Assessing a speaker’s voice quality for forensic purposes

Using the example of creaky voice and breathy voice

Katharina Klug

Doctor of Philosophy

University of York
Language and Linguistic Science

June 2023

Abstract

The research project explores ways to improve the assessment of voice quality (VQ) for forensic voice comparisons. Until today, a speaker's VQ is mainly assessed perceptually. However, the field has developed rapidly over the last two decades, prompting calls to objectify the analysis process by relying on voice acoustics instead. This poses a challenge as forensic audio recordings are degraded in several aspects.

The first study focuses on creaky voice (CV), which is particularly multifaceted in production and thus also in acoustics. Therefore, perceptually relevant categories must first be defined and tested before acoustic analysis can be conducted. A new CV classification scheme is conceptualised and tested. It is hypothesised that differences in speaker-specific CV spaces will facilitate speaker discrimination.

Using the example of breathy voice (BV), the second and fourth studies analyse the interplay between perception and acoustics. Spontaneous speech samples of BV speakers are compared with those of non-BV speakers under the studio condition and under the mobile phone condition. Under the studio recording condition, three parameters were found to correlate between perception and acoustics, i.e. *H1*-H2*, H1*-A1*, CPP*. Under the mobile recording condition, however, low frequency harmonics are attenuated and thus not meaningful. Therefore, the spectral tilt parameters of higher frequencies should be analysed instead.

The third study explores the suitability of f0 estimators with respect to *recording condition,* and *VQ*. Valid f0 estimation is required to obtain valid spectral slope measurements. The is explored using sustained cardinal vowels of one male and one female speaker in modal, breathy, and creaky VQ under two recording conditions (studio, mobile phone). Results allow for an informed decision which f0 estimator to use.

The research project sheds light on the needs and possibilities to refine VQ analysis for forensic application.

Contents

List of Figures

Introduction

Paper 1: Assessing creaky voice quality for forensic purposes

Paper 2: Analysing breathy voice in forensic speaker comparison using acoustics to confirm perception

Paper 3: Assessing the suitability of f0 estimators with respect to recording condition and voice quality

Paper 4: Analysis of breathy voice under mobile phone condition based on adequate f0 estimation

List of Tables

Paper 1: Assessing creaky voice quality for forensic purposes

Paper 2: Analysing breathy voice in forensic speaker comparison using acoustics to confirm perception

Paper 3: Assessing the suitability of f0 estimators with respect to recording condition and voice quality

Paper 4: Analysis of breathy voice under mobile phone condition based on adequate f0 estimation

Acknowledgements

Every doctoral thesis is a piece of teamwork. I am very proud of my team and extremely grateful for every single member. Here they are:

I am very grateful to my supervisors, Paul Foulkes, and Peter French, for teaching me how to become the best possible researcher. Thank you, Paul, for always being able to count on you and for your precision when I needed it most. Peter, you gave me courage when I was about to give up. I am very grateful for that.

Additionally, to my two official supervisors I am very proud to have two ‘undercover supervisors’: Christin Kirchhübel und Michael Jessen. Danke! Most of my studies are based on perceptual judgements. A massive thanks goes to my expert listeners Richard Rhodes, Jessica Wormald, Christin Kirchhübel, Paul Foulkes and Almut Braun. Phil Harrison mobile phone filtered the recordings for me – not one, not two, but three times. Cheers Phil, I owe you three pints! I had the privilege of working with Markus Niermann. Always a pleasure, Markus! Also, I am very proud to have had Vince Hughes as my TAP member. Ta Vince!

I am very grateful to Joshua Penney and Yen Shue. Without Joshua, I still wouldn’t be able to run REAPER in VoiceSauce. Thank you so much for providing me with the script and all the instructions (which Yen provided to you), and for all the clarifying emails. And whenever I’ve had problems with VoiceSauce, Yen has answered emails, replicated errors, and solved the problems.

I was very fortunate to be able to draw on already existing corpora. I can only imagine how much time and effort went into compiling the corpora. Therefore, I would like to applaud the wonderful people who shared their work with me: The DyViS (Nolan et al.), TUULS (Llamas et al.) and WYRED (Gold et al.) groups. William Haddican and Paul Foulkes for the Northern Englishes Corpus, Jessica Wormald for providing her Panjabi English Corpus, Christin Kirchhübel for the Deceptive Speech Corpus, Jeff Hemmen for the Cardinal Vowel Corpus, Ann Bradlow for the Speechbox ALLSTAR Corpus, David van der Vloed for FRIDA.

I would like to thank the following people for wonderful, in – depth discussions on voice quality: Pat Keating, John Esling, Katherine Dallaston and Kjartan Beier. I appreciate very much that you shared your expertise with me. Thank God there were also people who shared their knowledge of R and statistics with me: Amanda Cardoso, Márton Sóskuthy and Sonja Menges, I owe you a big thank you. John Esling, Anton Batliner, Scott Moisik and Christer Gobl were kind enough to give me permission to use illustrations from their work. I appreciate that very much.

Sarah Kelly and Joe Branson shared their home with me many times when I was in York. There are so many happy memories. Thank you, guys! And Kate Earnshaw and Wanda Mainka accompanied me throughout my promotional period, cheering me up whenever I needed someone to cheer me up, proofreading papers, formatting articles, processing data. A big thank you to Wanda for the final countdown. What a week…

I am extremely thankful for the comments and suggestions made by my truly remarkable examiners: Patricia Keating and Jessica Wormald. Thank you for making this thesis so much better.

And in the end, it’s only the ”people of the heart” who count. Thanks to my heart people Jona, Elli, and Karsten. Ich hab’ euch so lieb!

Authors declaration

I declare that this thesis is a presentation of original work and that I am the sole author unless otherwise stated in the statement of authorship preceding each article with more than one author.

Paper 2 has been published: Klug, K., Kirchhübel, C., Foulkes, P., & French, P. (2019). Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception. Proceedings of the 18th International Congress of Phonetic Sciences, Australia, 795-799.

Paper 1 has been accepted for publication: Klug, K., Kirchhübel, C., Foulkes, P., Braun, A., & French, P. (in press). Assessing creaky voice quality for forensic purposes. In Proceedings of the Aarhus International Conference on Voice Studies. Sciendo.

This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Signed: K. Klug

Date: 4 March 2024

Für meinen Opa

Introduction

1 The problem

The forensic sciences are in a state of flux. A paradigm shift is currently taking place that calls for an objectification of the analysis and interpretation of evidence (Morrison, 2022). This poses a challenge for many forensic sciences, including forensic speech science. Many features assessed in a forensic voice comparison, the most common task for forensic speech scientist (Foulkes and French, 2012), can neither be measured quantitatively nor interpreted logically correctly – yet. One of these features is voice quality.

2 Voice quality

A speaker's voice is not like a fingerprint: it is neither stable nor unique. Voice varies according to the speaker's age and state of health, his/her emotional state, and the conversational situation. Nevertheless, studies have shown that human listeners recognise familiar voices very well even in cases of emotional and speaking style mismatch (e.g. Park et al., 2016). This is due to "the characteristic auditory colouring of an individual speaker's voice" (Laver, 2009, p.1), namely voice quality (VQ).

When it comes to describing the difference between two speakers with the same accent, VQ is one, if not *the* key parameter, as it is determined by anatomical and habitual conditions. Therefore, it is an extremely useful feature for forensic voice comparisons. In a study conducted by (Gold and French, 2011, p.301), 36 forensic caseworkers filled out a questionnaire to gain insight into their working practices when conducting forensic voice comparisons. One of the findings is the widespread practice of analysing a speaker's VQ. 94% of the forensic phoneticians surveyed reported to analysing VQ. VQ is even considered by 32% of the respondents as the parameter with the greatest potential to discriminate between speakers (ahead of accent variants and vowel formants, both with 28% each).

In recent years the number of research projects on VQ, particularly on laryngeal VQ settings, has steadily increased. Especially the group around Kreiman and Keating at University of California (UCLA) is leading the way. The current project focusses on creaky and breathy VQ, which are reported to be the most common laryngeal voice qualities in Southern British English speakers (San Segundo et al., 2018), and very prevalent in German speakers (Kluge et al., 2019). Creaky and breathy voice quality contrast with each other in terms of glottal constriction: non-constricted breathy voice at one end and constricted creaky voice at the other end (Esling et al., 2019).

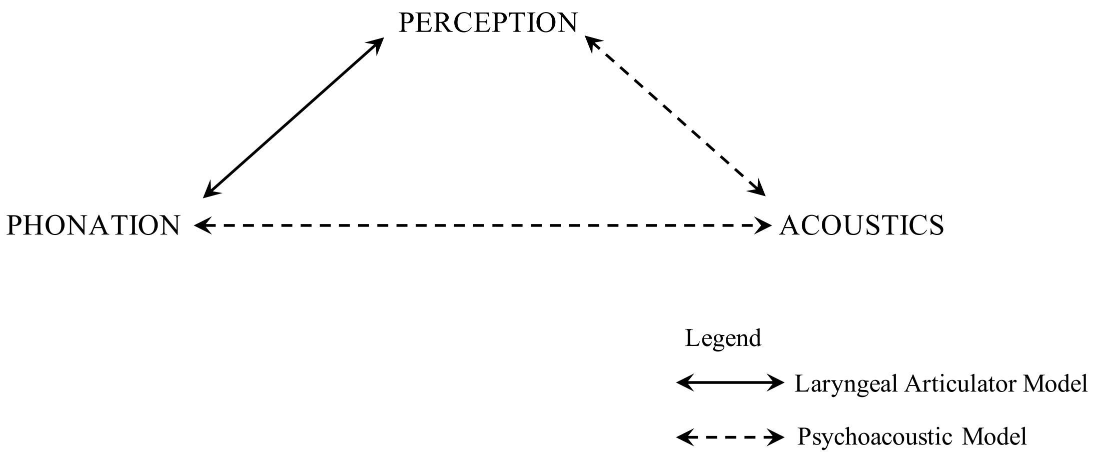


Figure 1 *Levels of voice analysis and models investigating the link between the levels.*

Laryngeal VQ may be assessed from three different angles: (1) in terms of the speaker's physiological phonation settings (phonation), (2) in terms of the acoustic effects of adopting the particular organic settings (acoustics), and (3) in terms of the auditory effects of adopting the settings (perception). The relationship between these levels appears to be extremely complex and often ambiguous. Many researchers from different disciplines, and thus with various perspectives on voice, investigate these relationships to obtain a clearer picture of the correlating parameters. Two main models exist exploring the link between the three levels from different perspectives: (1) the *Laryngeal Articulator Model* by Esling and colleagues (see Esling et al., 2019) which is grounded on the work of Laver (see Laver, 2009), and (2) the *psychoacoustic model* by Kreiman and colleagues (see e.g. Zhang et al., 2013; Zhang, 2016; Kreiman et al., 2014, 2021). The *Laryngeal Articulator Model* explores the relationship between phonation and perception based on visual evidence of the 'laryngeal articulatory mechanism'. The *psychoacoustic model*, on the other hand, seeks to decode the correlation between acoustics and perception, and between acoustics and phonation (see Figure 1). The insight from both models would allow for decoding the phonation–perception–acoustic-triangle.

The summary in bullet points given below for the voice qualities of interest are largely based on very detailed publications by luminaries of voice quality, namely Gordon and Ladefoged (2001); Laver (2009); Esling et al. (2019); Garellek (2019, 2022).

2.1 Creaky and breathy voice quality

The summaries below for the voice qualities of interest, i.e. breathy voice, and creaky voice, are largely Creaky voice (CV)

Creaky voice is exceptionally multidimensional. The listed aspects of perception, phonation, and acoustics are therefore unlikely to apply to all possible examples of creaky voice but serve as an initial overview. Paper 1 discusses the different modes of creaky voice in detail.

2.1.1 CV perception

Over the years various metaphorical CV descriptions have been used to describe the essential perceptual characteristics of CV. This is exemplified in the following list:

- "a rapid series of taps, like a stick being run along a railing" (Catford, 1964, p.32, cited by Laver, 2009, p.122)
- "a train of discrete excitations or pulses produced by the larynx" (Hollien and Wendahl, 1968, p.506)
- "popping", "frying", "ticking", "rasping" (Moore and von Leden, 1958, p.231)
- "popping of corn" (Henton and Bladon, 1988, p.10)
- "motor boat engine" (Blomgren et al., 1998, p.2650)
- "food cooking in a hot frying pan" (Ishi et al., 2008, p.47)
- "auditorily perceived as pulsatile, i.e., individual glottal cycles appear to be audible due to temporal segregation" (Devaraj and Aichinger, 2021, p.11)

These examples acknowledge the presence of *distinct glottal pulses*. Wendahl et al. (1963) and Coleman (1963), describe the highly damped vocal tract between glottal excitations as the source of individual pulse sensation. These energy losses in the vocal tract (Ishi et al., 2008) manifest themselves in dampened amplitudes between glottal excitations, i.e. a preceding glottal pulse has almost completely decayed before the next glottal pulse arrives (Coleman, 1963). Furthermore, sensation of distinct glottal pulses can also be triggered by aperiodic glottal pulses, often accompanied by very long pulses, i.e., low pitch, which is typically listed as characteristic of CV (e.g. Dallaston and Docherty, 2019; White et al., 2021). Human listeners are very sensitive to aperiodicity. Docherty et al. (1997) explain that the presence of only one or two slightly aperiodic pulses are sufficient for CV perception.

2.1.2 CV phonation

Following Esling et al. (2019), phonation is defined here as the result of involvement from three sets of folds: vocal folds, ventricular folds, and aryepiglottic folds. The phonation of CV seems to be not fully understood, probably as it may be highly variable. The characteristics typically described are listed below:

- Vocal folds have high adductive tension, high medial compression, but low longitudinal tension (Laver, 2009), they appear to be short and thick (Gerratt, 2001)
- Only a small section of the membranous glottis oscillates (Catford, 1964 cited by Esling et al., 2019), the cartilaginous glottis is fixed by lateral cricoarytenoid activity (Moisik, 2013)

- Glottal closing is abrupt, glottal closure is elongated (Moore and von Leden, 1958; Hollien, 1974; Childers and Lee, 1991; Gobl and Ní Chasaide, 2010)
- Larynx is constricted (Moore and von Leden, 1958, p.231) and elevated (Moisik, 2013)
- Ventricular folds are adducted and coupled with vocal folds, building a compressed structure (Hollien, 1974; Esling et al., 2019)
- Airflow rate is low due to strong adduction (Hollien, 1974; Gobl and Ní Chasaide, 2010)
- Different assumptions regarding the subglottal pressure: higher subglottal pressure than in modal voice reported by Murry (1971) and Hollien (1974), lower subglottal pressure reported by Blomgren et al. (1998)

2.1.3 CV acoustics

The following acoustic parameters are typically the sources of assessment. As with the phonatory characteristics, it is assumed that the acoustic characteristics do not apply to all CV modes but may differ due to differences in phonation.

- f0 ranges between 30.9 and 43.7 Hz (Michel, 1968), but not universally low (Gordon and Ladefoged, 2001)
- Glottal pulses are mostly reported to be aperiodic (Ishi et al., 2007, 2008; Redi and Shattuck-Hufnagel, 2001; Drugman et al., 2014; Gordon and Ladefoged, 2001), but could also be periodic (Garellek, 2022)
- Damping effects between glottal pulses (Drugman et al., 2014)
- Reduced overall acoustic intensity compared to modal voice (Gordon and Ladefoged, 2001)
- Formant bandwidths are narrow due to long closure phase (Gobl and Ní Chasaide, 2010)
- F1 is higher compared to modal voice (possibly due to raised larynx position, Kirk et al., 1993)
- Open quotient is decreased, i.e., very prominent H2 relative to H1 (Gordon and Ladefoged, 2001)
- Low-frequency harmonics have low amplitude levels due to low airflow (Gobl and Ní Chasaide, 2010)
- High-frequency harmonics have an equally high or higher amplitude than H1 (Gordon and Ladefoged, 2001; Laver, 2009; Garellek and Keating, 2011) due to a more abrupt glottal closure (Stevens, 1977 cited by Gordon and Ladefoged, 2001), thus spectral tilt parameters (H1–H2, H2–H4, H4–H2 kHz, H2 kHz–H5 kHz) are low for most creaky voice modes (Garellek, 2019)
- A1 is very prominent relative to H1 (Gordon and Ladefoged, 2001)
- Spectral noise parameters (HNR, and less clear CPP) are typically low due to irregularity (Garellek, 2019; Garellek and Keating, 2011)

2.2 Breathy voice (BV)

2.2.1 BV perception

For breathy VQ not many metaphorical descriptions exist.

- Similar to “sighing”, “voice mixed in with breath” (Catford, 1977, p.99 cited by Laver, 2009, p.132)
- Two components are audibly co – present: a friction component and a modal voice component, but the modal voice component is “markedly dominant” (Laver, 2009, p.134)
- Correlation with lowered larynx voice in terms of perception and physiology (Laver, 2009, p.31)

BV needs to be distinguished from *whispery voice,* which is described by (Catford, 1964, p.31) cited by (Laver, 2009, p.121) as “a relatively ‘rich’ hushing sound”, referring to the more turbulent and more prominent friction component in whispery voice (Esling et al., 2019).

2.2.2 BV phonation

BV phonation is typically described as “inefficient” because the vocal folds do not close completely (Laver, 2009). According to (Catford, 1977, p.99) cited by (Laver, 2009, p.132) the vocal folds “simply ‘flap in the breeze’ of the high velocity air-flow”.

- Cartilaginous glottis remains open as arytenoids remains apart, only the membranous glottis oscillates, the resulting glottal gap generates the friction component (Esling et al., 2019)
- Glottis and supraglottic space are unconstricted, resulting in a “more linear” airflow compared to whispery voice (Esling et al., 2019)
- Vocal folds have minimal adductive tension, weak medial compression, and rather low longitudinal tension (Laver, 2009), they appear to be short, relatively separate, and loose/thick (Esling et al., 2019)
- Pitch is typically low (Fairbanks, 1960, p.179 cited by Laver, 2009, p.133)
- Larynx is lowered (Laver, 2009)

2.2.3 BV acoustics

In contrast to CV, there is broad agreement with respect to BV acoustics.

- f0 is lower compared to modal voice (Fairbanks, 1960 cited by Laver, 2009; Gordon and Ladefoged, 2001)
- Reduced overall acoustic intensity compared to modal voice (Gordon and Ladefoged, 2001)
- F1 bandwidth is broad (Fant, 1972; Gordon and Ladefoged, 2001)
- F1 is lower compared to modal voice (possibly due to lowered larynx position, Thongkum, 1988)

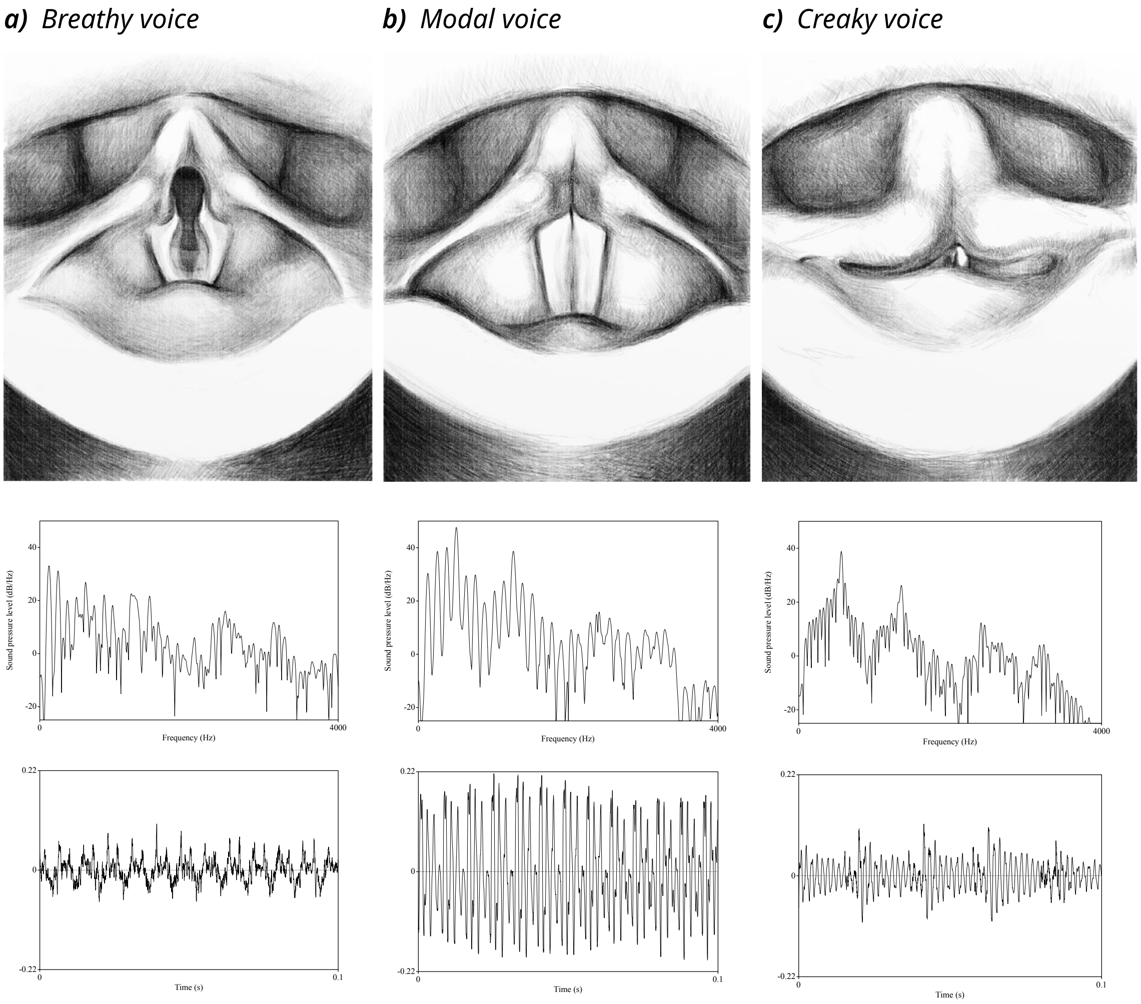


Figure 2 *Representatives of drawings of the larynx (first row), waveforms (second row), and harmonic spectra (third row) in the state of breathy voice (left), modal voice (middle), and creaky voice (right). (Drawings of the larynx are taken from Esling et al., 2019, p.45, 57, 65 with permission of the authors and Cambridge University Press. Waveforms and harmonic spectra are taken from the /schwa/ vowel of a male speaker from the cardinal vowel corpus of Hemmen, 2014.)*

- Open quotient is increased, i.e., very prominent H1 relative to H2 (Gordon and Ladefoged, 2001), higher H1-H2 compared to modal voice (Garellek, 2019, 2022)
- High-frequency harmonics have a lower amplitude than H1 (Gordon and Ladefoged, 2001) due to a less abrupt glottal closure (Stevens, 1977 cited by Gordon and Ladefoged, 2001), thus spectral tilt parameters (H2–H4, H4–H2 kHz, H2 kHz–H5 kHz) are higher compared to modal voice (Garellek, 2019)
- H1 is very prominent relative to A1 (Gordon and Ladefoged, 2001)
- Spectral noise parameters (HNR, CPP) are lower compared to modal voice (Garellek, 2019; Garellek and Keating, 2011) due to high frequency spectral noise (Gordon and Ladefoged, 2001)

Figure 2 shows examples of laryngoscopic drawings (taken from Esling et al., 2019) together with waveforms and harmonic spectra to illustrate physiological and acoustic differences (summarised above) between breathy voice and creaky voice compared to modal voice.

3 Forensic voice comparison

As it is not possible to obtain information on phonation directly in a forensic setting, forensic caseworkers need to rely on information gained from perception and acoustics.

Before 1990, VQ was assessed purely holistically, without relying on categories of a formal scheme (French, 2017). Since the 1990s, analysis has mostly been conducted using some kind of scheme. There are several existing schemes that provide impressionistic VQ categories. Gold and French (2011) found that 61%, of the 94% forensic caseworkers who reported analysing VQ, used a recognised VQ scheme to conduct the analysis. One such scheme is the Vocal Profile Analysis scheme (VPA), developed by Laver and colleagues in the early 1980s (Laver, 2009; Laver et al., 1981). The VPA assesses a speaker's overall habitual long-term VQ perceptually. Originally developed to meet the requirements of speech therapists (Mackenzie, 2005, p.295), the VPA appears to be suitable for forensic VQ analysis and is therefore often used in a modified form in casework.

However, as early as in the 1990s there were calls for VQ to be assessed on the basis of acoustic analysis (Nolan, 1991, p.490), and studies were conducted to explore for correlations between long-term VQs and signal acoustics (Jessen, 1997; Nolan, 1983).

4 Advantages and challenges of objectification

By relying on signal acoustics a method enables transparency and reproducibility, whereas a method based solely on human perception lacks transparency and the ability to independently reproduce the results. In addition, signal acoustics are resistant to cognitive bias (Morrison, 2022). However, there are several reasons why implementing an acoustic VQ analysis in a forensic application is challenging.

For some multidimensional VQ settings, such as creaky voice, perceptual relevant categories must first be defined before the search for correlating signal acoustics becomes meaningful. Besides, to the present day, the linkage between perception and acoustics does not seem to be fully understood – even with high-quality recordings. The poor recording quality typical of forensic evidential recordings is an additional obstacle. While the parameter Cepstral Peak Prominence (CPP) – a measure of periodicity – seems promising as it correlates with the percept breathiness (Hillenbrand et al., 1994; Barsties et al., 2017), it does not seem clearly distinguish breathy VQ from other less-periodic VQs (Fraile and Godino-Llorente, 2014), such as creaky voice (Garellek and Keating, 2011).

To assess within-speaker variation, large-scale studies, such as Park et al. (2016), are needed in which speakers are investigated under a variety of mismatch conditions, such as emotional mismatch, speaking style mismatch, as well as under non-contemporaneous conditions.

A further aggravating factor is technical mismatch and the limited knowledge we still have so far about its effect on signal acoustics. The influence of the mobile speech codec is just one example that produces noise when exploring the correlation between perception and acoustics. It also complicates plausibility checks of the acoustic measurements. The codec affects the signal in unpredictable manners as it constantly switches between different modes of operation (Guillemin and Watson, 2008). In addition, lost or corrupted speech frames are replaced based on information from previous frames and therefore partly contain synthetic speech that has been shown to degrade same-speaker comparisons and falsely improve different-speaker comparisons when conducting forensic automatic-speaker recognition (Nair et al., 2015).

To drive the paradigm shift, the linkage between long-term VQ perception and signal acoustics needs to be better understood, and the impact of technical factors, such as the mobile speech codec, needs to be factored into the analysis.

5 Thesis outline

The current thesis consists of four papers. Paper 1 explores creaky VQ and the possibility of describing it more precisely on the basis of perception. Paper 2 investigates whether signal acoustics can be used to underpin the perception of dominantly breathy VQ based on sonorants from spontaneous speech samples. While Paper 2 analysed the spontaneous speech samples under studio recording condition, Paper 4 examines the same material under the mobile recording condition. The methodology adopted in Paper 4 is based on findings from Paper 3, which investigates the impact of recording quality and VQ on the performance of f0 estimators. The differences between the results obtained when the analysis is based on the f0 estimator, which was found to be the most accurate in Paper 3, and the results obtained when f0 estimation is based on a less suitable f0 estimator are discussed. The conclusion chapter links the papers together and proposes a future process for assessing a speaker's VQ in forensic voice comparisons.

The thesis shows possibilities to improve the analysis of a speaker's VQ for forensic application. While many features analysed in a forensic voice comparison are nowadays investigated using the auditory-acoustic approach, the feature VQ is an exception as it still relies predominantly on a perceptual-guided approach. Forensic sciences should analyse and interpret data objectively (Morrison, 2022). For the feature VQ, this means that the relationship between perceptual aspects of VQ and signal acoustics must be better understood. It is equally important to be able to assess the possible effects of technical influences, e.g. mobile phone transmission.

References

Barsties, B., Maryn, Y., Gerrits, E., & De Bodt, M. (2017). The Acoustic Breathiness Index (ABI): A Multivariate Acoustic Model for Breathiness. *Journal of Voice* 31(4), 511.e11–511.e27. https://doi.org/10.1016/j.jvoice.2016.11.017

Blomgren, M., Chen, Y., Ng, M. L., & Gilbert, H. R. (1998). Acoustic, aerodynamic, physiologic, and perceptual proper- ties of modal and vocal fry registers. *Journal of the Acoustical Society of America* 103(5), 2649–2658.

Catford, J. C. (1964). Phonation types: The classification of some laryngeal components of speech production. In: Abercrombie, D., D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (eds.) *In honour of Daniel Jones: papers contributed on the occasion of his eightieth birthday 12 September 1961*, 26–37. London: Longmans, Green & Co.

Catford, J. C. (1977). *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press.

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America* 90(5) 2394–2410. http://dx.doi.org/10.1121/1.402044

Coleman, R. F. (1963). Decay characteristics of vocal fry. *Folia Phoniatrica et Logopaedica* 15(4), 256–263. http://dx.doi.org/10.1159/000262970

Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proceedings of the 19th International Congress of Phonetic Sciences, Australia* 532–536.

Devaraj, V., & Aichinger, P. (2021). Modelling of amplitude modulated vocal fry glottal area waveforms using an analysis-by-synthesis approach. *Applied Sciences* 11(5), 1990. http://dx.doi.org/10.3390/app11051990

Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L., & Walshaw, D. (1997). Descriptive adequacy in phonology: A variationist perspective. *Journal of Linguistics* 33(2), 275–310. http://dx.doi.org/10.1017/S002222679700649X

Drugman, T., Kane, J., & Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech and Language* 28(5), 1233–1253. http://dx.doi.org/10.1016/j.csl.2014.03.002

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). Voice quality: the laryngeal articulator model. Cambridge University Press. http://dx.doi.org/10.1017/9781108696555

Fairbanks, G. (1960). *Voice and articulation drill-book* (2nd edn), New York: Harper & Row.

Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly Progress and Status Report 2/3*, 28–52.

Foulkes, P. & French, J. P. (2012). Forensic Speaker Comparison: A Linguistic–Acoustic Perspective. In *The Oxford Handbook of Language and Law*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Fraile, R., & Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control* 14, 42–54. https://doi.org/10.1016/j.bspc.2014.07.001 French, 2017

French, J. P. (2017). A developmental history of forensic speaker comparison in the UK. *English Phonetics* 271–286. https://eprints.whiterose.ac.uk/117763/

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association* 41(2), 185–205. http://www.jstor.org/stable/44527030

Garellek, M. (2019). The phonetics of voice 1. In: W.F. Katz, P.F. Assmann (eds.) *Routledge Handbook of Phonetics* 75–106. Oxford: Routledge.

Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 101155.

Gerratt, B., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* 29(4), 365–381.

Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In: Hardcastle, W. J., J. Laver, & F. E. Gibbon (eds.) *The handbook of phonetic sciences*, 378–423. Wiley-Blackwell.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law* 18(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29, 383–406. https://doi.org/10.1006/jpho.2001.0147.

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language and the Law* 15, 193–218. https://doi. org/10.1558/ijsll.v22i1.17880

Hemmen, J. (2014). Vowel distance measures: performance and behaviour with regards to voice quality, gender, and intra-/inter-speaker factors. [Master thesis, University of York].

Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In: Hyman, L. M., & C. N. Li (eds.) *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, 3–29. London, New York: Routledge.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research* 37(4), 769–778. https://doi.org/10.1044/jshr.3902.311

Hollien, H., & Wendahl, R.W. (1968). Perceptual study of vocal fry. *The Journal of the Acoustical Society of America* 43(3), 506–509. http://dx.doi.org/10.1121/1.1910858

Hollien, H. (1974). On vocal registers. *Journal of Phonetics* 2(2), 125–143.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2007). Acoustic and EGG analysis of pressed phonation. *Proceedings of the 16th International Conference on Phonetic Sciences, Germany*, 2057–2060.

Ishi, C. T., Sakakibara, K. I., Ishiguro, H., & Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 47–56.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *International Journal of Speech, Language & The Law* 4(1), 84–103. https://doi.org/10.1558/ijsll.v4i1.84

Kirk, P. L., Ladefoged, J., & Ladefoged, P. (1993). Quantifying acoustic properties of modal, breathy, and creaky vowels in Jalapa Mazatec. In: Mattina, A. & Montler, T. (eds.) *American Indian linguistics and ethnography in honor of Lawrence C. Thompson*, 435–450. Ann Arbor, MI: University of Michigan.

Kluge, K., Müller, M., Dubielzig, C., Meinerz, C., & Masthoff, H. (2019). Distribution of Voice Quality Features in German – Preliminary results. Paper presented at the International Association for Forensic Phonetics and Acoustics, University of Huddersfield, UK.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens Spanish Journal of Speech Science.* 10.3989/loquens.2014.009

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *Journal of the Acoustical Society of America* 149, 457–465. https://doi.org/10.1121/10.0003331.

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 14, 139–155.

Laver, J. (2009). *The Phonetic Description of Voice Quality*. Cambridge University Press.

Mackenzie Beck, J. (2005). Perceptual analysis of voice quality: the place of vocal profile analysis. In Hardcastle, W. J., & Beck, J. (Eds.), *A Figure of Speech (Festschrift for John Laver)*, 285–322. Routledge.

Michel, J. F. (1968). Fundamental frequency investigation of vocal fry and harshness. *Journal of Speech and Hearing Research* 11(3), 590–594.

Moisik, S. R. (2013). The epilarynx in speech. [Doctoral dissertation, University of Victoria].

Moore, P., & von Leden, H. (1958). Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica* 10(4), 205–238.

Morrison, G. S. (2022). Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy*, 100270. https://doi.org/10.1016/j.fsisyn.2022.100270

Murry, T. (1971). Subglottal pressure and airflow measures during vocal fry phonation. *Journal of Speech and Hearing Research* 14(3), 544–551.

Nair, B. B., Alzqhoul, E. A., & Guillemin, B. J. (2015). Impact of frame loss aspects of mobile phone networks on forensic voice comparison. *International Journal of Sensor Networks and Data Communications* 4(2), 131–141. https://doi.org/10.4172/2090-4886.1000131

Nolan, F. (1983). The Phonetic Bases of Speaker Recognition. Cambridge University Press.

Nolan, F. (1991). Forensic phonetics. *Journal of Linguistics* 27(2), 483–493.

Park, S. J., Sigouin, C., Kreiman, J., Keating, P. A., Guo, J., Yeung, G., Kuo, F.Y., & Alwan, A. (2016). Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition. *Proceedings Interspeech, USA*, 1044–1048. https://doi.org/10.21437/Interspeech.2016-523

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29(4), 407–429.

San Segundo, E., Foulkes, P. French, P., Harrison, P., Hughes, V., Kavanagh, C. (2018). The use of the vocal profile analysis for speaker characterization: Methodological proposals. *JIPA* [online].

Stevens, K. (1977). Physics of laryngeal behavior and larynx modes, *Phonetica* 34, 264–79.

Thongkum, T. (1988). Phonation types in Mon-Khmer languages. In: O. Fujimura (ed.) *Vocal fold physiology: voice production, mechanisms and functions*, 319–334. New York: Raven Press.

Wendahl, R. W., Moore, G. P., & Hollien, H. (1963). Comments on vocal fry. *Folia Phoniatrica et Logopaedica* 15(4), 251–255. http://dx.doi.org/10.1159/000262969

White, H., Penney, J., Gibson, A., Szakay, A., & Cox, F. (2021). Optimizing an automatic creaky voice detection method for Australian English speaking females. *Proceeding of Interspeech, Czech Republic*, 1384–1388. http://dx.doi.org/10.21437/Interspeech.2021-711

Zhang, Z., Kreiman, J., Gerratt, B. R., & Garellek, M. (2013). Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. *Journal of the Acoustical Society of America* 133, 453–462. http://dx.doi.org/10.1121/1.4770235

Zhang, Z. (2016). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *Journal of the Acoustical Society of America* 139, 1493–1507. https://doi.org/10.1121/1.4944754.

Research Degree Thesis Statement of Authorship

University of York
York Graduate Research School

Candidate name	Katharina Klug	
Department	Language and Linguistic Science	
Thesis title	Assessing a speaker's voice quality for forensic purposes – Using the example of creaky voice and breathy voice	
Title of the work (paper/chapter)	Assessing creaky voice quality for forensic purposes	
Publication status	**Published**	
	Accepted for publication	x
	Submitted for publication	
	Unpublished and unsubmitted	
Citation details (if applicable)	Klug, K., Kirchhübel, C., Foulkes, P., Braun, A., & French, P. (in press). Assessing creaky voice quality for forensic purposes. In *Proceedings of the Aarhus International Conference on Voice Studies*. Sciendo.	
Description of the candidate's contribution to the work	Conceptualisation Data curation Formal analysis Investigation Methodology (lead) Writing – original draft	
Percentage contribution of the candidate to the work	80%	
Signature of the candidate	K. Klug	
Date (DD/MM/YY)	26/02/24	

Co – author contributions*

By signing this Statement of Authorship, each co – author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co – author	Christin Kirchhübel
Contact details of co – author	Soundscape Voice Evidence, Lancaster, UK ck@soundscapevoice.com
Description of the co – author's contribution to the work**	Investigation Writing – review and editing
Percentage contribution of the co – author to the work	5%
Signature of the co – author	
Date (DD/MM/YY)	26/02/2024

Name of co – author	Paul Foulkes
Contact details of co – author	Department of Language and Linguistic Science, University of York, UK paul.foulkes@york.ac.uk
Description of the co – author's contribution to the work*	Investigation Writing – review and editing
Percentage contribution of the co – author to the work	5%
Signature of the co – author	
Date (DD/MM/YY)	26/02/2024

Name of co – author	Almut Braun
Contact details of co – author	Bundeskriminalamt, Wiesbaden, Germany almut.braun@bka.bund.de
Description of the co – author's contribution to the work*	Investigation Writing – review and editing
Percentage contribution of the co – author to the work	5%
Signature of the co–author	Braun
Date (DD/MM/YY)	28/02/2024

Name of co – author	Peter French
Contact details of co – author	J P French International, Zurich, Switzerland Department of Language and Linguistic Science, University of York, UK peter.french@york.ac.uk
Description of the co – author's contribution to the work*	Methodology (supporting) Writing – review and editing
Percentage contribution of the co – author to the work	5%
Signature of the co – author	
Date (DD/MM/YY)	27/02/24

Copy and paste additional co – author panels as needed.

* Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

** The description of the candidate and co – authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co–authors may find it helpful to consider the CRediT (Contributor Roles Taxonomy) approach to recognising individual author contributions.

Assessing creaky voice quality for forensic purposes

Katharina Klug[1], Christin Kirchhübel[2], Paul Foulkes[1], Almut Braun[3], and Peter French[1,4]

[1] Department of Language and Linguistic Science, University of York, UK
[2] Soundscape Voice Evidence, Lancaster , UK
[3] Forensic Science Institute, Bundeskriminalamt, Wiesbaden, Germany
[4] J P French International, Zurich, Switzerland

katharinaklug@gmx.net, ck@soundscapevoice.com, paul.foulkes@york.ac.uk, almut.braun@bka.bund.de, and peter.french@york.ac.uk

Abstract: This study examines the multifaceted nature of creaky voice production (hereinafter CV), which results in diverse perception and acoustics. Therefore, CV categories need to be defined which are relevant for both perception and acoustics. Here, perceptually relevant CV categories are explored. A CV classification scheme is conceptualised and tested, which distinguishes between four CV modes (clean CV, harsh CV, breathy CV, and aperiodic creak) and acknowledges transitions into CV (e.g., modal voice – clean CV) as well as transitions between CV modes (e.g., harsh-breathy CV). The scheme is tested by four forensic speech scientists analysing spontaneous speech samples of six male English speakers. Results revealed that speakers vary with respect to the nature and frequency of preferred CV modes, i.e., CV space. We conclude that the nature of CV production may be useful to facilitate speaker discrimination.

Keywords: creaky voice, perceptual analysis, forensic voice comparison, speaker-discriminating potential

1 Introduction

Creaky voice (henceforth CV) is frequently used in many languages as a short-term voice quality setting to mark various functions, e.g.: (1) contrastive phonation types, (2) lexical tone and register, (3) intonation, (4) vowel-initial glottalisation, and (5) coda – /t/ glottalisation (see Garellek 2022 for a summary). Within forensic speech science, however, CV as a long-term setting is of particular interest. Dilley (1996) reported speakers of American English not only to differ in their use of CV as a short-term or long-term setting, but also in their characteristic acoustic CV preferences. This variation is assumed to be based on anatomical, physiological, and habitual idiosyncrasies (Redi & Shattuck-Hufnagel, 2001). The aim of the present study is to determine if "persistent creakers" (Henton & Bladon, 1988), who employ CV as a long-term voice quality setting, can be distinguished based on their particular CV production.

Forensic speech scientists analyse audio recordings related to legal (usually criminal) investigations. Frequently, the expert is asked to compare the speaker in an evidential recording with the recording of a known suspect, to address the likelihood of identity or non-identity. Analysing individual features like fundamental frequency, formants, rhythm, tempo, lexical choices, and voice quality, the level of similarity between the recordings is assessed, as well as the level of typicality in a defined reference population. Voice quality (henceforth VQ) is an important feature for the purpose of this task because of its potential power to discriminate between speakers (Gold & French, 2011). We consider that this speaker-discriminatory power can be further exploited by taking advantage of the multifaceted nature of CV through analysing it on a finer scale.

CV production is typically characterised by short and thick vocal folds (Gerratt & Kreiman, 2001), with only the membranous part of the glottis being able to vibrate (Ladefoged, 1971; Moisik, 2013a). Laver (2009) describes high adductive tension, high medial compression, and low longitudinal tension. The larynx is tensed (Moore & von Leden, 1958) and typically elevated (Moisik, 2013a), which causes the adducted ventricular folds to couple with the vocal folds (Hollien, 1974). The airflow rate is low due to strong adduction (Hollien, 1974; Gobl & Ní Chasaide, 2010).

Various metaphors were used over the centuries to describe the perception of CV: (Catford, 1964, p.32) refers to "a rapid series of taps, like a stick being run along a railing". (Moore & von Leden, 1958, p.231) compares it to "popping", "frying", "ticking" or "rasping" and (Blomgren et al., 1998, p.2650) associates it with a "motorboat engine". All these analogies describe individual glottal pulses which are separately perceivable. Thus, we consider the presence of *distinct glottal pulses* to be the main characteristic of CV.

1.1 Previous work

Some studies acknowledged the multifaceted nature of CV and suggested potential ways to differentiate between types or modes. The following studies – ordered by year of publication – were chosen to be most relevant for forensic application and should be considered when developing a classification scheme, designed to meet the requirements for forensic application.

The Vocal Profile Analysis protocol (VPA) developed by Laver (2009); Laver et al. (1981) has proven useful in analysing a speaker's VQ for forensic purposes and is used by forensic phoneticians in various adapted forms around the world (San Segundo, 2021). The VPA assesses a speaker's VQ using a componential approach by capturing individual VQ aspects including vocal tract features (e.g., lip and larynx position), overall muscular tension (i.e., vocal tract and laryngeal tension), and phonation features (e.g., creaky, and harsh voice). Each setting is assessed perceptually on a long-term basis using a 6-point scale, with a scalar degree of 4 to 6 being used for pathological voices. As for CV, according to Laver (2009) it is compatible with whispery voice, harsh voice, and falsetto and various combinations, e.g., harsh-whispery creaky voice.

Batliner et al. (1993) developed the 'MÜSLI' scheme (Münchner Schema für Laryngalisierungs-Identifikation) which differentiates between glottalisation, damping, diplophonia (i.e., amplitude variations), subharmonics (i.e., frequency variations) and aperiodicity.

Slifka (2006) studied CV production, which occurs at the end of utterances. It is characterised by irregular glottal pulses, abrupt glottal closing, and relatively rapid glottal opening. The vocal folds often lack proper glottal closure, resulting in increased airflow. Slifka (2006) reported that incidences of this irregular CV mode are also found in non-final position.

Ishi et al. (2008) highlighted the challenge to clearly distinguish between distinct and non-distinct glottal pulses, and thus the challenge to distinguish between presence and absence of 'vocal fry', a popular synonym for CV. They therefore introduced 'transitions' between vocal fry and modal voice, period-doubled voice, and harsh voice, recognising the phonation continuum. However, the authors did not study transitions systematically.

An example of a more systematic approach on the issue of transitions can be seen in Devaraj et al. (2023). Using synthesised vowels, the authors systematically varied fundamental frequency, open quotient, and amplitude quotient to assess the level of 'impulsiveness', i.e., the level of "temporal segregation of individual glottal pulses" (Devaraj et al., 2023, p.2) on a 7-point-Likert scale. They demonstrated that different graduations of vocal fry impulsiveness can be perceived, which seemed to correlate primarily with f0 and somewhat with open quotient. Therefore, transitions into CV should also be considered within a CV classification scheme.

Moisik (2013a) described two different CV modes based on laryngeal height: *raised CV* and *lowered CV*. Although CV is typically produced with a raised larynx position, CV can also be produced with a lowered larynx which prevents vocal-ventricular fold coupling and biases "phonation towards breathiness" (Moisik, 2013a, p.211). The longitudinal tension of the glottis and the aryepiglottic constriction above the glottis are considerably less than in raised CV.

Keating et al. (2023) described low f0 and irregular f0 to be sufficient to generate CV perception. They list glottal constriction as a further key property, which, on its own however, is not sufficient to generate creaky perception. All three key properties can occur in combination with each other as prototypical creak (low f0 + irregular f0 + constricted glottis), spread glottis creak (low f0 + irregular f0), and vocal fry (low f0 + constricted glottis). Furthermore, low f0 can occur together with multiple pulses.

Each of the classification approaches summarised above provided valuable perspectives on the multifaceted nature of CV and was taken into account when developing the proposed CV classification scheme.

1.2 Scope and Research Questions

The scope of the study is to explore the possibility of improving the analysis of CV by implementing a workable CV classification scheme with perceptually relevant categories. The approach is based on perceptual analysis, with visual signal characteristics used to complement the perceptual categories. Given the project's scope, the following research questions are posed:

RQ1

CV assessment: Is it possible to assess the *nature of CV production* using the proposed CV classification scheme? How *consistently do analysts* perform?

RQ2

CV production: Can *speakers be distinguished* by the nature of their CV production?

2 Method and Data

2.1 Method

The current study considers the presence of *distinct glottal pulses* as the main criterion of CV perception. Distinct glottal pulses can have two causes: (1) damping effects, i.e., amplitude attenuation, or (2) aperiodicity. (1) The vocal tract in CV is described to be highly damped between glottal excitations (Wendahl et al., 1963; Coleman, 1963). Possible causes for these damping effects are coupling effects between the vocal folds and the ventricular folds (Esling et al., 2019), a long glottal closure phase (Fant, 1979), or a slow glottal opening phase (Murry, 1971). Damping effects cause energy losses (Ishi et al., 2008), thus enabling the perception of individual pulses. (2) The perception of distinct glottal pulses, however, can also be generated by *aperiodic glottal pulses*, which often appear as long pulses, i.e., low pitch, typically listed as a CV characteristic, e.g., Dallaston & Docherty (2019); White et al. (2021). Human listeners are very sensitive to aperiodicity. Docherty et al (1997) revealed that the presence of only one or two slightly aperiodic pulses are sufficient for CV perception. These two causes for CV perception, i.e., damping effects and aperiodicity, form the core of the proposed CV classification scheme.

The CV scheme (Figure 1) is divided into a blue inner section surrounded by a grey outer section. The blue section displays the presence of distinct glottal pulses and thus CV, while the grey section displays adjacent non-creaky VQs which lack distinct glottal pulses. CV is separated into four CV modes. Three of them, i.e., *clean, harsh,* and *breathy CV*, produce CV by amplitude damping effects. *Aperiodic creak* (henceforth aperiodic C) is characterised by aperiodically spaced glottal pulses and thus segregated from the amplitude damped CV modes in the scheme. The CV modes are ordered according to perceptual proximity and/or distance. Aperiodic C is perceptually furthest from clean CV but may slightly overlap with harsh CV and breathy CV. Gerratt & Kreiman (2001) reported that aperiodic noisy voices perceptually overlap slightly with period-doubled voices. Aperiodic C results in a noisy and rough quality, comparable to harsh and breathy CV. Therefore, aperiodic C perceptually borders on harsh-breathy CV.

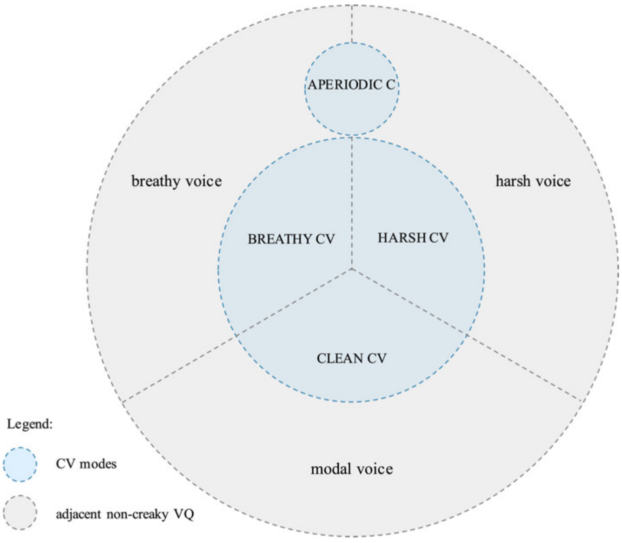


Figure 1 *Proposed CV classification scheme illustrating CV modes and adjacent non-creaky VQs*

Within the grey section, phonation types which are most closely related to CV are displayed, i.e., modal voice, harsh voice, and breathy voice. We follow Ishi et al. (2008) and Devaraj et al. (2023) in acknowledging the continuity of glottal pulse distinctiveness and thus between *CV and non-CV* and between *CV modes*, as indicated by the dashed lines.

Examples of waveforms and broad-band spectrograms (window length 0.03s) for each CV mode are shown below (Figure 2–Figure 5). They were selected based on 100% agreement between all analysts from the main study. In the absence of a clear example of breathy CV, the transition harsh-breathy CV is illustrated instead. Dashed lines mark the CV section.

2.1.1 Clean CV

Clean CV is characterised by distinct glottal pulses and the absence of additional perceptual characteristics. Clean CV can appear tense. The larynx is typically elevated (Moisik, 2013a). Figure 2 clearly shows the damping between the glottal pulses. In the waveform, excitation amplitudes are clearly separated

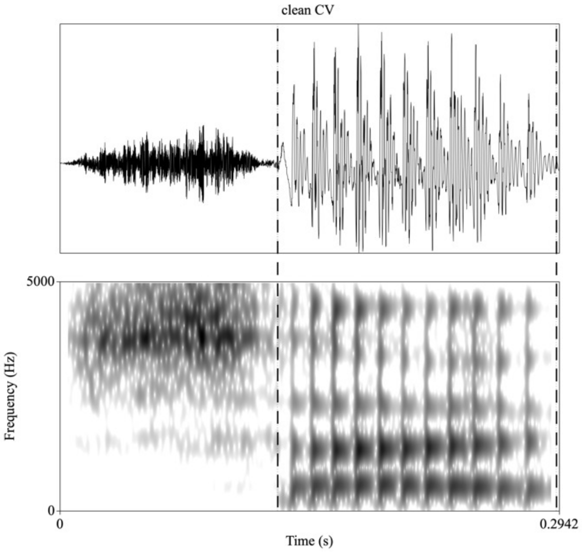


***Figure 2** Example for clean CV (speaker 004 <so>)*

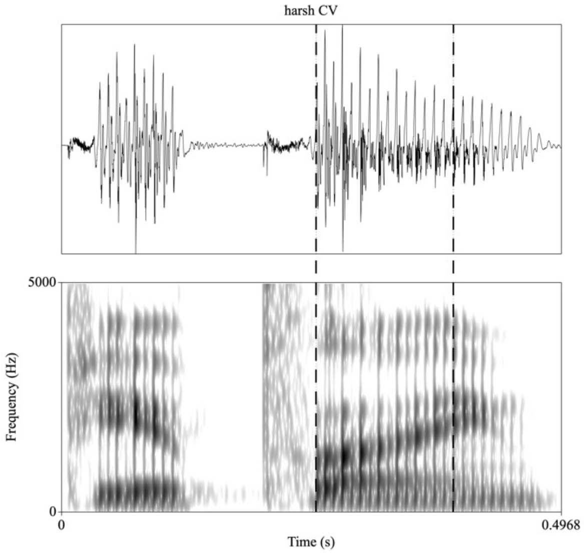


***Figure 3** Example for harsh CV (speaker 006 <be applying>)*

from each other by damped amplitudes. The same phenomenon can be seen in the spectrogram below: the dark vertical lines are followed by white lines, indicating the absence of energy. The vertical distinct lines in the spectrogram, which extend over the entire frequency range indicate the distinct glottal pulses (Laver, 2009).

2.1.2 Harsh CV

With *harsh CV*, the perception of distinct glottal pulses occurs simultaneously with the perception of harshness, i.e., roughness. Harsh voice quality is produced by coupling the vocal folds with either the ventricular folds or the aryepiglottic folds (Moisik, 2013b). This coupling effect results in amplitude modulations, subharmonics, or chaos (Anikin et al., 2021). Acoustically, subharmonics and amplitude modulations result in multiple pulses. Perceptually, either two simultaneous pitches (Kramer et al., 2013) or an indeterminate pitch is to be expected. Chaos is perceived as particularly harsh as the vocal folds oscillate highly irregularly, resulting in harmonic smearing and broadband noise (Anikin et al., 2021). The waveform of harsh CV in Figure 3 shows regular amplitude modulations resulting in two repetitive patterns. These amplitude modulations are also evident in the spectrogram as weak vertical lines between the strong vertical lines.

2.1.3 Breathy CV

Breathy CV is generated when distinct glottal pulses occur simultaneously with high frequency noise, and thus the impression of breathy/whispery voice quality that occurs together with CV. According to Klatt & Klatt (1990) the high frequency noise is particularly apparent in the frequency region around F3. The larynx is lowered (Moisik, 2013a). Previous studies use the following terminologies: *whispery CV* (Laver, 2009), *breathy-laryngealized mode of vibration* (Klatt & Klatt, 1990), *lowered/lax creaky voice* (Moisik, 2013a), or *spread glottis creak* (Keating et al., 2023). The example in Figure 4 displays harsh-breathy CV, the transitional mode between harsh CV and breathy CV.

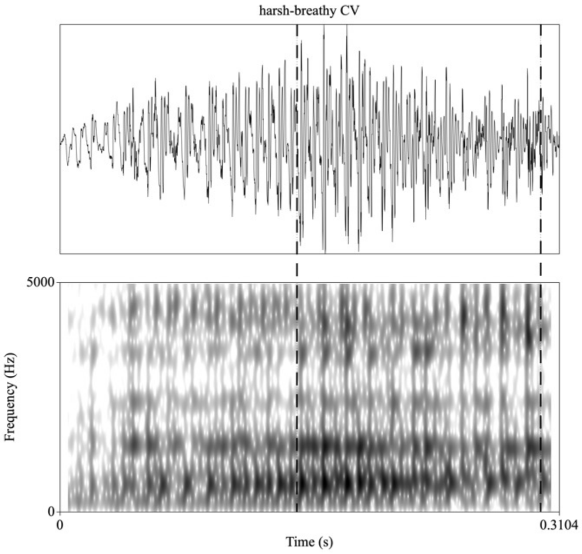


Figure 4 Example for harsh-breathy CV (speaker 005 <uh>)

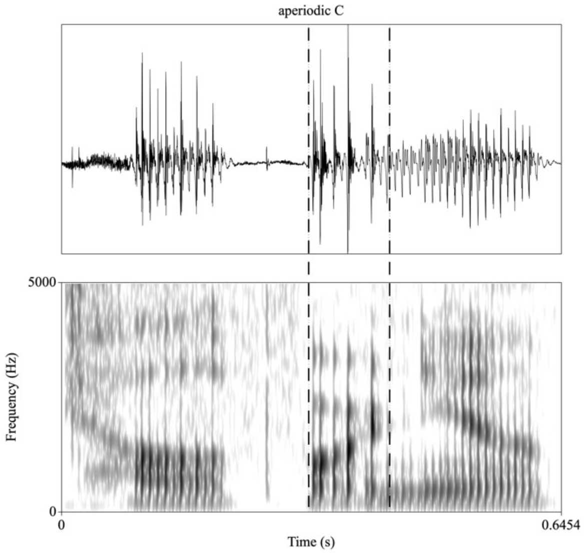


Figure 5 Example for aperiodic C (speaker 002 <California>)

Here, distinct glottal pulses occur simultaneously with high frequency noise and chaos. This becomes apparent by the less periodic waveform with irregular amplitude modulations. The spectrogram reveals less distinct vertical lines with considerable energy in between and less prominent formants, all of which indicates the presence of noise.

2.1.4 Aperiodic Creak

Aperiodic Creak is characterised by distinct glottal pulses resulting from aperiodic vocal fold oscillation rather than amplitude damping alone. Aperiodicity inherently prevents a clear pitch perception, and thus tends to be perceived as “noisy”. The lack of pitch is also reflected in the term *creak* rather than *creaky voice*. In contrast to the three CV modes which are based on long-term amplitude damping effects, *aperiodic C* occurs as a short-term setting. As stated above, human listeners are very sensitive to even slightly aperiodic pulses (Docherty et al, 1997), and thus very susceptible to aperiodic C. An example is shown in Figure 5. The aperiodicity becomes apparent through the irregular main excitation peaks in the waveform, which are also clearly visible as irregular occurring dark vertical lines in the spectrogram.

2.1.5 Adjacent voice qualities

The proposed VQs adjacent to CV are *modal voice*, *harsh voice*, and *breathy voice*. As described by Ishi et al. (2008), adjacent VQs are not clearly delineated from CV, either in terms of production or perception. The proposed adjacent VQs are hypothesised to share a continuum with CV. This continuum may result in analyst-specific thresholds when assessing presence or absence of distinct glottal pulses and could therefore affect the consistency between (and within) analysts when applying the CV classification scheme.

2.2 Data

A corpus of six speakers was compiled. The samples contained spontaneous speech of male English speakers, provided at 44.1 kHz sampling frequency and a bit depth of 16-bit. Each speaker was represented by three samples that were taken from the same recording, each about 30 seconds long. This resulted in a total of 18 samples. The recordings were obtained from five corpora of conversational speech, all involving UK or US English: WYRED (Gold et al., 2018), SpeechBox ALLSSTAR (Bradlow, n.d.), The Life Scientific (Al – Khalili, 2022), Northern Englishes (Haddican & Foulkes, 2017) and Deceptive Speech (Kirchhübel, 2013). The speakers were selected by the lead author to represent a wide range of CVs, including speakers who straddle the line between creaky and non-creaky VQ. Praat TextGrid files (Boersma & Weenink, 2023) were provided, specifying the syllables to be analysed. Only stressed syllables were chosen to exclude syllables phonated with low subglottal pressure. Filled particles *uh* and *um* were included too, as they are prone to creaky phonation (Muhlack et al., 2023). 20 syllables were labelled within each sample, resulting in 60 syllables per speaker, 360 syllables per analyst and 1440 syllables in total.

2.3 Procedure

To test the proposed CV classification scheme, a focus group was formed, consisting of four forensic speech scientists (the first to fourth authors), selected for their interest in VQ and CV in particular. All focus group participants had similar training in perceptual VQ analysis, based on the application of the Vocal Profile Analysis by Laver et al. (1981) and regularly analysed VQ for forensic casework and/or VQ research. The testing procedure involved an alternation of individual analytical listening tasks followed by in – depth online discussion sessions to address challenges and evaluate results. For the pilot test, training material was provided to the focus group participants to ensure familiarisation with the proposed CV classification scheme. This included (1) a summary with basic perceptual characteristics of the three amplitude damped CV modes, i.e., clean CV, harsh CV, and breathy CV, and (2) three short sample recordings selected by the first author to represent each of the CV modes. Aperiodic C was not included as it was only introduced after the pilot test. In the subsequent online discussion session, difficulties were discussed which provided helpful insight for the design of the main study.

In this paper the main study is discussed. Within the main study the participants were instructed to follow a strict assessment method, dividing the classification process into the following steps:

1. DistGloPuls: Is the syllable produced with distinct glottal pulses?
   - yes
   - no → NO CV
2. OVQ (Other Voice Quality): If the syllable is not produced with DistGloPuls, please state how you would characterise the syllable’s main VQ instead. (e.g., modal voice, harsh voice, breathy voice)
3. f0A/AmD: If the syllable is produced with DistGloPuls, are these generated due to *frequency aperiodicity* (f0A) or due to *amplitude damping* (AmD)?
   - f0A → APERIODIC CV
   - AmD
4. FurChar: If the DistGloPuls are due to AmD, are there any *further characteristics*, i.e., *multiple pulses* (MulPul), *high frequency noise* (Noi)?
   - no → CLEAN CV
   - MulPul → HARSH CV
   - Noi→ BREATHY CV

Table 1 Flowchart for assessing CV modes

Distinct glottal pulses	Absent	NO CV				
	Present	Frequency aperiodicity	APERIODIC CV			
		Amplitude damping	Further characteristics	Absent	CLEAN CV	
				Present	Multiple pulses	HARSH CV
					High frequency noise	BREATHY CV

5. CVmode: Depending on the entries of the tiers above, the respective *CV mode* can be determined (see Table 1). As VQ is continuous, ‘in-between’ transitional modes are also possible (e.g., harsh-clean CV, clean CV – modal voice).
6. Notes: Any *comments* or *questions* can be entered here.

The flowchart in Table 1 illustrates the stepwise classification process. This fine-grained procedure was designed to test the workability of the CV classification scheme in the first instance. Thus, a highly controlled context was chosen where samples were ’pre-processed’, i.e., analysts were directed to examine isolated syllables specifically chosen by the experimenter. If the scheme proves feasible in this context, then there are strong arguments for further exploring its feasibility in a less controlled (and forensically more realistic) environment, i.e., by applying it to samples of speech which are not pre-processed.

Calibration for the main study was carried out by providing sample recordings for each proposed CV mode together with a description of the expected characteristics visible in waveforms, broad-band and narrow-band spectrograms and FFT spectral slices (see Appendix). Analysis was based primarily on perceptual cues. Visual signal inspection could be used to corroborate perception.

2.4 Data analysis

For visualisation purposes, the syllable classifications provided by the four analysts (e.g., harsh CV) were transferred into x – y-coordinates. Using Microsoft Excel the CV classification scheme was placed on a grid so that all syllable classifications made for each speaker could be plotted into the scheme. By using a bubble chart, the frequency of occurrences of each classification was represented through bubble size. Colour was used to represent the ratings of individual analysts. Figure 7 shows the resulting bubble charts, which allow for comparisons between analysts within a chart, and between speakers across charts.

Quantitative analysis was performed by calculating the frequency with which each CV mode was assigned per speaker, both per analyst and across analysts (Table 2).

Agreement between analysts was assessed in two ways: (1) Agreement in classifying presence or absence of distinct glottal pulses (DistGloPuls) was calculated using Gwet’s

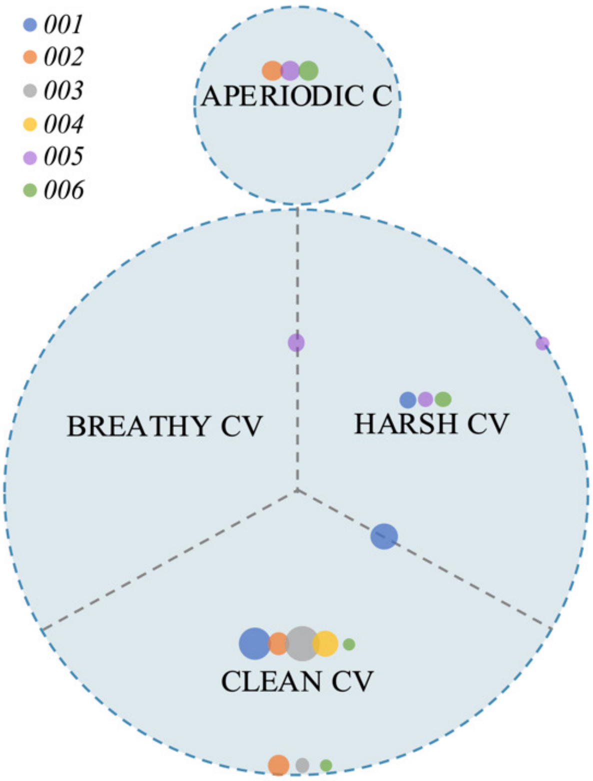


Figure 6 *Preferred CV modes per speaker across all analysts*

agreement coefficient (AC2). (2) Agreement in classifying the phonation mode of individual syllables was analysed using a distance measure calculated for each analyst pair.

3 Results and Discussion

The results are presented in terms of (1) CV assessment, and (2) CV production. Table 2, Figure 6, and Figure 7 illustrate the results.

Table 2 lists the *preferred CV modes* for each speaker. Speaker *001*, for example, preferred three CV modes – clean CV, harsh CV, and clean-harsh CV. The percentages illustrate the proportions with which particular CV modes were assigned. Thus, 41% of speaker *001's* syllables were rated as clean CV when averaged across the four analysts. To be classified as a *preferred CV mode,* two conditions must be met: (1) the CV mode occurs in at least 10% of all CV ratings *across all analysts* (see *% across analysts*, 1st row in Table 2), and (2) the same CV mode occurs in at least 10% of all CV ratings *by at least two analysts* to ignore analyst-specific outliers (see *% per analysts*, 2nd row in Table 2, percentages of individual analysts separated by slashes).

Figure 6 displays the same information as in Table 2 *(% across analysts*, 1st row) visually within the CV classification scheme. Each speaker is represented by a different colour. Only the part of the scheme covering CV is shown, while adjacent VQs are ignored.

The bubble charts in Figure 7 illustrate the ratings of *each analyst* for each of the six speakers separately. Here, different colours represent different analysts.

3.1 CV assessment

To assess the level of consistency between analyst, we need to differentiate between two aspects: (1) determination if CV is present, i.e., whether a syllable can be characterised by distinct glottal pulses (DistGloPuls), and (2) assessment of phonation mode.

3.1.1 DistGloPuls

The threshold between distinct and non-distinct glottal pulse oscillation is continuous. Therefore, we distinguish between three perceptual gradations to specify DistGloPuls: presence, absence, and inconclusiveness. Inconclusiveness of DistGloPuls was inferred when analysts indicated uncertainty when assessing DistGloPuls by either adding a question mark or indicating “slightly” or “borderline” in the comment box.

Analyst agreement on DistGloPuls was calculated using the Gwet’s agreement coef-

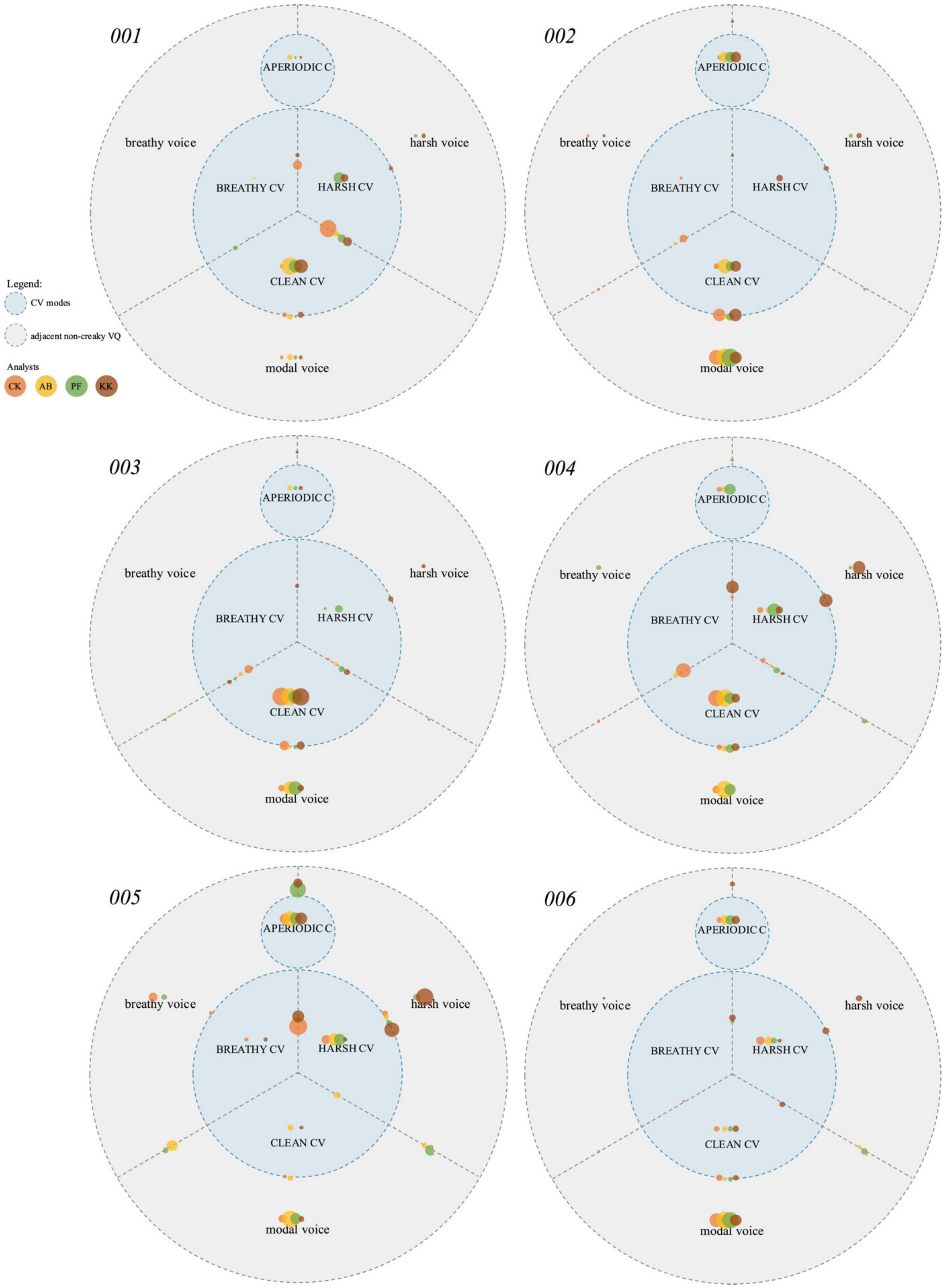


***Figure 7** Ratings of all labelled syllables per speaker and per analysts*

Table 2 *Preferred CV modes per speaker. Numbers indicate the percentage with which the classification was assigned from all CV ratings made, both across all analysts and per analyst (separated by slashes).*

	001	002	003	004	005	006
Preferred CV mode	clean CV	clean CV	clean CV	clean CV		clean CV
% across analysts	41	32	66	36		14
% per analyst	4/74/41/46	13/57/33/25	64/79/58/63	43/68/24/14		17/13/13/14
	harsh CV				harsh CV	harsh CV
	11				19	21
	0/0/32/14				14/30/38/3	38/33/17/6
		aperiodic C			aperiodic C	aperiodic C
		27			31	33
		6/35/43/25			19/42/43/26	17/46/50/23
	clean-harsh CV				harsh-breathy CV	
	29				25	
	68/9/18/20				54/0/10/26	
		clean CV-modal voice	clean CV-modal voice		harsh CV-harsh voice	clean CV-modal voice
		29	11		16	14
		55/5/23/33	18/5/6/13		5/6/10/40	25/8/13/11

ficient (AC2) for more than two analysts from the *irrCAC* package (Gwet, 2019) in Rstudio (R Core Team, 2021). The calculation was based on the ordinal raw ratings for each syllable and each analyst. Percent agreement reached AC2=0.56 for all analysts, indicating *moderate* agreement (according to Landis & Koch, 1977), while individual analyst pairs ranged between AC2=0.48 (*moderate*) and AC2=0.61 (*substantial*). Note that this scale indicates that the degree of agreement was well above chance level. Overall, the analysts in the current study seem to be in broad agreement when assessing the presence of distinct glottal pulses.

3.1.2 Phonation mode

The agreement between analysts when assessing the phonation mode of each syllable was calculated for each analyst pair by determining the perceptual distances between each syllable's ratings. Distances were determined as follows: (1) the distance between two amplitude damped CV modes (e.g., clean CV and harsh CV) was defined as 1 distance. The non-creaky VQs were treated similarly, e.g., modal voice and breathy voice were considered as 1 distance point apart. (2) As aperiodic C is perceptually further away from clean CV than it is from harsh and breathy CV, disagreement between aperiodic C and clean CV was classified to be greater (distance 2) than between aperiodic C and harsh or breathy CV (distance 1). (3) When analysts differed in their classification only by an adjacent transitional mode (e.g., clean CV vs. clean CV – modal voice) the distance between both classifications was rated to be 0.5. (4) The largest possible perceptual distance within the proposed CV classification scheme was considered between aperiodic C and modal voice and rated as 3 distances.

For four speakers, namely *001, 002, 003* and *005*, all analyst pairs classified the same syllables with an average distance of 0.5, which corresponds to distance category (3) above, i.e., the distance between clean CV and clean-harsh CV (see *% per analyst* of analysts 1 and 2 for speaker *001* in Table 2), or clean CV and clean CV – modal voice (speaker *002* in Table 2). For these four speakers, the individual analyst pairs ranged between 0.05 distances, which is almost perfect, to 0.9.

For speakers *004* and *006* analysts showed lower values of agreement when classifying the phonation mode. Across all analysts pairs the averaged distance was 0.9 with individual analyst pairs ranging between 0.4 and 1.3 distances.

Across all speakers each analyst pair showed similar agreement performances with distance values ranging between 0.6 and 0.7 distances. Consequently, no analyst pair was in greater agreement across all speakers than all other analyst pairs.

3.1.3 Further insights

The focus group discussion sessions revealed that apparent disagreement between analysts regarding DistGloPuls and phonation mode did usually not stem from differences in perception, but rather from the way in which the perception was classified. Often this was due to analyst-specific thresholds.

3.2 CV production

It can be seen from Figure 6 and Figure 7, that breathy CV featured much less frequently compared to the other CV modes. Indeed, it did not qualify as a preferred CV mode for any of the speakers. This may suggest that breathy CV is more unusual compared to the other CV modes.

In contrast, clean CV featured much more extensively among the speakers analysed. Although clean CV appears to be a more frequent and therefore a more typical CV mode, this does not mean that two speakers cannot be distinguished from each other just because they use clean CV as their preferred CV mode. This is because, generally, speakers have more than one preferred CV mode and that combination of preferred CV modes differs between speakers. For example, while speakers *001* and *002* both have clean CV as a preferred CV mode, they differ in terms of additional preferred CV modes, with speaker *001* using harsh CV and speaker *002* preferring aperiodic C. In fact, it is only speaker *003* (grey in Figure 6), whose preferred CV modes only revolve around clean CV. Whether a lack of diversity of CV modes within a speaker can be considered more unusual than a speaker displaying a variety of CV modes warrants further investigation.

Speaker *005* (purple) was the only speaker who did not prefer clean CV at all but instead spanned the upper-right corner of CV modes, i.e., aperiodic C and all modes which include some degree of harshness. This may suggest that the absence of clean CV is more unusual and therefore has a higher speaker idiosyncratic value.

Some pairs of speakers, such as *003* and *005*, can be clearly distinguished from each other. Other speakers' CV spaces overlap to a greater degree, e.g., *003* vs. *004* (grey vs. yellow), *002* vs. *006* (orange vs. green), and *001* vs. *004* (blue vs. yellow). Despite the occurrence of overlap, none of the six speakers share the same CV space, i.e., there is no complete overlap between any of the speakers with respect to the nature and frequency of their preferred CV modes.

Figure 7 supports the results by also including the syllables that were classified as non-creaky but directly bordered the speaker's CV spaces. This is most evident in speakers *002* and *005*.

4 Conclusion

Based on spontaneous speech samples of six male English speakers, the present study explored the possibility to assess the multi-faceted nature of CV further by subdividing it into four main CV modes, i.e., clean CV, harsh CV, breathy CV, and aperiodic C. Explicit recognition was given to transitional modes such as harsh-breathy CV, or harsh CV—harsh voice. Looking back at the research questions, we conclude as follows.

RQ1

CV assessment: Is it possible to assess the *nature of CV production* using the proposed CV classification scheme? How *consistently do analysts* perform?

Analysts showed moderate agreement when assessing the presence/absence of distinct glottal pulses, which was defined as the main characteristic of CV in the current study. When assessing the nature of the CV modes, the analysts showed strong agreement on four out of the six speakers using the proposed CV classification scheme. Discussion sessions revealed that disagreement between analysts usually did not stem from different perceptions; instead, often the same perception was just differently classified. Particular attention should therefore be paid to the calibration of the analysts, e.g., by providing more detailed training materials or conducting regular calibration sessions.

RQ2

CV production: Can *speakers be distinguished* by the nature of their CV production?

Overall, the speakers in the current study differed in the nature with which they produced CV. While some of the speakers could be more easily distinguished than others, even those with a higher level of similarity still did not fully overlap in their preferred CV space. Further, some CV modes or combination of CV modes appear to be more unusual than others. The results suggest that the fine-grained assessment of CV in a forensic voice comparison could help to discriminate between speakers. Thus, the speaker-discriminating potential of CV production should be further explored by applying the proposed CV classification scheme to a larger set of speakers. Analysing a larger sample would shed more light on the typicality of CV modes and CV spaces. Further testing would also reveal whether the proposed CV classification scheme captured the most important dimensions of CV production or whether revision is needed.

Successful implementation of the method in casework practice requires the provision of training material that enables calibration and thus harmonises analyst-specific threshold values. The training material should include sample recordings from various speakers which exemplify (1) thresholds for distinct glottal pulse presence, and (2) highlight expected perceptual characteristics of potential CV modes. In addition, visual cues from waveforms and spectrograms should be made explicit which may be used to corroborate perception. The analysis process also requires simplification, as it is too time-consuming for a casework application – comparing one speech sample with another speech sample may already take up to 15 hours.

A future aim could be to determine the nature of CV not just based on perceptual-visual analysis but to also use acoustic information to corroborate perception. A combined auditory-acoustic approach adopts the “complementary strengths of the two approaches”

(Nolan, 1997, p.765). To incorporate acoustic analysis, it needs to be discovered which acoustic measurements correlate with the perceptually relevant CV categories. Approaches such as Devaraj et al. (2023) and Keating et al. (2023) provide suggestions for relevant acoustic parameters which could be explored in this respect. Implementing an auditory-acoustic approach for assessing laryngeal VQ such as CV may further increase transparency and comparability of these assessments.

Acknowledgments

We would like to thank Patricia Keating, Jessica Wormald and one anonymous reviewer for valuable comments on previous versions of the paper.

References

Al – Khalili, J. (Host). (2022, Oct 18). A passion for fruit flies [Audio podcast]. In: *The life scientific*. BBC radio 4.

Anikin, A., Pisanski, K., Massenet, M., & Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proc Royal Society B* 288(1954)., 20210872.

Batliner, A., Burger, S., Johne, B., & Kießling, A. (1993). MÜSLI: a classification scheme for laryngealizations. *Proc ESCA Workshop on prosody, Lund, Sweden*, 176–179.

Blomgren, M., Chen, Y., Ng, M. L., & Gilbert, H. R. (1998). Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *Journal of the Acoustical Society of America* 103(5), 2649–2658.

Boersma, P. & Weenink, D. (2023). Praat: doing phonetics by computer (Version 6.3.08) [Computer software].

Bradlow, A.R. (n.d.) ALLSSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings. [Data Collection].

Catford, J. C. (1964). Phonation types: The classification of some laryngeal components of speech production. In: Abercrombie, D., D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (eds.) *In honour of Daniel Jones: papers contributed on the occasion of his eightieth birthday 12 September 1961*, 26–37. London: Longmans, Green & Co.

Coleman, R. F. (1963). Decay characteristics of vocal fry. *Folia Phoniatrica et Logopaedica* 15(4), 256–63.

Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proc 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 532–536.

Devaraj, V., Roesner, I., Wendt, F., Schoentgen, J., & Aichinger, P. (2023). Auditory Perception of Impulsiveness and Tonality in Vocal Fry. *Applied Sciences* 13(7): 4186

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24(4), 423–444.

Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L., & Walshaw, D. (1997). Descriptive adequacy in phonology: A variationist perspective. *Journal of Linguistics* 33(2) 275–310.

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). Voice quality: the laryngeal articulator model. Cambridge University Press.

Fant, G. (1979). Glottal source and excitation analysis. *STL-Quarterly Progress and Status Report* 1, 85–107.

Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 101155.

Gerratt, B., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* 29(4), 365–381.

Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In: Hardcastle, W. J., J. Laver, & F. E. Gibbon (eds.) *The handbook of phonetic sciences*, 378–423. Wiley-Blackwell.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18(2), 293–307.

Gold, E., Ross, S., & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proc 19th Interspeech. Hyderabad, India*, 2748–2752.

Gwet, K. L. (2019). irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package (Version 1.0) [Computer Software]

Haddican, W., & Foulkes, P. (2017). A comparative study of language change in Northern Englishes. [Data Collection]. Colchester, Essex: ESRC.

Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In: Hyman, L. M., & C. N. Li (eds.) *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, 3–29. London, New York: Routledge.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics* 2(2), 125–143.

Ishi, C. T., Sakakibara, K. I., Ishiguro, H., & Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 47–56.

Keating, P. A., Garellek, M., Kreiman, J., Chai, Y. (2023). Acoustic properties of subtypes of creaky voice. *Journal of the Acoustical Society of America* 153(3), Article 297.

Kirchhübel, C. (2013). The acoustic and temporal characteristics of deceptive speech. [Doctoral dissertation, University of York].

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87(2), 820–857.

Kramer, E., Linder, R., & Schönweiler, R. (2013). A study of subharmonics in connected speech material. *Journal of Voice* 27(1), 29–38.

Ladefoged, P. (1971). Preliminaries to linguistic phonetics. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1), 159–174.

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. M. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 14, 139–155.

Laver, J. (2009). The phonetic description of voice quality. Cambridge University Press.

Moisik, S. R. (2013). The epilarynx in speech. [Doctoral dissertation, University of Victoria].

Moisik, S. R. (2013). Harsh voice quality and its association with blackness in popular American media. *Phonetica* 69(4), 193–215.

Moore, P., & von Leden, H. (1958). Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica* 10(4), 205–238.

Muhlack, B., Trouvain, J., & Jessen, M. (2023). Distributional and Acoustic Characteristics of Filler Particles in German with Consideration of Forensic-Phonetic Aspects. *Languages* 8(2), Article 100.

Murry, T. (1971). Subglottal pressure and airflow measures during vocal fry phonation. *Journal of Speech and Hearing Research* 14(3), 544–551.

Nolan, F. (1997). Speaker recognition and forensic phonetics, In: Hardcastle, W. J. & J. Laver (eds) *The handbook of phonetic sciences*, 744–767 Blackwell.

R Core Team (2021). R: A language and environment for statistical computing (Version 4.1.0) [Computer software].

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29(4), 407–429.

San Segundo, E. (2021). International survey on voice quality: Forensic practitioners versus voice therapists. *Estudios de fonética experimental* 9–34.

Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice* 20(2), 171–186.

Wendahl, R. W., Moore, G. P., & Hollien, H. (1963). Comments on vocal fry. *Folia Phoniatrica et Logopaedica* 15(4), 251–255.

White, H., Penney, J., Gibson, A., Szakay, A., & Cox, F. (2021). Optimizing an automatic creaky voice detection method for Australian English speaking females. In: *Proc 22nd Interspeech, Brno, Czech Republic*, 1384–1388.

Research Degree Thesis Statement of Authorship

University of York
York Graduate Research School

Candidate name	Katharina Klug	
Department	Language and Linguistic Science	
Thesis title	Assessing a speaker's voice quality for forensic purposes – Using the example of creaky voice and breathy voice	
Title of the work (paper/chapter)	Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception	
Publication status	**Published**	x
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	
Citation details (if applicable)	Klug, K., Kirchhübel, C., Foulkes, P., & French, P. (2019). Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception. In Proceedings of the 18th International Congress of Phonetic Sciences. Melbourne: Australasian Speech Science and Technology Association Inc(pp. 795–799).	
Description of the candidate's contribution to the work	Conceptualisation Data curation Formal analysis Investigation Methodology (lead) Writing – original draft	
Percentage contribution of the candidate to the work	85%	
Signature of the candidate	K. Klug	
Date (DD/MM/YY)	24/06/23	

Co — author contributions*

By signing this Statement of Authorship, each co — author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co — author	Christin Kirchhübel
Contact details of co — author	Soundscape Voice Evidence, Lancaster, UK ck@soundscapevoice.com
Description of the co — author's contribution to the work**	Investigation (supporting) Methodology (supporting) Writing – review and editing
Percentage contribution of the co — author to the work	5%
Signature of the co—author	
Date (DD/MM/YY)	26/06/2023

Name of co — author	Paul Foulkes
Contact details of co — author	Department of Language and Linguistic Science, University of York, UK paul.foulkes@york.ac.uk
Description of the co — author's contribution to the work*	Investigation (supporting) Methodology (supporting) Writing – review and editing
Percentage contribution of the co — author to the work	5%
Signature of the co—author	
Date (DD/MM/YY)	26/6/23

Name of co – author	Peter French
Contact details of co – author	J P French Associates, York, UK Department of Language and Linguistic Science, University of York, UK peter.french@york.ac.uk
Description of the co – author's contribution to the work*	Methodology (supporting) Writing – review and editing
Percentage contribution of the co – author to the work	5%
Signature of the co–author	
Date (DD/MM/YY)	26th June 2023

Copy and paste additional co – author panels as needed.

* Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

** The description of the candidate and co – authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co–authors may find it helpful to consider the CRediT (Contributor Roles Taxonomy) approach to recognising individual author contributions.

Analysing breathy voice in forensic speaker comparison using acoustics to confirm perception

Katharina Klug[1], Christin Kirchhübel[2], Paul Foulkes[1], and Peter French[1,3]

[1] *Department of Language and Linguistic Science, University of York, UK*
[2] *Soundscape Voice Evidence, Lancaster, UK*
[3] *J P French Associates, York, UK*

kk667@york.ac.uk, ck@soundscapevoice.com, paul.foulkes@york.ac.uk, and peter.french@york.ac.uk

Abstract: This study explores the interplay between perception and acoustics, focussing on breathy voice. Using perceptual analysis, four forensic speech analysts rated 22 spontaneous speech samples with regard to whether they were breathy or non-breathy. The voices rated to be the most extreme on the breathy/ non-breathy continuum were then analysed acoustically. Spectral slope and additive noise characteristics were obtained from vowels and sonorant consonants using VoiceSauce. Significant correlations were found between the perception of breathiness and three acoustic measures, namely the intensity difference between the lowest two harmonics, the intensity difference between the lowest harmonic and the harmonic closest to the first formant, and cepstral peak prominence. Our results confirm that the findings from previous studies in relation to non-spontaneous speech are also applicable to spontaneous speech samples. Further, there appears to be no detriment when broadening the sample to include sonorant consonants as well as vowels.

Keywords: voice quality, breathy voice, forensic speaker comparison

1 Introduction

Voice quality (VQ) can be a highly individual marker of a speaker's voice, due both to its anatomical and habitual origin (Laver et al., 1981). It is widely considered an important variable for forensic speech science (Gold & French, 2011), a field that seeks to identify features with power to discriminate between individuals.

Owing to its multidimensional nature (Beck, 2005), VQ is generally analysed perceptually. However, there are ongoing efforts to utilise acoustic analysis in order to confirm perceptual VQ judgements.

Perceptual judgements are inherently subjective, as standards and thresholds may vary from rater to rater (Kreiman et al., 2005). Also, individual raters' standards often lack stability. For example, the range of voices presented in one rating session can cause a drift in VQ judgement when re – rating the same voice (Gerratt et al., 1993).

Nonetheless, the perceptual approach does have advantages over acoustic approaches. Perceptual analysis enables a holistic assessment of a speaker's overall VQ, including aspects of respiration, phonation, and articulation (Beck, 2005). In contrast, acoustic measurements can only provide information on very specific VQ aspects, e.g. the open glottal quotient. Furthermore, the human ear is capable of capturing fine-grained VQ differences even under less than optimal conditions (Köster & Köster, 2004). Typically, the recordings which forensic speech experts face are of poor technical quality and contain only partially analysable speech. This may limit or even prevent accurate acoustic measurements. A reduction in low-frequency energy, for example, which is common for telephone-transmitted speech, often affects the frequency of the first formant (Byrne & Foulkes, 2004). Therefore, further research is needed into how stable various acoustic parameters are within speakers, and across different transmission channels and speaking styles.

Given that both perceptual and acoustic approaches have strengths and weaknesses – it is preferable to combine the strengths of both approaches to assess the multidimensionality of VQ. This would position VQ assessments in line with other variables commonly assessed in forensic casework using a combined auditory-acoustic approach (e.g. vowels and consonants).

This study focuses on breathy voice, to assess the extent to which it is possible to combine auditory-perceptual and acoustic approaches.

2 Acoustics of Breathy Voice

2.1 Spectral slope parameters

Breathiness arises from incomplete or non-simultaneous glottal closure resulting in a higher open quotient. This, in turn, yields a strong first harmonic amplitude (H1, Klatt & Klatt, 1990) and a steep spectral slope (Hillenbrand et al., 1994).

Previous studies analysed spectral slope using harmonic-based and formant-based measurements. Harmonic-based measurements obtain amplitude differences between (1) the first and the second harmonic (H1–H2), (2) the second and fourth harmonic (H2–H4), or (3) the fourth harmonic and the harmonic closest to 2 kHz (H4–H2K). Formant-based measurements calculate the difference in amplitude between H1 and the harmonics closest to the first three formants (A1, A2, A3, i.e. H1–An).

2.2 Additive noise parameters

In breathy voice high frequency aspiration noise is generated via a persistent glottal gap, causing a decrease in additive noise measurements (Hillenbrand et al., 1994). Respiration noise is commonly measured by obtaining harmonic-to-noise ratio (HNR) and cepstral peak prominence (CPP). HNR displays the amplitude difference between harmonic and noise energy (de Krom, 1993) and decreases in breathy voice, e.g. Garellek (2012). HNR can be measured for various frequency bands, typically 0–500/1500/2500 Hz (labelled HNR05/15/ 25).

CPP is a measure of cepstral peak amplitude relative to the overall amplitude. As a measure of periodicity, it is helpful in detecting less periodic signals, e.g. mid and high frequency ranges in breathy voice due to aspiration noise (Hillenbrand et al., 1994). Lower CPP values correlate with breathy phonation due to the low-intensity higher harmonics, e.g. Wayland & Jongman (2003).

3 Corpus Data

3.1 Previous studies

Studies of the modal/breathy distinction have investigated either contrastive phonation types of various languages (Garellek et al., 2013; Garellek & Keating, 2011; Keating et al., 2011) or pathological voices, e.g. Alpan et al. (2009). Non-pathological voices and non-contrastive uses have been neglected. Furthermore, most studies have based their analysis on sustained vowels, e.g. Alpan et al.

(2009); Hillenbrand et al. (1994), vowels from isolated words, e.g. Garellek et al. (2013); Garellek & Keating (2011); Wayland & Jongman (2003), read speech, e.g., Hillenbrand et al. (1994), or synthetically manipulated stimuli, e.g. Garellek et al. (2013); Klatt & Klatt (1990). Studies of spontaneous speech are rare (San Segundo et al., 2018). However, Zraick et al. (2005) reports significant differences in perceptual judgements due to shorter vowel duration and assimilation processes found in spontaneous speech.

3.2 Current study

A selection of non-pathological breathy voice samples was compiled using six corpora of spontaneous conversation from male speakers of British English (Gold et al., 20118; Haddican & Foulkes, 2017; Kirchhübel, 2013; Llamas et al., 2016-19; Nolan et al., 2009; Wormald, 2016). The samples were all provided at 44.1 kHz frequency and 16-bit resolution sampling. The first author chose 22 voices based on auditory-perceptual analysis, aiming to reflect a natural mixture along the breathy/non-breathy continuum. Approximately three minutes of speech was extracted from each sample. Using Audacity (version 2.1.2, Audacity Team, 2017) the maximum intensity level was equalised across samples (max. amplitude -1.0 dB, remove DC offset, centre on 0.0 vertically).

4 Methodology

4.1 Auditory-perceptual investigation

A survey was conducted to generate perceptual ratings for breathiness. Four listeners were engaged, all experts in forensic speech analysis, involved in training and research on VQ. All regularly use the same analysis scheme – a modified Vocal Profile Analysis (VPA, Laver, 1980) – to rate VQ in forensic casework.

Using the survey tool Qualtrics, the participants were provided with the 22 voices in random order. For each sample they were asked: *'Would you mark breathiness as a dominant feature of this speaker's voice?'* Three answer choices were given together with a comment box: *'(1) Yes, (2) No, it is present but not dominant, (3) No, it is absent'*. The survey took 30–40 minutes. The listeners were allowed to listen to the samples as often as they liked and could leave and resume the survey at any point. They used closed-cup headphones in a quiet environment.

The voices which were rated by all four listeners to be the most extreme on both ends of the breathy/ non-breathy continuum were chosen for acoustic analysis. 8 voices qualified: 5 were rated as dominantly breathy and 3 as non-breathy. The between-rater consistency for these 8 voices was established. Cohen's Kappa (κ) was calculated for each rater pair using RStudio (Version 1.1.463 RStudio Team, 2016). Table 1 shows that all pairs of raters reached at least 'moderate' agreement. Two pairs (1–2, 2–4) reached 'substantial' agreement, and one pair (1–3) obtained 'almost perfect' agreement ($\kappa = 0.86$).

Table 1 *Between-rater agreement in the perception survey (Cohen's Kappa for rater-pairs; weights: equal; subjects: 8; raters: 2)*

Rater $-$ pair	Kappa	z	p $-$ value	Agreement
1–2	0.65	2.83	0.005	substantial
1–3	0.86	2.66	0.008	almost perfect
1–4	0.50	2.19	0.029	moderate
2–3	0.56	2.83	0.005	moderate
2–4	0.75	2.19	0.029	substantial
3–4	0.43	2.19	0.029	moderate

4.2 Acoustic investigation

Acoustic analysis was carried out on the selected voices. Generally, acoustic analysis of phonation is based on vocalic segments only (Garellek & Keating, 2011; Hillenbrand et al., 1994), as vowels by nature contain source-specific information. However, forensic recordings might be of very short duration and therefore can be restricted in terms of analysable speech available. Therefore, the data used in the present study included all sonorants (vowels, glides [j, w], liquids [l, r] and nasals [m, n, ŋ]), as they all contain glottal source characteristics. Sonorants were manually segmented using oscillographic, spectrographic and perceptual-impressionistic information and labelled on a segment-by-segment basis using Praat textgrids (Boersma & Weenink, 2017). When comparing breathy voices with non-breathy voices, initial visual examination suggests that sonorant consonants and vowels behave similarly in terms of central tendency. Accordingly, in the following acoustic investigation vowels and sonorant consonants were combined.

4.2.1 Measurement procedure

VoiceSauce (version 1.31 Shue et al., 2011) was used to take measurements from labelled segments. Default settings were applied: 0.96 pre-emphasis, 25 ms window length, measurements at 1 ms frame shift. The lower F0 range was adjusted to 40 Hz to capture potential intermittent low frequency creak components. The maximum measureable F0 limit was set to 300 Hz. To prevent formants from boosting nearby harmonic amplitudes, the formant-corrected harmonic amplitude measurements (Iseli et al., 2007) implemented in VoiceSauce were obtained and marked by an asterisk (e.g. H1*–H2*). All measurements were averaged across each labelled sonorant. Thus, there were 680–1149 averaged measurements per speaker.

4.2.2 Hypotheses

Table 2 summarises the acoustic measurements taken. We predicted the voices rated as dominantly breathy to show steeper spectral slope and lower additive noise. Furthermore, we predicted H1*–H2* to be most useful, as it is a rough indicator of open quotient (Stevens & Hanson, 1995).

4.2.3 Data analysis

We generated boxplots for each acoustic parameter using RStudio (Version 1.1.473, RStudio Team, 2016), and we used the lme4 package (Bates et al., 2011) to perform linear mixed effects analyses on the relationship between perceptual ratings and acoustic parameters taken from all sonorants. VQ classification (breathy/non-breathy) was entered into each model as a fixed factor. Speaker-specific variation was accounted for by including by$-$speaker random slopes. The alpha level was set at $p < 0.05$.

Table 2 *Predicted effects for spectral slope and additive noise measurements in breathy voice.*

Measure	Parameter	Predicted Effect	Prev. Studies
Spectral slope	H1*–H2*	non-breathy < breathy	Klatt & Klatt (1990)
	H2*–H4*		Garellek et al. (2013); Kreiman et al. (2011)
	H4*–H2K*		Kreiman et al. (2011)
	H1*–An*		Garellek & Keating (2011); Wayland & Jongman (2003)
Additive noise	CPP	non-breathy > breathy	Hillenbrand et al. (1994)
	HNR		Garellek (2012)

5 Results

Figure 1 illustrates the results from the acoustic analysis of all sonorants. Overall, clear differences can be seen between the speakers rated to be dominantly breathy (in grey) and those rated to be non-breathy (in white). Indeed, the distributions for each speaker reveal very little overlap between breathy and non-breathy VQ for H1*–H2*, H1*–A1* and, in particular, CPP.

Table 3 shows the results from the linear mixed effects modelling for all sonorants for each acoustic parameter. Confirming the impressions gained from Figure 1, significant differences were found in H1*–H2* (p<0.05), H1*–A1* (p<0.05) and CPP (p<0.01). Results close to significance were obtained for HNR05 (p=0.097) and HNR15 (p=0.077). The other measures did not show significant differences between breathy and non-breathy ratings based on all sonorants.

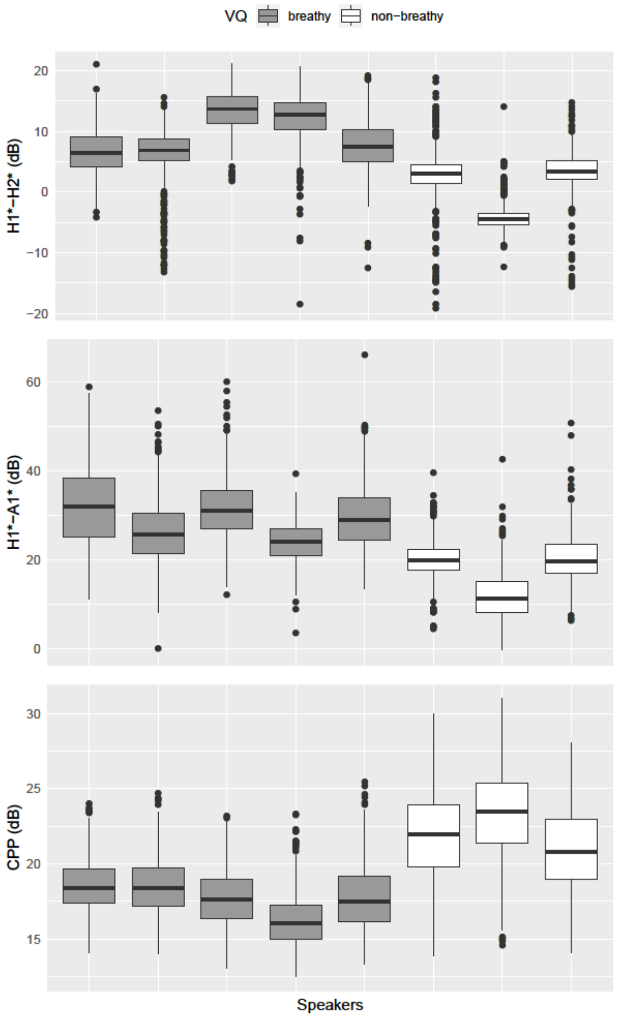


Figure 1 *Boxplots for acoustic parameters revealing the clearest differences comparing breathy with non-breathy voices (H1*–H2*, H1*–A1*, CPP).*

Table 3 *Estimate, standard error estimates (SE), t statistics, Satterthwaite approximated degrees of freedom (df) and predicting VQ classification (Pr(>|t|)) for each model (acoustic parameter). All models included by – speaker random slopes.*

Model	Estimate	SE	t	df	Pr(>|t|)
H1*–H2*					
(Intercept)	9.20	1.47	7.27	4.00	0.003
non-breathy	–8.51	2.84	–3.00	3.50	**0.047***
H2*–H4*					
(Intercept)	8.44	1.27	7.73	4.00	0.003
non-breathy	–0.01	1.72	–0.01	5.94	0.994
H4*–H2K*					
(Intercept)	5.27	0.97	5.42	3.98	0.007
non-breathy	–1.17	2.10	–0.55	3.11	0.719
H1*–A1*					
(Intercept)	28.78	1.70	17.92	4.00	0.000
non-breathy	–11.15	3.21	–3.47	3.35	**0.034***
H1*–A2*					
(Intercept)	31.70	1.40	22.73	4.00	0.000
non-breathy	–12.04	4.75	–2.53	2.39	0.107
H1*–A3*					
(Intercept)	25.19	1.34	18.82	4.01	0.000
non-breathy	–9.73	5.13	–1.89	2.30	0.182
HNR05					
(Intercept)	8.03	2.47	3.25	4.00	0.031
non-breathy	9.73	4.44	2.19	3.81	0.097
HNR15					
(Intercept)	17.73	1.94	9.13	4.00	0.001
non-breathy	7.57	2.93	2.24	4.88	0.077
HNR25					
(Intercept)	21.94	1.70	12.93	4.00	0.000
non-breathy	5.28	3.01	1.75	3.88	0.157
CPP					
(Intercept)	17.78	0.41	43.27	4.00	0.000
non-breathy	4.19	0.79	5.31	3.52	**0.009****

6 Discussion

The present study confirms the capability of two low frequency spectral slope parameters (H1*–H2*, H1*–A1*) and one additive noise parameter (CPP) to distinguish auditory-impressionistic judgements of dominantly breathy voices from non-breathy voices. These results are in line with previous studies (Alpan et al., 2009; Garellek et al., 2013; Garellek & Keating, 2011; Hillenbrand et al., 1994; Keating et al., 2011; Wayland & Jongman, 2003), but extended the findings from elicited speech to spontaneous speech samples. Mid-to-high frequency spectral slope parameters have previously been found to support the perception of breathiness (H2*–H4*: Garellek et al., 2013; Kreiman et al., 2011, H4*–H2K*: Kreiman et al. 2011, H1*–A2*: Garellek & Keating 2011 and H1*–A3*: Wayland & Jongman 2003). This was not the case here.

Given the more complex nature of spontaneous speech and the weaker intensity of sonorant consonants the results of the present study are promising.

7 Conclusion and future work

Our results indicate that the perception of breathy VQ in spontaneous speech is captured mainly in a steep spectral slope of low frequency ranges (H1*–H2*, H1*–A1*) and in a low cepstral peak prominence (CPP). This outcome lays open the potential to formally adopt the combined auditory-acoustic approach for the assessment of VQ when breathiness is involved.

The auditory-perceptual approach is still the ‘gold standard’ in VQ analysis (San Segundo et al., 2018), which we do not want to challenge. However, our results demonstrate that there is potential for perceptual analysis to be corroborated by acoustic measurements. This would most likely have a positive effect on within-rater and between-rater consistency. Including sonorant consonants increases the practicability of this analysis in the forensic setting as speech samples are often short.

It remains to be examined how the measurements investigated here perform in recordings of poorer quality. Work in progress will test the effect of a mobile-landline telephone filter on acoustic measurements to assess the robustness under forensically realistic conditions.

Acknowledgements

We thank our two additional expert listeners Richard Rhodes and Jessica Wormald. Further thanks go to Amanda Cardoso and Márton Sóskuthy for assistance with statistical analysis.

References

Alpan, A., Schoentgen, J., Maryn, Y., Grenez, F., Murphy, P. (2009). Cepstral analysis of vocal dysperiodicities in disordered connected speech. Proc. 10th Interspeech Brighton, 959–972.

Audacity Team (2017). Audacity (R): Free Audio Editor and Recorder [Computer application]. Ver- sion 2.1.2. released 25/11/2017. URL: http://www.audacityteam.org

Bates, D., Maechler, M., Bolker, B. (2011). lme4. R package version 0.999375-38.

Beck, J. M. (2005). Perceptual analysis of voice quality: the place of vocal profile analysis. In: Hardcastle, W. J., Mackenzie Beck, J. (eds)

A figure of speech. A festschrift for John Laver, 285–322. New York: Routledge.

Boersma, P., Weenink, D. (2017). Praat [computer program]. Version 7.0.22 (released 15/11/2017) URL: http://www.praat.org/

Byrne, C., Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. International Journal of Speech Language and the Law 11.1, 83–102.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Lang. Hear. Res. 36.2, 254–266.

Fraile, R., Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. Biomedical Signal Processing and Control 14, 42–54.

Garellek, M., Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. JIPA 41.2, 185–205.

Garellek, M., Keating, P., Esposito, C. M., Kreiman, J. (2013). Voice quality and tone identification in White Hmong. Journal of the Acoustical Society of America 133, 1078–1089.

Garellek, M. (2012). The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. Journal of Phonetics 31, 152–161.

Gerratt, B., Kreiman, J., Antonanzas-Barroso, N., Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. J. Speech Lang. Hear. Res. 36, 14–20.

Gold, E., French, P. (2011). International practices in forensic speaker comparison. International Journal of Speech, Language and the Law 18.2, 293–307.

Gold, E., Ross, S., Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. Proc. 19th Interspeech Hyderabad, 2748–2752.

Haddican, W., Foulkes, P. (2017). A comparative study of language change in Northern Englishes. [Data Collection]. Colchester, Essex: ESRC. URL: http://reshare.ukdataservice.ac.uk/851013/

Hillenbrand, J., Cleveland, R. A., Erickson, R. L. (1994). Acoustic correlates of breathy voice quality. J. Speech Lang. Hear. Res. 37, 769–778.

Iseli, M., Shue, Y.-L., Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. Journal of the Acoustical Society of America 121.4, 2283–2295.

Keating, P., Esposito, C., Garellek, M., Khan, S., Kuang, J. (2011). Phonation contrasts across languages. Proc. 17th ICPhS Hong Kong, 1046–1049.

Kirchhübel, C. (2013). The acoustic and temporal characteristics of deceptive speech, Doctoral dissertation, University of York.

Klatt, D. H., Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87, 820–857.

Kreiman, J., Vanlancker-Sidtis, D. Gerratt, B. R. (2005). Perception of voice quality. In: Pisoni, D. B., Remez, R. E. (eds.) *Handbook of speech perception*, 338–362. Oxford: Blackwell.

Kreiman, J., Garellek, M., Esposito, C. M. (2011). Perceptual importance of the voice source spectrum from H2 to 2 kHz. Journal of the Acoustical Society of America 130, 2570.

Köster, O., Köster, J. P. (2004). The auditory-perceptual evaluation of voice quality in forensic speaker recognition. The Phonetician 89, 9–37.

Laver, J., Wirz, S., Mackenzie, J., Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. Edinburgh University Department of Linguistics Work in Progress 13, 139–155.

Laver, J. (1980). The phonetic description of voice quality. Cambridge: Cambridge University Press.

Llamas, C., Watt, D., French, J. P. 2016–19. The use and utility of localised speech forms in determining identity: forensic and sociophonetic perspectives. ESRC ES/M010883/1

Nolan, F., McDougall, K., de Jong, G., Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech, Language and the Law 16.1, 31–58.

RStudio Team (2016). RStudio: Integrated development for R. RStudio, Inc., Boston, MA URL: http://www.rstudio.com/

San Segundo, E., Foulkes, P. French, P., Harrison, P., Hughes, V., Kavanagh, C. (2018). The use of the vocal profile analysis for speaker characterization: Methodological proposals. JIPA [online].

Shue, Y.-L., Keating , P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. Proc. 18th ICPhS Hong Kong, 1846–1849.

Stevens, K. N., Hanson, H. M. (1995). Classification of glottal vibration from acoustic measurements. In: Fujimura, O., Hirano, M. (eds.) *Vocal fold physiology: Voice quality control*, 148–180. San Diego: Singular.

Wayland, R., Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. Journal of Phonetics 31.2, 181–201.

Wormald, J. (2016). Regional variation in Panjabi-English, Doctoral dissertation, University of York.

Zraick, R. I., Wendel, K., Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. Journal of Voice 19.4, 574–581.

Research Degree Thesis Statement of Authorship

University of York
York Graduate Research School

Candidate name	Katharina Klug	
Department	Language and Linguistic Science	
Thesis title	Assessing a speaker's voice quality for forensic purposes – Using the example of creaky voice and breathy voice	
Title of the work (paper/chapter)	Assessing the suitability of f0 estimators with respect to recording condition and voice quality	
Publication status	**Published**	
	Accepted for publication	
	Submitted for publication	
	Unpublished and unsubmitted	x
Citation details (if applicable)		
Description of the candidate's contribution to the work	Conceptualisation Data curation Formal analysis (supporting) Investigation Methodology (equal) Writing – original draft	
Percentage contribution of the candidate to the work	90%	
Signature of the candidate	K. Klug	
Date (DD/MM/YY)	26/02/24	

Co – author contributions*

By signing this Statement of Authorship, each co – author agrees that:

(i) the candidate has accurately represented their contribution to the work;

(ii) if required, permission is granted for the candidate to include the work in their thesis (note that this is separate from copyright considerations).

Name of co – author	Markus Niermann
Contact details of co – author	Forensic Science Institute, Bundeskriminalamt, Wiesbaden, Germany markus.niermann@bka.bund.de
Description of the co – author's contribution to the work**	Formal analysis (lead) Methodology (equal) Software Writing – review and editing
Percentage contribution of the co – author to the work	10%
Signature of the co–author	Markus Niermann
Date (DD/MM/YY)	26/02/2024

Copy and paste additional co – author panels as needed.

* Note that where a paper has multiple authors, the statement of authorship can focus on the key contributing/corresponding authors.

** The description of the candidate and co – authors contribution to the work may be framed in a manner appropriate to the area of research but should always include reference to key elements (e.g. for laboratory-based research this might include formulation of ideas, design of methodology, experimental work, data analysis and presentation, writing). Candidates and co–authors may find it helpful to consider the CRediT (Contributor Roles Taxonomy) approach to recognising individual author contributions.

Assessing the suitability of f0 estimators with respect to recording condition and voice quality

Katharina Klug[1], and Markus Niermann[2]

[1] *Department of Language and Linguistic Science, University of York, UK*
[2] *Forensic Science Institute, Bundeskriminalamt, Wiesbaden, Germany*

Abstract: This exploratory study aims to test the suitability of fundamental frequency (f0) estimators with respect to recording condition and voice quality, which are hypothesised to negatively affect the validity of f0 measurements. In voice studies, f0 estimators are often blindly trusted because the extraction methods are difficult to understand. Valid f0 measurements are required to obtain valid spectral tilt measurements, because voice analysis programs as *VoiceSauce* (Vicenik et al., 2023) locate harmonics based on f0. Controlled productions of sustained cardinal vowels phonated in modal voice, breathy voice, and creaky voice are used to assess the validity of six f0 estimators: Praat Autocorrelation, Praat Cross-Correlation, REAPER, Snack, Subharmonic-to-harmonic ratio and STRAIGHT. The f0 estimators are tested under two recording conditions (studio, mobile phone). The results allow an informed choice of f0 estimators for future voice studies.

Keywords: f0 estimators, fundamental frequency, pitch, recording condition, voice quality

1 Introduction

The acoustic assessment of voices is becoming increasingly important in various disciplines dealing with speech in general and voice in particular. Signal processing, speech and language pathology, language acquisition, automatic speech and speaker recognition and forensic speech science are examples of disciplines that would benefit from a better understanding of the factors that influence voice acoustics. One of the crucial factors is the performance of f0 estimators, which often form the basis for further analysis steps. VoiceSauce, for example, is a voice analysis program (Shue, 2010) that performs automatic voice measurements. The f0 values estimated here are used to determine the location of the harmonics by looking for the highest amplitude within 10% of the estimated multiples of f0. Invalid f0 measurements therefore lead to the selection of spectral peaks that are not the actual harmonics, which in turn results in incorrect assumptions regarding the signal's spectral tilt (Shue et al., 2011, p.1846).

Most f0 estimators are expected to provide valid f0 values for modal voice samples from studio recordings. However, we hypothesise that voice qualities deviating from neutral phonation as well as degraded audio recording conditions cause problems. Voice qualities that deviate from modal voice are characterised by "confounding variables" that can affect the performance of f0 estimators. Breathy voice samples are characterised by high frequency noise and low-amplitude harmonics reflected in a low harmonic-to-noise ratio Wayland et al. (henceforth HNR 1997). Creaky voice (hereon CV) has a multifaceted nature (see, e.g., Keating et al., 2023a; Klug et al., in Press) that can cause various problems in f0 estimation. Klug et al. (in Press) distinguish between CV modes which are caused by amplitude damping effects (clean, harsh, and breathy CV), and aperiodic creak which is caused by aperiodicity of the glottal pulses. While aperiodicity is a challenge for f0 estimation in aperiodic creak, additional features of the amplitude damped CV modes can cause problems, e.g., multiple pulses in harsh CV or high-frequency noise (and thus low HNR) in breathy CV.

In contrast to studio recordings, the quality of a mobile phone filtered recording is degraded in several respects. The transmitted frequency range is band-pass limited, i.e., spectral components below and above this range are so weak that they are no longer detectable and components near the band-pass cut-off frequencies are attenuated. In addition, the codec has an influence on the speech signal. Speech codecs compress the digitized speech signal for a transmission at a lower data rate, often accepting quality degradations. Other factors such as indoor locations, tall buildings nearby or a moving signal, e.g., when a call is made from a moving car, further degrade the signal. The specific frequency range that is transmitted varies depending on mobile device, network provider and operating mode, which is determined by the channel conditions such as the number of simultaneous users. Codecs constantly switch between these operating modes which differ with respect to signal compression and transmitted frequency range (Cox et al., 2009, p.106). Using the example of the GSM-AMR narrowband codec, a long-established codec in the mobile network, the lower frequency limit for all modes is 100 Hz, while the upper frequency limit varies between the modes and ranges between 2.8k Hz and 3.6k Hz (Guillemin & Watson, 2008, p.201). All codecs either correct speech frames that have been damaged or lost during transmission or replace them by repetitions or extrapolations of previous speech frames (ibid.).

2 F0 estimators

F0 estimators may operate in different domains. The f0 estimators of interest in the current study perform either in the time domain, in the frequency domain, or in both domains (see Table 1). In the following, the operating principles of the f0 estimators are presented in a highly condensed and simplified way to derive hypotheses about their performance with respect to voice quality and recording condition. To facilitate readability, the names of the f0 estimators are abbreviated throughout the paper. Abbreviations are based on those used in VoiceSauce: PRAAT Autocorrelation – pF0ac, PRAAT Cross-Correlation – pF0cc, REAPER – rF0, Snack – sF0, Subharmonic-to-Harmonic-Ratio – shrF0, and STRAIGHT – strF0.

Table 1 *Overview of f0 estimators tested, sorted by working domain.*

Domain	F0 estimators			
Time	pF0ac	pF0cc	rF0	sF0
Frequency		shrF0		
Time-frequency		strF0		

2.1 pF0ac

pF0ac – Praat Autocorrelation – estimates the local f0 and evaluates the local harmonic strength to determine the harmonic-to-noise ratio as an indicator of the degree of periodicity (Boersma, 1993, p.98). It is a subtype of cross-correlation, as the windowed signal is cross-correlated with itself at various time lags. Since the correlation is greatest when the time lag is ‘0’, i.e., when the signal is correlated with the exact copy of itself, the lag with the next largest correlation peak is considered the f0.

To prevent the autocorrelation from declining, the autocorrelation function of the windowed signal is divided by the autocorrelation function of the window (ibid., 100). This processing step prevents octave jump errors. pF0ac spans the analysis window over three periods. The length of these three periods is determined by the pitch floor specified.

During post-processing, the global path finder searches for the global best path through the f0 candidates in each analysis frame. The path with the lowest cost of voiced/unvoiced decisions and octave jumps is preferred. pF0ac is the default method for pitch analysis in Praat (Boersma and Weenink, 1992–2023b).

2.2 pF0cc

The pF0cc – Praat Cross-Correlation – determines acoustic periodicity by forward cross-correlation (Boersma and Weenink, 1992–2023b). The cc function implemented in Praat basically corresponds to the RAPT algorithm, here sF0 (see explanation below, Boersma, 2020). pF0cc has a better temporal resolution than pF0ac as the analysis window spans only one actual period. It is implemented in Praat mainly for experimental purposes or for particularly short time windows (Boersma and Weenink, 1992–2023b). However, surprisingly, it is implemented as default method within VoiceSauce when picking the Praat f0 estimator.

2.3 rF0

The Robust Epoch And Pitch EstimatoR – REAPER – was developed by Talkin (2015) as a speech processing system. Epochs are time instants that indicate zero crossing points in the waveform which are assumed to represent glottal closure (Murty & Yegnanarayana, 2008, p.1602). The underlying EpochTracker detects the glottal closure instants (GCI), from which f0 is derived. This procedure roughly resembles the manual approach for determining f0.

After applying a high-pass filter to remove direct current (DC) bias and low-frequency noise, various features are extracted, e.g., graded GCI candidates. The GCI candidates are scored with respect to the preceding and following pulses and transformed into transition costs between successive periods in the signal. Finally, the best path of the GCI candidates is sought by starting at the end of the signal and tracing the path back through the signal. The period candidates with the lowest cost are considered the best. The output consists of f0 estimates and the estimated GCI location.

Epochs are also used by the SoE algorithm (strength of excitation) in VoiceSauce, which indicates the “relative amplitude of impulse-like excitation in each pitch pulse” (Keating et al., 2023b). Although both algorithms define epochs in the same way, they use different approaches to determine them. rF0 performs an adaptive LPC analysis to subtract the vocal tract from the signal and uses backtracking to find the optimal epoch track. Thus, each epoch largely depends on its predecessors and successors The SoE algorithm, on the other hand, uses a rigid filter function to derive to zero crossing points and is less dependent on adjacent zero crossing points.

2.4 sF0

sF0 – Snack – applies the RAPT algorithm which is released in the *ESPS* package (*Entropic Speech Processing System*) of the Snack library. It was developed by Talkin (1995).

sF0 operates a normalised cross-correlation function (NCCF) in two passes, using two differently sampled versions of the speech signal: (1) the version with the original sampling rate and (2) a version with a significantly reduced sampling rate. The reduced version (2) is used for the NCCF of the first pass to roughly search for locations of local maxima. The second pass uses the version with the original sampling rate (1) and performs the NCCF only in the vicinity of the already found local maxima to determine frequency and amplitude estimates (ibid., 508). In the process, f0 candidates are assigned to each frame or alternatively are hypothesised to be voiceless. Based on transition cost, octave-jump cost, and voicing-state transition costs, the best possible f0 path is determined.

2.5 shrF0

The shrF0 estimator – Subharmonic-to-Harmonic-Ratio – was developed by Sun (2000, 2002) with the aim of handling voices with alternating pulse cycles, i.e., amplitude and frequency modulations. Two basic requirements shaped the development: (1) the f0 estimator should reflect human perception, and (2) the magnitude of the subharmonics relative to the harmonics determines the degree of cycle alternation.

According to (Sun, 2002, p.135f.), shrF0 is less prone to pitch doubling and pitch halving errors. It is the only f0 estimator tested in the current study that operates in the frequency domain based on spectrum compression. This involves transforming the linear frequency scale into a logarithmic frequency scale and interpolating the results. To obtain the *sum of the subharmonic amplitude (SS)* and the *sum of the harmonic amplitude (SH)*, the spectrum is shifted on the logarithmic frequency abscissa at even orders for *SH* and respectively at odd orders for *SS* and added together. This process resembles spectral compression. The position of the global maximum and the next local maximum, i.e., SH and SS, are located. The relative magnitude of the two maxima determines which one is considered f0. Dividing these two summation values yields the subharmonic-to-harmonic-ratio (henceforth SHR).

Sun & Xu (2002) found pitch perception to correlate with SHR. If SHR is below the threshold value of 0.2, subharmonics have no influence on perception. With a SHR value above 0.4, the subharmonics are perceived as f0. The range between 0.2 and 0.4 seems to be perceptually ambiguous (ibid., p. 333). Accordingly, the shrF0 estimator uses a default threshold value of 0.4. Subharmonics exceeding this threshold value are considered F0 candidates.

Due to interpolation and harmonic summation, the method produces higher fine error rates, i.e., errors below 20% (Sun, 2000, p.678).

2.6 strF0

strF0 – STRAIGHT – examines f0 detection in the two domains, time and frequency, and thus takes another perspective on signal periodicity. In VoiceSauce (Vicenik et al., 2023), the XSX version from the VOCODER package TANDEM-STRAIGHT is implemented. STRAIGHT is a speech synthesis tool and was originally developed for perception-oriented speech research (Kawahara et al., 2012, p.387). TANDEM-STRAIGHT is the latest version.

Essentially, the f0 algorithm splits a sound into source and filter information and is therefore called *eXcitation Structure eXtractor* (XSX) (Fujimura et al., 2009, p.136). In this process, a temporally stable power spectrum (TANDEM spectrum) is divided by its corresponding spectral envelope (STRAIGHT spectrum). What remains is a power spectrum that only contains periodicity information, the fluctuation spectrum (Kawahara et al., 2008, p.3934). The f0 candidates generated in this way are evaluated by a set of periodicity detectors distributed over the specified frequency range. The periodicity detectors estimate the salience of each f0 candidate. Thus, multiple, coexisting quasi-periodic frequency candidates and transitions between harmonics and subharmonics are generated, which are considered to reflect pitch perception (Fujimura et al., 2009).

(Kawahara et al., 2012, p.387) revealed that even fast temporal f0 variations can be extracted, as the fundamental period is updated at each glottal cycle. Since even single, repetitive events are captured (Fujimura et al., 2009, p.169), strF0 is found to reliably represent transient phenomena, e.g., on- and offset. strF0 does not make voiced/unvoiced decisions and therefore yields f0 candidates even for voiceless or aperiodic segments (Penney et al., 2020, p.3243).

3 Data

A corpus compiled by Hemmen (2014) was used to analyse the acoustics. The original corpus contained cardinal vowel productions from five speakers producing four VQs: modal voice, breathy voice, creaky voice, and nasal voice (henceforth MV, BV, CV, NV). A subset of the corpus could only be used. The recordings of two male speakers contained clipping. These speakers were excluded because clipping distorts the signal's spectrum (Bie et al., 2015; Boersma & Weenink, 2023) and thus has an unpredictable effect on acoustic analysis. A further female speaker was eliminated as her creaky voiced sample was judged by the lead author (KK) to be not creaky but harsh. As nasal voice is a velopharyngeal rather than a laryngeal VQ feature, the results for nasal voice are not reported in this paper. This leaves one male and one female speaker producing primary and secondary cardinal vowels (plus [ə]). Cardinal vowel 12 [œ] was not recorded in the original corpus. This left 16 sustained vowels, which were each repeated three times. The total 48 token per speaker per VQ result in 144 token per speaker, which leads to a total of 576 tokens for two recording conditions and two speakers.

The data was analysed in two recordings conditions: studio and mobile. The mobile condition is generated by transmitting the recordings through an actual GSM mobile network. This was done in two recording sessions (1st rec session male speaker, June 2022/ 2nd rec session female speaker, February 2023). Differences in the equipment used between the two recording sessions were small and are specified in brackets (1st/2nd recording session). The individual recordings per speaker in studio quality were concatenated into one sound file. This enabled transmission within a single mobile phone call. The concatenated studio recording was replayed from a laptop using the audio software Sound Forge (version 9.0e/Sound Forge Pro version 11.0) via an audio interface (M-Audio M – Track Hub USB) connected via a

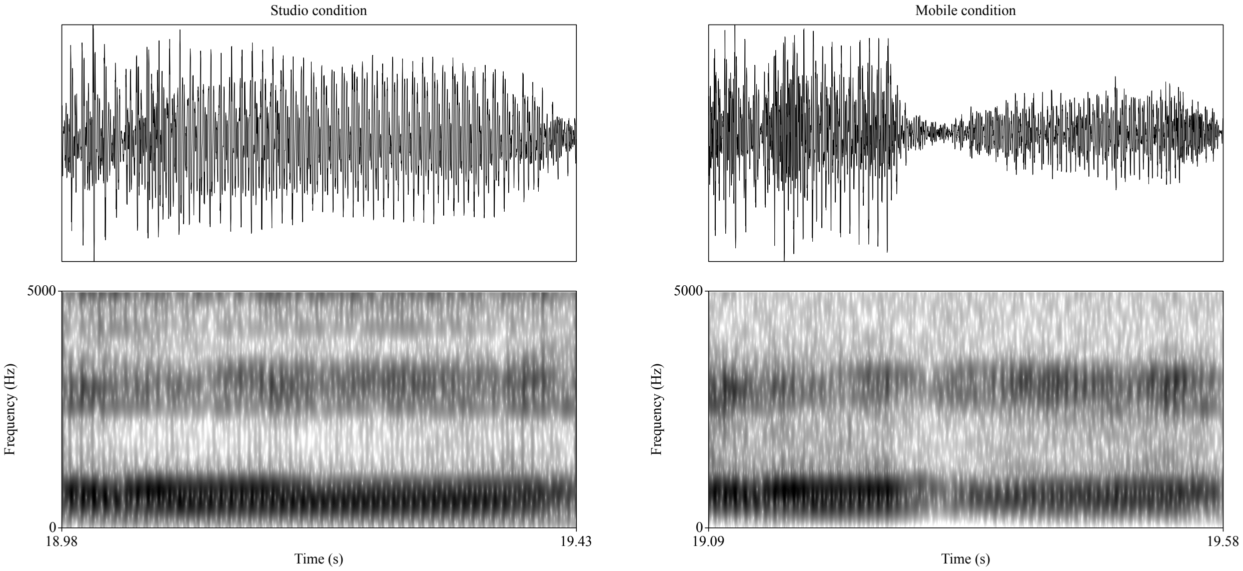


***Figure 1** The effects of the codec (male speaker, modal voice, cardinal vowels 13(2))*

cable to the input of a smartphone audio interface (Tascam iXZ) which was connected to the headset port of a mobile phone (Samsung Galaxy S8). A call was in progress between the mobile phone and a landline telephone (Audioline AUB 1) which was connected to a telephone balance unit (Prospect TC – 30). The output of the balance unit was routed via a mixer (Behringer Ultralink Pro) to increase the signal level before going to the input of an audio interface (M-Audio Delta 66) in a desktop computer. The mobile phone transmitted speech was recorded on this computer using Sound Forge (version 9.0e). This approach allowed the speech signal to be transmitted through the mobile phone network instead of producing an acoustic copy through the loudspeaker, which degrades acoustic detail. The transmitted frequency range is approximately 100 Hz to 3.6k Hz, although the frequencies below 400 Hz are highly attenuated. The impact of the codec is demonstrated in Figure 1 and Figure 2. Figure 1 displays differences in the waveform and the spectrogram between the studio and the mobile recording condition. In the mobile condition, the waveform is strongly attenuated at some points and less periodic. The spectrogram shows indistinct spectral energies with scarcely prominent formants.

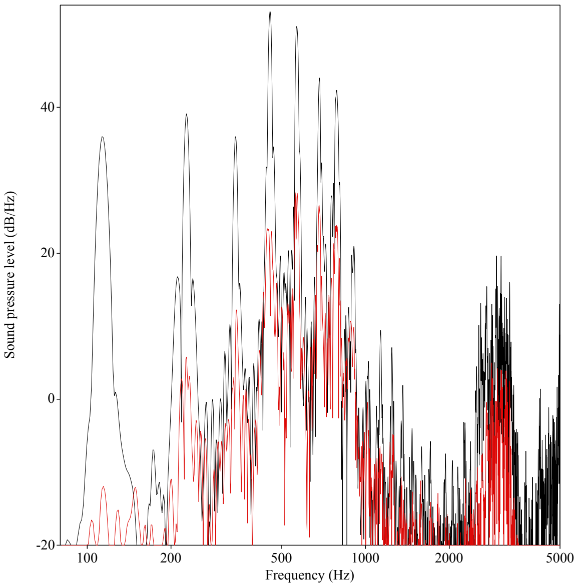


***Figure 2** Comparing LTAS on a logarithmic scale of the studio (black) and the mobile recording condition (red) (male speaker, modal voice, cardinal vowel 13(2))*

Figure 2 compares the long-term average spectra on a logarithmic scale of the studio recording (in black) with those of the mobile recording (in red). The lower frequency limit shows a gradual increase in amplitude, i.e., the channel is minimally stimulated around 100 Hz, but stays highly attenuated up to about 400 Hz and slightly attenuated up to about 500 Hz. The upper frequency limit ends quite abruptly around 3.6k Hz.

The two speakers differ in their production of creaky voice. The female speaker mainly produces aperiodic creak, which is characterised by aperiodically spaced glottal pulses. The male speaker, on the other hand, produces periodic clean CV and periodic harsh CV. While clean CV is characterised by perceptually distinct glottal pulses only, harsh CV is characterised by perceptually distinct glottal pulses together with harshness generated by amplitude modulations and/or subharmonics.

4 Hypotheses

Based on the f0 estimators' operating principles outlined in section two, the following hypotheses are derived.

pF0ac is described to be "accurate, noise-resistant and robust" (Boersma and Weenink, 1992–2023b). Since the method is said to measure HNR reliably, pF0ac is expected to perform well on VQs with low HNR. Those are VQs where the noise component is high and the amplitude difference between harmonics and noise is low (De Krom, 1993). Breathy voice and modes of creaky voice are characterised by a low HNR. However, (Sukhostat & Imamverdiyev, 2015, p.411) report pF0ac (and pF0cc) to be less robust for speech signals characterised by low f0. Creaky voice often occurs in connection with low f0 (Dallaston & Docherty, 2019). Therefore, the performance of pF0ac for creaky voice samples is difficult to predict.

Since pF0cc corresponds to the method of sF0, it is assumed that the two f0 estimators perform approximately the same. pF0cc was developed to detect frequency and amplitude modulations that are characteristic of harsh voice (not considered here) and certain modes of creaky voice. Therefore, pF0cc may be more reliable for creaky voice samples. However, as stated above, this hypothesis contradicts the findings from (Sukhostat & Imamverdiyev, 2015, p.411), that pF0cc is error-prone for low-frequency signals. Consequently, the performance of pF0cc on creaky voice samples is also highly questionable.

REAPER is said to be particularly suitable for estimating f0 from creaky voiced signals. It is designed for application on studio-quality recordings but is reported to be "fairly robust to recording quality" (Talkin, 2015). However, phase distortion negatively affects the performance of the EpochTracker. Phase distortions can be caused by non-specialised recording equipment and filter effects (Ó Cinnéide, 2012). Therefore, the mobile condition, which can result in phase distortions, can be challenging.

sF0 is reported to be robust to f0 height and noise condition and is suitable for "peculiar voice[s] and recording conditions" (Talkin, 1995, p.508). Therefore, sF0 is hypothesised to outperform other f0 estimators under mobile condition and when VQ differs from modal voice.

Keelan et al. (2010) reported that pF0ac, pF0cc and sF0 perform worse on female voices than two other algorithms in their study not considered here (SWIPE and SHS). They show that pF0ac, pF0cc and sF0 produce more pitch halving errors for female speakers. In the current study, pF0ac and pF0cc are therefore expected to produce more errors in female voices. Whether sF0 is prone to errors in female speakers is questionable, as (Talkin, 1995, p.508) reports robustness for f0 height.

The shrF0 estimator is also hypothesised to outperform other f0 estimators for voices characterised by amplitude and/or frequency modulations, i.e., for harsh creaky voice. The shrF0 is assumed to produce errors in speakers with SHR values between 0.2 and 0.4, as also perception is unambiguous within this range (Sun & Xu, 2002). Since the shrF0 method contains interpolation and summation processes, the performance of the f0 estimator is hypothesised to be less precise. Therefore, we expect fewer octave errors for shrF0, but more small-scale errors.

Fujimura et al. (2009) have shown that strF0 accurately determines f0 in multiple-pulsed voices, which may also be characteristic for harsh creaky voice. Furthermore, strF0 is expected to provide more reliable results in syllable on- and offset due to its ability to capture fast temporal f0 variations. However, as strF0 is designed for high-quality input signals, the mobile condition is expected to cause problems. In addition, strF0 requires segment labelling, as it does not perform voicing decisions. Thus, breathy voices, characterised by too little voicing as well as aperiodic creak will be prone to voicing decision errors.

Table 2 summarises the hypotheses regarding the performance of the f0 estimators examined in the current study. A plus sign ‘+’ indicates that the f0 estimator is hypothesised to outperform other estimators for the independent variable indicated, e.g., VQ. In contrast, a minus sign ‘−’ predicts that the independent variable in question will challenge the respective f0 estimator. A question mark ‘?’ indicates that there is conflicting information about the f0 estimator’s performance, e.g., ‘+’ for low HNR voices but ‘−’ for low-frequency voices as in creaky voice for pF0ac. All signs in brackets () refer to a specific mode of creaky voice, i.e., harsh creaky voice (according to Klug et al., in Press).

Table 2 *Summary of hypotheses regarding the performance of tested f0 estimators. (Abbreviations: BV–breathy voice, CV–creaky voice, Mob–mobile recording condition, f–female speaker, plus sign–good performance, minus sign–bad performance, question mark–unknown performance, brackets–performance of specific CV mode)*

f0 estimators	**Independent variables**			
	VQ		**Rec cond**	**Speaker sex**
pF0ac	+BV	?CV		–f
pF0cc		?CV		–f
rF0		+CV	−Mob	
sF0	+BV	+CV	+Mob	?f
shrF0		(+)CV		
strF0	–BV	?CV	–Mob	

In addition to the strengths and weaknesses of specific f0 estimators under degraded recording conditions, also the potential impact of a missing fundamental frequency on f0 detection in general should be addressed. Figure 3 shows the effect of a highpass-filter in which f0 is missing on the time and frequency domain. The normative power spectrum (top) compares the power spectrum of the original signal (dashed red line) with the power spectrum of the highpass-filtered signal (solid blue line). It can be seen that an absent fundamental frequency (here 100 Hz) has a negligible effect on (1) the amplitude or (2) the frequency of higher harmonics. We therefore assume that state-of-the-art f0 estimators also operate reliably even in the limited frequency range of a mobile filtered recording.

The centre image in Figure 3 shows the differences in the time domain. Although

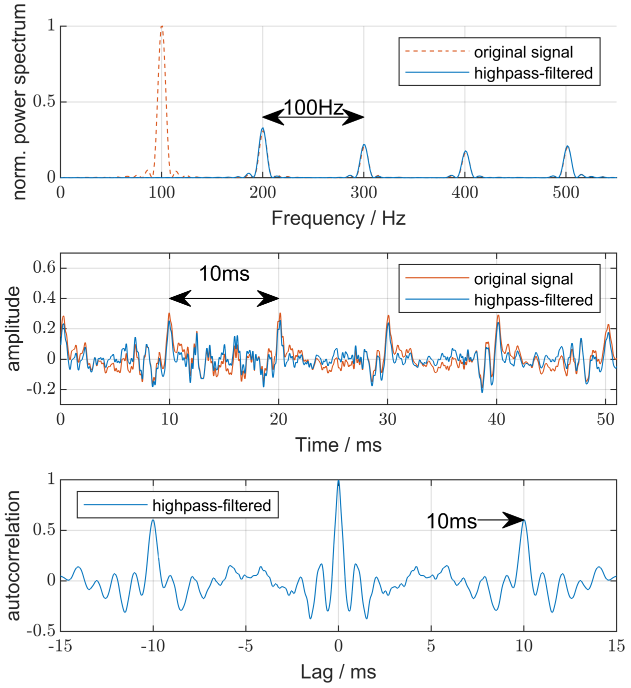


Figure 3 *Comparison of an original unfiltered signal, f0 = 100 Hz (dashed red line) and a highpass-filtered signal, f0 missing (solid blue line), in the frequency domain (top), time domain (centre) and its autocorrelation function (bottom)*

the complex waveforms differ between the original and the highpass-filtered signal, the peaks occur every 10ms, indicating that the time domain is not affected by a missing fundamental frequency. This conclusion is supported by the autocorrelation function, depicted in the bottom image, which shows the second highest peak at a time lag of 10ms for the highpass-filtered signal. Thus, the limited frequency range in a mobile filtered recording should also not affect the performance of f0 estimators operating in the time domain.

5 Methodology

5.1 Pre-processing

Not all f0 estimators perform voicing decisions, e.g., strF0. Therefore, the vowels in the signal were segmented using Praat TextGrids (Boersma and Weenink, 1992–2023b) in order to base acoustic analysis on vowels only, A Praat script was used to delete the pauses between the vowels, leaving the cardinal vowels only.

Segmentation was performed separately for both recording conditions, as the codec also affects the time signal, resulting in different vowel lengths (see Figure 1). This approach also allows to test correlations between the performance of f0 estimators and vowel characteristics (i.e., vowel height, vowel backness, lip rounding).

5.2 Manual assessment of ground truth f0

To assess the performance of the f0 estimators, the f0's ground truth was determined manually using the vowel-only recordings in the two recording conditions (studio, mobile). To do so, a Praat pitch object was created per recording which generated f0 candidates and suggested an f0 path. The pitch object enables the user to manually correct the path by either choosing different candidates or devoice frames, i.e., measure points. The lead author (KK) went through each recording period-by-period, manually determining each period's length to derive f0 (f=1/T). The f0 path was manually corrected accordingly. If none of the suggested f0 candidates within the Praat pitch object corresponded to the manually measured period, the measure point was devoiced and thus removed from analysis. This left only reliable measurements for analysis. In the creaky voice samples of the two speakers, the f0 was rarely ever below 30 Hz. Since the lower f0 limit for the f0 estimators for the creaky voice samples was set at 30 Hz, these sections were devoiced to match the frequency range specified for the f0 estimators. The manually corrected f0 path was exported as list and imported in VoiceSauce using the *manual data input* option. Using the option *resample to length*, the manually determined measure points were interpolated with the data from the f0 estimators.

For two recordings it was difficult to determine the ground truth f0: the *creaky voice* sample for the female speaker and the *breathy voice* sample for the male speaker. The *creaky voice* sample was characterised by a high level of aperiodicity and thus included sections of highly variable period lengths with barely two periods of the same length. The *breathy voice* sample showed a tendency towards whispery voice which distinguishes from breathy voice by laryngeal constriction (Esling et al., 2019, p.58). In whispery voice, f0 estimators may confuse non-harmonic friction energy with harmonics (Laver et al., 1982). Furthermore, the degree of voicing was at some points very low, resulting in sections of rather breath, respectively whisper, which lacked a clear f0. As the two recordings suffered further from the mobile transmission, two versions of ground truth f0s were produced for the mobile condition. Ground truth 1 (henceforth gt1) contained the results of a very precise pulse-by-pulse measurement strategy. It thus captured the aperiodic f0 values for the creaky voice sample of the female speaker and the barely voiced sections in the breathy voiced sample of the male speaker. Ground truth 2 (henceforth gt2), in contrast, was generated according to perception. If a stretch of speech was perceived as noisy or voiceless rather than voiced due to aperiodicity or the lack of voicing, the corresponding frames were devoiced. This led to substantially less f0 values in the gt2 which could be used for analysis.

5.3 Acoustic analysis

VoiceSauce (Shue, 2010) was used to conduct the acoustic analysis. The software automatically outputs a wide range of voice measurements (e.g., f0, formants, spectral tilt measurements, CPP, additive noise measurements). VoiceSauce is increasingly used in voice analysis. It proved useful for the current analysis as it already includes four f0 estimators, namely strF0, sF0, pF0, and shrF0, the latter two having been implemented after the first release (ibid.). The pF0 is implemented as cross-correlation, i.e., pF0cc. To also determine the performance of pF0ac, the source code was changed accordingly by changing `vars.F0Praatmethod` in line 48 from 'cc' to 'ac' in the *vs_Initialize.m* file (Y. Shue, personal communication, Dec 17 2020).

Additionally, rF0 (Talkin, 2015) was implemented as it is repeatedly reported to outperform f0 estimation on creaky voices (Doreen, 2017; Dallaston & Docherty, 2019; Szakay & Torgersen, 2019; Penney et al., 2020)).

F0 measurements were taken every 10ms. Within VoiceSauce most f0 estimators do not use a fixed window length but operate pitch synchronously. Only sF0 uses a fixed window size of 25ms. Pre-test showed that a wide frequency range that was set constant across all voice qualities (30–300 Hz for the male speaker, 30–500 Hz for the female speaker) led to higher error rates for most f0 estimators, especially for shrF0. rF0, in contrast, was not affected by a wider frequency range. Specific frequency ranges per speaker and VQ were evaluated to reduce the number of f0 detection errors. The *two-pass detection process* (De Looze & Hirst, 2008; Hirst & de Looze, 2021) was used to find the most suitable frequency range. In the first pass Praat's recommended range for a

Table 3 *Specified f0 ranged per speaker and VQ based on the two-pass detection process (De Looze & Hirst, 2008)*

VQ	Male (Hz)	Female (Hz)
Breathy voice	80–200	130–300
Modal voice	80–200	160–370
Creaky voice	30–130	30–430

male voice (75–300 Hz), respectively a female voice (100–500 Hz) was used (Boersma and Weenink, 1992–2023b) to detect the speakers' broad f0 range. For creaky voice the lower frequency range was adjusted to 30 Hz to enable the detection of very low-pitched creaky sections, which may be as low as 38.1/43.3 Hz (Keating & Kuo 2012, studying male/female American English speakers). From the resulting f0 range the first and the third quartile values (25% and 75%) were multiplied by a coefficient, i.e., 0.75 and 1.5. The results obtained yielded the defined frequency range that was used in the second pass to generate the final f0 estimates. Table 3 lists the specified f0 ranges per speaker and VQ. All data points were used for analysis, resulting in 2403–4060 measure points per VQ condition.

5.4 Statistical analysis

All data points for which the ground truth data and/or the f0 estimator data returned the value '0' and were thus rated to be voiceless are excluded from the analysis. The performance of each f0 estimator was compared to the f0 ground truth (gt1). This allows us to evaluate the performance of each f0 estimator. These comparisons were made with respect to the following independent variables: *voice quality* and *recording condition*. As only one male and one female speaker were analysed, limited conclusions on *speaker sex* could be drawn. The same applies to *vowel characteristics*, i.e., vowel height, vowel backness, and lip rounding, given the small sample size.

Two error measures were used, indicating the difference between the ground truth f0 and the specific f0 estimator in Hz. The *Mean Error Absolute* (henceforth *MEA*) is defined as

$$\mathit{MEA} = \frac{1}{K}\sum_{k=0}^{K-1}\left|\hat{f}\left(k\right) - f_{gt}\left(k\right)\right| \,,$$

where $\hat{f}\left(k\right)$ denotes the f0 estimate at frame k, $f_{gt}(k)$ the ground truth at frame k, and K the total number of frames. It specifies the distance between these two frequencies without weighting the error. Thus, few large errors are not punished more than many small errors. We favoured this error measure over its weighted counterpart, the *root mean square error*, because we consider octave errors to be less serious than errors that are not harmonic-related. For voice qualities characterised by multiple pulses, such as harsh creaky voice, uncertainty in octave height even reflects perception as the amplitudes of $f_0/2$ and f_0, or f_0 and $2f_0$ may be equally high in magnitude. To reflect this view, a new error measure is introduced by the second author (MN), which does ignore all errors related to octave: *Octave-Corrected Mean Error Absolute* (henceforth *OMEA*)

$$\mathit{OMEA} = \frac{1}{K}\sum_{k=0}^{K-1}\left|e\left(k\right)\right| \,,$$

where $e(k)$ is defined as specified below, omitting k for readability:

$$e = \min\left(\left|\hat{f} - \frac{f_{gt}}{4}\right|, \left|\hat{f} - \frac{f_{gt}}{3}\right|, \dots \right.$$
$$\left. \dots, \left|\hat{f} - 3f_{gt}\right|, \left|\hat{f} - 4f_{gt}\right|\right)$$

As explained in section 4.3 the performance of the f0 estimators for two voice qualities in mobile condition are compared to two different ground truths (gt1, gt2), i.e., the creaky voice sample of the female speaker and the breathy voice sample of the male speaker.

6 Results

The Tables 4–7 summarise the most important findings arranged by independent variables, shaded in dark grey. The most striking results are shown in bold. For each tested (combination of) independent variable(s), the three best performing f0 estimators are listed.

Table 4 summarises the differences in *recording condition* across both speakers and across all three VQs. In studio condition pF0cc, sF0 and pF0ac performed best, deviating between 4.1 and 4.8 Hz from the f0 ground truth. The same f0 estimators outperformed in the mobile condition deviating on average between 3.7 and 5.6 Hz. Surprisingly, the best f0 estimator, pF0cc, produced a lower error rate in the mobile condition compared to the studio condition, although the studio condition was hypothesised to be less error-prone.

Table 4 *Best performing f0 estimators in studio and mobile recording condition (across the two speakers and all VQs).*

Recording condition	Best f0 estimator	f0 MEA (Hz)
Studio	pF0cc	4.1
	sF0	4.5
	pF0ac	4.8
Mobile	pF0cc	3.7
	pF0ac	4.8
	sF0	5.6

By adding the independent variable *speaker sex* (see Table 5), it became clear that the f0 estimators produced substantially higher errors for the female speaker (5.7-6.8/ 4.6-7.8 Hz) than for the male speaker (1.2-1.8/ 0.9-1.6 Hz) in both recording conditions.

Table 6 reveals the most valid f0 estimators in studio and mobile *recording condition* with respect to *speaker sex* and *VQ.* It shows that the female's high MEA values across all VQs in Table 5 were due to the poor f0 performances of the creaky cardinal vowels. While the other VQs deviated from the ground truth f0 by a maximum of 1.6 Hz, the best f0 estimator for CV, rF0, yielded a MEA of 6.3 Hz in the studio and 4.2 Hz in the mobile condition. The error rate of the second best f0 estimator pF0cc of the female's CV sample was much higher in both recording conditions (26.1 Hz/ 19.9 Hz). These high MEA values were probably due to the specific mode in which the female speaker produced CV. While the male speaker used periodic *clean CV* and periodic *harsh CV*, according to the classification scheme proposed by Klug et al. (in Press), the female speaker produced *aperiodic creak*.

Table 5 *Best performing f0 estimators in studio and mobile recording condition with respect to speaker sex (across all VQs).*

Recording condition	Speaker sex	Best f0 estimator(s)	f0 MEA (Hz)
Studio	male	sF0	1.2
		strF0	1.5
		pF0ac/pF0cc	1.8
	female	**pF0cc**	**5.7**
		pF0ac	**6.7**
		sF0	**6.8**
Mobile	male	sF0	0.9
		pF0ac	1.4
		pF0cc	1.6
	female	**pF0cc**	**4.6**
		pF0ac	**6.4**
		sF0	**7.8**

Table 6 Best performing f0 estimator in studio and mobile recording condition with respect to speaker sex and VQ.

Recording condition		Studio		Mobile	
Speaker sex	**VQ**	**Best f0 estimator(s)**	**f0 MEA (Hz)**	**Best f0 estimator(s)**	**f0 MEA (Hz)**
male	BV	sF0	1.5	sF0	1.4
		pF0ac	1.8	pF0ac	2.8
		strF0/pF0cc	2.0	pF0cc	4.1
	MV	sF0	0.7	sF0	0.7
		pF0ac	0.8	pF0ac	0.8
		strF0/pF0cc	0.9	pF0cc	0.9
	CV	sF0	1.4	sF0	1.2
		strF0	1.6	pF0ac/pF0cc	1.6
		rF0	2.2	rF0	3.0
female	BV	sF0	1.6	sF0	1.6
		strF0	2.0	pF0ac	3.1
		pF0ac	2.2	pF0cc	3.5
	MV	pF0ac/strF0	1.1	pF0ac	1.1
		pF0cc	1.2	pF0cc	1.2
		shrF0	1.8	strF0	1.6
	CV	**rF0**	**6.3**	**rF0**	**4.2**
		pF0cc	**26.1**	**pF0cc**	**19.9**
		sF0	**29.9**	**sF0**	**25.2**

As aperiodicity is characterised by extremely variable pulse lengths, the female CV sample was especially challenging for f0 estimation and thus produced considerably higher error rates. The fact that the MEA of rF0 in the mobile condition was lower compared to its MEA in the studio condition was somewhat surprising and cannot be explained.

As hypothesised, MV produced the lowest error rates across the tested VQs for *speaker sex* and *recording condition*. The lowest MEA diverged from the ground truth by only 0.7 Hz for the male speaker (sF0) and by 1.1 Hz for the female speaker (pF0ac/strF0). With the exception of rF0, all tested f0 estimators performed well on MV in studio condition, producing MEA values of 2 Hz maximum.

The mobile recording condition did not pose problems for the best f0 estimators listed in Table 6 but yielded equally low or even slightly lower error rates. However, this did not apply to all f0 estimators tested. strF0 produced considerably higher MEA values for all three VQs produced by the male speaker in the mobile condition than in the studio condition. For the high pitched VQs of the female speaker the error rate of strF0 was surprisingly low in the mobile condition, namely 1.6 Hz in MV and 5.2 Hz in BV.

Overall, the best f0 estimator for the male speaker in both recording conditions was sF0. For the female speaker, the result was more divers, but seems to be driven by VQ rather than by the recording condition: pF0ac for MV, sF0 for BV, and rF0 for CV.

Looking at the OMEA measure, which ignores octave jump errors, it becomes apparent that octave errors mainly occurred in the female's CV samples of most f0 estimators tested, especially in shrF0 (see Table 7, also deducible from Figure 5).

As expected, the performance of most f0 estimators improved when evaluated against the ground truth 2 (gt2), which devoiced sec-

Table 7 Performance of f0 estimators for CV of the female speaker, performance under studio recording condition is compared to gt1 only, comparison under mobile recording condition is compared to gt1 (shaded white) and to gt2 (shaded grey)

Recording condition	f0 estimators	F0 MEA (Hz)	F0 OMEA (Hz)
Studio	pF0ac	34.4	3.9
	pF0cc	26.1	4.0
	rF0	6.3	2.8
	sF0	29.9	3.8
	shrF0	75.6	10.5
	strF0	42.0	8.3
Mobile	pF0ac	33.3	3.4
		14.2	1.3
	pF0cc	19.9	2.8
		4.1	1.4
	rF0	4.2	2.5
		2.9	1.9
	sF0	25.2	3.0
		10.7	1.5
	shrF0	92.2	9.9
		83.5	8.8
	strF0	59.0	8.5
		63.9	8.5

tions with apparently low degree of voicing in the male's BV sample and the female's CV sample in the mobile recording condition. This perception-based approach removed challenging frames from the analysis. Especially for the CV sample of the female speaker the error rate lowered for five out of the six f0 estimators (see entries in Table 7 shaded in grey).

No clear effect was found when assessing a correlation between the performance of f0 estimators and vowel characteristic, i.e. vowel height, vowel backness, and lip rounding. The neutral vowel [ə] seems to cause the least challenge for f0 estimators for both speakers in both recording conditions. However, especially for this independent parameter the sample size was way too little to allow for conclusion.

Figure 4 and 5 illustrate the performances of all f0 estimators for the male speaker (Figure 4) and the female speaker (Figure 5). The f0 estimators are arranged vertically, while the voice qualities (ordered by recording condition) are arranged horizontally. The error metric MEA is indicated above each plot. The thick blue line illustrates the f0 ground truth 1, while the thin red line represents the performance of each f0 estimator. Deviations from the f0 ground truth occur mostly as octave jump errors visible as large orange spikes. Smaller spikes indicate small-scale errors.

Some errors only became apparent through the visualization, which were not detected on the basis of the error metrics alone. For example, strF0 produced striking errors for the CV samples of both speakers, but only in the mobile condition. The f0 estimator predominantly outputs f0 values around 50 Hz without much variation, while the ground truth was (1) higher and (2) more variable for both speakers. This applies even more to the female speaker than to the male speaker. We suspect that the electrical network frequency, which varies between 49.5 and 50.5 Hz in the UK, was misinterpreted as f0, as it is very prominent in the mobile recording condition. Interestingly, none of the other f0 estimators tested showed similar tendency. Since the lower frequency limit for MV and for BV was considerably higher, i.e., 80/130/160 Hz, the mains hum for these voice qualities could not be misinterpreted as f0.

In addition, strF0 produced fewer errors in the mobile recording condition when the samples were generally high in pitch. More specifically, while the error metric for the female's MV and BV samples in the mobile condition produced an MEA of 1.6 Hz and

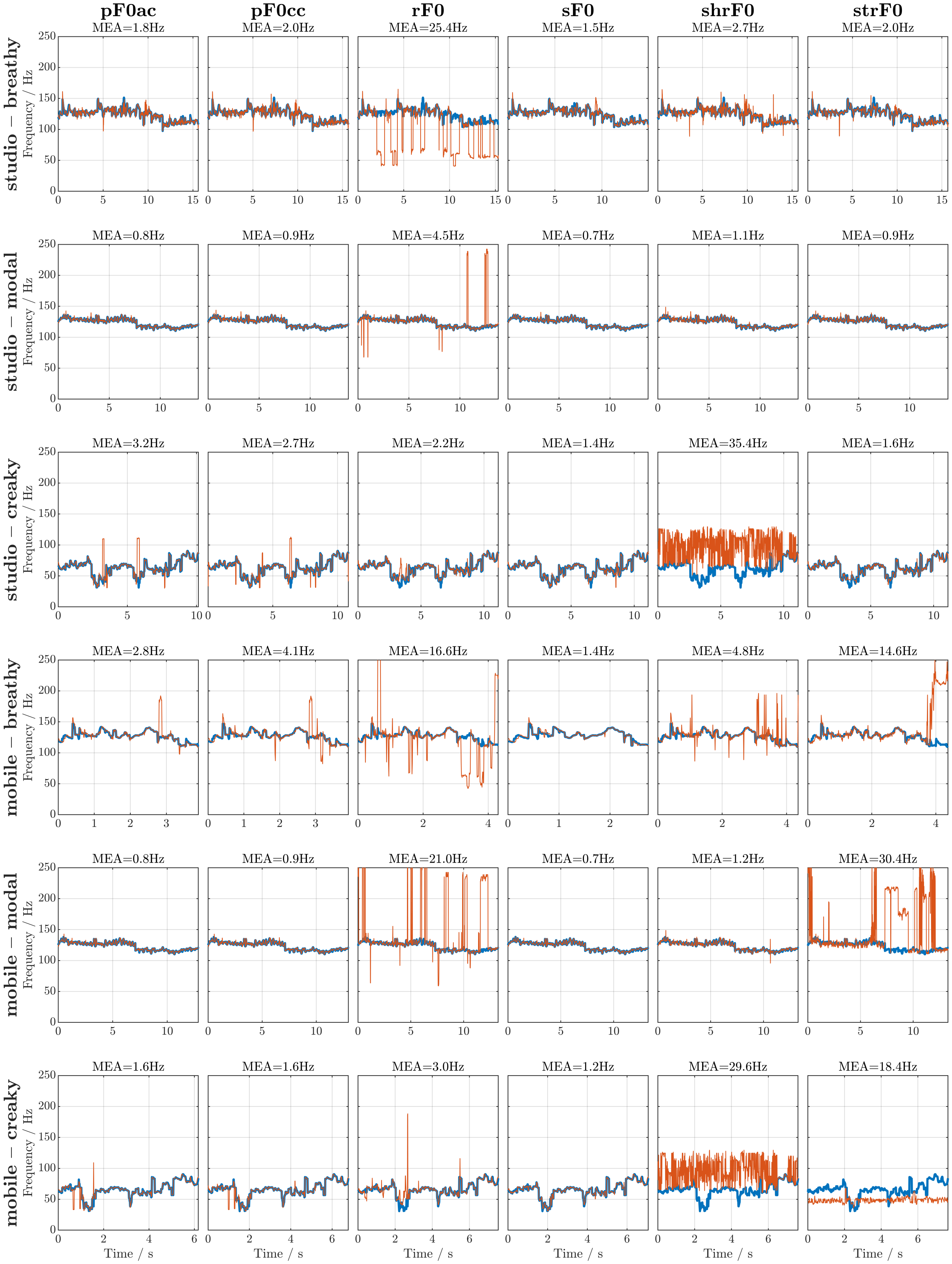


Figure 4 *Performance of the individual f0 estimators for the male speaker across all cardinal vowels in both recording phone conditions. The thick blue line illustrates f0 ground truth 1, the thin red line shows the performance of each f0 estimator. The name of each f0 estimator and the error metric MEA are shown above each plot.*

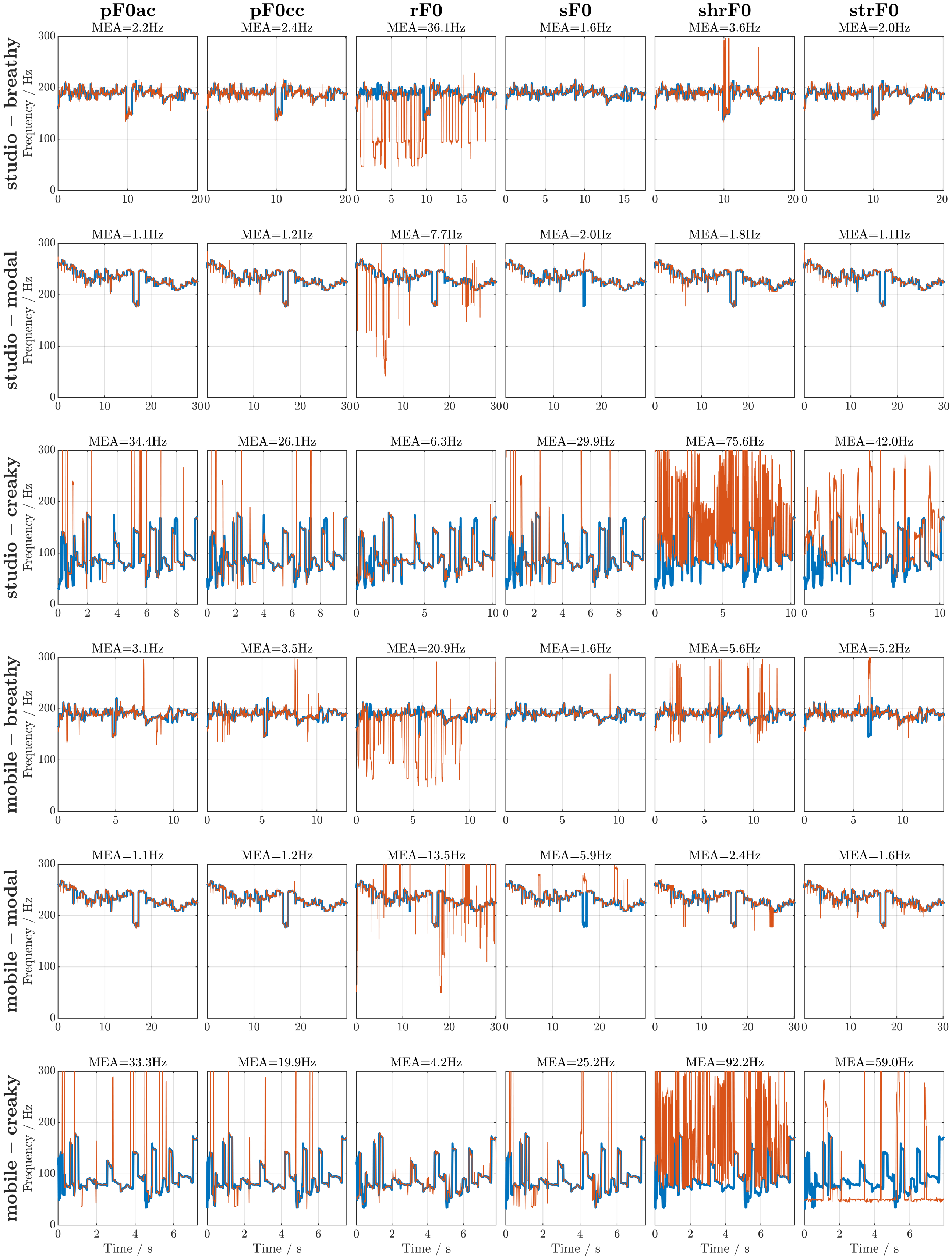


Figure 5 *Performance of the individual f0 estimators for the FEMALE SPEAKER across all cardinal vowels in both recording phone conditions. The thick blue line illustrates f0 ground truth 1, the thin red line shows the performance of each f0 estimator. The name of each f0 estimator and the error metric MEA are shown above each plot.*

5.2 Hz respectively, the male's MV and BV samples had much higher error rates, i.e., 30.4/14.6 Hz. However, for CV, which is produced at a very low f0 by both speakers, the MEA was higher for the female speaker (59.0 Hz) than for the male speaker (18.4 Hz). From this we conclude that a low f0 in the mobile condition leads to difficulties with strF0.

7 Discussion

The current explorative study assessed the performance of six f0 estimators with respect to *recording condition*, *voice quality*, and *speaker sex.* Results suggest that for MV in the studio recording condition, almost all tested f0 estimators showed very low error rates, with the exception of rF0. However, when estimating f0 in samples containing BV or CV, caution is needed in the choice of the f0 estimator. In both recording conditions, the f0 of aperiodic CV samples were best determined using rF0. The f0 of BV was best captured in both recording condition using sF0. Generally, the mobile phone filter degraded the performance of most of the f0 estimators tested, but not of all, e.g., sF0 for the male speaker. The shrF0 estimator produced very poor results when detecting the f0 of CV samples for the two speakers in the two recording conditions, although the speakers produced different modes of CV. We therefore do not recommend using shrF0 for f0 estimation of CV. rF0, on the other hand, should be avoided for the f0 determination of MV and BV. As the study only included two speakers, the conclusions about the impact of *speakers' sex* are extremely vague and should be treated with caution. However, strF0 appeared to cope better with high-pitched voices than with low-pitched voices in the mobile recording condition.

Harsh voice was not considered as VQ in the current study. Many f0 estimators, focused specifically on amplitude and frequency modulations which may appear in voices, rated to be harsh. Future studies should expand the VQ dimensions and include harsh and whispery voice to assess the effects of laryngeal irregularities and laryngeal constriction on the performance of f0 estimators.

Furthermore, the sample size should be considerably increased, to expand the explorative nature of the current study and to confirm results.

By using controlled speech in the form of cardinal vowel productions which were recorded in studio quality and transmitted in the same manner, the study provided an important baseline for establishing causality between the performance of the f0 estimators and the independent variables tested. In this way, factors causing noise such as diversity of speech sounds, recording device and transmission mode could be controlled. The results obtained provide an essential starting point for further exploring the performance of f0 estimators under conditions reflecting forensic casework.

As previously shown by Keelan et al. (2010), specifying the frequency range for each recording substantially improved the performance of all tested f0 estimators. Pre-tests showed that a constant large frequency range (male speaker: 30–300 Hz, female speaker: 30–500 Hz) led to substantially larger error rates, especially for the mobile condition. BV was affected the most. Here, the best f0 estimator for the male speaker, pF0cc, yielded an MEA of 11 Hz, while the same f0 estimator based on the specified frequency range deviated only 4.1 Hz from the ground truth f0.

Given the very low upper frequency limit transmitted in our samples which were mobile phone filtered (only 3.6k Hz), we speculate that the GSM-AMR narrowband codec was used in the two recording sessions to transmit the studio recordings through the actual GSM mobile network. Since more modern networks and codecs are available for mobile telephony today, the use of the GSM-AMR narrowband codec was probably related to the hardware used, i.e., the conventional landline telephone. So far, it cannot be ruled out that a bandpass-limited, strongly degraded recording (with weak harmonics and a pronounced noise floor) may cause problems due to the limited frequency range when estimating f0. However, tests revealed that the f0 can still be detected even when absent, both in the time and frequency domain. We therefore hypothesise that it is not the limited frequency range that causes problems for some f0 estimators under mobile recording conditions, but other aspects of mobile telephony, e.g. the codecs.

The present study shows the importance of choosing the f0 estimator wisely, considering the speaker's dominant voice quality and the recording condition of the voice sample to be analysed. This allows the most valid f0 path to be determined, which is used to obtain valid harmonic measurements required to assess precise spectral tilt characteristics. Thus, f0 estimation forms the base to search for correlations between perception and signal acoustics. It is hypothesised that previous studies which could not find such correlations may have suffered from invalid f0 estimates.

7 Acknowledgements

We would like to thank Joshua Penney and Yen Shue for their help with setting up VoiceSauce and implementing the REAPER f0 estimator. Further acknowledges go to Phil Harrison for mobile filtering the recordings and to Patricia Keating and Jessica Wormald for feedback on a previous version of the paper.

References

Bie, F., Wang, D., Wang, J., & Zheng, T. F. (2015). Detection and reconstruction of clipped speech for speaker recognition. *Speech Communication* 72, 218– 231. https://doi.org/10.1016/j.specom.2015.06.008

Boersma, P., & Weenink, D. (1992–2023b). Praat Manual. Amsterdam: University of Amsterdam, Phonetic Sciences Department. https://www.fon.hum.uva.nl/praat/manual/ [Accessed 17 Jun 2023].

Boersma, P. & Weenink, D. (2023). Praat: doing phonetics by computer (Version 6.3.08) [Computer software]. http://www.praat.org/

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences 17*, 97–110.

Boersma, P. (2020). Pratt Users List [Online]. Pitch detection AC vs. CC method. https://groups.io/g/Praat-Users-List/topic/pitch_detection_ac_vs_cc/78829266 [Accessed 12 Jun 2023].

Cox, R. V., Neto, S. F. D. C., Lamblin, C., & Sherif, M. H. (2009). Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]. *IEEE Communications Magazine* 47(10), 106–109. https://doi.org/10.1109/MCOM.2009.5273816

Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proceedings of the 19th International Congress of Phonetic Sciences, Australia*, 532–536.

De Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, Hearing Research* 36(2), 254–266. https://doi.org/10.1044/jshr.3602.254

De Looze, C. & Hirst, D. J. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Proceedings of 4th International Conference on Speech Prosody, Brazil*, 135–138.

Doreen, K. (2017). *Fundamental frequency distributions of bilingual speakers in forensic speaker comparison* [Master thesis, University of Canterbury].

Esling, J.H., Moisik, S.R., Benner, A., & Crevier-Buchman, L. (2019). Voice quality: the laryngeal articulator model. Cambridge University Press. http://dx.doi.org/10.1017/9781108696555

Fujimura, O., Honda, K., Kawahara, H., Konparu, Y., Morise, M., & Williams, J. C. (2009). Noh voice quality. *Logopedics Phoniatrics Vocology* 34(4), 157–170.

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language & The Law* 15, 193–218. https://doi.org/10.1558/ijsll.v22i1.17880

Hemmen, J. (2014). *Vowel distance measures: performance and behaviour with regards to voice quality, gender, and intra – /inter-speaker factors.* [Master thesis, University of York].

Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. In: Knight, R.-A. & Setter, J (Eds) *Cambridge Handbook of Phonetics 1*, 336–361. Cambridge University Press.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, USA*, 3933–3936. https://doi.org/10.1109/ICASSP.2008.4518514

Kawahara, H., Morise, M., Nisimura, R., & Irino, T. (2012). Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation. *Proceedings of the Interspeech, USA*, 386–389. https://doi.org/10.21437/Interspeech.2012-139

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America* 132(2), 1050–1060. https://doi.org/10.1121/1.4730893

Keating, P. A., Garellek, M., Kreiman, J. & Chai, Y. (2023a). Acoustic properties of subtypes of creaky voice. *Journal of the Acoustical Society of America* 153(3), 297.

Keating, P., Kuang, J., Garellek, M., & Esposito, C. M. (2023b). A cross-language acoustic space for vocalic phonation distinctions. *Language* 99(2), 351–389.

Keelan, E., Lai, C., & Zechner, K. (2010). The importance of optimal parameter setting for pitch extraction. *Proceedings of the 160th Meeting of the Acoustical Society of America* 11(1), 060004.

Klug, K., Kirchhübel, C., Foulkes, P., Braun, A., & French, P. (in press). Assessing creaky voice quality for forensic purposes. In: *Proceedings of the Aarhus International Conference on Voice Studies*. Sciendo.

Laver, J., Hiller, S., & Hanson, R. (1982). Comparative performance of pitch detection algorithms on dysphonic voices. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* 7, 192–195.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *Transactions on Audio, Speech, and Language Processing* 16, 1602–1613.

Penney, J., Cox, F., & Szakay, A. (2020). Glottalisation, coda voicing, and phrase position in Australian English. *Journal of the Acoustical Society of America* 148(5), 3232–3245. https://doi.org/10.1121/10.0002488

Shue, Y.L. et al. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the 17th International Congress of Phonetic Science, Hong Kong*, 1846–1849.

Shue, Y. L. (2010). *The voice source in speech production: Data, analysis and models.* [Doctoral dissertation, University of California]. http://www.phonetics.ucla.edu/voice-project/Publications/shue_dissertation.pdf

Sukhostat, L., & Imamverdiyev, Y. (2015). A comparative analysis of pitch detection methods under the influence of different noise conditions. *Journal of Voice* 29(4), 410–417. https://doi.org/10.1016/j.jvoice.2014.09.016

Sun, X., & Xu, Y. (2002). Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice 16*(4), 443–459. https://doi.org/10.1016/S0892-1997(02)00119-4

Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. *Proceedings of the Sixth International Conference on Spoken Language Processing*, 676–679.

Sun, X. (2002). Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, USA*, 333–336.

Szakay, A., & Torgersen, E. (2019). A re – analysis of F0 in ethnic varieties of London English using REAPER. *Proceedings of the International Congress of Phonetic Sciences, Australia*, 1675–1878.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B. & Paliwal, K. K. (Eds.) *Speech Coding and Synthesis*, 497–518. Elsevier Science B.V.

Talkin, D. (2015). *REAPER: Robust Epoch And Pitch EstimatoR*. Retrieved from https://github.com/google/REAPER

Vicenik, C., Lin, S., Keating, P., & Shue, Y. (2023). Online documentation for VoiceSauce. http://www.phonetics.ucla.edu/voicesauce/documentation/index.html.

Wayland, R., Gargash, S., & Jongman, A. (1997). Acoustic and perceptual investigation of breathy voice. *The Journal of the Acoustic Society of America* 97, 3364. https://doi.org/10.1121/1.413011

Ó Cinnéide, A. (2012). Phase distortion robust voice source analysis. [Doctoral thesis, Dublin Institute of Technology].

Analysis of breathy voice under mobile phone condition based on adequate f0 estimation

Katharina Klug[1]

[1] *Department of Language and Linguistic Science, University of York, UK*

kk667@york.ac.uk

Abstract: In a previous study, Klug et al. (2019) investigated the relationship between the perception of ‘breathiness’ and signal acoustics using spontaneous speech samples from male speakers rated as dominantly breathy and dominantly non-breathy. Under studio recording conditions, three acoustic measures were found to differ significantly between breathy and non-breathy speakers, namely H1*-H2*, H1*-A1*, and CPP.

The current study examines the same data under mobile phone recording conditions – typical of evidential recordings submitted for forensic speaker comparison – to determine whether acoustics survive the mobile phone transmission (Byrne & Foulkes, 2004; Guillemin & Watson, 2008)).

VoiceSauce (Shue et al., 2011) was used to generate the speech acoustics. A pre-test showed that the software’s default f0 estimator, STRAIGTH (Kawahara et al., 1999), may not be the most suitable f0 estimator for mobile phone transmitted recordings. The main study therefore relies on the results of Klug & Niermann (2024), who found that the Snack f0 estimator (Talkin, 1995) performs best for breathy voice quality under mobile recording condition.

As mobile phone transmission attenuates frequencies up to about 400Hz, all spectral tilt measurements in relation to H1 and H2 are considered to be invalid. Four acoustic parameters showed systematic differences between breathy and non-breathy speakers under the mobile recording condition, namely CPP, HNR05, and the spectral tilt parameters A1-A3, and H4*-A2*.

The results indicate that an informed choice of the f0 estimator to be used is necessary to generate meaningful acoustics.

1 Introduction

In a forensic voice comparison, voice quality (henceforth VQ) settings are usually assessed auditorily, due to the lack of knowledge about which acoustic features are still reliable indicators of VQ characteristics even in degraded forensic recordings. Acoustic correlates are usually obtained from high-quality studio recordings that are based on a very clean phonetic input, such as sustained vowel phonation or lists of minimal pair (Hillenbrand et al., 1994; Wayland & Jongman, 2003; Garellek & Keating, 2011). In a forensic context, however, the voice samples to be analysed are degraded. Therefore, the findings from high-quality studio recordings cannot be uncritically transferred to forensic case material.

Hughes et al. (2019) and Chan (2023) investigated the evidential value of laryngeal VQ acoustics using on semi-automatic speaker recognition systems (henceforth SASR). Both studies relied on the same methodology using large sample sizes of forensic-realistic recordings. Interestingly, they arrived at different conclusions. Hughes et al. (2019) found that the tested laryngeal VQ parameters do indeed capture speaker-specific information, as speaker-discrimination performance improved – even under the mobile recording condition. Chan (2023), in contrast, reported that the analysed laryngeal VQ acoustics carried only limited speaker-discriminatory power.

The current study highlights two possible factors that could be responsible for this discrepancy in. results. Both studies included spectral tilt parameters of low frequency ranges, which may be attenuated in mobile phone transmitted recordings. Furthermore, they relied on the STRAIGHT f0 estimator which was found by Klug & Niermann (2024) to be not suitable for mobile transmitted speech. Both factors can have a major influence on the results.

In a previous study (Klug et al., 2019) it was investigated whether perceived breathy voice quality can predict signal acoustics from spontaneous speech samples – focusing on recordings in studio quality (henceforth studio condition). The follow-up experiment presented here addresses the same research question under the mobile phone recording condition (henceforth *mobile condition*): Do signal acoustics differ between speakers rated to be dominantly breathy from dominantly non-breathy speakers after recordings were transmitted through a mobile phone network? Thus, can we confirm our perception of a breathy VQ with the help of signal acoustics – even under forensically realistic recording conditions? Particular attention is paid to the potentially limiting factors, the choice of the f0 estimator and the low-frequency spectral tilt parameters.

2 Acoustics of breathy voice quality

2.1 Studio condition

Breathy VQ is characterised by the absence of a complete glottal closure, and thus by a high open quotient. In studio condition this is reflected acoustically by a prominent first harmonic (Bickley, 1982; Ladefoged, 1983), high frequency aspiration noise (Klatt & Klatt, 1990) reflected in lower values of spectral noise parameters, e.g., Harmonic-to-noise ratio, and a lower degree of signal periodicity measured by Cepstral Peak Prominence (Hillenbrand et al., 1994) and spectral tilt (Hillenbrand & Houde, 1996).

2.2 Mobile condition

However, if a speech signal is transmitted through a mobile phone network, the speech signal will be heavily affected so that the features found under studio recording conditions may no longer be valid. To allow for wireless transmission, the information transfer rate is drastically reduced. To do so the networks apply speech codecs which compress the speech signal. Guillemin & Watson (2008) describe the process to be “lossy” and “highly synthetic”. A well-established codec is the GSM-AMR codec, where GSM stands for *Global System Mobile Communication network* and AMR for *Adaptive Multi-Rate speech codec*. It operates in eight different modes which differ in their source coding bit rates. All modes show extensive bandwidth limitations. They share the same low frequency limit, which ranges around 100Hz and differ in the upper frequency limit which varies between 2,8 and 3,6kHz (Guillemin & Watson, 2008, p. 201). Although these threshold frequencies may be transmitted in the signal, the relative amplitude gradually increases, respectively decreases. While the upper frequency end seems to be characterised as a sharp roll-off, the lower frequency end appears to roll-off gradually. Within this roll-off frequency range, the harmonics' relative amplitude is attenuated and thus invalid. Section 3 discusses the mobile bandpass filter in more depth.

Further limitations are expected which affect the signal acoustics. Byrne & Foulkes (2004) reported that the accuracy of F1 estimation is severely compromised. Guillemin & Watson (2008) found "white islands", i.e., missing energy, in the spectrogram in certain frequency ranges as well as differences in spectral detail. These effects may further influence the measurement of the spectral tilt and spectral noise parameters. In addition, the absence of high frequencies that are not transmitted through the mobile phone filter could hinder the detection of high-frequency noise.

3 Corpus Data

The same corpus as in Klug et al. (2019) was used. The corpus is composed of recordings of existing corpora containing high-quality spontaneous speech samples from British male speakers, digitised at a sampling frequency of 44.1kHz and a resolution of 16-bit (Gold et al., 2018; Haddican & Foulkes, 2017; Kirchhübel, 2013; Llamas et al., 2016–19; Nolan et al., 2009; Wormald, 2016). The original recordings are referred to as studio condition throughout the paper.

The mobile condition was generated by transmitting the recordings through a real GSM mobile network in August 2018. To do so the original recordings were concatenated into one sound file using Praat (Boersma & Weenink, 2023) to transmit the entire corpus within a single mobile call. The concatenated studio recording was replayed from Marantz PMD670 connected via a cable to the input of a smartphone audio interface (Tascam iXZ) which was connected to a mobile phone (Samsung Galaxy S8). A call was in progress between the mobile phone and a landline telephone (Audioline AUB 1) which was connected to a telephone balance unit (Prospect TC – 30). The output of the balance unit was routed via a mixer (Behringer Ultralink Pro) to increase the signal level before going to the input of an audio interface (M-Audio Delta 66) in a desktop computer. The mobile phone filtered speech was recorded on the computer using Sound Forge (version 9.0e). The detectable transmitted frequency range is approximately between the cut-off frequencies 100Hz and 3.6kHz. Spectral components in the vicinity of the cut-off frequencies are partly considerably attenuated, leaving a presumably *meaningful frequency range* between 400Hz and 3.2kHz. The evaluation of what is considered a meaningful frequency range is based on visual inspection of the long-term average spectra (henceforth LTAS) of all eight speakers by taking a rather conservative view. The estimated frequency values serve as a rough indication. No systematic, objective analysis was carried out to define the meaningful frequency range. Figure 1 shows exemplary the LTAS of speaker *5* phonating the word /no/ under the studio condition (in black) and under the mobile condition (in red). The lower and upper cut-off frequencies, around 100Hz and 3.6kHz, are separated by grey dashed lines, the meaningful frequency range between 400Hz and 3.2kHz is marked by black dashed lines.

The limited frequency range transmitted in the recordings suggests that the GSM-

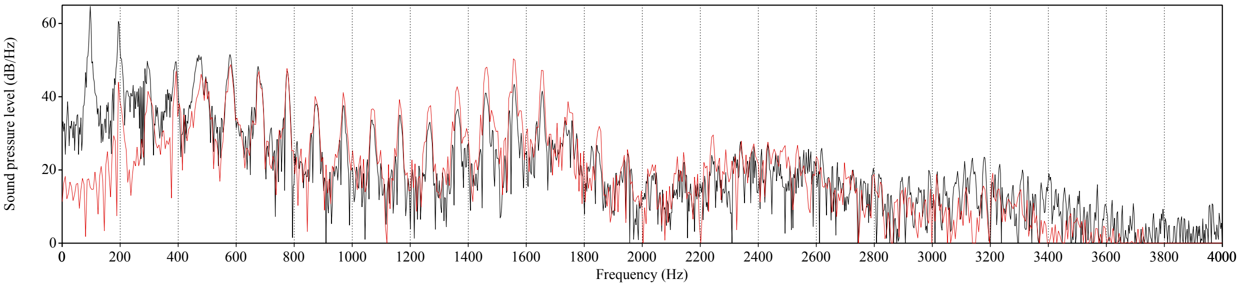


***Figure 1** Comparison between the LTAS of speaker 5 phonating /no/ under the studio condition (in black) vs. the mobile condition (in red).*

AMR narrowband codec was used. This was probably due to the conventional landline telephone which was used. Wideband audio algorithms transmit frequencies between 50 and 7000Hz, while fullband covers the entire frequency range that humans perceive, i.e., 20–20.000Hz (Cox et al., 2009; Gibson, 20016).

4 Methodology

4.1 Auditory-perceptual investigation

The methodology of Klug et al. (2019) was replicated using the mobile phone recordings. Using Qualtrics, a survey was conducted to obtain perceptual judgements on 22 speakers in the mobile condition– eight months after the survey had been conducted with the recordings in studio quality. The same four forensic speech scientists as in Klug et al. (2019) rated the speakers by answering the following question: “Would you mark breathiness as a dominant feature of this speaker’s voice?”. Three answer choices were given:

- Yes.
- No, it is present but not dominant.
- No, it is absent.

In addition, a comment box was provided for each recording. This was frequently used by participants to justify ratings, specify intermittent presence, or indicate the scalar degree of breathiness or other dominant phonatory settings.

The same eight speakers as in the studio condition were rated at the two ends of the breathiness scale: five dominantly breathy speakers and three dominantly non-breathy speakers. Thus, these speakers qualified for the acoustic analysis. Perhaps due in part to confirmation bias (Kahneman, 2011, p. 81), only three speakers were rated to have a dominantly non-breathy VQ independent of recording condition. Confirmation bias describes the tendency of the system, here perception, to accept the suggestion made uncritically. By phrasing the research question “Would you mark breathiness as a dominant feature of this speaker’s voice?”, the analysts’ prior assumptions were fixed on breathiness. This might have prevented the unbiased perception of non-breathiness. The outcome of the perceptual evaluation was thus an unbalanced number of speakers at the two ends of the breathiness scale, i.e., five breathy vs. three non-breathy speakers.

Using Gwet’s AC2 by running the {irrCAC} package (Gwet, 2019) in R (R Core Team, 2021), the agreement coefficient for each

Table 1 Between-rater agreement based on assessment of the speakers to be analysed, in mobile condition. (Gwet's AC2 for rater-pairs; pa – percent agreement, pe – percent chance agreement, Coeff. val – agreement coefficient estimate-AC2, Agreement conclusion is based on the scale of Landis & Koch, 1977)

Rater-pair	pa	pe	Coeff. val	p − value	Agreement
1–2	0.88	0.48	0.76	0.002	substantial
1–3	0.79	0.45	0.62	0.080	substantial
1–4	0.88	0.43	0.78	0.012	substantial
2–3	0.88	0.55	0.72	0.007	substantial
2–4	0.92	0.51	0.83	0.0004	almost perfect
3–4	0.88	0.51	0.75	0.008	substantial

Table 2 Within-rater agreement comparing each rater's assessment of the eight speakers to be analysed in studio condition with the assessment in mobile condition.

Rater	pa	pe	Coeff. val	p − value	Agreement
1	0.83	0.45	0.70	0.040	substantial
2	0.92	0.51	0.83	0.0004	almost perfect
3	1	0.47	1	0	perfect
4	0.88	0.49	0.75	0.019	substantial

rater pair was determined based on the ratings of the eight speakers to be analysed (Table 1). The level of between-rater agreement ranged from "substantial" to "almost perfect" according to the scale of Landis & Koch (1977). Table 2 lists the within-rater agreement coefficients when comparing the ratings under studio condition with the ratings under mobile condition. All raters showed a very high within-rater consistency when rating breathiness under mismatched recording condition.

4.2 Acoustic investigation

4.2.1 Measurement procedure

Table 3 Speaker-specific f0 range (based on the two-pass detection process from de Looze & Hirst, 2008).

Speaker	Frequency range (Hz)
1	70–160
2	70–160
3	90–220
4	100–240
5	60–150
6	80–190
7	70–180
8	80–180

VoiceSauce (Shue et al., 2011) was used to perform automatic measurements of the segmented sonorants (i.e., vowels, glides [j, w], liquids [l, r] and nasals [m, n, ŋ]) for the eight speakers to be analysed. In addition, the vowel-only subset was examined to exclude the potentially negative impact of non-vocalic segments. The f0 measurements were taken with a frame shift of 1ms, most other estimates were taken pitch synchronously (harmonics over a three-cycle window and estimates for energy, CPP and HNR over a five-cycle window). To reduce the number of f0 errors, the frequency range was specified per speaker according to the two-pass detection process described by de Looze & Hirst (2008) and Hirst & de Looze (2021). The first pass is used as a rough f0 estimation using the default pitch range of Praat for male speakers, i.e., 75–300Hz. From the first-pass-results the first and third quartiles (25 and 75%) were multiplied by a coefficient (0.75 and 1.5 respectively), to determine the speaker-specific frequency range. The second pass, i.e., the actual voice analysis, was subsequently carried out based on these speaker-specific frequency ranges which are summarised in Table 3.

In a pre-test, VoiceSauce's standard f0 estimator STRAIGHT (strF0, Kawahara et al., 2008, 2012) was used to estimate the f0. However, as VoiceSauce bases its harmonic estimation on the f0 estimates, valid f0 estimation is crucial. According to the results gained from Klug & Niermann (2024) the Snack f0 estimator (sF0) of the Snack library (Talkin, 1995) is hypothesised to perform most accurate under the mobile condition and breathy VQ. Thus, the main test relied on the f0 estimator Snack (Sjölander, 2004). Both studies used the default formant estimator Snack.

4.2.2 Data analysis

For analysis all data points per speaker were used, varying between 41.764 and 89.692. Statistical analysis was conducted using R (R Core Team, 2021).

To investigate the relationship between the perceptual VQ ratings (breathy voice vs. non-breathy voice) and the acoustic parameters, linear mixed effect analysis was performed using the *lme4* package (Bates et al., 2015). Each of the *acoustic parameters* functioned as *dependent variable.* For each potential acoustic parameter, a separate model was generated. *VQ* was entered as *fixed effect* into each model. *Speaker* was added as *random effect*, accounting for differences between speakers.

Following the approach of Winter (2013), a likelihood ratio test was subsequently performed for each acoustic parameter tested in order to determine p $-$ values. For this purpose, two models were generated for each acoustic parameter: (1) a full model including the *fixed effect VQ*, and (2) a null model omitting the fixed effect. Likelihood ratio tests were performed using the `anova()` function (Chambers & Hastie, 1992) to test whether the likelihood of the two models differed significantly. If the models proved to be significantly different, the result could be attributed to the fixed effect, i.e. VQ. The alpha level was set at $p < 0.05$.

4.2.3 Acoustic parameters

Table 4 shows all acoustic parameters which were tested for correlations with the percept breathiness. As the validity of the formant measurements has not been tested, potentially incorrect formant measures would negatively affect many parameters. To minimise the negative impact, all harmonic amplitude parameters and thus all spectral tilt parameters were modelled twice: (1) on the basis of the uncorrected harmonic amplitudes, and (2) on the basis of the formant-corrected harmonic amplitudes. If formants are measured correctly, the formant-corrected version, marked with an asterisk *, minimises the boosting effect of nearby formants (Iseli et al., 2007).

As low frequency harmonics are highly attenuated and are thus not meaningfully assessing the speaker’s spectral tilt in low frequency ranges, only the frequency range from H4 onwards will be assessed for differences in spectral tilt characteristics.

4.2.4 Hypotheses

It is hypothesised that an interaction between breathy voice and signal acoustics can also be found even under the mobile condition. This hypothesis is based on the results from the auditory-perceptual investigation (section 4.1). The forensic speech scientists reliably distinguished between dominantly breathy and non-breathy speakers even under the mobile recording condition, evident in the high level of within-rater-agreement when comparing the performance under the two recording conditions, studio vs. mobile (see Table 2).

Furthermore, the audio compression employed in mobile phone transmission is reported to impact CPP and HNR (Cavalcanti et al., 2023). Thus, the spectral noise param-

***Table 4** Acoustic parameters tested. Harmonic amplitude and spectral tilt parameters were used in uncorrected and formant-corrected (*) version.*

Acoustic parameter	Explanation
Harmonic amplitude	
H4	Amplitude of the 4th harmonic
H2K	Amplitude of the harmonic closest to 2kHz
A1/A2/A3	Amplitude of the harmonic closest to the 1st/ 2nd/ 3rd formant
Spectral tilt	
H4-A1/A2/A3/H2K	Difference in amplitude between the minuend and the subtrahend
H2K-A1/A2/A3	
A1-A3/A1-A2/A2-A3	
Spectral noise	
SHR	Subharmonic-to-Harmonic Ratio (Sun, 2002, 2000)
HNR05/15/25/35	Harmonics-to-Noise Ratio (de Krom, 1993) within the frequency ranges 0–500Hz/ 0–1500Hz/ 0–2500Hz/ 0–3500Hz
CPP	Cepstral Peak Prominence across the entire frequency range (Hillenbrand et al., 1994)
Energy	
Energy	Root Mean Square energy

eters are hypothesised to show a less clear correlation with the percept breathiness under the mobile recording condition compared to correlation found under the studio recording condition. would be less clear under the mobile condition.

Formant estimation in nasals is error-prone due to the presence of “nasal formants” (Styler, 2017). Shue et al. (2011) point out that F1 estimation is less accurate in breathy vowels. Also, CPP is sensitive to articulatory variation, e.g. nasalance (Madill et al., 2019). Therefore, all measures that refer to formant frequencies, e.g., A1, as wel as formant-corrected measures, and CPP could be negatively affected when the analysis is based on all sonorants. It is therefore hypothesised that the vowel-only subset shows a clearer correlation.

5 F0 estimation

The lack of correlation between the percept of breathy voice and speech acoustics in recordings which have been transmitted through a mobile phone network, led to the study of Klug & Niermann (2024). They tested the performance of six f0 estimators under two recording conditions (studio vs. mobile) and under three voice qualities (modal voice, breathy voice, and creaky voice) based on sustained cardinal vowel production of one male and one female speaker. The following f0 estimators were examined: Praat Autocorrelation (Boersma, 1993), Praat Cross-Correlation (Boersma & Weenink, 2023), REAPER (Talkin, 2015), Snack (Talkin, 1995), Subharmonic-to-harmonic ratio (Sun, 2000, 2002), and STRAIGHT (Kawahara et al., 2008, 2012). To assess the performance of the f0 estimators, the automatic measurements were compared with the manually obtained f0 ground truth data. For the breathy voiced cardinal vowels of the male speaker under the studio condition the Snack f0 estimator was found to be most accurate, deviating from the manually assessed f0 ground truth on average by only 1.5Hz (see Table 5). Under the mobile condition also the Snack f0 estimator revealed the most accurate f0 estimation, deviating by 1.4Hz from the f0

***Table 5** Performance of f0 estimators based on breathy cardinal vowels of a male speaker (data taken from Klug & Niermann 2024).*

f0 estimator	Mean error absolute (Hz)	
	Studio condition	Mobile condition
Snack	1.5	1.4
Praat ac	1.8	2.8
STRAIGHT	2.0	14.6
Praat cc	2.0	4.1
SHR	2.7	4.8
REAPER	25.4	16.6

ground truth. The default f0 estimator within VoiceSauce, STRAIGHT, however, produced a considerably higher mean error absolute, i.e., 14.6Hz. Therefore, f0 estimation in the main study is based on Snack. Snack is very critical in accepting frames as voiced, STRAIGHT in contrast, does not perform voicing decision at all and even output data for voiceless and aperiodic frames. Thus, for the main study considerably fewer datapoints were included in the analysis.

It should be highlighted that accurate harmonic estimation not just depends on accurate f0 estimation, but also on accurate formant estimation. This may impact parameters such as A1, A2, and A3 which search for the harmonic closest to the respective formant. Additionally, the formant correction algorithm from Iseli et al. (2007) is dependent on accurate formant estimation. The current study does not make any efforts to assess the accuracy of formant estimation. Thus, results are reported for both, the formant corrected and uncorrected harmonic amplitudes.

6 Results

6.1 Pre-test

The pre-test is based on VoiceSauce's default f0 estimator STRAIGHT (henceforth strF0). Linear models were constructed for each of the acoustic parameters as dependent variable. Only the model for CPP turned out to be significant ($\chi^2(1) = 6.65$, $p = 0.0099$**), lowered by 1.7 Hz $\pm\, 0.5$ (standard errors). Figure 2 displays the difference between breathy and non-breathy speakers using a boxplot. Surprisingly, none of the harmonic amplitudes and thus none of the spectral tilt parameters showed a significant difference between the breathy and the non-breathy speakers.

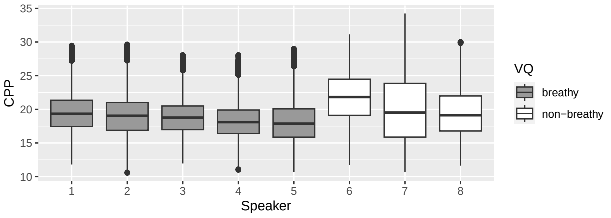


__Figure 2__ Boxplot comparing CPP as a function of breathy perception. Data relies on all sonorants.

Basing the model construction on the vowel-only dataset, a slightly higher significance level was found ($\chi^2(1) = 7.03$, $p = 0.008$**), lowered by 1.9 Hz $\pm\, 0.6$ (standard errors).

6.2 Main study

In contrast to the pre-test, the main study is based on the Snack f0 estimator. Results are visualised using boxplots in Figure 3. The mixed model estimate, the standard error, and the t $-$ value for each dependent variable which significantly affected the percept breathiness can be obtained from Table 6. Table 7 comprises the results from the likelihood ratio test, i.e. the Chi-Square value, degrees of freedom, and the p$-$value^2, using the `anova()` function.

Using the Snack f0 estimator, correlations between the percept breathiness and two spectral tilt parameters, as well as two spectral noise parameters are found under the mobile recording condition. The clearest effect is found for CPP, which significantly predicts the percept breathiness ($\chi^2(1) = 12.27$, $p = 0.00046$***) decreasing by about 3dB±0.6 (standard errors) compared to speakers rated as dominantly non-breathy. HNR also turned out to be significant for two frequency ranges: HNR05 ($\chi^2(1) = 7.35$, $p = 0.0067$**) and HNR15 ($\chi^2(1) = 3.99$, $p = 0.046$*). As

Table 6 *Predicted effects based on linear effect analysis for spectral tilt and spectral noise parameters for the percept breathiness under mobile condition based on all sonorants.*

	Model	Estimate	Std. Error	t value
Sonorants	**CPP**			
	breathy	19.62	0.34	
	non-breathy	2.99	0.56	5.39
	HNR05			
	breathy	30.06	1.16	
	non-breathy	6.60	1.90	3.47
	HNR15			
	breathy	29.97	1.18	
	non-breathy	4.38	1.93	2.27
	A1-A3			
	breathy	20.57	1.09	
	non-breathy	-4.75	1.77	-2.38
	H4*-A2*			
	breathy	7.15	0.97	
	non-breathy	-4.20	1.58	-2.66
	A1*-A2*			
	breathy	-0.96	1.27	
	non-breathy	-4.10	2.07	-1.98
	H2K-A1			
	breathy	-27.34	1.26	
	non-breathy	3.88	2.06	1.89
Vowels-only	**CPP**			
	breathy	20.30	0.34	
	non-breathy	3.03	0.55	5.48
	HNR05			
	breathy	30.03	1.21	
	non-breathy	6.80	1.97	3.44
	HNR15			
	breathy	28.22	1.20	
	non-breathy	4.99	1.96	2.55
	A1-A3			
	breathy	19.62	1.08	
	non-breathy	-3.92	1.76	-2.23
	H4*-A2*			
	breathy	5.79	1.20	
	non-breathy	-4.29	1.96	-2.19
	A1*-A2*			
	breathy	-3.43	1.24	
	non-breathy	-2.92	2.02	-1.44
	H2K-A1			
	breathy	-26.22	1.31	
	non-breathy	3.03	2.15	1.41

Table 7 *Results from the likelihood ratio using the anova() function for spectral tilt and spectral noise parameters. The full model (xxx.md), including the fixed effect VQ, is compared with the null model (xxx.nl), without the fixed effect.*

	Model	Chisq	Df	Pr(>Chisq)	
Sonorants	CPP.nl CPP.md	12.27	1	0.00046	***
	HNR05.nl HNR05.md	7.35	1	0.0067	**
	HNR15.nl HNR15.md	3.99	1	0.046	*
	A1-A3.nl A1-A3.md	5.13	1	0.024	*
	H4*-A2*.nl H4*-A2*.md	5.07	1	0.024	*
	A1*-A2*.nl A1*-A2*.md	3.19	1	0.074	
	H2K-A1.nl H2K-A1.md	2.95	1	0.086	
Vowels-only	CPP.nl CPP.md	12.46	1	0.00042	***
	HNR05.nl HNR05.md	7.27	1	0.007	**
	HNR15.nl HNR15.md	4.75	1	0.029	*
	A1-A3.nl A1-A3.md	3.87	1	0.049	*
	H4*-A2*.nl H4*-A2*.md	3.77	1	0.052	
	A1*-A2*.nl A1*-A2*.md	1.85	1	0.17	
	H2K-A1.nl H2K-A1.md	1.78	1	0.18	

the low frequency harmonics, H1 and H2, are attenuated in the mobile condition, the most common spectral tilt measures, H1-H2 and H1-A1, will not produce any meaningful output. Instead, higher frequency spectral tilt measures were examined. The following occurred to be significant: A1-A3 ($\chi^2(1) = 5.13$, $p = 0.024$*), and H4*-A2* ($\chi^2(1) = 5.07$, $p = 0.024$*). Two further spectral tilt measures showed a tendency towards significance, i.e., A1*-A2* ($\chi^2(1) = 3.19$, $p = 0.074$), and H2K-A1 ($\chi^2(1) = 2.95$, $p = 0.086$).

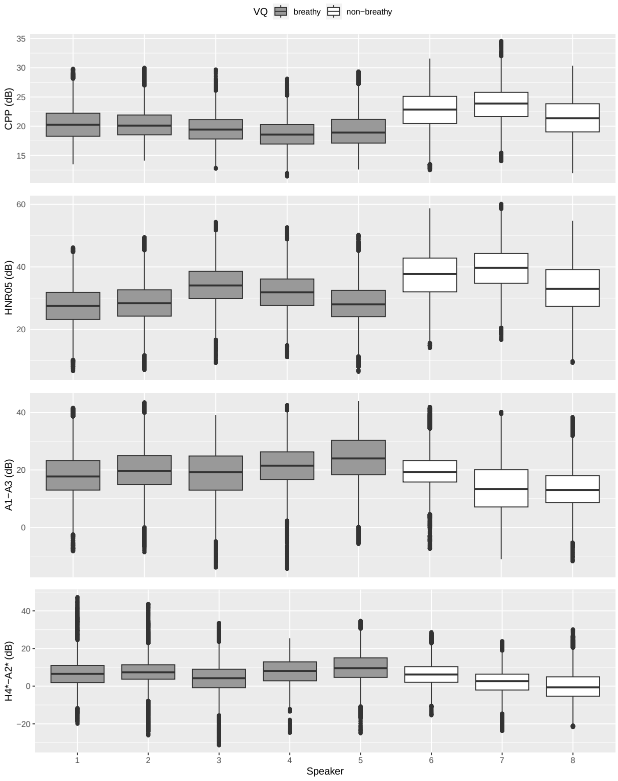


__Figure 3__ Boxplots for acoustic parameters revealing the clearest differences between dominantly breathy and non-breathy speakers in mobile filtered recordings, i.e. CPP, HNR05, A1-A3, and H4-A2**

Surprisingly, when the models are constructed based on the vowel-only subset, higher p$-$values are found for the spectral tilt parameters: A1-A3 ($\chi^2(1) = 3.87$, $p = 0.049$*), H4*-A2* ($\chi^2(1) = 3.77$, $p = 0.052$). The parameters which showed a tendency towards significance now clearly lacked significance, i.e., A1*-A2* ($\chi^2(1) = 1.85$, $p = 0.17$), and H2K-A1 ($\chi^2(1) = 1.78$, $p = 0.18$). In contrast, the spectral noise parameters, CPP, HNR05, and HNR15, revealed similarly low or slightly lower p $-$ values when analysis was based on vowels only: CPP ($\chi^2(1) = 12.46$, $p = 0.00042$***), HNR05 ($\chi^2(1) = 7.27$, $p = 0.007$**), and HNR15 ($\chi^2(1) = 4.75$, $p = 0.029$*).

Figure 3 illustrates the key dependent variables predicted by breathy VQ based on all sonorants, namely CPP, HNR05, A1-A3, and H4*-A2*.

7 Discussion

Klug et al. (2019) found that the percept breathiness can predict the already established acoustic parameters H1*-H2*, H1*-A1*, and CPP also in spontaneous speech samples relying on not just vowels but all sonorants, when the data is available in studio quality condition. However, when the data has been transmitted through a mobile phone filter, another set of acoustic parameters need to be relied on as low frequency components are highly damped in the mobile condition. The current study revealed that the spectral noise parameters CPP and HNR05 are still the best acoustic indicators for breathy VQ. Additionally, spectral tilt parameters in higher frequency ranges can be used as a further indicator of breathiness. In the current study A1-A3 and H4*-A2* distinguished between speakers rated to be dominantly breathy and dominantly non-breathy.

While the spectral noise parameter CPP is independent of an accurate f0 estimation, HNR as well as all spectral tilt parameters strongly depend on accurate f0 estimation. The pre-test in the current study proved the magnitude of the influence of inaccurate f0 estimation. Using strF0, no relationship was found between the percept breathiness and HNR parameters and spectral tilt parameters under the mobile condition. Klug & Niermann (2024), however, found the Snack f0 estimator to perform best when the analysed speech (1) is characterised by breathiness and (2) was transmitted through a mobile network. Thus, when f0 estimation was based on the Snack f0 estimator instead, dominantly breathy speakers were characterised by significantly lower HNR05 and HNR15, as well as significantly higher A1-A3 and H4*-A2*. Furthermore, A1*-A2* and H2K-A1 showed a trend towards significance when analysis was based on all sonorants. Surprisingly, when the analysis was limited to vowel-only segments, no substantial advantage was found.

It would be interesting to see if applying the current findings to Chan (2023) would lead to another outcome, namely that the laryngeal VQ parameters *have* a speaker-discriminatory value. As the study of Hughes et al. (2019) already found the evidential strength of laryngeal VQ parameters, an even clearer effect is hypothesised when relying on a more suitable F0 estimator and including spectral tilt parameters in higher frequency ranges, as opposed to parameters involving H1 and H2.

The significance level for HNR05 was substantially lower compared to HNR15. This may indicate that the breathy voice speakers are mainly characterised by weak low frequency harmonics. Aspiration noise, on the other hand, is expected in higher frequency ranges (Klatt & Klatt, 1990) and should therefore be reflected in HNR15, HNR25 and HNR35.

Although CPP has been shown to be robust to mobile phone transmission, CPP alone should not be used to assess the relationship with a speaker's VQ. As a measure of cycle periodicity, it does not appear to distinguish between different causes of aperiodicity (i.e. laryngeal frication or laryngeal irregularity). Therefore, it may not be particularly useful for assessing the relationship with specific aspects of VQ (Fraile & Godino-Llorente, 2014), as it might yield similar results for a speaker characterised by e.g. aperiodic creak as for a speaker characterised by e.g. breathy VQ. Potentially the sub-band cepstral approach by Clermont et al. (2016) and Kinoshita et al. (2022) would allow for a more specific comparison in the cepstral domain by comparing cepstral distances. This approach may enable the differentiation between specific deviations from modal VQ. So far, CPP should not be used as a single acoustic correlate of breathiness.

A look at the comments provided by the four forensic speech scientists who took part in the survey to generate perceptual judgements, shed light on how the analysed voices were perceived. In general, participants provided comments to specify the speaker's VQ. Speakers 1 and 2 received very few comments, which could indicate that these are clear examples of dominantly breathy speakers. For speaker 1 only one participant confirmed the dominantly breathy VQ. Speaker 2 received one comment that also creaky voice is present, although breathy voice dominates. The further three speakers, which were rated to be dominantly breathy, received much more divers comments. Speaker 3 was described as a "different kind of breathiness [...], more persistent/consistent characteristic", and as "tense larynx which overshadows for me". Furthermore, he was described as "whispery and breathy". The speakers 4 and 5 seem to stand out. The comments for speaker 4 contain "unusual phonation", "aging effects", i.e. "tremor", "glottal leakage", "marked phonatory irregularities". "Whispery voice" is stated by three out of four participants. Speaker 5 received the following comments "seems like lowered voice – might expect increased breathiness", "I think his speaking style makes it sound more breathy [...] – impression of sotto voce voice". Here, the speaker seems to produce breathy VQ as a stylistic element, by lowering his voice. It can be concluded that the group of speakers rated to be dominantly breathy seems to contain very different modes of breathy VQ. Two speakers were described to produce further non-modal VQ settings, such as creaky voice or tense larynx voice in addition to dominantly breathy VQ. Three speakers were characterised by intermittent whispery VQ. Also, breathy VQ seem to be produced very differently, e.g., as a stylistic element as in speaker 5 or due to ageing effect such as speaker 4. The group of dominantly non-breathy speakers, in contrast, received very consistent comments. Breathiness is described to be slightly and intermittently present, rather than a dominant VQ feature. All speakers are characterised by a dominantly creaky VQ – although special attention was paid to exclude marked and extreme examples of creaky voices when creating the corpus.

8 Future work and conclusion

There is considerably need for future work to gain a better understanding of the relationship between perception and acoustics, not only for breathy VQ, but for laryngeal VQs in general. Assessing the discriminatory potential of VQ acoustics employing the semi-automatic speaker recognition method seems to be a convenient approach, especially in the context of forensic application, as the discipline is currently undergoing a paradigm shift (see Morrison, 2022).

As mentioned in Section 5, future studies need to investigate which formant estimator is best suited for which VQ and recording condition. This will allow for an informed decision on which formant estimator to use. A valid assumption about the properties of the spectral tilt parameters highly depends on valid formant estimation, as harmonic amplitudes near formants (A_n) and amplitude corrections are only meaningful if formant analysis is meaningful. The study shows that the impact of technical aspects can be substantial, especially on spectral tilt measurements.

The current study revealed that using sonorants rather than vowel-only segments does not deteriorate the predictability of acoustic parameters. This is particularly beneficial for the analysis of forensic audio recordings, which may be short and therefore lack sufficient vocalic segments. Special care should be taken to use an f0 estimator that is suitable for the speaker's dominant VQ and the recording condition at hand. When the data has been transmitted through a mobile phone network it should be avoided to analyse low frequency harmonics as well as spectral tilt parameters which involve low frequency harmonics. So far, it is not recommended to rely on only one single acoustic parameter, especially not if it is CPP.

The comments of the forensic speech scientists who assessed the speakers perceptually indicate that the speakers analysed form a diverse group of dominantly breathy speakers. Thus, the findings are particularly promising that acoustics can be used to distinguish dominantly breathy speakers from dominantly non-breathy speakers – even in degraded mobile recording condition using sonorants from spontaneous speech samples.

Acknowledgements

Many thanks go to Christin Kirchhübel, Jessica Wormald, Richard Rhodes, and Paul Foulkes for providing perceptual judgements. Thank you, Phil Harrison, for mobile phone filtering the studio recordings. Furthermore, I would like to express my gratitude to Patricia Keating and Jessica Wormald for their valuable comments on a previous version of the paper.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48. https://doi.org/10.48550/arXiv.1406.5823

Bickley, C. (1982). Acoustic analysis and perception of breathy vowels. *MIT Speech Communication Working Papers* 1, 71–81.

Boersma, P., & Weenink, D. (2023). Praat Manual. Amsterdam: University of Amsterdam, Phonetic Sciences Department. https://www.fon.hum.uva.nl/praat/manual/ [Accessed 17 Jun 2023].

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17, 97–110.

Byrne, C., & Foulkes, P. (2004). The 'Mobile Phone Effect' on vowel formants. *International Journal of Speech, Language & The Law* 11(1), 83–102. https://doi.org/10.1558/ijsll.v11i1.83

Cavalcanti, Englert, M., Oliveira, M., & Constantini, A. C. (2023). Microphone and Audio Compression Effects on Acoustic Voice Analysis: A Pilot Study. *Journal of Voice* 37(2), 162–172. https://doi.org/10.1016/j.jvoice.2020.12.005

Chambers, J. M., & Hastie, T. J. (1992). Statistical Models in S, Wadsworth & Brooks/Cole.

Chan, R. K. (2023). Evidential value of voice quality acoustics in forensic voice comparison. *Forensic Science International*, 348, 111725.

Clermont, F., Kinoshita, Y., & Osanai, T. (2016). Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation. *Proceedings of the 16th Australasian International Conference on Speech Science & Technology, Australia*.

Cox, R. V., Neto, S. F. D. C., Lamblin, C., & Sherif, M. H. (2009). Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]. *IEEE Communications Magazine* 47(10), 106–109. https://doi.org/10.1109/MCOM.2009.5273816

de Krom, G. (1993). A cepstrum-based technique for determining a harmonic-to-noise ratio in speech signals. *J. Sp Hear. Res.* 36, 254–266.

de Looze, C. & Hirst, D. J. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *In Proceedings of 4th International Conference on Speech Prosody, Brazil*, 135–138.

Fraile, R., & Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control* 14, 42–54. https://doi.org/10.1016/j.bspc.2014.07.001

Garellek, M., Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *JIPA* 41(2) 185–205.

Gibson, J. D. (2016). Speech Compression. *Information (Basel)* 7(2), 32. https://doi.org/10.3390/info7020032

Gold, E., Ross, S., & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proceedings of Interspeech. India*, 2748–2752. http://dx.doi.org/10.21437/Interspeech.2018-65

Gwet, K. L. (2019). irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package (Version 1.0). https://CRAN.R-project.org/package=irrCAC

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language & The Law* 15, 193–218. https://doi.org/10.1558/ijsll.v22i1.17880

Haddican, W., & Foulkes, P. (2017). A comparative study of language change in Northern Englishes. [Data Collection]. Colchester, Essex: ESRC. http://reshare.ukdataservice.ac.uk/851013/

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research* 39(2), 311–321.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research* 37(4), 769–778. https://doi.org/10.1044/jshr.3902.311

Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. In: Knight, R.-A. & Setter, J (Eds.) *Cambridge Handbook of Phonetics 1*, 336–361. Cambridge University Press.

Hughes, V., Cardoso, A., Foulkes, P., French, J.P., Harrison, P., & Gully, A. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences, Australia*.

Iseli, M., Shue, Y. L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America* 121(4), 2283–2295. https://doi.org/10.1121/1.2697522

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kawahara, H., Morise, M., Nisimura, R., & Irino, T. (2012). Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation. *Proceedings of the Interspeech, USA*, 386–389. https://doi.org/10.21437/Interspeech.2012-139

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, USA*, 3933–3936. https://doi.org/10.1109/ICASSP.2008.4518514

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A. (1999). Restructuring speech representations using a pitch- adaptive time-frequency smoothing and an instantaneous- frequency based F0 extraction. *Sp. Comm.* 27, 187–207.

Kinoshita, Osanai, T., & Clermont, F. (2022). Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment. *Journal of Phonetics* 94, 101177. https://doi.org/10.1016/j.wocn.2022.101177

Kirchhübel, C. (2013). The acoustic and temporal characteristics of deceptive speech. [Doctoral dissertation, University of York]. https://core.ac.uk/download/pdf/19496297.pdf

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America* 87(2), 820–857.

Klug, K., Kirchhübel, C., Foulkes, P., & French, P. (2019). Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception. *Proceedings of the 18th International Congress of Phonetic Sciences, Australia*, 795–799.

Klug, K., Niermann, M. (2024). Assessing the suitability of F0 estimators with respect to recording condition and voice quality. *Unpublished manuscript*.

Ladefoged, P. (1983). The linguistic use of different phonation types. In: D. M. Bless & J. H. Abbs (Eds.) *Vocal fold physiology: Contemporary research and clinical issues*, 351–260. San Diego: College-Hill Press..

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1), 159–174. https://doi.org/10.2307/2529310

Llamas, C., Watt, D., French, J. P. (2016–19) The use and utility of localised speech forms in determining identity: forensic and sociophonetic perspectives. ESRC ES/M010883/1

Madill, Nguyen, D. D., Yick-Ning Cham, K., Novakovic, D., & McCabe, P. (2019). The Impact of Nasalance on Cepstral Peak Prominence and Harmonics-to-Noise Ratio. *The Laryngoscope* 129(8), E299–E304. https://doi.org/10.1002/lary.27685

Morrison, G. S. (2022). Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy* 100270. https://doi.org/10.1016/j.fsisyn.2022.100270

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & The Law* 16(1), 31–57.

R Core Team (2021). R: A language and environment for statistical computing (Version 4.1.0) [Computer software]. https://www.R-project.org/

Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proc. 18th ICPhS Hong Kong*, 1846–1849.

Sjölander, K. (2004). Snack sound toolkit. KTH Stockholm, Sweden. http://www.speech.kth.se/snack.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America* 142(4), 2469–2482. https://doi.org/10.1121/1.5008854

Sun, X. (2002). Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, USA*, 333–336.

Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. *Proceedings of the Sixth International Conference on Spoken Language Processing*, 676–679.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B. & Paliwal, K. K. (Eds.) *Speech Coding and Synthesis*, 497–518. Elsevier Science B.V.

Talkin, D. (2015). REAPER: Robust Epoch And Pitch EstimatoR. Retrieved from https://github.com/google/REAPER

Wormald, J. (2016). Regional variation in Panjabi- English, Doctoral dissertation, University of York.

Wayland, R., Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics* 31(2), 181–201.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv*:1308.5499. http://arxiv.org/pdf/1308.5499.pdf

Conclusion

Using the examples of two laryngeal voice qualities, i.e., creaky voice and breathy voice, the current thesis addresses the main research question how voice quality (VQ) analysis can be improved for forensic application. Specifically, the thesis explores the following research questions:

Paper 1 CV assessment: Is it possible to assess the *nature of CV production* based on the proposed CV classification scheme? How *consistently do analysts* perform?

CV production: Can *speakers be distinguished* by the nature of their CV production?

Paper 2 Can the perception of a dominantly breathy voice be corroborated using signal acoustics when the analysis is based on sonorants from spontaneous speech samples *in studio recording condition*?

Paper 3 Which impact do *recording quality* and *voice quality* have on the performance of f0 estimators?

Paper 4 Can the perception of a dominantly breathy voice be corroborated using signal acoustics when the analysis is based on sonorants from spontaneous speech samples, even after the speech has been transmitted through a *mobile filter*?

Which impact does the choice of the f0 estimator have?

Paper 1 examines the multifaceted nature of creaky VQ. The research questions relate (1) to the assessment of creaky VQ, and (2) to the speaker-discriminatory potential of CV production. A perceptually orientated CV classification scheme is proposed, which distinguishes between four main CV modes, namely clean CV, harsh CV, breathy CV, and aperiodic Creak. The scheme is designed to meet the requirements of forensic caseworkers by proposing a small number of perceptually relevant CV categories which are hypothesised to be still perceptually distinguishable in poor quality recordings, typical for forensic recordings. Using spontaneous speech samples of six male English speakers, four analysts applied the proposed CV classification scheme by rating pre-labelled syllables. Across all rated syllables the analysts agreed moderately when assessing the presence of distinct glottal pulses, which was defined as the main characteristic of CV perception. When assessing the syllable's specific CV mode, the analysts showed a high level of agreement for four of the six speakers, differing in average by only 0.5 perceptual distances. For the remaining two speakers, the analysts performed less consistently, differing almost twice as much, i.e., by an average of 0.9 perceptual distances. To compare: the distance between e.g. clean CV and harsh CV or breathy CV was defined to be 1. The between-analyst variation for these two speakers seems rather high; however, during discussion sessions it became obvious that syllables which were rated differently usually stem from similar perceptions which were differently classified. This finding emphasises the importance of calibration procedures, which should be carried out either by providing more detailed training material and/or by conducting calibration sessions.

From the six speakers analysed, there is no complete overlap with respect to the speakers' preferred CV space. Although some speakers overlap to some extent in their CV space, most speakers generally produce more than one CV mode and thus differ in the combination of preferred CV modes. Large scale studies are needed to explore the typicality of CV modes. For example, the speakers of this small-scale study produced clean CV more frequently than breathy CV; however, the sample size is way too small to allow for conclusion. The study highlights that the multifaceted nature of CV may be useful to discriminate between speakers in a forensic voice comparison. A finer-grained CV classification approach enables to assess a speaker's individual CV space. Furthermore, defining perceptually relevant CV modes is needed to explore acoustic correlates of each CV dimension. Thus, the study forms the prerequisite for assessing CV acoustically.

Paper 2 and Paper 4 focus on breathy VQ. While Paper 2 focusses on sonorants from spontaneous speech samples in studio recording quality, Paper 4 analyses the same recordings after being mobile phone transmitted. Breathy voice seems to be less-dimensional compared to creaky voice. Following Gauffin (Lindblom, 2009, see) and Esling et al. (2019), breathy voice ranges on two dimensions, i.e., (1) the continuum of glottal openness to distinguish breathy voice from modal voice and phonation modes in between, and (2) the continuum of epilaryngeal constriction to distinguish breathy voice from whispery voice (Esling et al., 2019). It was therefore assumed that there are clear correlations between acoustics and perception. The contribution of Paper 2 to VQ research is to investigate whether the expected correlation still holds when the analysis is based on sonorants from spontaneous speech samples, rather than relying on sustained vowel productions, minimal pair lists, or read speech.

Expert listeners rated 22 voices with respect to absence/ presence of dominantly breathy VQ. The voices at the extremes were used for acoustic analysis. The results turned out to be very promising. Three of the acoustic parameters which were reported in the literature to indicate breathy VQ turned out to significantly distinguish the group of speakers rated to be dominantly breathy from the group of dominantly non-breathy speakers, namely H1*-H2*, H1*-A1*, and CPP. The results support the application of the combined auditory-acoustic analysis approach for the assessment of dominantly breathy VQ and thus harmonise the feature with most of the established parameters that are typically analysed in a forensic voice comparison, e.g. f0, formants. In this way, the purely perceptual VQ analysis approach could be confirmed by the signal acoustics. However, the study only supports the approach for high-quality recordings in which there may not be enough vowel sounds due to the brevity of the recording.

As high-quality recordings are rare in a forensic setting, Paper 4 investigated the same research question under the influence of the mobile phone filter. Using the same corpus mobile filtered recording quality, correlations were examined between the precept breathiness and signal acoustics. A pre-test revealed no correlation with any glottal source parameter, i.e., spectral tilt parameters such as H1*-H2*, and H1*-A1*. Only CPP still served as an indicator of breathy VQ. This result was very surprising given that the same group of expert listers as in Paper 2 rated the same voices as dominantly breathy and dominantly non-breathy. Therefore, it was assumed that at least some signal characteristics related to breathiness must still be present in the mobile filtered signal. The hypothesis was put forward that the applied f0 estimator performed inaccurately, which had a negative effect on the HNR parameters and the harmonic estimation and consequently on all spectral tilt parameters.

Hence, Paper 3 assessed the influence of recording quality and voice quality on the accuracy of f0 estimators. Six f0 estimators were examined: Praat Autocorrelation, Praat Cross-Correlation, REAPER, Snack, Subharmonic-to-harmonic ratio and STRAIGHT using the VoiceSauce (Shue et al., 2011) software. The *recording conditions* (studio recording condition and mobile filtered recording condition), the *voice qualities* (breathy voice, modal voice, and creaky voice) as well as *speaker sex* were tested as independent variables – the later only to a very limited extent, as the sample only contained one male and one female speaker. The ground truth f0 of the cardinal vowel samples were manually determined based on each period's length. The output of the tested f0 estimators were compared to the ground truth f0 data. Each f0 estimator's performance was evaluated based on the error measure, mean error absolute (MEA). Results showed that modal voice samples in studio condition did not cause problems for most tested f0 estimators. However, deviations from high quality recordings and modal VQ challenges f0 estimators to different degrees due to differences in operation mode. Aperiodic creak, for example, is most accurately captured by REAPER, while for breathy voice, irrespective of recording condition, Snack outperformed the remaining f0 estimators. STRAIGHT, the default f0 estimator in VoiceSauce, should be avoided when analysing mobile filtered recordings. I conclude that f0 estimators should be wisely chosen when the recording is not high-quality and when the voice to be analysed deviates from modal VQ as the impact of an inaccurate f0 estimation is considerable. The study provides all results to allow researchers to make an informed decision which f0 estimator to use for which material.

The findings from Paper 3 were incorporated into the main study of Paper 4. By reanalysing the data based on the most accurate f0 estimator for breathy voice under the mobile recording condition, the Snack f0 estimator, correlations between the percept breathiness and signal acoustics were found. The clearest correlation still held CPP, but also the lowest harmonic-to-noise parameter, HNR05, turned out to be significant. The mobile phone bandpass filter strongly attenuates the most important low-frequency harmonics up to about 400 Hz, harmonic amplitudes for the first to the fourth harmonic were considered not meaningful. Thus, mid- to high-frequency spectral tilt parameters were inspected for correlation with breathy voice perception. The following turned out to be significant, A1-A3 and H4*-A2*. Relying on the vowel-only subset when analysing the signal acoustics did not result in a clearer correlation. The correctness of the formant estimation was not checked in the study. Correct formant estimation is another major influencing factor that has a direct effect on the spectral tilt measurements. Further studies are needed to investigate the impact of recording quality and voice quality on formant estimation in order to make informed decisions about which formant estimator is most appropriate for which signal.

Perception is still the "gold standard" in the assessment of voice quality (see Oates, 2009), not only when assessing pathological voices. Even in forensic voice comparisons, the perceptual approach is the leading method (San Segundo et al., 2018). The papers 1, 2, and 4 included some kind of perceptual VQ assessment of expert listeners. Overall, the groups of expert listeners performed consistently when rating a speaker's VQ with respect to breathiness and creakiness. For non-pathological voices – which are usually analysed in forensic voice comparison cases – Kreiman et al. (1993) found that the evaluation of specific VQ aspects is relatively similar between listeners and relatively stable within listeners. The present thesis also found overall high between- and within-listener agreement. However, the groups of expert listeners represent a very homogeneous group, as all experts had received similar training in VQ analysis and were colleagues in casework and/or research projects. It can therefore be assumed that the expert listeners are better calibrated than a more heterogenous group with experts from different laboratories and/or universities from different countries. There are many further factors that can influence the between-analyst agreement when assessing the VQ perceptually. The terminology used may not be universally applied, individual internal standards – although relatively similar and stable for non-pathological voices – may differ when assessing the degree of specificity. Experts may focus on different sections within a recording when rating a speaker's long-term VQ. Differences exist in terms of the VQ protocols applied and VQ training received and thus the requirements for calibration may differ considerably.

Thus, the ideal future for voice quality assessment should take an auditory-acoustic approach. As Nolan (1997) stated almost thirty years ago, the auditory-acoustic approach combines the strengths of both approaches, which is particularly important in teh evaluation of degraded audio recordings. if the recording quality allows for it. Using acoustics to confirm the perception of a speaker's dominant VQ can be either implemented into the traditional auditory-acoustic approach or applied within the forensic semi-automatic speaker recognition approach (FSASR, Drygajlo et al., 2015). Both approaches contribute to objectifying the to date less objective process of VQ analysis by enabling transparency and replicability of findings. Additionally, when FSASR is applied, the conclusion framework will be logically correct.

Using synthesised stimuli, Kreiman et al. (2021) found that four harmonic source parameters are needed to accurately model a speaker's voice, i.e., H1-H2, H2-H4, H4-H2K, and H2K-H5K. In addition, previous studies already proofed the relevance of further parameters, namely f0, formant frequencies and bandwidths, intensity, as well as the inharmonic voice source. This proposed psychoacoustic model is supposed to capture a voice as well as Laver's Vocal Profile Analysis scheme (Laver et al., 1981); however, by quantifying a speaker's overall voice quality rather than individual VQ setting as breathiness or creakiness. In contrast to individual VQ settings which provide meaningful acoustic parameters which correlate with perception, as the Paper 2 and Paper 4 applied, the psychoacoustic model outputs "complete integral pattern" (Kreiman et al., 2021, p.464) which are not necessarily perceptually meaningful. Two main caveats prevent the psychoacoustic model from application in forensic voice comparisons. (1) Three parameters out of the "four-piece source spectral model" are not meaningfully available when a recording has been transmitted using the GSM-AMR narrowband codec which is used in mobile telephony. Thus, the harmonic source parameter H4-H2K becomes especially important for forensic application, although the low-frequency spectral tilt parameters, up to H4 are reported to be the most sensitive in characterising the overall pulse shape and thus the overall voice quality (Garellek et al., 2016). (2) The model is developed to describe steady-state phonation rather than a speaker's "long-term-average tendencies" (Mackenzie Beck, 2007, p.5) derived from spontaneous speech samples. We need to define the relevant parameters of the psychoacoustic model for distinguishing voices in degraded bandpass limited audio recordings. Further studies searched for parameters with the highest speaker-discriminatory potential. Lee et al. (2019) reported high harmonic amplitudes (H2 kHz*-H5 kHz) and CPP to account for most variance between speakers using high quality recordings. Jessen (2023) only found the CPP parameter to have a speaker-discriminating potential analysing authentic forensic recordings. More studies are needed which assess large sample sizes of forensic realistic recordings.

As already stated throughout the thesis, the impact of technical aspects on the assessment of voice quality is immense. Thus various technical conditions need to be analysed systematically in match and mismatch condition. Nash (2019) and Van der Vloed et al. (2020) investigated the impact of acoustic mismatch (e.g., net speech duration, signal-to-noise ratio, reverberation, frequency bandwidth and transcoding) and mismatch in recording device on the discrimination performance using FASR systems. Similar investigations on the effects of technical mismatch on VQ acoustics are needed to gain a better understanding of the magnitude of the influencing factors. However, as technology is developing rapidly it is hypothesised that the next generation of mobile communication, i.e., Enhanced Voice Services (Super Wideband and Fullband), will become standard within the next couple of years. These are designed to transmit speech efficiently in high quality covering the entire frequency range that humans perceive, i.e., 20–20.000 Hz (Bruhn et al., 2015), enabling the analysis of all four parameters of the psychoacoustic model (Kreiman et al., 2021). Furthermore, the handling of frame losses is reported to be improved in challenging channel conditions. However, even if the unlimited frequency range will be transmitted, impact on the signal acoustics would still need to be investigated, such as handling of frame loss, background noise, and silence frames as well as potential impact of dynamically changing bit rates.

The research project sheds light on the necessity and possibilities of refining the assessment of VQ for forensic applications. VQ is known to be an extremely useful speaker discrimination feature. To keep pace with developments, the analysis needs to be underpinned by signal acoustics to prevent the feature from being considered inadequate due to a purely perceptual analysis technique. This requires a better understanding of the relationship between perception and signal acoustics in degraded recordings.

References

Bruhn, S., Pobloth, H., Schnell, M., Grill, B., Gibbs, J., Miao, L., Järvinen, K., Laaksonen, L., Harada, N., Naka, N., Ragot, S., Proust, S., Sanda, T., Varga, I., Greer, C., Jelínek, M., Xie, M., & Usai, P. (2015). Standardization of the new 3GPP EVS codec. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 5703–5707. https://doi.org./10.1109/ICASSP.2015.7179064

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition. Verlag für Polizeiwissenschafthttps://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf [accessed 29 Jun 2023]

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019) Voice quality: the laryngeal articulator model. Cambridge University Press.

Garellek, M., Samlan, R., Gerratt, B. R., & Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America* 139(3), 1404–1410.

Jessen, M., Konrat, C., & Horn, J. (2023). Voice comparison analysis of forensic recordings using the voicesauce program. *Proceedings of the 20th International Congress of Phonetic Sciences, Czech Republic*, paper 973.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research* 36(1), 21–40.

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America* 149(1), 457–465.

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 14, 139–155.

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America* 146(3), 1568–1579. https://doi.org/10.1121/1.5125134

Lindblom, B. (2009). Laryngeal mechanisms in speech: The contributions of Jan Gauffin. *Logopedics Phoniatrics Vocology*, 34(4), 149–156.

Mackenzie Beck, J. (2007). Vocal profile analysis scheme: A user's manual. Edinburgh: Queen Margaret, University College–QMUC, Speech Science Research Centre.

Nash, J. (2019). The effect of acoustic variability on automatic speaker recognition systems. [PhD dissertation, University of York]. https://etheses.whiterose.ac.uk/27337/ [accessed 29 June 2023]

Nolan, F. (1997). Speaker recognition and forensic phonetics, In: Hardcastle, W. J. & J. Laver (eds) *The handbook of phonetic sciences*, 744–767 Blackwell.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatrica et Logopaedica* 61(1), 49–56

San Segundo, E., Foulkes, P. French, P., Harrison, P., Hughes, V., Kavanagh, C. (2018). The use of the vocal profile analysis for speaker characterization: Methodological proposals. *JIPA* [online].

Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the 18th International Congress of Phonetic Sciences, Hong Kong*, 1846–1849.

Van der Vloed, D., Kelly, F., & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA: A forensically realistic database. *Proceedings of The Speaker and Language Recognition Workshop, Japan*, 402–407. http://dx.doi.org/10.21437/Odyssey.2020-57 [accessed 29 Jun 2023]

Appendix

Creaky Voice – Sample Acoustics

- Acoustic signal characteristics can be used supportively to corroborate perception
- Extracts are spontaneous speech samples of male British English speakers
- For each speaker, two syllables are presented side by side to assess the nature and relative prominence of spectral characteristics.
  - vowel phonated in (near) modal voice as reference basis
  - vowel phonated in CV
- Following speech acoustics are provided:
  - WF – waveform
  - NBS – narrow-band spectrogram
  - BBS – broad-band spectrogram
  - S – averaged FFT spectrum (vowel portion used to create the spectrum is marked by the dashed line in waveform and spectrograms)
  - frequency range: 0–4 kHz

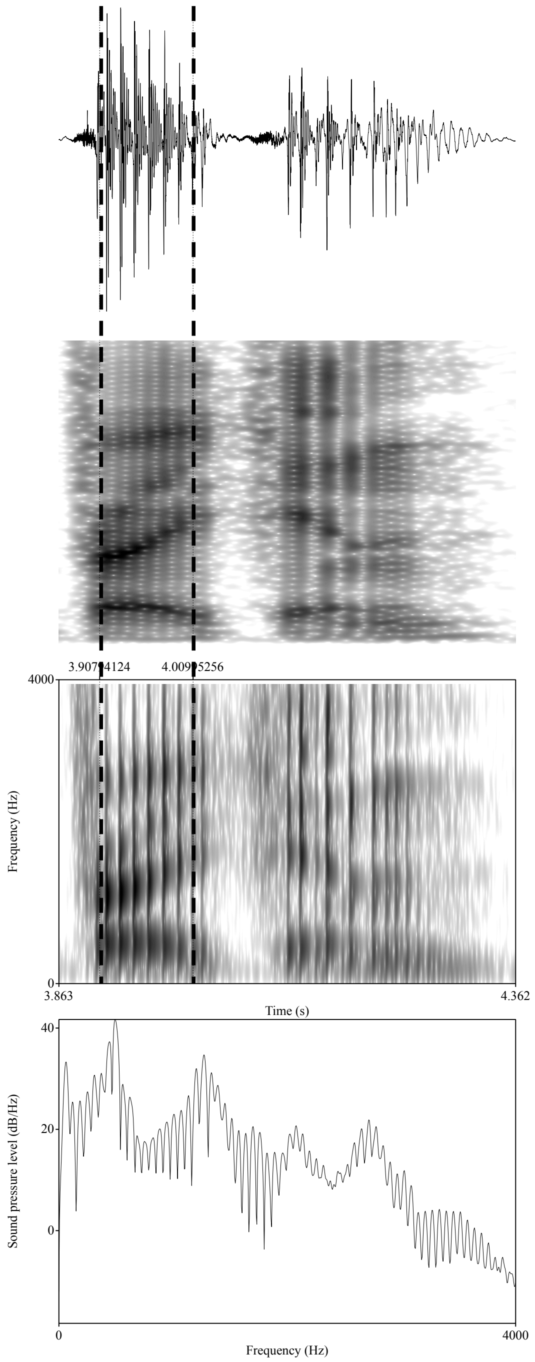
Clean CV

Speaker N26

Modal voice

'Yay'

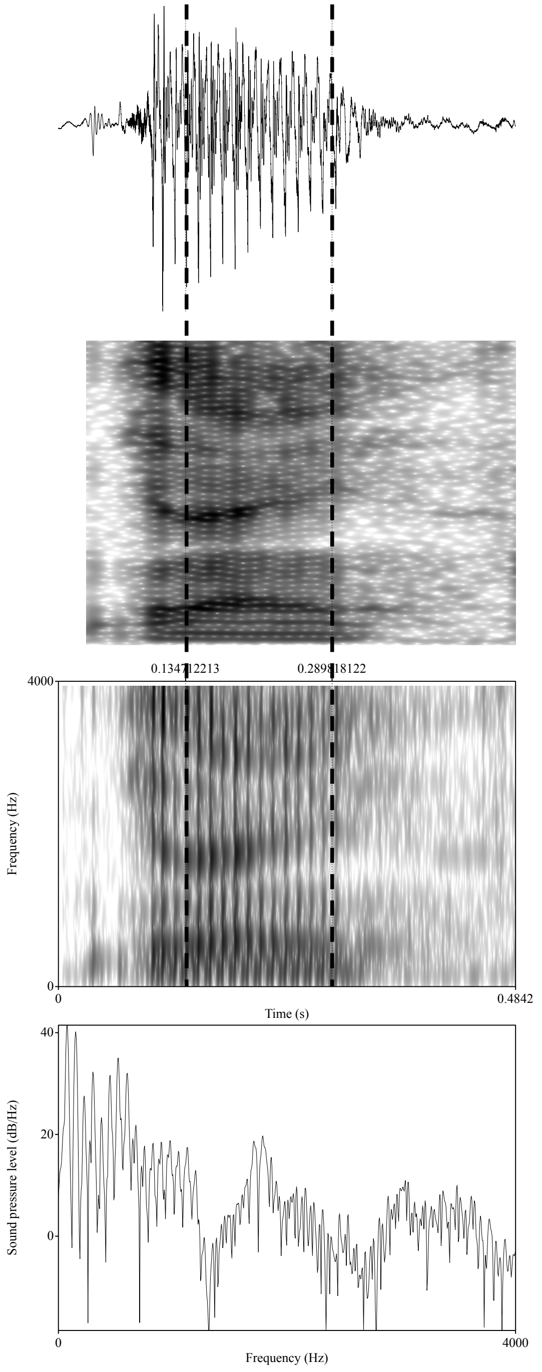
Sec. 0.1



Clean CV

'*broken*'

Sec. 3.86



WF: damping between glottal pulses

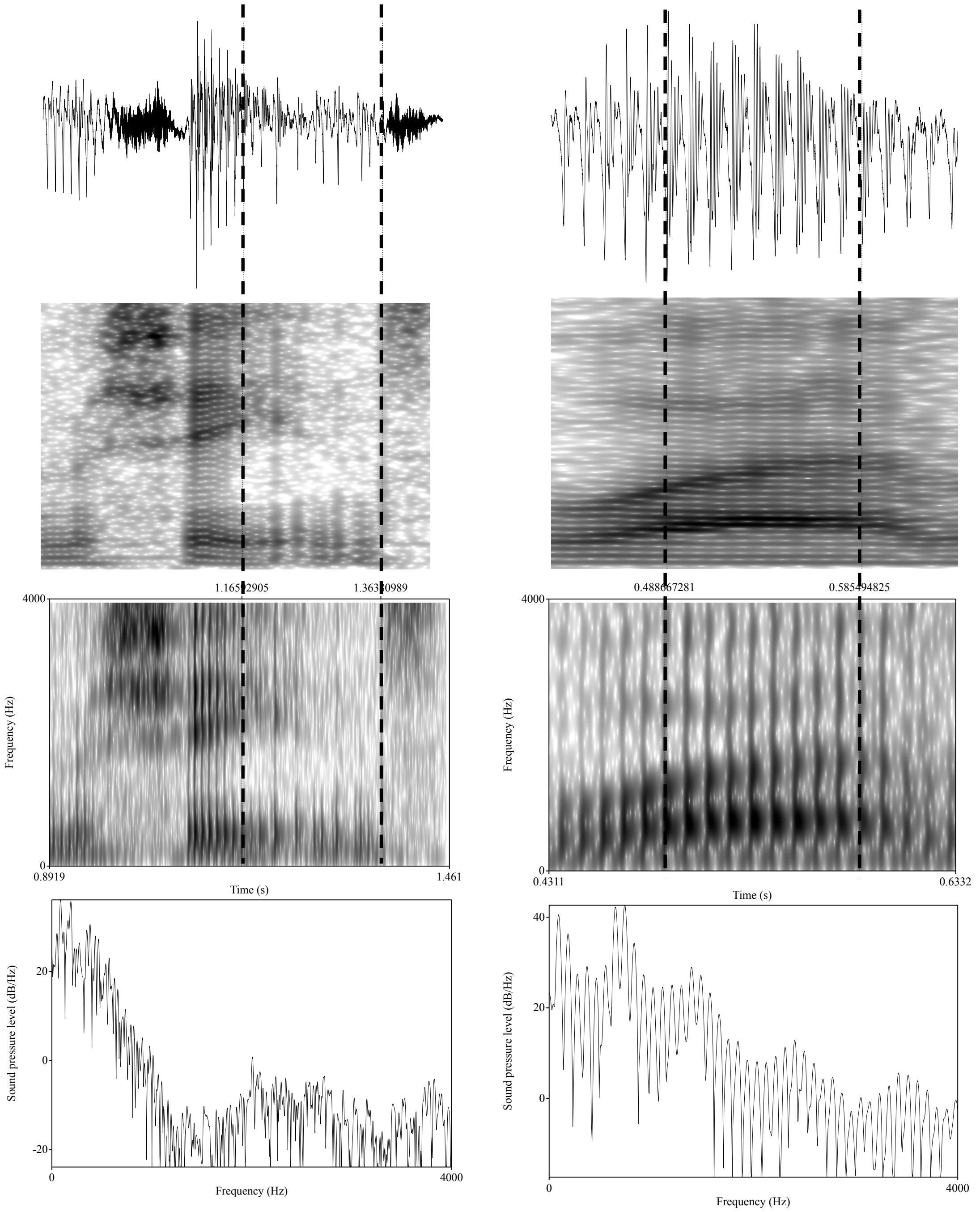
NBS: blurry vertical striations

BBS: very distinct vertical lines over the whole frequency range

S: distinct harmonics over the whole frequency range

Aperiodic C

Speaker DA	
Modal voice	**Aperiodic C**
'wear'	*'machines'*
Sec. 0.42	Sec. 0.88

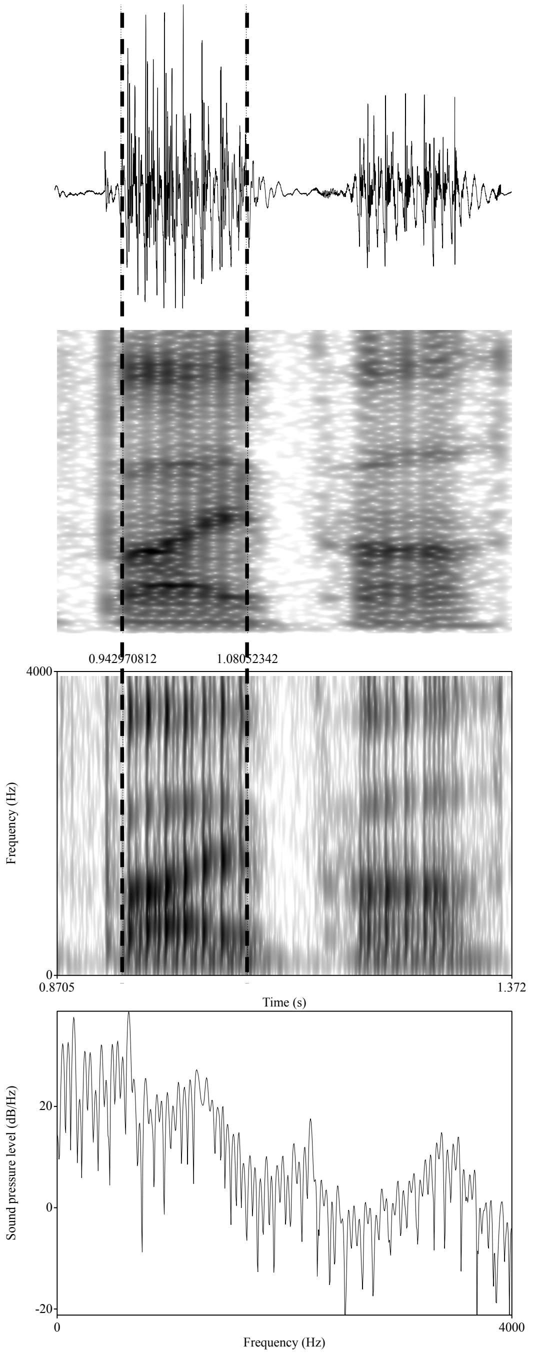


WF: irregular spaced main excitation peaks
BBS: randomly alternating vertical lines, not spanning the whole frequency range
S: interharmonic noise, hardly detectable harmonics

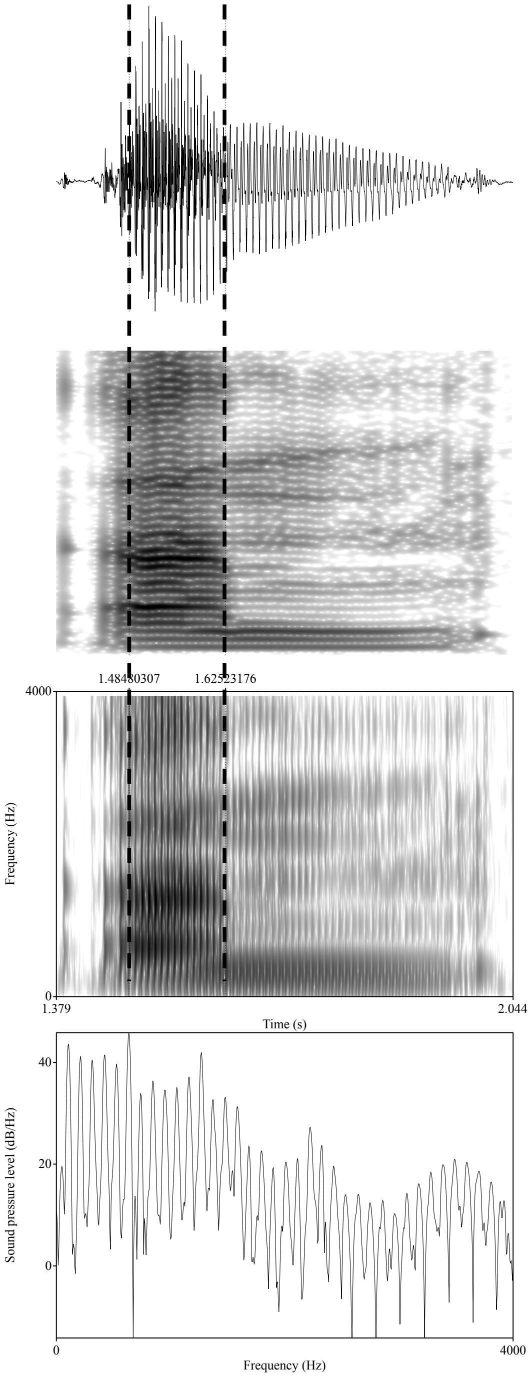
Harsh CV

Speaker N113

Modal voice
'um'
Sec. 1.37



Harsh CV
'bypass'
Sec. 0.89



WF: three repeating patterns
BBS: shallower vertical lines between main glottal pulses
S: harmonics with varying magnitude occur alternately

Breathy CV

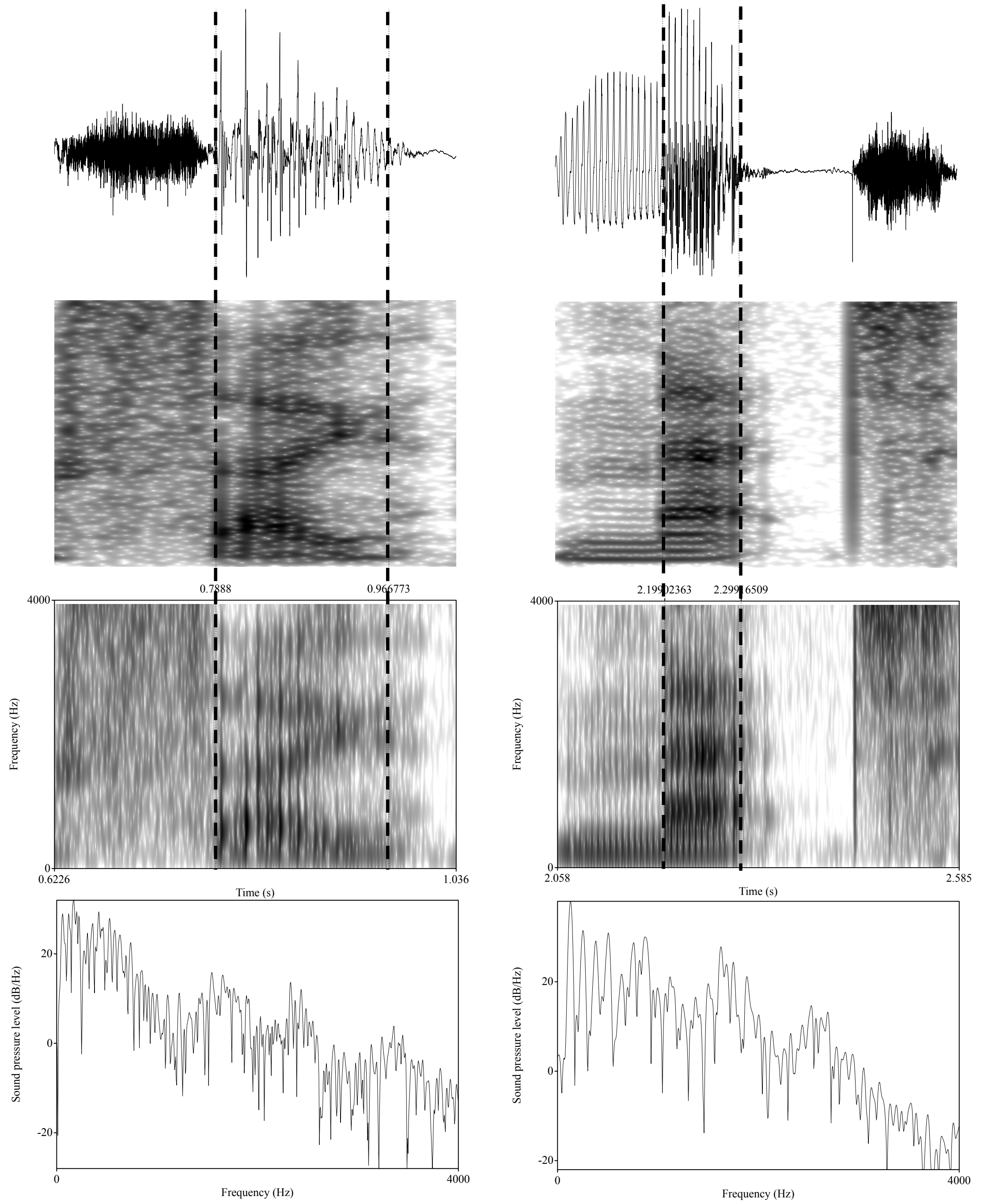
Speaker KN

Modal voice

'met'
Sec. 2.04

Breathy CV

'so'
Sec. 0.62



WF: aperiodically spaced main excitation peaks, irregular amplitude modulation
BBS: less distinct vertical lines, more distinct on formant frequencies
S: high frequency noise, less distinct harmonics