codec compressions. Part of the reason for these similar effects might be found in the phonological distinction made by the MFA i.e. the Montreal Forced Aligner. However, these two fricatives are also distinct from the other sounds as they are the only two with a negative skewness mean value in the WAV baseline. This suggests that the codec compression result in more energy below the mean for segments, which already have more energy centred at this point prior to compression. This is an unexpected finding, and it will require further research contrasting segments specifically by this feature potentially on a continuum from 0 to establish a more exact reason for this.

The kurtosis values showed some high positive values already in the WAV baseline, which would suggest a peaked rather than flat distribution. However, it has been pointed out by McCarthy, how this only holds true if the distribution is not skewed (McCarthy 2019). As it has been established in the previous paragraph, and as it is typical for spectra, the distributions are indeed skewed. It might also be added here that Maniwa and Jongman reported similarly high kurtosis values in initial and final position (calculated as percentage of the full segment) of the fricatives in their results (2009) . Hence, this is not a completely novel finding and might be related to the place of measurement within the segment, which in this thesis was the central 20 ms frame.

Another explanation for these high values might be found in the use of multitaper analysis. As mentioned, the application of multiple tapers smooths the spectrum, resulting in a potentially more peaked spectrum and in that way positive and higher kurtosis values. For these reasons, these values should be evaluated with these caveats in mind, and not taken as unquestionable evidence for the acoustic effects of the codec compressions alone.

With that said, the effects on kurtosis were again both codec and segment dependent, and presented both increases and decreases. The smallest effect in the low bitrate was found for /ð/ in the AMR-WB compression by an increase of 0.20 percent, while the largest effect was in the MP3 compression by an increase of 48.73 percent for /s/. The effects between these two values showed decreases also ranging above 40 percent. In the average bitrate, the smallest and largest effects were both decreases. The smallest effect was found for /ð/ in the AMR-WB compression with a decrease by 0.14 percent, while the largest effect was 13.43 percent for /s/ in the Opus compression. In the high bitrate the least affected segment was /z/ in the MP3 compression with a decrease of 0.56 percent, while the largest effect was to /ʃ/ in the AMR-WB compression.

Lastly, similarly to kurtosis, frequency peak is potentially not truly representative of the acoustic characteristics of the investigated fricatives. This is because these fricatives often have energy above or around the 8 kHz limit that is used here. The frequency peak is not an accurate representation of the acoustic content of the segment, but rather an illustration of the impact of the down-sampling. However, this is a finding in itself as it illustrates the importance of considering the relation between sampling rate and choice of acoustic measures. Moreover, as with kurtosis, the results are codec, bitrate, and as would be expected, segment dependent.

In the low bitrate the smallest and largest effect on frequency peak were found in the AMR-WB and the MP3 compressions. This was a decrease by 0.30 percent for /ʃ/ using AMR-WB and a decrease by 21.22 percent for /θ/ using MP3. In the average bitrate, the smallest effect was found in the MP3 compression, where /ʃ/ lowered by 0.27 percent, while the largest effect was found for /θ/ in the Opus compression with a decrease of 25.57 percent. In the high bitrate, the smallest change was for /z/ in the Opus compression with a decrease of 0.16, while the largest change appeared in the AMR-WB compression with a decrease for /θ/ by 14.54 percent.

The introduction mentioned how the three investigated codecs can be divided into two main groups, namely speech codecs and psychoacoustic codecs. The first two studies showed how this distinction is indeed relevant for the observed acoustic effects on the investigated fricatives. All three codecs affected the acoustic quality of the input signal significantly, but MP3 (psychoacoustic codec) turned out in the low bitrate, to reduce the signal substantially in comparison to the two speech codecs.

This was evident both from significant changes to the spectral measures, but particularly from visual inspection of the spectrographic representations of the investigated fricatives. Here, the MP3 codec clearly excluded parts of the speech signal in the low bitrates from the transmission, and at times introduced a complete drop in amplitude around 5 kHz. In comparison, both AMR-WB and Opus across bitrates largely maintained speech relevant content with the main visible reduction being to the frequencies around the upper frequency limit as well as an overall decrease in intensity. It is therefore clear that whether the speech codecs perform better in terms of maintaining the acoustic quality of fricatives is bitrate dependent. In low bitrates, this means that the speech codecs will produce relatively more accurate and reliable spectral measures.

In addition, this is interesting because the low MP3 bitrate was not the lowest of the three applied bitrates i.e. AMR-WB 6.6 kbps, Opus 16 kbps and MP3 12 kbps. Hence the effect of the MP3 codec further suggests that bitrate on its own cannot be taken as a predictor of acoustic integrity and quality, but has to be evaluated individually for each codec type. This is substantiated by the fact that the greatest effects for AMR-WB and Opus across the spectral measures were not consistently found in the low bitrates. It was previously mentioned that the goal with newer technology is to achieve better quality at lower bitrate, which further emphasises the point not to take bitrate as a direct correlate of quality.

The allophonic distinction between /f/ and [fʲ] made by the MFA turned out not to be affected equally by the codec compressions. It should be kept in mind that the number of tokens of [fʲ] were smaller than the number of tokens of /f/, and that the effects were not as prevalent in the live transmission.

But, together with the fact that a number of tokens had formant structures carried over from the preceding or following vowels, this suggest that the codec compressions are on some level sensitive to context. Moreover, Ohala has described how the allophonic distinctions and palatalisation are dependent on specific segmental contexts e.g. how /s/ becomes /ʃ/ in specific contexts (e.g. 1994). The fact that preceding and following segment was a fixed effect in the mixed effects modelling, further emphasise this and calls for more research on this specifically.

In summary, the codec compressions overall affect /ʃ/ the least and /θ/ the most, while the remaining segments tend to fall in between these two segments in terms of the magnitude of the effects. However, the effects are both codec and bitrate dependent, which is illustrated by the variation in which codecs produce the largest and smallest effects on the fricatives.

6.1.1 1 kHz tokens

The 1 kHz tokens, were the tokens which were expected to have been most affected by the codec compressions, which was expected to be seen as large reductions in intensity and potentially entire segments not encoded by the codecs. These tokens were identified as tokens, which had a CoG value above or below 1 kHz in either baseline or codec compression, which then moved to either below or above this limit as a consequence of the codec compression. Tokens, which had values below 1 kHz in both WAV and codec compression were also included in these datasets, but only the pairs, which showed the movement from above to below 1 kHz or vice versa was analysed spectrographically. The methodological implications and relevance of separating these tokens from the main dataset is discussed at the end of this section.

Firstly, more tokens were indeed moved to the 1 kHz dataset in the lower bitrates in comparison to the higher bitrates, which indicate some tendency for greater effects depending on the bitrate. This is also supported by the results from the main dataset.

This leads to the second point, which is the relevance of creating this dataset and analysing these segments separately. It was expected that the tokens in this dataset would potentially show a high level of reduction in intensity and potentially complete non-encoding. A few instances of non-encoding did occur. This was particularly of the voiceless plosive production of /ð/, this is suggested to be because the codecs mistook the burst and breathiness for noise. In that way, plosives appear to be substantially affected by the codec compression, which encourages future research on digital transmission and speech to expand to other consonants in addition to fricatives. The substantial reductions in intensity were mainly to /f/ and /θ/.

It is also observed, how some of the tokens which increased from a CoG value below 1 kHz in the WAV files to above in one of the codec compressions do so because of insertion of frequency information, which was not present in the original WAV files. These findings are all relevant and have implications for linguistic research, however most of the CoG mean values for these tokens were already below 1 kHz in the original WAV files. It was only for AMR-WB that this was not always the case. Most of the tokens in the dataset also turned out to be instances of /ð/ in initial position of the word *the*, which had a CoG mean below 1 kHz in both WAV and codec compression. This indicates that it is potentially relevant to separate these tokens to avoid skewing the dataset and assess whether these might be cases of voiced approximant realisations of this segment.

As a consequence, it might be that the 1 kHz tokens actually behaved similarly to the tokens in the main dataset, and that the codecs had similar effects on other tokens, which did not fit the 1 kHz criteria. Additional, qualitative analysis of the segments in the main dataset is required to establish this more clearly.

In addition, together with [fʲ], /ʃ/ had no tokens in the 1 kHz tokens dataset. The reason for this finding is suggested to be the codec compressions overall tendency to recognise and encode high intensity frequency more easily. Both [fʲ] and particularly /ʃ/ were relatively high in intensity and have in that way been less affected by the compression than e.g. /θ/ which is lower in intensity and therefore more prone to the effects of the codec compression. As mentioned in the theoretical background, this is also likely related to the fact that the codec compression aims to remove background noise, which potentially matches the lower intensity frequencies of /θ/ (3GPP 2020).

The findings of the live transmission study support this hypothesis, and indicate that the distinction in intensity might be a more determining factor for the effect of the codec compressions in comparison to the distinction between voiced and voiceless segments. This will be addressed in more detail in section 6.2 below.

Specifically for live transmission and the 1 kHz tokens, not enough tokens were found to match the 1 kHz criteria for it to be meaningful to analyse these segments separately. However, large reductions in intensity and frequency information were still observed in the different live transmission conditions. This suggests as mentioned above that the 1 kHz criteria might not be the most appropriate criteria to identify these tokens. For further research, it is suggested based on the effects found on intensity to include measures of amplitude and single out the large reductions and non-encodings based on this rather than CoG.

It was of general interest whether the fricatives became more or less alike especially for CoG, which has previously been found to distinguish sibilants from non-sibilants (e.g. Maniwa, Jongman, and Wade 2009; Jongman, Wayland, and Wong 2000; Shadle and Mair 1996).

In the average bitrate, especially in the 1 kHz tokens, it was noted how /f/, /θ/, /ʃ/ and/or /ð/ became almost identical in terms of CoG mean values. This is key because the distinction between sibilants and non-sibilants from the spectral moments is in that way limited in codec compressed speech.

6.2 The effects of live transmission

The second research question expanded on mainRQ1 by investigating the same effects, but in live transmission under three conditions. These conditions were: a) a recording of the signal directly transmitted from one phone and recorded from the receiving phone via a cable (i.e. Direct condition);

b) a recording of the signal played through a KEMAR dummy head without background noise (i.e. No noise condition); and c) a recording again using the KEMAR dummy head, but this time with background noise played from an ambisonic speaker array to imitate real life conditions (i.e. With noise condition). All internet and data connections were turned off on the phones, which meant that they were accessing the AMR-WB 3G network.

A few general points on this research are relevant to mention before engaging with the results. Live transmission adds a number of variables, which cannot be controlled in the current setup. This includes e.g. available bitrate, variation in network access, and hardware i.e. the technology, loud speaker and microphone installed in the transmitting and receiving phone. The hardware was controlled to a certain extent by using the same mobile phone model for both the transmitting and receiving device. However, it cannot be guaranteed that the components are identical, and it means that findings cannot necessarily be generalised to other mobile phones without more experiments.

Despite the fact that these variables cannot be controlled, it is essential to do this research as it illustrates the effects of the codec compression closer to a real-life scenario. The study in Chapter 5 used transmission via the AMR-WB network, and used a smaller dataset than in the two previous chapters.

Another aspect, which set the live transmission study apart from the two previous studies, was the fact that input WAV files had a sample rate of 44.1 kHz. This was the case because the study aimed to replicate real life speech, where this was not possible in the previous studies due to technical requirements of the codec compressions.

However, the aim was still to be able to compare the results across chapters, and therefore the analysis band in the live transmission study was kept between 500 Hz and 8 kHz as in the previous studies. For future research, the 44.1 kHz files might be analysed and down-sampled prior to transmission, but for this thesis, comparability was prioritised. This, along with other methodological decisions, influences the ecological validity of the experiment in Chapter 5. Some of these decisions include the laboratory setup as well as the use of a KEMAR dummy head and replay of the read speech rather than live speech of participants. These decisions were made in light of the number of variables and factors already introduced simply by transmitting the signal live via mobiles phones. With the knowledge from the current research, it will be possible in future research to more accurately tease apart the effects of the transmission and any additional factors e.g. speaker or less controlled levels of background noise.

Thus, the study and its results are meant to give an indication of the potential effects, which might be expected in live transmission during every day phone conversations, while inspiring further and more detailed analysis on the topic.

One of the results, which applies across segments, is again the question of bandwidth. The bandwidth was not assessed using a white noise signal as in the two previous studies, but an indication of the available frequency range can be taken from the maximum frequency peak values. In the WAV files, the highest maximum value was 8,010 Hz. This specific value was an artefact of the multitaper analysis, but was either the same in the different conditions for the live transmission, or it was lowered.

The fact that in all conditions one segment i.e. /s/ maintained the 8.010 Hz value suggests that the bandwidth has not changed from the original WAV file and that good network access has been available. Furthermore, it is interesting to note that the effect on the maximum values for frequency peak were highly segment dependent. An example of this is how the maximum values for /s/ were unaffected by live transmission, while /f/, [fʲ], and /θ/ all have a maximum value around 8 kHz in the WAV baseline, but have these values decrease by around 1 to 2 kHz. In that way, these results indicate that the frequencies in the higher end of the spectrum are affected differently depending not only on whether the fricative is voiced or voiceless, but potentially also intensity level and other acoustic characteristics.

Finally, the live transmission created a time discrepancy with the original WAV files, which meant that the forced aligned TextGrids based on the WAV files, if applied directly to the live transmitted files, would not be accurate. The analysis in Chapter 5 accounted for this discrepancy and ensured correct measurements, which is important to do in any linguistic research working with a similar type of dataset to avoid incorrect measurements.

It was expected that live transmission particularly with the introduction of background noise would show substantially greater effects than what was found in the controlled conditions. This was because live transmission introduces a number of additional variables to the encoding process, including the hardware and the ANC i.e. Active Noise Cancellation in mobile phones. The background noise was expected to enhance the effects as more of the frication and aperiodic frequency content of the fricatives were expected to be mistaken for and not encoded. The results overall, but not consistently, confirmed this prediction as the Direct condition at times presented the largest effects on the mean values for the spectral measures. Overall, CoG, SD, and frequency peak consequently lowered the spectral measures. Skewness showed variation between decreases and increases, while kurtosis more consistently, but not solely, showed increases following the live transmission.

In summary, the smallest change for CoG was an 8 percent decrease of /ʃ/ in the Direct condition, while the largest effect was a 41 percent decrease for /θ/ in the condition with background noise.

SD was overall the measure which was influenced the most by the live transmission. This meant that the largest effect was found for [fʲ] as a decrease of 44 percent in the Direct condition. For all the conditions for [fʲ], as well as /f/, all the effects were decreases by around 40 percent. The smallest effect for SD was a 10 percent decrease for /z/ in the condition with background noise.

The only measure and segment that showed no changes following the live transmission was skewness for [fʲ] in Direct condition. In contrast, the largest effect observed for skewness was a 246 percent decrease for /z/, also in the Direct condition. Changes over 200 percent were also found for /s/ in the Direct condition as well as for /θ/ in the condition with background noise.

For kurtosis, the smallest effect is found for /ð/ with a decrease by 16 percent, while the largest change is again found for /θ/, which increased by 183 percent. Finally, the frequency peak presented the smallest effects for /ʃ/, which across all the conditions lowered by 2 percent, while /θ/ and /f/ presented the largest effects with a decrease of 35 percent in the condition with background noise.

These results, as expected, suggest that live transmission generally has greater effects on these fricatives, in comparison to what was observed in the controlled conditions even using the low bitrates. This is apart from skewness, which appears more affected in the controlled conditions. /ʃ/ is still the least affected and /θ/ among the most affected.

It was unexpected that the Direct condition for any of the measures would produce some of the most substantial effects. The reason for this will require further research, but underlines how the many technical aspects of digital transmission as well as experimental setup affect the results.

In contrast to the results from the two previous studies, in the live transmission the prediction that /f/ and [fʲ] would be affected equally by the codec compression was borne out. In that way suggesting that live transmission made the codec compression less sensitive to the allophonic distinction. /θ/, /f/ and [fʲ] were overall the segments most affected by live transmission in terms of CoG and SD. They were aslo the tokens where the condition with background noise had the most substantial effect. This follows the prediction that the relatively low intensity of these three voiceless fricatives would make them more prone to acoustic reductions in live transmission, while being mistaken for background noise.

The analysis of spectrograms, waveforms and spectra revealed three main tendencies in the live transmitted data. Firstly, the reduction in intensity and frication across the segments again appeared following all the codec compressions, but more substantially than in the controlled conditions of the first two studies. Whether this reduction has a perceptual relevance in terms of distinguishing the fricatives will require further research, because the reduction appears to happen across the signal, and thus, the relative difference in intensity between segments is potentially maintained.

Secondly, the influence of position in the segment as well as segmental context appeared more influential on the effect of the codec compressions. This was seen as the formant structure of preceding and following sounds were often carried over into the fricative. This will be addressed further in section 6.3, which looks at the role of voicing and intensity.

Thirdly, the analysis of the spectra showed how live transmission both smoothed and added variation to the spectrum in comparison to the WAV files and more so than what was observed in the codec compressions under the controlled conditions in Chapters 3 and 4. The smoothing is to be expected from the multitaper analysis, but since both the WAV files and digitally transmitted files were analysed this way, it cannot be the only explanation. This is also evident from the fact that it is segment dependent and happening to different degrees than in the controlled studies, which suggests a clear difference between the controlled conditions and the live transmission. Moreover, the variations introduced by the transmission suggest that the codec compressions are affecting the amplitude differently across the frequency range. It could for example be seen how bands of frequencies e.g. from 3 to 4 kHz were left completely silent by the transmission. The fact that e.g. the drops in amplitude observed in the spectra could often also be observed in the FFT generated spectrograms, suggests that the effects are in fact due to the transmission rather than the multitaper analysis.

In sum, these findings show how live transmission produces more substantial and different effects both quantitatively, based on the spectral measures, and qualitatively based on e.g. spectrographic analysis in comparison to the results found in the controlled conditions. This means that the effects of the codec compression alone cannot be taken as representative of what is experienced in regular phone conversations using the AMR-WB 3G network. Part of the effects observed are also a consequence of the hardware used i.e. the Samsung Galaxy phones, which means that replicating the study with different phones is likely to produce some difference in results. In that way, the results warrant more research into this type of digitally transmitted speech regardless of the increased number of variables in play.

6.3 Voicing and intensity matter

It was initially predicted that due to especially the VAD i.e. Voice Activity Detection, which works primarily based on criteria of pitch and tone (3GPP 2020b) that the codec compression would be sensitive to phonetic voicing. In other words, it would be sensitive to the level of periodicity of the waveforms of the speech sounds and in this case the fricatives.

The results show how voicing does appear to make a difference to the effects of the codec compressions, but in different ways than what was initially predicted. In the controlled conditions in Chapter 3 and 4 for /s/ and /z/ across codecs and bitrates, the effects on these two sounds, based on the spectral measures, were not identical, but most of the time followed a similar pattern. This would suggest that voicing plays a minor role, however, the forced aligner i.e. the MFA made a phonological rather than a phonetic distinction between the two sounds. This means that a number of the tokens of /z/ were in fact phonetically voiceless. It is unclear how many tokens of /z/ were in fact voiceless, and it is in that way not possible to assess the exact impact of this fact on the results. However, since differences are still observed between the two sounds and the spectrographic examples did show voiced realisations, the majority of the segments have likely not been voiceless. It is recommended to always manually assess these segments and the performance of the forced aligner.

The spectrographic analysis also revealed how voicing does in fact appear to play a part in the codec compression, but in relation to the preceding and following sounds. A number of examples were observed in the controlled conditions using especially the low bitrate, where the formant structure from the preceding or following sound was replicated in the fricative following the codec compression. It is unclear, whether the formant structure observed to carry over from the following sound is low level structure already present in the fricative in the WAV file, which is then enhanced, or the codec compression pre-empt the voicing and/or insert the formant structure. The latter could potentially be happening if the transition between the fricative and the vowel are encoded in the same 20 ms. frame.

This tendency was even more prevalent in live transmission with the AMR-WB in Chapter 5, where it also became clear that /f/, [fʲ], and /θ/ generally grouped together in terms of effects of the transmission, while /s/ and /z/ did the same. From the spectrograms, it was clear that the reason for this was likely the intensity level of these sounds. /f/, [fʲ], and /θ/ were generally low in intensity already in the WAV files, and more substantially reduced in exactly intensity with at times little frequency information maintained from the WAV baseline following transmission. The reverse could be observed for /s/ and /z/, where /s/ even at points increased in intensity.

It should be noted here that another difference between the voiceless and voiced segments are their amount of low and high frequency content. The voiced segments will be centred lower in the spectrum than the voiceless segments. It was clear that across codecs, bitrates and conditions, the main effects were to the higher frequency content, which for this reason means that the voiceless segments would be affected more than the voiced segments.

Thus, the results show how intensity of the fricatives play a key role in the effect of the codec compressions, and how likely the fricatives are to be mistaken for noise and not be encoded fully by the codecs. In addition, voicing also plays a role, but in relation to preceding and following sounds, which means that further research is needed in relation to context. From this thesis, the results indicate that more formant structure and lower frequency content might be expected for fricatives following codec compression especially in low bitrates and live transmission.

6.4 TH-fronting, /s/-retraction, and other dialectal phenomena

The final point, relating back to mainRQ1 and the acoustic implications of the codec compression using different bitrates in controlled conditions, relates to how similar the various segments become following the codec compressions.

As mentioned in the introduction in Chapter 1, the dialectal features predicted to be affected by or occur due to the codec transmission are not actually produced by the speakers, but rather mechanically mimicked by the codec compressions. In other words, a sociolinguistic feature such as TH-fronting might not be present in the original input WAV signal, and will continue not to be so, but during the encoding, /f/ and /θ/ are made similar by the compression. This leads to the listener potentially perceiving TH-fronting to be present, as it is acoustically mimicked by the codec compression without actually being present in the produced speech input. This is important, when considering how codec compression and digital transmission of speech potentially vehicle diffusion of sound changes, which will be considered in more detail in section 6.7.3.

As it will be evident from this section, it will not always be possible to conclude whether an observed effect is enough to e.g. constitute a specific dialectal phenomenon. This is primarily due to a lack of acoustic research on these variables as well as the fact that the methodologies in fricative research vary substantially (see section 2.2 in chapter 2), which means that different studies provide different spectral values for these segments (e.g. Maniwa, Jongman, and Wade 2009; Shadle and Mair 1996).

In that way, it should be kept in mind that just because these fricatives become more similar in terms of e.g. CoG mean value, it does not mean that they are in fact identical or it would not be possible to distinguish them perceptually or based on other acoustic measures.

Throughout this thesis, the main interest has been the difference between /f/ and /θ/ and between /s/ and /ʃ/. This is because the distinction between these sounds is related to the dialectal phenomena of TH-fronting and /s/-retraction. TH-fronting is primarily investigated as a perceptual feature and primarily assessed qualitatively (e.g. Wood 2003; Stuart-Smith, Timmins, and Tweedie 2007; Stuart-Smith et al. 2013; Levon and Fox 2014), while /s/-retraction has been found to correlate with lower CoG values for /s/ (e.g. Baker, Archangeli, and Mielke 2011; Stuart-Smith et al. 2019).

Firstly, in terms of TH-fronting none of the codec compressions result in all spectral measures becoming identical for /f/ and /θ/, while the two segments are also often similar already in the WAV baseline e.g. a difference in CoG mean value between the two of roughly 140 Hz to 155 Hz.

More similarity is observed in terms of CoG and, in certain cases, the linear predictions do indeed suggest that the two are identical based on this value. This is particularly in the low bitrate for AMR-WB and MP3 and in the average bitrate for AMR-WB and Opus, where the two segments only vary in CoG mean values by 35 Hz to 71 Hz. Overall, it appears that the lower the bitrate, the less the distance between /f/ and /θ/ becomes in terms of CoG. In addition, the spectrographic analysis revealed how /f/ and /θ/ were often strongly reduced in intensity especially in live transmission. This will inevitable make the sounds less distinct, but whether this is enough for it to be perceived as TH-fronting will require further research because the spectral measures are generally not found to be distinguishing features between non-sibilants.

For /s/-retraction, the main distinguishing feature is a lowered CoG relative to /ʃ/. /s/ is lowered across the bitrates, while /ʃ/ is generally the least affected segment. Nevertheless, the difference between the two remains above or around 1 kHz in all the codec compressions and bitrates, which suggest that the two are still clearly distinguishable despite the effects on /s/ being significant. In order to establish the potential occurrence of this feature more exactly, research is needed because /s/-retraction is found to be speaker specific and determined by the relative distance between /s/ and /ʃ/ (e.g. Baker, Archangeli, and Mielke 2011; Stuart-Smith et al. 2019). Specifically Baker et al. states 'A retracted /s/ is characterized by a centroid frequency that is 75.5% of the distance from /s/ to /ʃ/' (2011, 360). From the CoG values obtained in the present study, where /s/ is one of the segments showing the smallest effects of the codec compression, while maintaining a distance of over 1 kHz to /ʃ/, /s/-retraction is unlikely.

Comparing these findings to the live transmission results, the findings specifically in relation to TH-fronting and /s/-retraction vary from the observations in the controlled conditions. The CoG for /s/ is still lowered in all the live transmitted conditions, but the difference in mean CoG between /s/ and /ʃ/ are greater in live transmission than what was the case in the controlled conditions. In that way, based on the mean CoG values, the data do not suggest that /s/-retraction is prominent in live transmission using the AMR-WB codec.

Until now, the main question has been whether the digital transmission made /f/ and /θ/ more similar, which would potentially mean the occurrence of what might be perceived as TH-fronting.

In contrast to the results in the controlled conditions, apart from a few exceptions, the live transmission made /f/ and /θ/ more distinct based on the spectral measures, in comparison to the WAV baseline. CoG and SD illustrate the varying tendencies across conditions, where for CoG the largest distance between the two segments are found in the With noise condition, while for SD the smallest difference is found in the With noise condition. Thus, live transmission rather than increasing the amount of TH-fronting potentially limits it.

Another sociolinguistic variable, which has not previously been mentioned, but which turned out to potentially be relevant based on the results is /s/-fronting. This feature is a known idiosyncratic feature used e.g. to indicate group identity. /s/-fronting acoustically results in a higher frequency peak than non-fronted /s/(Levon and Holmes-Elliott 2013). The frequency peaks reported in this thesis were, because of down-sampling, not truly representative, which means that the results cannot be seen as the most accurate reflection of the codec compressions influence on this feature. However, the fact that the frequency peak for /s/is consistently lowered by the codec compressions suggest that /s/-fronting, similar to TH-fronting in live transmission might be limited in codec compression. This will require further research both acoustically as well as perceptually.

Lastly, a few other linguistic phenomena are worth mentioning, which has been beyond the scope of this thesis, but which based on the results are likely to also be affected by the codec compression and digital transmission. This includes e.g. /h/-dropping defined by Milroy as 'variable loss of [h] in stressed syllables initially before vowels' (1992, 197). [h] was not included in the analysis for reasons discussed in section 6.7 on methodological considerations below. With that said, [h] is a voiceless glottal fricative and characterised by its aperiodicity and particularly low intensity (Laufer 1991). The results of this thesis have shown the greatest effects on the low intensity fricatives e.g. /f/ and /θ/, which suggest that a phoneme like [h] would also be affected, and likely mistaken for noise. This would in turn lead to the transmission potentially mimicking /h/-dropping.

Additional perspectives are whether e.g. /ð/ and [v] become more similar, which is a known feature of African American Vernacular English (Thomas 2007), and whether the plosive-like structure observed for /ð/ and the effects of the codec compression might be linked to TH-stopping (Drummond 2018).

6.5 Forensic phonetics and perception

For forensic phonetics, these findings are particularly relevant, and whereas for other linguistic fields of research it may be possible to avoid these types of recordings and comparison between high quality studio recordings and codec compressed recordings, this is often not possible in forensic phonetic casework.

Here, the forensic phonetician is forced to work with the provided sound files and potentially incriminating recordings (Jessen 2018; Hughes et al. 2020). From this thesis, it is clear that the effects on the acoustic characteristics of fricatives are varied and highly dependent on segment, codec, and bitrate as well as whether the signal is live transmitted or not, where the latter will be the case in incriminating sound files.

The tendencies observed in the controlled conditions are similar to, but not identical to, what was observed in the live transmitted speech. Therefore, a high-quality recording cannot just be codec compressed and completely imitate what might be observed in the live transmitted recording. Therefore, exact guidelines on how to compensate for the effect of digital transmission cannot be given based on the current research. However, if the effects described here are kept in mind, and the recordings are considered in terms of the type of codec and bitrate and assessed both quantitatively as well as qualitatively, as is common practice (see Foulkes and French 2012) most effects will be clear to the phonetician.

Related to forensic phonetic casework, an equally important aspect of this type of speech is its perception. Whether the acoustic changes observed in this thesis, are enough to affect perception will require additional research. However, as any perception is based on the acoustic information available, the two are inherently linked.

It is important to note that it is not possible to make any conclusions on perception based on the research presented in this thesis. However, it is a key aspect for any future research and for a number of the questions raised by the results. Therefore, it will briefly be considered here.

This project has shown significant changes to the spectral moments, added formant structure, and substantial reductions in frequency information and intensity. From this, it might be suggested that the perception is likely also affected because of the changed and limited acoustic information. The perception might in that way be more reliant on segmental context and the acoustic cues between the fricatives and the following segment, both of which are well-known feature in the subconscious process of phonemic restoration (see Kashino 2006; R. M. Warren and Obusek 1971; Samuel 1987). Thus, in forensic phonetic analysis, if it is the case that a significant amount of understanding and interpreting digitally transmitted speech is reliant on perceptual mechanisms such as phonemic restoration and top-down information e.g. context, this could potentially lead to biased and/or incorrect interpretations (see Fraser 2003; Fraser, Stevenson, and Marks 2011; Fraser 2018).

Moreover, this is a fact, which is essential to investigate further, because the quality of the signal especially for people with hearing impairments, or non-native speakers, are essential, because they are more dependent on the acoustic information for correct perception. This will be addressed further in section 6.7 on interdisciplinary perspectives.

6.6 Methodological considerations

This section will consider some of the methodological aspects and choices in this thesis, and how these might be expanded or improved in future research.

Firstly, the speakers in the corpus, which was used throughout this thesis, were of different ages and five dialectal backgrounds. This means that a certain amount of sociolinguistic variation including idiosyncratic variations were present in the dataset. It was beyond the scope of this thesis to engage with these as the aim was a baseline of the effects of the codec compressions and digital transmission. With that said, these variations are important and the speaker and dialect specific effects of the digital transmission are a key aspect for further research.

Secondly, and also related to the corpus, is the fact that the thesis investigated a selected set of English fricatives and excluded [v], [h], and [ʒ]. [v] and [h] due the large amount of variation in the production of these two sounds (Johnson 2011), which would have been beyond the scope of this thesis to account for. [ʒ] was not included because of the low number of tokens of this fricative in the dataset. For future research, this means that these fricatives are still to be investigated. In addition, the fact that all the fricatives were English, does not mean that the results are not relevant for other languages, but simply that the effect needs to be investigated across languages and phonemic inventories.

Another aspect, which was also touched upon earlier in this thesis, is the use of read speech in comparison to spontaneous speech. The read speech gives a more controlled dataset, which was preferable for a baseline.

However, future research should include and investigate real life spontaneous speech because features such as the Lombard effect and elisions, known to occur in spontaneous speech, are not present in read pre-recorded files (see Nakamura, Iwano, and Furui 2008; Mehta and Cutler 1988). It should be kept in mind that using real life participants and spontaneous speech introduce a number of variables, which will vary dynamically and speaker specifically across the recording. Thus, before this type of research is undertaken, it is recommended to investigate and measure dynamic features in codec compressed speech under controlled conditions using Generalised Additive Model Models i.e. GAMs (Wood 2017). This includes e.g. spectral measures but beyond the central frame of the segments, and measures of e.g. changes in amplitude dynamically across the segments.

For comparative reasons and to single out the effect of the codec compression rather than bandwidth specific features, the original high quality studio WAV files were down-sampled to 16 kHz in the studies in Chapters 3 and 4. In future, it would be useful to compare the original 44.1 kHz WAV files to the down-sampled 16 kHz WAV files using the same spectral measures. This would indicate any changes to the sound files caused by the initial down-sampling and get a better indication of a real life scenario, while singling out the effect of the transmission.

Related to bandwidth, the multitaper analysis caused some unexpected maximum values for the peak frequencies in the live transmission. Here the 44.1 kHz files were analysed, but with an upper frequency limit for the analysis at 8 kHz, which caused the values to partly be an artefact of the frequency bins in the multitaper analysis. These values were primarily used to assess whether the peak was in fact above 8 kHz, which would indicate a good network access and access to a good bitrate quality. Therefore, this was not critical to the present project, but should be addressed in any future potential replication of this research.

The present study only investigated live transmission using one of the codec compressions, namely AMR-WB with a limited number of speakers, in only one location with one type of background noise, and using only one phone model. In that way, further research is needed for a) other codecs e.g. Opus; b) more speakers as well as speakers recorded live; c) different locations with different types and levels of background noise; and d) different types of hardware i.e. phone models and computers.

Overall, with research into these various aspects of live transmission, it would be possible to draw more definite conclusions on the linguistic implications of codec compression and digital transmission on speech, as they have previously been discussed in this chapter. In addition, it would allow more accurate improvements to the various speech codecs from a linguistic perspective, which would both improve perception and user experience of the various speech technologies.

Lastly, the number of spectrographic examples used for this study were randomly generated and limited in number, which was also the reason the examples were not of the same segments across the chapters. The latter meant that the examples were not as readily comparable, but on the other hand illustrated a wider variation of the potential effects of the codec compression and digital transmission. The decision to only investigate a selection of spectrograms was made considering the main objective of the thesis was the quantitative and inferential analysis. However, for future research more spectrographic and qualitative analysis might prove extremely useful in clarifying the acoustic implications of digital transmission, not only on fricatives. This is illustrated by the vast amount of phonetic variation observed during the spectrographic analysis, especially for the voiced fricatives, which broadly underlines the importance of this type of qualitative analysis.

6.7 Future and interdisciplinary perspectives

Some future perspectives for research have already been provided throughout this chapter. However, this section will focus on broader perspectives and more directly address the interdisciplinary relevance of this thesis. These are all also more directly linked to mainRQ3, which was concerned with how the results might inform linguistic research, but also how the results might be useful beyond linguistics.

6.7.1 Remote data collection

One of the potential implications and future perspectives of the research done in this thesis, which were mentioned in the initial chapters related to remote data collection. There are two primary types of remote data collection, where either the data is transmitted and then recorded from the receiving device (e.g. Labov 2000; Smorenburg and Heeren 2019), or where the recording is administered via a digital connection, but the recording is made locally to the participant on a separate device monitored by the participant (e.g. Leemann et al. 2020).

For the former i.e. scenario a, the speech signal is affected by both hardware and transmission, whereas in the latter i.e. scenario b, the signal is not affected by transmission, but rather the hardware and any potential apps used to store and sent the collected data.

The research in this thesis has been concerned with scenario a, and more specifically digital transmission.

The live transmission study in Chapter 5 was done using the AMR-WB codec, and showed substantial changes especially to the voiceless and less intense fricatives. What is particularly interesting in this perspective, is that these effects were observed in all three conditions, including the condition, where the signal was recorded directly via a cable plugged into the receiving mobile phone's combined head- and microphone socket. This means that even without background noise or the influence from the surrounding environment and loudspeaker of the receiving phone, the signal, and here the fricatives specifically, were still affected. Thus, the effects of live transmission cannot be expected to be limited by controlling this factor.

Whether these effects are of a magnitude to affect the research findings will depend on the specific research in question. For example, if the aim of the research is to investigate spectral characteristics, changes of over 40 percent to the spectral measures as they were observed here, will have a substantial impact on the validity of the results. On the other hand, if the goal of the research is to investigate perceptual effects or related to broader e.g. phonological categories, the effects are potentially less influential. However, this will require further and more specific research on e.g. the perception of live transmitted speech to establish for sure.

The AMR-WB codec is used for mobile phone communication, which means that the effect of digital transmission e.g. via Skype of Zoom, where Opus is used (Zoom North America 2020; S 2018), is still to be investigated. From the controlled conditions, it was clear that if the quality is good i.e. the bitrate is high, the effects are limited for Opus and depending on the segment, in some instances not significant. In that way, based on this thesis, this would suggest that Opus compressed data is potentially usable for remote data collection. However, this is keeping in mind that the results do not include the effects of hardware e.g. computer microphone and loudspeaker. In addition, Opus uses VoIP i.e. Voice Over Internet Protocol during live transmission, which introduces another set of factors including network access, sophisticated noise reduction, and package loss (see Chakraborty, Misra, and Prasad 2019). Thus, even though Opus performed well under the controlled conditions, this cannot be taken to mean that it will perform equally well during live transmission.

Consequently, whether remote data collection overall is advisable using digitally transmitted speech over long distances requires a larger study including different hardware and codecs. Until such research is available, based on this thesis, remote data collection in scenario a, is not recommended if the purpose is spectral analysis.

Another branch of remote data collection, which is related to the MP3 codec and does not include actual transmission i.e. scenario c, is the download of material e.g. from films or interviews from websites such as YouTube (Chen, Hurley, and Karim 2005). This is a perspective, which has not received much attention throughout this thesis, but which is nevertheless important if data is downloaded and analysed e.g. for sociolinguistic purposes.

The results of the effects of the MP3 compression were as mentioned very dependent on bitrate, and particularly the low bitrate presented substantial and significant changes to the fricatives. However, in the average and high bitrates, the effects were far more limited. Thus, if only looking at the MP3 compression alone, it suggests that files can be used if the bitrate is of a certain level, which should be assessed by the research before the data is analysed.

However, it is also possible that the files are double compressed, which also relates back to scenario b, where the files are stored and potentially sent via an app. Double compression, as the name suggests, means that the signal has been compressed more than once. In forensic contexts, this is important because the compression makes it possible to edit a file, and via a second compression hide the edits (Wang et al. 2014). For this reason research has also been conducted on how to detect this and ensure the authenticity of MP3 compressed files (e.g. Wang et al. 2014).

More concretely, in the example of YouTube, the video or audio, which is downloaded for example for sociolinguistic research, has potentially already been compressed once before by the person who uploaded the file. This both means that the person might have edited the file, but also that the compression has happened twice and affected the acoustic composition of the files, which is the reason this can be detected via complex algorithmic analysis (Yan et al. 2018).

As a result, the effects of the MP3 compression are complex to assess in the perspective of remote data collection. Regardless, it again underlines how researchers need to assess the technological aspects of the data critically e.g. in terms of bitrate and third party apps used for e.g. storage as these all apply different levels of compression. In the perspective of this thesis, this means that the effects observed on the MP3 compressed files might be different or enhanced in these remote data collection scenarios.

6.7.2 Charisma and gender

A field of research, which span both industry and linguistics, is charisma studies. In this field, studies have been conducted on how digital transmission of speech affects the level of perceived charisma of a speaker, which has been correlated with clarity of speech as well as CoG measures (e.g. Siegert and Niebuhr 2021; Niebuhr, Skarnitzl, and Tylečková 2018). This thesis has clearly shown how e.g. intensity and spectral characteristics are reduced for these fricatives, which is likely to affect clarity of speech. This in turn would affect the level of perceived charisma and emphasise the relevance of segment specific analysis e.g. on consonants (see Siegert and Niebuhr 2021).

Female speakers have also been shown to be more affected by the digital transmission when investigating perceived charisma, which is likely based on the acoustic implications of the digital transmission. In addition, the effects of the digital transmission are also expected to be more prominent for female speakers (Niebuhr and Siegert 2021). This is potentially because female speakers generally have a higher F0, and more energy centred in the higher frequencies e.g. the spectral peak for /s/ for female English speakers have been found to be as high as 12 kHz (Stuart-Smith et al. 2019). If this is the case, children are, for the same reason, also likely to be more affected by the transmission than male speakers are.

Again, with the increased amount of remote work and digitally transmitted speech in the workplace, this perspective only becomes more relevant.

6.7.3 Diffusion of sound changes

From the perspective of future research for sociolinguistics, the potential of diffusion of sound changes via digitally transmitted speech is still to be considered. In Chapter 1 and 2, it was presented how potential acoustic artefacts of the digital transmission might diffuse across and between speech communities via social networks. Unfolding this with the results of this thesis in mind, requires some consideration of the different approaches to social networks and diffusion of sound changes.

In brief, the traditional social network theories as described by Milroy and Llamas (2013) include some level of face-to-face interaction as a requirement for the ties between networks to work as vehicles for sound changes. In newer research, Stuart-Smith et al. have shown how linguistic features can spread via TV (2013). This was to limited extents, but the fact that this type of diffusion of variants are even possible underlines how interaction is not a requirement. Moreover, it illustrates how digital transmission, which include actual social interaction, also has the potential to diffuse variants and sound changes.

In addition, researchers have developed models that encompass this fact by investigating the influence of neutral change i.e. change based on frequency of occurrence (Kauhanen 2016; Bermúdez-Otero 2017). In addition, Sayers suggested the mediate innovation model in which the digital devices function as the medium for the sound changes and variants to travel between social networks (Sayers 2014). In the networks, the changes are then diffused via ties as initially suggested by Milroy and Llamas.

With the increased amount of digitally transmitted speech, the exposure to this type of speech is increased (2014). Thus, mobile phones and computers have the potential to diffuse sound changes following the neutral and mediate models (International Telecommunication Union (ITU) 2022a; 2022b; Hargrave 2020).

This is in two ways. Firstly, the digital transmission might work as a vehicle to spread already existing dialectal or sociolinguistically unique variants between social networks regardless of geographic proximity. Secondly, the codec compression and transmission might produce or mimick sociolinguistic variants, which is then diffused. Only the latter has been touched upon in this thesis.

The results have shown some substantial acoustic changes to the fricatives in question, but has not been able to confirm conclusively whether specific dialectal phenomena are mimicked by the digital transmission. However, the effects were at times substantial enough for the fricatives not to be encoded, which means the effects are of a magnitude to be sociolinguistically relevant. To what extent the digital transmission has been or will be the vehicle for specific sound changes, is a question that must be addressed in more large-scale sociolinguistic studies.

6.7.4 Speech and hearing technology

The linguistic knowledge on the implications of digital transmission, is also useful in the development of various speech technologies including speech codecs, but also speech to text technologies, where the speech recognition technology needs to be sensitive to the acoustic input to enable the correct transcription (see Liu et al. 2020). In that way, by better understanding the acoustic implications on the individual segments, it will be possible to improve the technology and provide more accurate results. An example of this could be the codecs' sensitivity to the acoustically inherent intensity of the segments, where the low intensity segments such as /f/ and /θ/ were substantially more affected that e.g. /s/. From this knowledge, it would be possible to adjust the algorithms to be more sensitive to this specific aspect, rely more on information from the transition into the following sound or simply make the user aware that clarity of speech is essential particularly related to these type of sounds.

Another perspective can be made to hearing aids and cochlear implants. These also use digital technology to translate and amplify the sound waves produced by a speaker via a digital device directly to the user, and implement noise reduction technology including noise reduction algorithms (see Dhawan and Mahalakshmi 2016; Chung, Zeng, and Waltzman 2004). For example, for cochlear implants wireless remote microphone technology is developed to communicate directly with the user and eliminate extraneous noise (Wesarg et al. 2020). Thus, these technologies have similarities to what is found in the digitally transmitted speech described here by also involving a selection process of frequency information. In addition, acoustics and perception are inherently linked as the acoustic input is the determining factor for what is perceived. In that way, knowledge about the acoustic effects of technological developments such as codecs can help clarify how and why perception might be affected or improved.

Thus, applying the same type of analysis frame to better understand what acoustic information is essential for accurate perception will be beneficial for the research and development of hearing aid and cochlear implant technology.

6.8 Conclusion

In conclusion, one of the main motivations for this thesis was to give a better understanding of digitally transmitted speech by investigating the specific acoustic effects on fricatives, while aiming to provide advice for linguists working with or considering working with this type of speech.

Firstly, the combination of the significant findings for the spectral moments as well as frequency peak together with the observed spectrographic changes suggest that the effects of codec compression particularly in the low and average bitrates have an impact across fields of linguistics. This is because digital transmission makes otherwise reliable acoustic measures less reliable in this type of codec compressed and digitally transmitted speech. An example is how the spectral measures including CoG, which have been found to distinguish e.g. sibilants from non-sibilants, have been found to be significantly lowered and make the fricatives almost identical e.g. in the average bitrate in comparison to high quality WAV files. Moreover, this thesis has shown how across bitrates, Opus generally performed the best based on the magnitude of the changes for any of the spectral measures. MP3 showed very similar effects to Opus in the average and high bitrate, but was the worst performing codec in the low bitrate. This leaves AMR-WB, which in the average and high bitrates was the worst performing codec, but in the low bitrate performed better than the MP3.

Finally, the three studies conducted as part of this thesis all point to the same two main conclusions to keep in mind when engaging with digitally transmitted speech. These are: a) direct comparison between high quality studio recordings and codec compressed speech should be avoided if the aim is not the comparison between them; and b) the effect of the codec compressions are both segment and bitrate dependent, increased in live transmission and must be evaluated based on these parameters.

More data is needed to conclude exactly how the codec compression might be compensated for in linguistic research, but the conclusions from the previous Chapters are repeated here: it is essential to be cautious when using digitally transmitted speech for data-collection especially for acoustic and segmental purposes (e.g. Sanker et al. 2021, Leemann et al. 2020, Siegert and Niebuhr 2021). Taken together, this thesis has illustrated the interdisciplinary relevance of understanding the acoustic implications of digital transmission on fricatives as well as other speech sounds.

Appendix A

a.1 The Chicken Little story (Best et al. 2012)

One day Chicken Little was pecking at the earth under the acorn tree, looking for tasty worms, when he felt a sharp WHACK on his fluffy head.

“Ouch! Oh dear! Disaster!” cried Chicken Little, who was a bit of a drama queen. “The sky is falling!” His head hurt, and he could feel a big, painful bump on it.

“I’d better warn everyone!” he squawked. And off he raced, in a panicked cloud of fluff.

He found Plucky Ducky doing backstroke in the pond.

“Get out of the bath! Alert the authorities!” shrieked Chicken Little. “The sky is falling!”

“Ahoy there!” quacked Plucky Ducky, puffing out her chest. “Stay calm. I’ll save you. I’m not scared of anything. Remember that time that rabid dog–”

“No time to waste! Or bathe!” shrieked Chicken Little. “A chunk of sky just bruised my head. It’s starting to throb – look at this bump!”

Chicken Little wasn’t playing around: Plucky Ducky could see the swelling under his feathers.

“Oh boy,” she said. “Never fear, I’m no coward. I’ll get us out of this alive. Come on! We’d better grab the others.”

Hurrying down the path, the pair met Weepy Sheepy, chewing gloomily on some clover.

“Run!” screeched Chicken Little. “Mass evacuation, the sky is falling. Look at my poor head!”

Weepy Sheepy burst into tears. “Please don’t raise your voice at me,” he said, sobbing. “I’m feeling a bit delicate.”

“Have courage,” said Plucky Ducky sympathetically, “I’ll protect you. No need to nominate me for a medal or anything, this is just my method of operation. Join us! Make haste!”

“Oh help, I’m scared,” wailed Weepy Sheepy, trotting after them. “Has anyone got a tissue?”

Down in the woods they found Perky Turkey, picking flowers.

“Hi-ho!” grinned Perky Turkey, who was always happy. “Hello all! What a lovely day.”

“It’s not a lovely day at all!” screamed Chicken Little, hopping around madly. “It’s the apocalypse! The sky is falling!”

“What fun!” said Perky Turkey cheerfully. “You’re so sweet, dear Chicken Little, always off having adventures. You really brighten things up.”

“It’s an international disaster!” bawled Weepy Sheepy. “We’re all doomed.”

“Goodness, things can’t be that bad,” smiled Perky Turkey. “Soothe yourselves. Let’s look on the bright side.”

“There is no bright side,” squeaked Chicken Little. “I’m saying the whole sky’s about to crash down.”

“Be brave! Stay tough! We will endure,” assured loyal Plucky Ducky. “I’m coordinating the evacuation. You’d better come with us.”

“Sure,” said Perky Turkey joyously. “Count me in, it sounds like a hoot.” And off they went. Down by the bog, they found Biggy Piggy lying in a bath of frothy mud.

“The sky is falling!” screeched Chicken Little. “A big piece just hit me in the head!”

Biggy Piggy snorted. “Rubbish. You’re always panicking, you paranoid fowl. Apparently you got hay-fever once, and told everyone it was bird flu.”

“But it’s true – honestly, it’s the end of the world,” wept Weepy Sheepy. He was so upset he could barely breathe.

Biggy Piggy scowled. “Pathetic! Don’t snivel, you big, immature wuss. You’re always having a blub. Another false alarm. This is getting beyond a joke.” Weepy Sheepy’s lip began to tremble.

“I’m afraid they’re right, my friend,” said Plucky Ducky. “But fear not. My plan is to relocate everyone to a safe underground bunker, with enough food to last for months.”

Biggy Piggy struggled to sit up; he wasn’t exactly lithe. “Food for months?” he grunted. “Hmm, I do like my grub. Fine, I’ll come, if these two bright sparks don’t annoy me.”

“See!” said Perky Turkey, clapping her wings. “Every cloud has a silver lining!”

“You can shut your beak too,” glowered Biggy Piggy as they all ran away down the lane.

Beside the flowing brook, the animals found Groggy Froggy sitting blearily on a log.

“Evacuate! Evacuate!” squealed Chicken Little. “The sky is falling!”

Groggy Froggy winced. “Some sympathy please – could you keep the noise down?” he said. “I’ve got the worst hangover. I think I’ve poisoned my liver.”

Plucky Ducky stepped up. “Pull yourself together, Froggy. We’ve come to save you.”

“The world’s ending,” sniffled Weepy Sheepy. “It’s Doomsday.”

Groggy Froggy looked confused. “Am I dreaming?” he asked woozily. “I thought today was Thursday. Where am I?” He couldn’t remember anything.

Biggy Piggy stamped his trotter, in a huff. “The brain of this frog is pickled,” he snapped. “He belongs in rehab. Let’s leave him here to sober up, and get to that bunker in time for supper.”

“No,” said Plucky Ducky, loyal as ever. “We’re all in this together. Enough talk, Froggy – it’s join us or perish, you have no choice.”

Flopping down beside them, Groggy Froggy gave a loud, boorish burp. “All right,” he said vaguely. “But can we stop at the pub? Alcohol is great for stress.”

Further down the lane they found Gloaty Goaty.

“The sky is falling!” screeched Chicken Little. “The universe is exploding! The end is near!”

Gloaty Goaty sniggered. “Oh, terrific,” he said. “You’re having another panic attack. You’re so entertaining when you get in a flap.”

“I’m not playing around!” yelled the frazzled Chicken. “Look at this bump!” He gave his head a rub.

“Now now,” gobbled Perky Turkey happily, as blithe as ever. “It might be the Apocalypse, but I have faith that everything will be just fine. Let’s keep our chins up.”

Gloaty Goaty laughed. “Sure. And how many chins have you got, turkey? I can count five – the royal flush!”

Plucky Ducky waggled a wing with authority. “Behave, Goaty! No time for sniping. I’m coordinating this rescue operation. Join us!”

Gloaty Goaty snickered. “Bravo! What a pithy summary. You’re hilarious. Are you our saviour, then?”

“Some people call me a hero for my efforts,” admitted Plucky Ducky, bowing modestly, “but I think that’s going a bit overboard.”

Biggy Piggy rolled his eyes. The delay was making him seethe. “Can we go? I’m starving.”

“I’m really dehydrated,” pleaded Groggy Froggy. “I need to procure some liquid refreshment.”

"Oh boy, you nutcases make me feel normal," crowed Gloaty Goaty. "I always enjoy a good disaster. Sign me up for the Apocalypse club."

At the end of the lane, the animals saw a big white van parked at the crossroads ahead. Beyond it stood a powerfully built man in blood-stained white cloth overalls, squinting at a map.

"Civil emergency!" blared Chicken Little. "Alert the media! Call a lawyer!"

The man peered at them short-sightedly. He'd lost his glasses, but he knew a chicken when he saw one. And a sheep. And a goat. Oh yes, he knew his animals alright.

"What's all this palaver?" said the man with the van. "Looks like someone left the farm gate open."

"The sky is falling!" gasped Chicken Little.

"Is it now?" said the man thoughtfully, rubbing his chin. "Can I do anything to help?"

"We're heading for shelter," said Plucky Ducky. "There's no time to waste."

"Aha," said the man slyly. "Shelter, you're saying? Can I offer you a lift? I know a safe place. It's not far away."

"Um… You know the way to the bunker?" asked Plucky Ducky apprehensively, suddenly realising she hadn't really thought this through.

"Well sure, it's like a bunker," said the man, his voice smooth. "It's definitely secure. And I know a shortcut that'll get you there in a jiffy. How about hopping in the back of my van?"

As the nervous animals crowded together, Perky Turkey looked at the vehicle. It had pretty lettering painted along one side.

She spelled the word under her breath: "B – U – T – C – H – E…" The last letter was hidden behind the man. The turkey froze, her mouth agape.

"Guys, look!" she hissed urgently, pointing at the ghastly word. "It's a ploy! He's trying to lure us to our death!"

One by one, the animals began mouthing the letters, and began to writhe and tremble in pure, genuine, heartfelt fear – all except Groggy Froggy, the little grub, who was only half-awake, with dribble on his chin.

"So what do you say?" said the man, peering at them. "Shall we all head off?"

But the animals were frozen in terror. They cowered together, hardly dared to breathe.

"Don't move," whispered Plucky Ducky. "It's a trap!"

Chicken Little shook so hard his beak began to rattle. Then, with one especially violent tremble, something popped out of his frothy feathers and fell onto the road with a CLUNK.

The animals stared down at it. There on the road lay a big brown acorn. They looked at Chicken Little. The bump on his head was gone.

At last, they realised: the sky was not falling at all. But now they were in real trouble.

"You silly, thick-headed chicken," growled Biggy Piggy. "I really loathe you bird-brained types!"

The man squinted down. "What is that thing?" he asked suspiciously.

Plucky Ducky thought fast. "It's a hand grenade!" she shouted, grabbing the acorn with one webbed foot. "A highly accurate and lethal miniature explosive! You'll be safer behind me, animals!"

The man stepped back. "Don't throw it at me," he said. "I'm just trying to help." He peered at the object. Maybe they were bluffing. But it did look like a tiny hand grenade.

"Get back! We see your evil plot!" yelled Plucky Ducky. "Vamoose, or I'll blow you to smithereens!"

The man hesitated. These animals were crazy – better safe than sorry. He jumped in his van, revving up the engine.

"Next time I'll bring my chopper and put you all on the slab! I'll make mincemeat out of the lot of you!" he yelled furiously as he sped away down the path leaving a puff of smog.

As the animals watched him go, they sighed with relief.

Weepy Sheepy began to sob. "What a nightmare!" he howled. "I think I'm having a nervous breakdown. I need a hug."

"Ahem. No need to reward me for saving your lives," declared Plucky Ducky proudly. "Please don't bother. But I suppose I can't avoid getting that medal now."

"Oh, joy!" said Perky Turkey. "Wasn't I saying everything would turn out for the best?"

Gloaty Goaty gave a gleeful laugh.

"Can you believe that guy?" he said. "Humans! They're just SO gullible!"

Appendix B

b.1 words added to the Montreal Forced Aligner’s English UK dictionary

Word	**Transcription**
ahoy	ə h ɔj
apprehensively	æ p ɹ ɪ h ɛ n s ɪ v ʎ i
bawled	b ɒː l ə d
blearily	b ʎ ɪ ɹ ə ʎ i
brained	b ɹ ej n d
burp	b ɜː p
frazzled	f ɹ æ z ɬ̩ ə d
gloaty	ɡ l əw t i
gloomily	ɡ l ʉ mʲ ə ʎ i
glowered	ɡ l əw ə d
gobbled	ɡ ɒ b l ə d
mouthing	m aw ð I ŋ
nutcase	n ɐ t cʰ ej s
panicking	pʰ æ ɲ ɪ k I ŋ
revving	ɹ ə v I ŋ
sheepy	ʃ i p ə
shrieked	ʃ ɹ iː k ə d
sniffled	s ɲ ɪ f ʎ ə d
sniggered	s ɲ ɪ ɡ ə d
snorted	s n ɒː t ə d
sightedly	s aj ʔ ɪ d ʎ i
squeaked	s k w iː k ə d
squealed	s k w iː ɫ ə d
squinted	s c w ɪ n t ə d
squinting	s c w ɪ n t I ŋ
vamoose	v æ m u s
waggled	w æ ɡ ə ɫ d
wailed	w ej l ə d
winced	w ɪ n s ə d
woozily	w ʉː z i ʎ i

Reference List

3GPP. 2007. ‘Technical Specification Group Services and System Aspects; ANSI-C Code for the Adaptive Multi Rate - Wideband (AMR-WB) Speech Codec’. Technical specification TS 26.173. V7.0.0. France.

———. 2018. ‘Technical Specification Group Services and System Aspects; Mandatory Speech CODEC Speech Processing Functions; AMR Speech CODEC; General Description’. Technical specification 26.071 V15.0.0. France. https://www.3gpp.org/ftp/Specs/archive/26_series/26.071/26071-f00.zip.

———. 2020a. ‘3GPP TS 26.094’. Technical specification 16.0.0. Technical Specification Group Services and System Aspects; Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Voice Activity Detector (VAD). France: 3GPP.

———. 2020b. ‘3GPP TS 26.194’. Technical specification 16.0.0. Technical Specification Group Services and System Aspects; Speech Codec Speech Processing Functions; Adaptive Multi-Rate - Wideband (AMR-WB) Speech Codec; Voice Activity Detector (VAD). France.

———. 2022. ‘Performance Characterization of the Adaptive Multi-Rate Wideband (AMR-WB) Speech Codec (Version 17.0.0).’ Specification TR 26.976. France.

Alzqhoul, Esam A. S., Balamurali Nair, and Bernard J. Guillemin. 2012. ‘Speech Handling Mechanisms of Mobile Phone Networks and Their Potential Impact on Forensic Voice Analysis’. In *13th Australian International Conference on Speech Science & Technology*. Sydney.

Atal, B.S. 2006. ‘The History of Linear Prediction’. *IEEE Signal Processing Magazine* 23 (2): 154–61. https://doi.org/10.1109/MSP.2006.1598091.

Ayyad, Hadeel Salama, B. May Bernhardt, and Joseph P. Stemberger. 2016. ‘Kuwaiti Arabic: Acquisition of Singleton Consonants: Kuwaiti Arabic: Acquisition of Singleton Consonants’. *International Journal of Language & Communication Disorders* 51 (5): 531–45. https://doi.org/10.1111/1460-6984.12229.

Bailey, George. 2016. ‘Automatic Detection of Sociolinguistic Variation Using Forced Alignment’. *U. Penn Working Papers in Linguistics* 22 (2).

Bailey, George, Stephen Nichols, Danielle Turton, and Maciej Baranowski. 2022. ‘Affrication as the Cause of /s/-Retraction: Evidence from Manchester English’. *Glossa: A Journal of General Linguistics* 7 (1). https://doi.org/10.16995/glossa.8026.

Baker, Adam, Diana Archangeli, and Jeff Mielke. 2011. ‘Variability in American English S-Retraction Suggests a Solution to the Actuation Problem’. *Language Variation and Change* 23 (3): 347–74. https://doi.org/10.1017/S0954394511000135.

Bauer, Laurie. 2002. *An Introduction to International Varieties of English*. Edinburgh: Edinburgh University Press.

Behrens, Susan J., and Sheila E. Blumstein. 1988. ‘Acoustic Characteristics of English Voiceless Fricatives: A Descriptive Analysis’. *Journal of Phonetics* 16: 295–98.

Bellard, Fabrice, and FFmpeg Team. 2000. ‘FFmpeg’.

Bermúdez-Otero, Ricardo. 2017. ‘Individual Differences and the Explanation of Sound Change’. *Semantic Scholar*.

Besette, Bruno, and Redwan Salami. 2002. ‘The Adaptive Multirate Wideband Speech Codec (AMR-WB)’. *IEEE Transactions on Speech and Audio Processing* 10 (8): 620–36. https://doi.org/1063-6676/02$17.00.

Best, C. Shaw, G. Docherty, B. Evans, P. Foulkes, and J. Hay. 2012. ‘The You Came to Die?! Corpus. (ARC Discovery Project DP120104596).’

Blacklock, Oliver S. 2004. ‘Characteristics of Variation in Production of Normal and Disordered Fricatives, Using Reduced-Variance Spectral Methods’. PhD, England: University of Southhampton.

Boersma, P., and D. Weenik. 2021. ‘Praat: Doing Phonetics by Computer’.

Busso, Carlos, Sungbok Lee, and Shrikanth S. Narayanan. 2007. ‘Using Neutral Speech Models for Emotional Speech Analysis’. In , 2225–28. Belgium: ISCA. https://doi.org/10.21437/Interspeech.2007-605.

Byrne, Catherine, and Paul Foulkes. 2004. ‘The “mobile Phone Effect” on Vowel Formants’. *Speech, Language and the Law*, 83-102, 11 (1).

Chakraborty, Tamal, Iti Saha Misra, and Ramjee Prasad. 2019. *VoIP Technology: Applications and Challenges*. https://doi.org/10.1007/978-3-319-95594-0.

Chambers, J.M., and T.J. Hastie. 1992. ‘Statistical Models in S’. R package. Wadsworth & Brooks/Cole.

Chen, Steve, Chad Hurley, and Jawed Karim. 2005. 'YouTube'. Online video platform. YouTube.Com. 14 February 2005. https://www.youtube.com/.

Chung, King, Fan-Gang Zeng, and Susan Waltzman. 2004. 'Using Hearing Aid Directional Microphones and Noise Reduction Algorithms to Enhance Cochlear Implant Performance'. *Acoustics Research Letters Online* 5 (2): 56–61. https://doi.org/10.1121/1.1666869.

Cockos Incorporated. 2023. 'Reaper - Digital Audio Workstation'. English. New York: Cockos Incoporated. https://www.reaper.fm/.

Dan, Cudjoe. 2014. 'Review of Generations and Physics of Cellphone Technology'. *International Journal of Information Science*, 7.

Decker, Paul De, and Jennifer Nycz. 2011. 'For the Record: Which Digital Media Can Be Used for Sociophonetic Analysis?' *University of Pennsylvania Working Papers in Linguistics* 17 (2).

Del Tredici, Marco, and Raquel Fernández. 2018. 'The Road to Success: Assesing the Fate of Linguistics Innovations in Online Communities'. In *27th*, 1591–1603. Santa Fe. https://www.aclweb.org/anthology/C18-1135.pdf.

Dhawan, Ritwik, and P. Mahalakshmi. 2016. 'Digital Filtering in Hearing Aid System for the Hearing Impaired'. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 1494–97. Chennai, India: IEEE. https://doi.org/10.1109/ICEEOT.2016.7754932.

Drummond, Rob. 2018. 'Maybe It's a Grime [t]Ing: TH -Stopping among Urban British Youth'. *Language in Society* 47 (2): 171–96. https://doi.org/10.1017/S0047404517000999.

Eckert, Penelope. 2012. 'Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation'. *Annual Reviews of Linguistics* 41 (87): 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.

Ellis, Stanley. 1994. 'The Yorkshire Ripper Enquiry: Part I'. *International Journal of Speech Language and the Law* 1 (2): 197–206. https://doi.org/10.1558/ijsll.v1i2.197.

ETSI. 1992. 'Voice Activity Detection'. Recommendation GSM 06.32 V3.0.0. https://www.etsi.org/deliver/etsi_gts/06/0632/03.00.00_60/gsmts_0632sv030000p.pdf.

Fant, Gunnar. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton de Gruyter.

Foulkes, Paul, and Peter French. 2012. 'The Oxford Handbook of Language and Law'. In *Forensic Speaker Comparison: A Linguistic–Acoustic Perspective*, edited by Lawrence M. Solan and

Peter M. Tiersma, 558–72. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041.

Fraser, Helen. 2003. ‘Issues in Transcription: Factors Affecting the Reliability of Transcripts as Evidence in Legal Cases’. *International Journal of Speech, Language and the Law - Forensic Linguistics* 10 (2): 203–26. https://doi.org/10.1558/sll.2003.10.2.203.

———. 2018. ‘Forensic Transcription: How Confident False Beliefs about Language and Speech Threaten the Right to a Fair Trial in Australia’. *Australian Journal of Linguistics* 38 (4): 586–606. https://doi.org/10.1080/07268602.2018.1510760.

Fraser, Helen, Bruce Stevenson, and Tony Marks. 2011. ‘Interpretation of a Crisis Call: Persistence of a Primed Perception of a Disputed Utterance’. *International Journal of Speech Language and the Law* 18 (2): 261–92. https://doi.org/10.1558/ijsll.v18i2.261.

Freeman, Valerie, and Paul De Decker. 2021. ‘Remote Sociophonetic Data Collection: Vowels and Nasalization over Video Conferencing Apps’. *The Journal of the Acoustical Society of America* 149 (2): 1211–23. https://doi.org/10.1121/10.0003529.

French, Peter, Philip Harrison, and Jack Windsor Lewis. 2006. ‘The Yorkshire Ripper Hoaxer Trial’. *The International Journal of Speech, Language and the Law* 13 (2): 255–73. https://doi.org/10.1558/ijsll.2006.13.2.255.

Galov, Nick. 2023. ‘15 Skype Statistics, Facts & Trends in 2023’. WebTribunal. 20 May 2023. https://webtribunal.net/blog/skype-statistics/.

Garg, Vijay K. 2007. ‘CHAPTER 4 - An Overview of Digital Communication and Transmission’. In *Wireless Communications & Networking*, edited by Vijay K. Garg, 85–122. The Morgan Kaufmann Series in Networking. Burlington: Morgan Kaufmann. https://doi.org/10.1016/B978-012373580-5/50038-7.

Garrido, Iván, Susana Lagüela, Stefano Sfarra, and Pedro Arias. 2020. ‘Development of Thermal Principles for the Automation of the Thermographic Monitoring of Cultural Heritage’. *Sensors* 20 (12): 3392–3412. https://doi.org/10.3390/s20123392.

Gayer, Marc, Markus Lohwasser, and Manfred Lutzky. 2003. ‘Implementing MPEG Advanced Audio Coding and Layer-3 Encoders on 32-Bit and 16-Bit Fixed-Point Processors’. *Journal of the Audio Engineering Society*, 7.

Goode, B. 2002. ‘Voice over Internet Protocol (VoIP)’. *Proceedings of the IEEE* 90 (9): 1495–1517. https://doi.org/10.1109/JPROC.2002.802005.

Gordon, M., and K. Sands. 2002. ‘A Cross-Linguistic Acoustic Study of Voiceless Fricatives’. *Journal of International Phonetic Association* 32 (2): 141–74. https://doi.org/10.1017/S002510030200102.

GRAS Sound & Vibration. 2023. ‘GRAS 45BC KEMAR Head & Torso with Mouth Simulator, Non-configured’. GRAS an axiometrix Solutions Brand. 2023. https://www.grasacoustics.com/products/head-torso-simulators-kemar/kemar-non-configured/product/749-45bc.

Guillemin, Bernard J., and Catherine I. Watson. 2006. ‘Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification’. In *11th Australian International Conference on Speech Science & Technology*, 483–88. Auckland University: New Zealand.

———. 2008. ‘Impact of the GSM Mobile Phone Network on the Speech Signal: Some Preliminary Findings’. *Journal of Speech, Language and the Law* 15 (2): 193–218. https://doi.org/10.1558/ijsll.v15i2.193.

Hargrave, Sean. 2020. ‘In the Grip of a Climate Crisis, Demand for Video Calls Is Soaring’. *Wired UK*, 27 January 2020. https://www.wired.co.uk/article/video-conferencing-era.

Harrison, Philip. 2022. ‘Spectral Analysis Script’. MatLab. Natick, Massachusetts, United States: The MathWorks Inc.

Herre, Jürgen, and Sascha Dick. 2019. ‘Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review’. *Applied Sciences* 9 (14): 22. https://doi.org/10.3390/app9142854.

Hughes, Vincent, Philip Harrison, Paul Foulkes, Peter French, and Amelia J Gully. 2020. ‘EFFECTS OF FORMANT ANALYSIS SETTINGS AND CHANNEL MISMATCH ON SEMI-AUTOMATIC FORENSIC VOICE COMPARISON’. In *19th*, 5. Melbourne.

International Telecommunication Union (ITU). 2022a. ‘Statistics’. ITU.Int. 2022. https://www.itu.int:443/en/ITU-D/Statistics/Pages/stat/default.aspx.

———. 2022b. ‘Time Series of ICT Data for the World, by Geographic Regions, by Urban/Rural Area and by Level of DevelopMent’. ITU World Telecommunication/ICT Indicators database. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ITU_regional_global_Key_ICT_indicator_aggregates_rev1_Jan_2022.xlsx.

Jamieson, Donald G., Vijay Parsa, Moneca C. Price, and James Till. 2002. 'Interaction of Speech Coders and Atypical Speech I: Effects on Speech Intelligibility'. *Journal of Speech, Language, and Hearing Research* 45 (3): 482–93. https://doi.org/10.1044/1092-4388(2002/038).

Janevski, Toni, Scott Markus, and Milan Jankovic. 2017. *Quality of Service Regulation Manual*. Switzerland: International Telecomminication Union (ITU).

Jessen, Michael. 2008. 'Forensic Phonetics: Forensic Phonetics'. *Language and Linguistics Compass* 2 (4): 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x.

———. 2018. 'Forensic Voice Comparison'. In *Handbook of Communication in the Legal Sphere*, edited by Jaqueline Visconti, 219–55. Berlin: Walter de Gruyter.

Johnson, Keith. 2011. *Acoustic and Auditory Phonetics*. 3rd ed. Oxford: Wiley-Blackwell.

Jongman, Allard, Ratree Wayland, and Serena Wong. 2000. 'Acoustic Characteristics of English Fricatives'. *The Acoustical Society of America* 108 (3): 1252–63. https://doi.org/S0001-4966(00)02909-X.

Kashino, Makio. 2006. 'Phonemic Restoration: The Brain Creates Missing Speech Sounds'. *Acoustical Science and Technology* 27 (6): 318–21. https://doi.org/10.1250/ast.27.318.

Kauhanen, Henri. 2016. 'Neutral Change'. *Journal of Linguistics* 53: 327–58. https://doi.org/10.1017/S002222671000141.

Koenig, Bruce E. 2009. 'Forensic Authentication of Digital Audio Recordings'. *J. Audio Eng. Soc.* 57 (9): 34.

Kong, Eun Jong, and Jieun Kang. 2021. 'Age and Gender Differences in the Spectral Characteristics of Korean Sibilants*'. *Phonetics and Speech Sciences* 13 (1): 37–44. https://doi.org/10.13064/KSSS.2021.13.1.037.

Kottayi, Sivadasan, Raed Althomali, T. M. Thasleema, and N. K. Narayanan. 2016. 'Active Noise Control for Creating a Quiet Zone around Mobile Phone'. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, 0073–0077. Melmaruvathur, Tamilnadu, India: IEEE. https://doi.org/10.1109/ICCSP.2016.7754434.

Künzel, H. J. 2001. 'Beware of the "Telephone Effect": The Influence of Telephone Transmission on the Measurement of Formant Frequencies'. *Forensic Linguistics* 8 (1): 80–99.

Kuznetsova, A., PB Brockhoff, and RHB Christensen. 2017. ‘lmerTest Package: Tests in Linear Mixed Effects Models’. *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13.

Labov, William. 2000. ‘The Telsur Project at the Linguistics Laboratory’. Database. Www.Ling.Upenn.Edu/Phonoatlas. 2000. http:// www.ling.upenn.edu/phonoatlas.

Ladefoged, Peter. 1996. *Elements of Acoustic Phonetics*. 2. ed. Chicago, Ill: University of Chicago Press.

Laufer, Asher. 1991. ‘The “Glottal Fricative”’. *Journal of International Phonetic Association* 21 (2): 91–93.

Leemann, Adrian, Péter Jeszenszky, Carina Steiner, Melanie Studerus, and Jan Messerli. 2020. ‘Linguistic Fieldwork in a Pandemic: Supervised Data Collection Combining Smartphone Recordings and Videoconferencing’. *Linguistics Vanguard* 6 (s3): 20200061. https://doi.org/10.1515/lingvan-2020-0061.

Lenth, Russell. 2020. ‘Emmeans: Estimated Marginal Means, Aka Least-Squares Means’. R package. https://CRAN.R-project.org/package=emmeans.

Levon, Erez, and Sue Fox. 2014. ‘Social Salience and the Sociolinguistic Monitor: A Case Study of ING and TH-Fronting in Britain’. *Journal of English Linguistics* 43 (3): 185–217.

Levon, Erez, and Sophie Holmes-Elliott. 2013. ‘East End Boys and West End Girls: /S/-Fronting in Southeast England’. *University of Pennsylvania Working Papers in Linguistics* 19 (2).

Liu, Yuchen, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. ‘Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding’. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05): 8417–24. https://doi.org/10.1609/aaai.v34i05.6360.

Loeffler, John. 2021. ‘The History Behind the Invention of the First Cell Phone’. 24 January 2021. https://interestingengineering.com/the-history-behind-the-invention-of-the-first-cell-phone.

Lombard, E. 1911. In *Annales Des Maladies De L’oreille, Du Larynx, Du Nez Et Du Pharynx*, 37:101–19. France.

Luo, Shan. 2020. ‘Articulatory Tongue Shape Analysis of Mandarin Alveolar–Retroflex Contrast’. *The Journal of the Acoustical Society of America* 148 (4): 1961–77. https://doi.org/10.1121/10.0002111.

Maher, Robert C. 2018. *Principles of Forensic Audio Analysis*. Modern Acoustics and Signal Processing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99453-6.

Maniwa, Kazumi, Allard Jongman, and Travis Wade. 2009. 'Acoustic Characteristics of Clearly Spoken English Fricatives'. *The Journal of the Acoustical Society of America* 125 (6): 3962–73. https://doi.org/10.1121/1.2990715.

MathWorks Inc. 2010. 'MATLAB.' Natick, Massachusetts, United States.

McAuliffe, Michael, Michaela Socolof, Elias Stengel-Eskin, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. 'Montreal Forced Aligner'. Computer program. http://montrealcorpustools.github.io/Montreal-Forced-Aligner/.

McAuliffe, Michael, and Morgan Sonderegger. 2022. 'Mfa_english_uk_mfa_dictionary_2022'. {English (UK) MFA Dictionary. https://mfa-models.readthedocs.io/pronunciation dictionary/English/English (UK) MFA dictionary v2_0_0a.html.

Mehta, Gita, and Anne Cutler. 1988. 'Detection of Target Phonemes in Spontaneous and Read Speech'. *Journal of Language and Speech* 31 (2): 135–56. https://doi.org/10.1177/002383098803100203.

Milroy, Jim. 1992. 'Middle English Dialectology'. In *The Cambridge History of the English Language*, edited by Norman Blake, 2:156–206. Cambridge: Cambridge University Press.

Milroy, Lesley, and Carmen Llamas. 2013. 'Social Networks'. In *The Handbook of Language Variation and Change*, edited by J.K. Chambers and Natalie Schilling-Estes, 2nd ed., 409–27. West Sussex: John Wilay & Sons, Inc.

Milroy, Lesley, and James Milroy. 1985. 'Linguistic Change, Social Network and Speaker Innovation'. *Language in Society* 21 (2): 339–84. https://doi.org/10.1017/S0022226700010306.

———. 1992. 'Social Network and Social Class: Toward and Integrated Sociolinguistic Model'. *Language in Society* 21: 1–26.

Morrison, Geoffrey Stewart. 2016. 'INTERPOL Survey of the Use of Speaker Identification by Law Enforcement Agencies'. *Forensic Science International*, 9.

Muñoz-Mulas, Cristina, Rafael Martínez-Olalla, Pedro Gómez-Vilda, Agustín Álvarez-Marquina, and Luis Miguel Mazaira-Fernández. 2013. 'Gender Detection in Running Speech from Glottal and Vocal Tract Correlates'. In *Advances in Nonlinear Speech Processing*, edited by

Thomas Drugman and Thierry Dutoit, 25–32. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-38847-7_4.

Nakamura, Masanobu, Koji Iwano, and Sadaoki Furui. 2008. ‘Differences between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance’. *Computer Speech & Language* 22 (2): 171–84.

Niebuhr, Oliver, and Ingo Siegert. 2021. ‘Case Report: Women, Be Aware That Your Vocal Charisma Can Dwindle in Remote Meetings’. *Frontiers in Communication* 5. https://www.frontiersin.org/articles/10.3389/fcomm.2020.611555/full.

———. 2023. ‘A Digital “Flat Affect”? Popular Speech Compression Codecs and Their Effects on Emotional Prosody’. *Frontiers in Communication* 8. https://www.frontiersin.org/articles/10.3389/fcomm.2023.972182.

Niebuhr, Oliver, Radek Skarnitzl, and Lea Tylečková. 2018. ‘The Acoustic Fingerprint of a Charismatic Voice - Initial Evidence from Correlations between Long-Term Spectral Features and Listener Ratings’. In *9th International Conference on Speech Prosody 2018*, 359–63. ISCA. https://doi.org/10.21437/SpeechProsody.2018-73.

Nolan, Francis. 2001. ‘Speaker Identification Evidence: Its Forms, Limitations, and Roles’. In *Law and Language: Prospect and Retrospect*, 19. Levi, Finland.

Ohala, John. 1994. ‘Hierarchies of Environments for Sound Variation’. *Acta Linguistica Hafniensia* 27 (37): 371–82.

Öhman, Lisa, Anders Eriksson, and Pär Anders Granhag. 2010. ‘Mobile Phone Quality VS. Direct Quality: How the Presentation Format Affects Earwitness Accuracy’. *The European Journal of Psychologyl Applied to Legal Context* 2 (2): 161–81.

Pharao, Nicolai, and Marie Maegaard. 2017. ‘On the Influence of Coronal Sibilants and Stops on the Perception of Social Meanings in Copenhagen Danish’. *Linguistics* 55 (5). https://doi.org/10.1515/ling-2017-0023.

Prerau, MJ, RE Brown, JM Ellenbogen, and PL Patrick. 2017. ‘Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis’. *Physiology (Bethesda)* 32 (1): 60–92.

Raake, Alexander. 2006. *Speech Quality of VoIP: Assessment and Prediction*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470033005.

Radial Engineering. 2023. ‘X-Amp’. Radial Engineering. 2023. https://www.radialeng.com/product/x-amp.

RCore Team. 2020. ‘R: A Language and Environment for Statistical Computing’. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2019. ‘RStudio: Integrated Development for R’. Boston, MA: RStudio Inc. http://www.rstudio.com/.

Rumsey, Francis., McCormick, Tim,. 2009. *Sound and Recording*. Elsevier.

Rutter, Derek R. 1987. *Communicating by Telephone*. Edited by Michael Argyle. Great Britain: Pergamon press.

S, Susan. 2018. ‘What`s the Strategy about SILK Codec on Skype’. Answer. *Microsoft Community*. https://answers.microsoft.com/en-us/skype/forum/all/whats-the-strategy-about-silk-codec-on-skype/5135db3a-58df-4c74-a39d-207a0cf8fdb5.

Samuel, Arthur G. 1987. ‘Lexical Uniqueness Effects on Phonemic Restoration’. *Journal of Memory and Language* 26 (1): 36–56. https://doi.org/10.1016/0749-596X(87)90061-1.

Sanker, Chelsea, Sarah Babinski, Roslyn Burns, Marisha Evans, Jeremy Johns, Juhyae Kim, Slater Smith, Natalie Weber, and Claire Bowern. 2021. ‘(Don’t) Try This at Home! The Effects of Recording Devices and Software on Phonetic Analysis’. *Language* 97 (4): e360–82. https://doi.org/10.1353/lan.2021.0075.

Sauter, Martin. 2010. *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. John Wiley & Sons.

Sayers, Dave. 2014. ‘The Mediated Innovavtion Model: A Framework for Researching Media Influence in Language Change’. *Journal of Sociolinguistics* 18 (2): 185–212.

Schutze, Stephan. n.d. *St Kilda Road Traffic*. B-Format (AmbiX). Vol. Track 61. St Kilda Road: Urban.

Shadle, C.H. 1985. ‘The Acoustics of Fricative Consonants’. 506. Cambridge: Massachusetts Institute of Technology. http://asa.scitation.org/doi/10.1121/1.393552.

Shadle, C.H., and S.J. Mair. 1996. ‘Quantifying Spectral Characteristics of Fricatives’. In *ICSLP ’96*, 3:1521–24. Philadelphia, PA, USA: IEEE. https://doi.org/10.1109/ICSLP.1996.607906.

Siegert, Ingo, and Oliver Niebuhr. 2021. ‘Speech Signal Compression Deteriorates Acoustic Cues to Perceived Speaker Charisma’. In *Tagungsband Der 32. Konferenz*, 1–10.

Smorenburg, Laura, and Willemijn Heeren. 2019. ‘The Distribution of Speaker Information in Dutch Fricatives /s/ and /x/ from Telephone Dialogues’. *The Journal of the Acoustical Society of America* 147 (2): 949–60.

Son, Rob J. J. H. van. 2005. ‘A Study of Pitch, Formant, and Spectral Estimation Errors Introduced by Three Lossy Speech Compression Algorithms’. *ACTA ACUSTICA UNITED WITH ACUSTICA* 91: 771–78.

Soundfield by RØDE. 2023. ‘Ambisonic Sound Library’. Ambisonic Sound Library. 2023. https://library.soundfield.com/.

Strevens, Peter. 1960. ‘Spectra of Fricative Noise in Human Speech’. *Language and Speech* 3 (1): 32–49.

Stuart-Smith, Jane, Gwilym Pryce, Claire Timmins, and Barrie Gunter. 2013. ‘Television Can Also Be a Factor in Language Change: Evidence from an Urban Dialect’. *Language* 89 (3): 501–36. https://doi.org/10.1353/lan.2013.0041.

Stuart-Smith, Jane, Morgan Sonderegger, Rachel Macdonald, Jeff Mielke, Michael McAuliffe, and Erik Thomas. 2019. ‘LARGE-SCALE ACOUSTIC ANALYSIS OF DIALECTAL AND SOCIAL FACTORS IN ENGLISH /S/-RETRACTION’. In , 1273–77. Melbourne.

Stuart-Smith, Jane, Clair Timmins, and Fiona Tweedie. 2007. ‘Talkin’ Jockney’? Variation and Change in Glaswegian Accent’. *Journal of Sociolinguistics* 11 (2): 221–60.

Tagliamonte, Sali A. 2014. ‘Situating Media Influence in Sociolinguistic Context’. *Journal of Sociolinguistics* 18 (2): 223–32.

Tan, Lizhe, and Jean Jiang. 2019. ‘Chapter 10 - Waveform Quantization and Compression’. In *Digital Signal Processing*, 3rd ed., 475–527. Cambridge, Massachusetts: Academic Press.

Teac Europe GmbH. 2022. ‘Tascam iXZ | Mic/Guitar Interface for Smartphones And Tablet Computers’. Tascam Europe. 22 December 2022. https://tascam.eu/en/ixz.

The 3rd Generation Partnership Project. 2021. ‘About 3GPP’. 3gpp.Org. 2021. https://www.3gpp.org/about-3gpp.

Thomas, Erik R. 2007. ‘Phonological and Phonetic Characteristics of African American Vernacular English: Phonological and Phonetic Characteristics of AAVE’. *Language and Linguistics Compass* 1 (5): 450–75. https://doi.org/10.1111/j.1749-818X.2007.00029.x.

Torreira, Francisco. 2012. ‘Investigating the Nature of Aspirated Stops in Western Andalusian Spanish’. *Journal of the International Phonetic Association* 42 (1): 49–63. https://doi.org/10.1017/S0025100311000491.

Triton. 2022. ‘Choosing Audio Bitrate Settings’. Https://Tritondigitalcommunity.Force.Com. 2022. https://tritondigitalcommunity.force.com/s/article/Choosing-Audio-Bitrate-Settings?language=en_US.

Trousdale, Graeme. 2010. *An Introduction to English Sociolinguistics*. Edinburgh: Edinburgh University Press.

Trudgill, Peter. 1972. ‘Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich’. *Language in Society* 1 (2): 179–95. https://doi.org/10.1017/S0047404500000488.

———. 2014. ‘Diffusion, Drift, and the Irrelevance of Media Influence’. *Journal of Sociolinguistics* 18 (2): 214–22.

Valin, JM., K. Vos, and T. Terriberry. 2012. ‘Definition of the Opus Audio Codec’. RFC6716. RFC Editor. https://doi.org/10.17487/rfc6716.

Vodafone. 2023. ‘Vodafone – Our Best Ever Network | Now With 5G’. Vodafone. 2023. https://www.vodafone.co.uk/?icmp=uk~1_consumer~topnav~logo~vodafone_logo&linkpos=topnav~1.

Wang, Tianzhuo, Xiangwei Kong, Yanqing Guo, and Bo Wang. 2014. ‘Exposing the Double Compression in MP3 Audio by Frequency Vibration’. In *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, 450–54. Xi’an, China: IEEE. https://doi.org/10.1109/ChinaSIP.2014.6889283.

Warren, Richard M. 1970. ‘Perceptual Restoration of Missing Speech Sounds’. *Science* 167 (3917): 392–93. https://doi.org/10.1126/science.167.3917.392.

Warren, Richard M., and Charles J. Obusek. 1971. ‘Speech Perception and Phonemic Restorations’. *Perception & Psychophysics* 9 (3): 358–62. https://doi.org/10.3758/BF03212667.

Warren, Tom. 2020. ‘Zoom Admits It Doesn’t Have 300 Million Users, Corrects Misleading Claims’. The Verge. 30 April 2020. https://www.theverge.com/2020/4/30/21242421/zoom-300-million-users-incorrect-meeting-participants-statement.

Weenik, D., and P. Boersma. 2004. ‘Sound: To LPC (Burg)...’ Fon.Hum.Uva.Nl/Praat. April 2004. https://www.fon.hum.uva.nl/praat/manual/Sound__To_LPC__burg____.html.

Wesarg, Thomas, Yvonne Stelzig, Dan Hilgert-Becker, Bjorn Kathage, Konstantin Wiebe, Antje Aschendorff, Susan Arndt, and Iva Speck. 2020. ‘Application of Digital Remote Wireless Microphone Technology in Single-Sided Deaf Cochlear Implant Recipients’. *Journal of the American Academy of Audiology* 31 (04): 246–56. https://doi.org/10.3766/jaaa.18060.

Whitrow, James. 2019. ‘What Is the Difference Between VoIP and VOLTE? | Commsplus’. 15 November 2019. https://www.commsplus.co.uk/blog/what-is-the-difference-between-voip-and-volte.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.

Wood, Elizabeth. 2003. ‘TH-Fronting: The Substitution of f/v for 9/0 in New Zealand English’.

Wood, Simon N. 2017. *Generalized Adaptive Models*. 2nd ed. New York: Chapman & Hall/CRC.

Xiph.Org Foundation. 2022. ‘Opusenc/Opusdec’.

Yan, Diqun, Rangding Wang, Jinglei Zhou, Chao Jin, and Zhiefeng Wang. 2018. ‘Compression History Detection for MP3 Audio Diqun Yan, Rangding Wang, Jinglei Zhou, Chao Jin and Zhifeng Wang’. *KSII Transactions on Internet and Information Systems* 12 (2). https://doi.org/10.3837/tiis.2018.02.007.

Yost, William A. 2015. ‘Psychoacoustics: A Brief Historical Overview’. *Acoustics Today* 11 (3): 46–53.

Zhang, Cuiling, Geoffrey Stewart Morrison, Ewald Enzinger, and Felipe Ochoa. 2013. ‘Effects of Telephone Transmission on the Performance of Formant-Trajectory-Based Forensic Voice Comparison – Female Voices’. *Speech Communication* 55: 796–813.

Zhao, Sherry Y. 2010. ‘Stop-like Modification of the Dental Fricative⁄ð⁄: An Acoustic Analysis’. *The Journal of the Acoustical Society of America* 128 (4): 2009–20. https://doi.org/10.1121/1.3478856.

Zoom North America. 2020. ‘EVOLVED RECORDING THE ZOOM H5 Flexible Audio for Video Perfection’. Zoom-Na.Com. 2020. https://www.zoom-na.com/products/field-video-recording/field-recording/h5-handy-recorder.