Data-driven Speech Intelligibility Enhancement and Prediction for Hearing Aids

Zehai Tu

Supervisors: Prof Jon Barker, Dr Ning Ma

Department of Computer Science
University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

Speech and Hearing Research Group July 2023

This thesis is dedicated to my parents for their endless love, support and encouragement.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this work are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Zehai Tu
July 2023

Acknowledgements

First of all, I must thank my supervisors Prof Jon Barker and Dr Ning Ma. Their help and advice ultimately lead to the completion of this thesis. They have continuously provided inspiration and feedback throughout my PhD research journey, and a lot of support to help me to survive the pandemic period. Also, thanks to Jon for advice on climbing and hiking, and thanks to Ning for the warm Chinese New Year treat.

I would also like to thank Prof Eleni Vasilaki for her help to get me the PhD position, and her support in the early stage of my PhD.

Many thanks to all my friends and colleagues in the Regent Court, especially all those past and present members of the Speech and Hearing group. Thanks to the happy moments spent together, and the tolerance for my endless complaints on everything. Without their support, I would not be able to complete the four-year journey.

Thanks to my climbing friends, the laughters and wounds in the Sheffield climbing gyms.

This PhD was supported by the Departmental Scholarship from the department of Computer Science at the University of Sheffield. I also want to acknowledge the financial support from the Clarity project, funded by UKRI.

Thanks to Shuangyunyi for making sure I finished this thesis. Thanks to Zhihao for discussions every now and then. Also thanks to all my hometown friends.

Finally, thanks to my parents for their unconditional love and support!

Abstract

Hearing impairment is a widespread problem around the world. It is estimated that one in six people are living with some degree of hearing loss. Moderate and severe hearing impairment has been recognised as one of the major causes of disability, which is associated with declines in the quality of life, mental illness and dementia. However, investigation shows that only 10-20% of older people with significant hearing impairment wear hearing aids. One of the main factors causing the low uptake is that current devices struggle to help hearing aid users understand speech in noisy environments. For the purpose of compensating for the elevated hearing thresholds and dysfunction of source separation processing caused by the impaired auditory system, amplification and denoising have been the major focuses of current hearing aid studies to improve the intelligibility of speech in noise. Also, it is important to derive a metric that can fairly predict speech intelligibility for the better development of hearing aid techniques.

This thesis aims to enhance the speech intelligibility of hearing impaired listeners. Motivated by the success of data-driven approaches in many speech processing applications, this work proposes the differentiable hearing aid speech processing (DHASP) framework to optimise both the amplification and denoising modules within a hearing aid processor. This is accomplished by setting an intelligibility-based optimisation objective and taking advantage of large-scale speech databases to train the hearing aid processor to maximise the intelligibility for the listeners. The first set of experiments is conducted on both clean and noisy speech databases, and the results from objective evaluation suggest that the amplification fittings optimised within the DHASP framework can outperform a widely used and well-recognised fitting. The second set of experiments is conducted on a large-scale database with simulated domestic noisy scenes. The results from both objective and subjective evaluations show that the DHASP-optimised hearing aid processor incorporating a deep neural network-based denoising module can achieve competitive performance in terms of intelligibility enhancement.

A precise intelligibility predictor can provide reliable evaluation results to save the cost of expensive and time-consuming subjective evaluation. Inspired by the findings that automatic speech recognition (ASR) models show similar recognition results as humans

in some experiments, this work exploits ASR models for intelligibility prediction. An intrusive approach using ASR hidden representations and a non-intrusive approach using ASR uncertainty are proposed and explained in the third and fourth experimental chapters. Experiments are conducted on two databases, one with monaural speech in speech-spectrum-shaped noise with normal hearing listeners, and the other one with processed binaural speech in domestic noise with hearing impaired listeners. Results suggest that both the intrusive and non-intrusive approaches can achieve top performances and outperform a number of widely used intelligibility prediction approaches.

In conclusion, this thesis covers both the enhancement and prediction of speech intelligibility for hearing aids. The proposed hearing aid processor optimised within the proposed DHASP framework can significantly improve the intelligibility of speech in noise for hearing impaired listeners. Also, it is shown that the proposed ASR-based intelligibility prediction approaches can achieve state-of-the-art performances against a number of widely used intelligibility predictors.

Table of contents

List of figures **xv**

List of tables **xix**

Nomenclature **xxi**

1 Introduction **1**
1.1 Mechanisms in hearing . . . 2
1.2 Declines in hearing ability . . . 3
1.3 A brief history of hearing aid development . . . 5
1.4 Motivations . . . 6
1.4.1 Data-driven speech intelligibility enhancement . . . 7
1.4.2 Intelligibility prediction . . . 8
1.5 Research questions . . . 9
1.6 Contributions . . . 10
1.7 Thesis Outline . . . 11

2 Background and Related Work **13**
2.1 Hearing loss compensation . . . 14
2.1.1 Linear amplification . . . 14
2.1.2 Nonlinear amplification . . . 15
2.1.3 Frequency lowering . . . 18
2.2 Speech denoising for hearing aids . . . 19
2.2.1 Beamformers . . . 19
2.2.2 DNN-based approaches . . . 21
2.3 Intrusive speech intelligibility prediction . . . 23
2.3.1 Acoustic representation-based approaches . . . 24
2.3.2 ASR-based approaches . . . 27
2.3.3 Hearing impairment intelligibility prediction . . . 28

2.4 Non-intrusive speech intelligibility prediction 31
2.4.1 Acoustic representation-based approaches 31
2.4.2 Pseudo reference-based approaches 32
2.4.3 Data-driven approaches 33
2.4.4 ASR-based approaches 34

3 Optimising Hearing Aid Fitting with the DHASP framework 35
3.1 Introduction 35
3.2 DHASP framework overview 37
3.3 Fitting optimisation for clean speech 38
3.3.1 Differentiable HASPI-based objective 38
3.3.2 Experiments 42
3.3.3 Results 44
3.4 Fitting optimisation for noisy speech 44
3.4.1 Differentiable MSBG model-based objective 45
3.4.2 Experiments 48
3.4.3 Results 49
3.5 Conclusions 53

4 Incorporating DNN-based Denoising into the DHASP framework 55
4.1 Introduction 55
4.2 Method 56
4.2.1 Denoising module 57
4.2.2 Amplification module 58
4.2.3 Hearing loss model 59
4.3 Experimental setup 59
4.3.1 Overview of CEC1 Database 60
4.3.2 System setup 61
4.3.3 Evaluation 62
4.3.4 Baselines 63
4.4 Results 65
4.4.1 Objective results 65
4.4.2 Subjective results 66
4.5 Conclusions 70

5 Intrusive Intelligibility Prediction with ASR Hidden Representations 71
5.1 Introduction 71

5.2 Similarities between ASR hidden representations . . . 72
5.2.1 DNN-based ASR model . . . 73
5.2.2 Hidden representations . . . 74
5.2.3 Similarity computation . . . 75
5.3 Experimental setup . . . 76
5.3.1 Databases . . . 76
5.3.2 ASR configuration . . . 76
5.3.3 Evaluation . . . 77
5.4 Monaural speech in SSN with normal hearing listeners . . . 77
5.4.1 Corpus description . . . 77
5.4.2 Baselines . . . 78
5.4.3 Results . . . 83
5.5 Processed binaural speech in domestic noise with hearing impaired listeners 84
5.5.1 Corpus description . . . 84
5.5.2 Baselines . . . 85
5.5.3 Results . . . 89
5.6 Conclusions . . . 91

6 Non-intrusive Intelligibility Prediction with Unsupervised ASR Uncertainty 95
6.1 Introduction . . . 95
6.2 Unsupervised ASR uncertainty estimation . . . 98
6.2.1 Sequence-level uncertainty estimation . . . 98
6.2.2 Token-level ASR posterior . . . 100
6.3 Experimental setup . . . 100
6.4 Monaural speech in SSN noise with normal hearing listeners . . . 101
6.4.1 Baselines . . . 101
6.4.2 Results . . . 103
6.5 Processed binaural speech in domestic noise with hearing impaired listeners 104
6.5.1 Baselines . . . 104
6.5.2 Results . . . 105
6.6 Conclusions . . . 111

7 Conclusions 113
7.1 Contributions . . . 114
7.1.1 Speech intelligibility enhancement for hearing aids . . . 114
7.1.2 Speech intelligibility prediction . . . 115
7.2 Limitations and future research . . . 117

References **119**

List of figures

2.1 Flow chart showing speech to the perception of a hearing impaired listener. 14

2.2 Speech loudness perception of the normal and reduced dynamic range, adapted from (Kuk, 1996). The arrows represent the loudness range of input speech, and the box areas represent the loudness dynamic range of listeners. Sub-figure (a) and (b) show the normal dynamic range and the reduced dynamic range perceiving loud, conversational and soft speech, respectively. Sub-figure (c) presents the linear amplification effect for the reduced dynamic range. The loud speech can be uncomfortably too loud for impaired hearing, despite the soft speech can be perceived. Sub-figure (d) shows the nonlinear dynamic range compression amplification that provides insertion gain adaptively, so that speech at various sound pressure levels can be perceived comfortably. 16

2.3 Presentations of three frequency lowering techniques. Spectrum of an unprocessed signal (a) and that processed by three frequency lowering techniques: (b) frequency compression, (c) frequency transposition, (d) frequency composition. U, S, and T represent unprocessed, source, and target frequency bands, respectively. 18

2.4 General approach of intrusive intelligibility prediction. 25

2.5 HASPI auditory model for envelope extraction. 29

2.6 Diagram of the MSBG hearing loss model. 30

2.7 Non-intrusive intelligibility prediction with estimated reference. 32

3.1 Overall workflow of DHASP. 37

3.2 Differentiable HASPI-based model. 39

3.3 Frequency responses of the DHASP optimised and the NAL-R fittings for standard audiograms. 43

3.4 HASPI intelligibility scores of the original, NAL-R processed, and DHASP processed signals. 43

3.5 Differentiable MSBG-based model. 45
3.6 Frequency responses of NAL-R fitting, custom fittings optimised with noisy data (Cn), custom fittings optimised with Wiener filtering enhanced noisy data (Cw), and the general fittings (G) for different hearing losses. 50

4.1 Overall workflow of DHASP including a denoising module and an amplification module. 56
4.2 Overall workflow of the two-stage optimisation for the denoising and the amplification modules. In the second stage, the denoising module can be jointly optimised together with the amplification module. 57
4.3 Structure of MC-Conv-TasNet 58
4.4 An example scene in CEC1 database. 60
4.5 Subjective evaluation correctness of the proposed approach and the CEC1 baseline, and average hearing thresholds of each listener. 67
4.6 Box plots of subjective evaluation correctness against better-ear SNRs of the unprocessed scenes. 68

5.1 A general framework for intrusive intelligibility prediction. The proposed approach in this chapter uses an ASR model as the representation extractor. 72
5.2 ASR architecture and hidden representations at three different levels. 74
5.3 Scatter plot of the listener word correctness scores (WCS) distribution in the evaluation set at different SNR levels. The opaqueness is correlated with the density. 78
5.4 Scatter plots of all intelligibility predictions on the Grid corpus evaluation set, along with the logistic fitting functions. 81
5.5 RMSE and NCC of the intelligibility predictors at different SNRs. The dashed lines represent the baseline approaches, and the solid lines represent the proposed ASR hidden representation-based approaches. 82
5.6 Intelligibility prediction for hearing impaired listeners with the MSBG hearing loss simulator. 85
5.7 Scatter plots of all intelligibility predictions on the CPC1 *closed-set* evaluation set, along with the logistic fitting functions. 87
5.8 Scatter plots of all intelligibility predictions on the CPC1 *open-set* evaluation set, along with the logistic fitting functions. 88
5.9 Listener- and system-wise average intelligibility with standard errors on the *closed-set*. 93

6.1 An ensemble of ASR models is used to estimate the uncertainty of a processed speech, which is then used for intelligibility prediction. 98
6.2 Scatter plots of all intelligibility predictions on the Grid corpus evaluation set, along with the logistic fitting functions. 102

List of tables

3.1 Hearing losses of the audiograms used in this study. 48
3.2 Evaluation scores of various fittings applied to noisy speech before and after the enhancement of Wiener filtering. 51

4.1 Overview of the systems submitted to CEC1. 63
4.2 CEC1 **objective** evaluation results. 65
4.3 CEC1 **subjective** evaluation results. 66

5.1 Evaluation results on the Noisy Grid corpus in terms of RMSE, NCC, and KT. The down arrow indicates the smaller the better, and the up arrows indicate otherwise. 80
5.2 Evaluation results on both CPC1 *closed-set* and *open-set* in terms of RMSE, NCC, and KT. 86
5.3 Evaluation results on the *closed-set* of decoder representations from different ASR models. 91
5.4 Listener- and system-wise evaluation results on the *closed-set* of predicted intelligibility. 92

6.1 Evaluation results on the Noisy Grid corpus in terms of RMSE, NCC, and KT. 103
6.2 Evaluation results on ASR ensembles trained by different training databases. 104
6.3 Evaluation between the listening results WCS and predicted intelligibility measures on CPC1 evaluation set. 106
6.4 Listener- and system-wise evaluation results on the *closed-set* of predicted intelligibility. 107
6.5 Evaluation results on the *closed-set* of different ensembles of ASR models trained on different databases, and with or without using MSBG hearing loss simulation. 108
6.6 The effect of tuning uncertainty estimation hyperparameters on system performance as measured by RMSE, NCC and KT. 110

Nomenclature

Acronyms / Abbreviations

ABECm Across-Band Envelope Correlation metric

AI Articulation Index

ASR Automatic Speech Recognition

CNN Convolutional Neural Network

CSII Coherence Speech Intelligibility Index

DHASP Differentiable Hearing Aid Speech Processing

DNN Deep Neural Network

DTW Dynamic Time Warping

ESTOI Extended Short-time Objective Intelligibility

HASPI Hearing Aide Speech Perception Index

HR Hidden Representation

IHC Inner-hair Cell

IIR Infinite Impulse Response

KT Kendall's Tau

MBSTOI Modified Binaural STOI

ModA average Modulation-spectrum Area

MVDR Minimum Variance Distortionless Response

NCC Normalised Cross Correlation

NCM Normalised Covariance Measure

NI-STOI Non-Intrusive STOI

NIC-STOI Non-Intrusive Codebook-based STOI

OHC Outer-hair Cell

rDRm reduced speech Dynamic Range measure

RMSE Root Mean Square Error

RMS Root Mean Square

SDR Signal-to-distortion Ratio

sEPSM speech-based Envelope Power-Spectrum Model

SII Speech Intelligibility Index

SNR Signal-to-noise Ratio

SRMR Speech to Reverberation Modulation energy Ratio

SRT Speech Reception Threshold

SSN Speech Shaped Noise

STI Speech Transmission Index

STOI Short-time Objective Intelligibility

WCS Word Correctness Score

WER Word Error Rate

Chapter 1

Introduction

Hearing impairment is a widespread problem across the world. It is estimated that 1.59 billion people are living with some degree of hearing impairment, and 430 million of them suffer from moderate or severe impairment (World Health Organization, 2021). Declines in hearing ability are associated with not only a decrease in life quality but also an increased risk of developing mental illness. Not being able to hear clearly or understand what others say is more than just an inconvenience: the impediment to daily communication leads to social isolation, which can increase the risk of cognitive decline and mortality (Loughrey et al., 2018). Furthermore, it has been recognised that hearing impairment may be the most important modifiable risk factor for dementia (Livingston et al., 2017).

Despite hearing impairment causing serious consequences, the difficulty of understanding speech in noise is yet to be adequately resolved. It is estimated that only 10-20% of older people with significant hearing impairment choose to wear hearing aids (Davis et al., 2016). One of the most important factors contributing to this very low uptake is that the benefit from hearing aids can be minimal in noisy environments (McCormack and Fortnum, 2013). '*I can hear you, but I can't understand you*' is one of the most common complaints from hearing aid users (Lesica, 2018). Despite restoring audibility to some degree, current hearing aids are often ineffective at restoring speech intelligibility in the presence of background noise. Therefore, hearing aids that can significantly improve the intelligibility of speech in noise are highly desirable. Additionally, a major challenge in developing such hearing aids is that the factors governing speech intelligibility are only poorly understood.

The focus of this thesis is enhancing speech intelligibility for hearing impaired listeners. Inspired by the recent success of data-driven approaches in many audio signal processing applications (Purwins et al., 2019), this work seeks to extend the applications of the data-driven methodology to intelligibility improvement for hearing aids. In addition, accurate

intelligibility prediction is also found to be crucial for the development of hearing aids, and it is thus studied in this thesis as well.

This chapter starts with a brief introduction to the mechanisms of hearing. The declines in hearing ability are then explained to motivate the usage of hearing aids for speech intelligibility improvement. In the following section, the development of hearing aids is presented. Afterwards, the motivations for the two major themes of work in this thesis, data-driven speech intelligibility enhancement and prediction, are established. Finally, the research questions, the contributions, and the outline of the thesis are presented.

1.1 Mechanisms in hearing

In order to better understand hearing impairment, it is necessary to first understand how a healthy auditory system processes acoustic signals. To 'hear', sound pressure waves need to be converted to electrical signals in the brain. This transduction is performed in a snail shell-like structure called the cochlea. This structure conducts the transformation from the mechanical signals that pass through the outer and middle ear to the electrical signals of auditory nerves which are then perceived by the brain. Specifically, the incoming sound waves lead to the vibration of the basilar membrane along the length of the cochlea. Following the vibration, the inner hair cells (IHCs) that are attached to the basilar membrane within the cochlea then release neurotransmitters onto auditory nerves to induce corresponding electrical signals.

A simplified understanding of what the human auditory system does in the mechanical to electrical signal transformation is a process consisting of amplification, dynamic range compression, and frequency analysis (Lesica, 2018). First, when the incoming sounds are too weak to vibrate the basilar membrane strongly enough to activate auditory nerve activities, outer hair cells (OHCs) can provide active amplification by reinforcing the passive movement of the basilar membrane. Second, the incoming sound is compressed because the OHC amplification decreases as the sound level increases. The compression ensures that a wide-spanning of sound levels, e.g., from normal breathing sounds at around 10 dB to chainsaw noises that can reach 120 dB, can be encoded with a limited dynamic range of auditory nerve activity. Third, the frequency selectivity is achieved by the structure of the cochlea itself. The spiral shape reduces the size of the cochlea, and the reduced diameter of coiling is reflected in the width of the basilar membrane, hence tuning the frequency range. Consequently, the basilar membrane vibration amplitudes and the subsequent auditory nerve activities in different cochlea areas reflect the energy at different frequencies in the incoming

sounds. Therefore, the electrical signal sent to the brain could be approximated as having undergone a frequency analysis.

It is insufficient to understand the auditory system performing only such simple frequency analysis, because the transformation is highly nonlinear. The amplification and compression themselves perform relatively simple processing and ideally can be replaced by the wide dynamic range compression, which is a common scheme of current hearing aids. Meanwhile, OHCs that modulate basilar membrane movement not only help activate their attached area but also influence other cochlea regions. As a result, the basilar membrane vibration does not simply reflect the energy at a certain frequency of an incoming sound. It is also dependent on the energy at other frequencies, which leads to nonlinear cross-frequency interactions. The auditory nerve activities can thus differ from those of the simple frequency analysis in various ways. One is that the energy at two frequencies in the incoming sounds can activate the vibration of the basilar membrane in an additional cochlea area, which is posited to be triggered by the third frequency that is not presented in the sounds. Another is the winner-take-all, i.e., the dominant frequency activation in a local basilar membrane area can suppress the amplification provided by OHCs at other frequencies with lower levels of stimulation. The resulting complex auditory nerve activity patterns are crucial for recognising speech in noise (Sachs et al., 1983; Sachs and Young, 1980). Due to the complex nonlinearity of the transformation, it is extremely difficult to thoroughly model the mechanisms in hearing, thus difficult to restore the transformation process when some part of the auditory system is dysfunctional.

1.2 Declines in hearing ability

Hearing impairment is usually categorised as one of three types: conductive, sensorineural, and mixed hearing impairment. Conductive hearing impairment involves a problem in the outer or middle ear, sensorineural hearing impairment involves a problem in the inner ear, and mixed hearing impairment is a combination of the previous two. Conductive hearing impairment can often be treated with surgical intervention or pharmaceuticals to partially or sometimes fully restore the hearing ability. For that reason, hearing aid users are usually suffering from sensorineural hearing impairment, which is caused by the degradation in the inner ear and is usually permanent. Therefore, the sensorineural hearing impairment will be the focus of the thesis.

There are a number of factors that can damage or lead to the dysfunction in the inner ear. Ageing is the major reason for sensorineural hearing impairment and age-related hearing loss has been projected to be one of the top leading causes of burden of disease by 2030

(Mathers and Loncar, 2006). As the inner ear structures usually degenerate over time, the hearing ability declines with ageing. Exposure to loud sounds is another factor causing sensorineural hearing impairment (Sliwinska-Kowalska et al., 2012). This includes exposure to loud noises, such as a gunshot; occupational noises such as construction or factory work; and recreational sounds such as listening to loud music. Ototoxic drugs can also lead to temporary or permanent dysfunction of the inner ear (Arslan et al., 1999). Diseases that can result in high fever may also damage the cochlea (Mateer et al., 2018). Additionally, genetic makeup can make a certain group of people more susceptible to ear damage from loud sounds or ageing (Willems, 2000).

The most obvious consequence of hearing impairment is the degradation of the amplification and compression in the cochlea. A typical symptom is a loss of hearing sensitivity, i.e., auditory nerves can no longer be triggered by weak sounds, while less auditory nerve activity is likely to be elicited by louder sounds than that in normal hearing. This degradation is usually caused by the dysfunction of OHCs, which provide active amplification and compression. This motivates the use of the wide dynamic range compression in hearing aids (Kates, 2008), which is designed to provide the amplification and compression that damaged OHCs can not provide anymore. The wide dynamic range compression will apply greater amplification to weak sounds than it does to stronger sounds. However, this strategy is insufficient to restore the intelligibility of speech in noise, as hearing impairment is more complex than that.

Hearing impairment can be described as a distortion of auditory neural activity patterns (Lesica, 2018). One major type of distortion is caused by the dysfunction of the highly nonlinear auditory processing. The nonlinear cross-frequency interactions in the cochlea can be lost to some degree, and these interactions are highly dependent on the OHCs and are crucial for speech perception in noisy environments (Recio-Spinoso and Cooper, 2013; Sachs et al., 1983; Sachs and Young, 1980). Therefore, the auditory nerve activity patterns from an impaired ear are very different from those from a normal ear, and fail to provide a sufficient basis to distinguish different sound sources and are less robust to recognise speech with background noises. Consequently, the loss of these interactions caused by the dysfunction of OHCs can lead to difficulty in understanding speech in noise. Another type of distortion is originated from the auditory nerves themselves (Liberman and Kujawa, 2017), which results in the degeneration of transmission from IHCs to the brain. Additionally, the hearing impairment can lead to impaired temporal processing by some measures, e.g. failure to detect the short pauses within a sound (Humes et al., 2010). The temporal processing in the auditory system is critical for both the ability to localise sound sources and the ability to recognise speech in noisy and reverberant environments (Marrone et al., 2008). Furthermore, the long-term influences of hearing impairment usually extend to the brain itself due to brain

plasticity (Tremblay and Miller, 2014). One example is that the brain gradually learns to enforce the inputs from the ear as the inputs are weakened with the development of the impairment, which can lead to improved discrimination of perceivable low-level sounds (Gourévitch et al., 2014). As a consequence of the complexity of the neural activity distortion, it is unlikely to fully restore normal auditory perception with hearing aid devices.

1.3 A brief history of hearing aid development

Due to the complicated neural activity distortion caused by hearing disability, conventional hearing aids are not really trying to fully restore normal auditory perception. A major interest in hearing aid studies has been to benefit hearing impaired listeners by improving speech intelligibility for better communication in daily environments. In this section, the history of hearing aids for intelligibility enhancement is briefly reviewed. It covers the early solution of amplification, recent attempts at restoring compression, and more advanced hearing aids with noise suppression. This advancement follows the developing knowledge of the declines in hearing ability, i.e., from the simple understanding of amplification and compression dysfunction in the cochlea, to auditory neural activity pattern distortion.

In the early days of active hearing aid development, the aim was to amplify the incoming sound so that hearing impaired listeners can hear and understand it. In 1898, the first portable hearing aid was invented by using a carbon microphone to turn a weak signal into a strong one with an electric current. This device, consisting of a separate microphone, headphones, amplifier, and battery, was bulky and difficult to use. Later in the early 1900s, vacuum tube hearing aids were developed and gained popularity in the market. These hearing aids leveraged telephone microphones to convert speech into electric signals and amplified by the valves, and delivered through the receivers. The vacuum tubes were then replaced by transistors in hearing aids, which were smaller, required less power and amplified signals with less distortion. In these early years, the key question was how best to amplify the incoming sound. The one-half gain rule, that is to amplify the sounds slightly less than one-half of the hearing thresholds, was proposed by Lybarger (1963) and provided a basis for linear amplification.

Later, the dysfunction of compression in the cochlea was tackled with dynamic range compression (Fowler, 1936). In the 1970s, microprocessor-based hearing aids were invented which enabled multi-channel processing of audio signals. Later in the late 1980s, the first fully digital hearing aids, which used microcomputers to control analogue amplifiers, filters, and limiters, were brought to the market. The digital multi-channel processing enabled the technique of wide dynamic range compression. This technique is to dynamically adjust

the amplification provided by hearing aids according to the levels of incoming sounds, i.e., to amplify less when the sound is loud and to amplify more when the sound is quiet. The amplification formulae from this stage were developed to be compatible with dynamic range compression to maximise listening comfort along with intelligibility.

Despite the success of hearing aids in improving intelligibility in quiet environments, they provided little improvement to speech intelligibility in noisy environments. The reason is that the distortion of auditory neural activity patterns could not be fully restored, and these patterns are crucial for human auditory systems to perform source separation as introduced in the previous section. As a result, modern hearing aids, which are more like high-tech hearing buds, offering Bluetooth connections to smartphones, and rechargeable batteries, always deploy a denoising module to suppress environmental interference. The noise suppression function has the potential to compensate for the dysfunction of source separation in the human auditory systems, and thus to improve the speech intelligibility of noisy speech for hearing impaired listeners.

In brief, the development of hearing aids was from amplification, compression, to noise suppression. The early-stage hearing aids aimed to restore audibility by amplifying incoming sounds. Later, compression techniques were focused to control excessive loudness that resulted from linear amplification. For modern hearing aid studies, speech in noise enhancement has been gaining more attention. With the advancement of hearing aid techniques, the ultimate goal of hearing aids has always been to enhance speech intelligibility for hearing impaired listeners.

1.4 Motivations

For the purpose of enhancing speech intelligibility for hearing impaired listeners, an ideal hearing aid is expected to provide adequate amplification and effective noise suppression. This thesis hence focuses on the optimisation of hearing aid amplification and denoising using data-driven approaches, which recently have achieved significant advancement in many speech processing tasks, such as speech recognition (Nassif et al., 2019), speech separation (Wang and Chen, 2018), etc.

During the development of hearing aid intelligibility enhancement algorithms, it has been found that an accurate intelligibility predictor can play a crucial role. A good intelligibility model can benefit the optimisation of both hearing aid fittings and noise suppression. Additionally, it can provide reliable performance evaluation of hearing aids. As subjective evaluation by human listening experiments can be quite expensive and time-consuming, accurate objective evaluation can help accelerate the development of new hearing aid models.

Therefore, novel intelligibility prediction approaches are also proposed and presented in this thesis.

1.4.1 Data-driven speech intelligibility enhancement

Typically, a hearing aid amplifies sounds according to a frequency-gain amplification table, which is fitted to the listener's audiogram, i.e., measurements of the levels at which pure-tones become audible at various frequencies. An appropriate amplification fitting is expected to not only make overall loudness comfortable but also make speech intelligible. The fitting process should be conducted by an audiology specialist, otherwise, the hearing aid may work poorly. This configuration process can take months and typically requires a number of return visits to an audiology clinic. Therefore, an automated fitting approach would be highly desirable as it can speed up the process and reduce costs.

Additionally, there is a need for scene-dependent fittings in spite of the remarkable success of current general hearing aid prescriptions. According to Kochkin (2010), the satisfaction of hearing aid users can vary a lot across a range of listening environments. For example, over 90% of hearing aid users are satisfied with the communication improvement in one-on-one situations, while less than 60% of users are satisfied in school or a classroom. Therefore, automated optimisation of hearing aid fittings for different noisy environments could help improve speech intelligibility. Also, whether the noise suppression feature provided by advanced hearing aids is turned on or not can be another important factor that can influence hearing aid fittings.

Recently, the emergence of the differentiable digital signal processing (Engel et al., 2020) provided an approach to the automated data-driven optimisation of parameterised speech processing models. In general, the performances of data-driven models are closely related to the data used for the optimisation. This enables the automated optimisation of customised scene-dependent hearing aid fittings, as the data generated in different listening environments can be different, and thus the optimisation produces different solutions. Additionally, the objective function (i.e., optimising target) is also crucial and determines the performance of the optimised models. With data-driven optimisation, the fittings can be optimised to directly maximise the intelligibility of hearing impaired listeners by introducing an intelligibility-based objective function. Motivated by this, this thesis explores the efficacy of data-driven optimisation for hearing aid fittings. It is not only the general fittings with respect to listeners' hearing abilities, but also fittings customised to various noisy environments that will be investigated.

Effective noise suppression is crucial for hearing aid speech intelligibility enhancement, as the source separation ability of hearing impaired listeners can be profoundly degraded.

In recent years, data-driven approaches with deep neural networks (DNNs) have brought a huge improvement to speech denoising (Luo and Mesgarani, 2019; Xu et al., 2014; Zhang et al., 2020). However, a limited number of works have been conducted for the purpose of hearing aid speech enhancement. This is due to the strict real-time requirement of hearing aids, i.e., the desirable latency for hearing aid processing can be as low as 5 to 6 ms (Stone et al., 2008). Moreover, many speech denoising works target improving the recognition accuracy of automated speech recognition systems, or improving the perception for normal hearing listeners. Few works have been conducted for the purpose of speech intelligibility improvement for hearing impaired listeners. Therefore, this thesis will further explore the efficacy of DNN-based noise suppression models in the case of hearing aid speech intelligibility enhancement.

1.4.2 Intelligibility prediction

Although the factors governing speech intelligibility have been studied since the 1920s (French and Steinberg, 1947), there is still much that is poorly understood. Most existing approaches predict intelligibility by measuring the signal-to-noise ratio (SNR) at modulation frequencies or the correlation within frequency bands between a high quality reference speech signal and the degraded speech. Although these approaches are suitable for speech with some types of degradation, e.g., stationary additive noise, reverberation, and clipping, they do not generally work well when speech is degraded by strong non-stationary noise or non-linear processing by, for example, Wiener filtering, DNNs (Gelderblom et al., 2017; Yamamoto et al., 2017).

It is common for speech intelligibility predictors to leverage additional speech-related information apart from the degraded speech signal itself, e.g., the transcription, and the corresponding clean reference speech signal. These methods are usually described as *intrusive* intelligibility prediction. On the contrary, *non-intrusive* intelligibility prediction, which *only* uses the degraded signal itself, has drawn increasing attention because of its application in realistic scenarios, where a reference signal or transcription can be difficult to access. A number of non-intrusive predictors heavily rely on environmental knowledge, such as room reverberant characteristics (Falk et al., 2010), therefore their application is limited. Another group of non-intrusive approaches essentially attempt to estimate reference signal features or transcriptions so that they can then follow the same procedures as the intrusive approaches (Andersen et al., 2017; Sørensen et al., 2017a). In addition, data-driven methods have been proposed, and they directly train predictors given a large amount of speech and its corresponding intelligibility data pairs (Andersen et al., 2018b; Zezario et al.,

2020). Consequently, the data quantity and quality will largely decide the performance of data-driven predictors.

In recent years, a number of works have take advantage of automatic speech recognition (ASR) models as intelligibility predictors, as ASR models could show similar patterns to humans in some speech recognition scenarios (Cooke, 2006; Schädler et al., 2015). Despite the potential of ASR-based intelligibility predictors, a limited number of works have demonstrated significant improvements over other existing intelligibility prediction approaches. This thesis follows the idea of using ASR models to model intelligibility, and explores novel approaches to take advantage of recent advanced ASR models to achieve more accurate speech intelligibility prediction, especially for the intelligibility of hearing impaired listeners.

1.5 Research questions

This thesis aims to provide insights from two aspects that are crucial to the development of hearing aids in terms of speech intelligibility enhancement, corresponding to the two aforementioned sections in the motivation: *data-driven hearing aid speech enhancement* and *speech intelligibility prediction*.

The research goal of data-driven hearing aid speech enhancement is to enable data-driven speech processing for hearing aids by introducing intelligibility objectives that model a wide range of hearing abilities. It is necessary to explore the optimised enhancement models for not only quiet speech but also speech in noisy environments. The research questions motivating this part of the work are:

- How well can data-driven optimised hearing aid fittings perform in terms of intelligibility improvement for speech in noisy and noise-free environments?
- Can the hearing aid fittings optimised for different noisy environments provide benefits over general fittings?
- How well can hearing aid speech enhancement models with a DNN-based denoising module perform in noisy environments?

The research goal of the chapters related to speech intelligibility prediction is to study how to take better advantage of ASR models for both intrusive and non-intrusive robust intelligibility prediction for a wide range of hearing abilities. As ASR models are optimised to understand speech, they are supposed to be able to extract important features for speech recognition. The efficacy of these ASR features for intelligibility prediction is worthy

of further investigation. In addition, current ASR-based intelligibility predictors rely on transcription, which could be difficult or expensive to achieve in some cases. Therefore, this research also aims to explore how to exploit ASR models for accurate non-intrusive intelligibility prediction. In conclusion, the following research questions are expected to be addressed:

- How well can the features extracted by ASR models perform in terms of robust intelligibility prediction?
- How can ASR models predict intelligibility non-intrusively, i.e., without using extra information like reference signals or transcription?

1.6 Contributions

The contributions from this work are listed as follows:

- In Chapter 3, a data-driven differentiable hearing aid speech processing framework is built to enable the automated optimisation of the enhancement models. The optimised hearing aid fittings could outperform well-recognised fittings for both quiet and noisy speech in terms of objective evaluation.
- In Chapter 4, a DNN-based hearing aid enhancement system is proposed for intelligibility enhancement in noisy environments. The system is proven well-performed by both objective and subjective evaluation.
- In Chapter 5, an ASR-based intelligibility predictor that takes advantage of hidden representations of the DNN-based ASR model is proposed. The proposed method is shown to outperform both widely used existing approaches and the transcription-based ASR predictor.
- In Chapter 6, the uncertainty of ASR models is proposed to use as an intelligibility predictor. The proposed approach does not need supervised optimisation with intelligibility labels, and can achieve performances approaching those of intrusive ASR predictors.

It is worth mentioning that the author has also helped the organisation of the Clarity project for machine learning challenges for hearing aid processing. A full list of publications from the author's PhD research is listed here:

1. Y. Pan, M. Bahman, **Z. Tu**, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification". In *Interspeech*, Shanghai, China, 2020.

2. **Z. Tu**, N. Ma, J. Barker, "DHASP: Differentiable hearing aid speech processing". In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 2021.

3. **Z. Tu**, N. Ma, J. Barker, "Optimising Hearing Aid Fittings for Speech in Noise with a Differentiable Hearing Loss Model". In *Interspeech*, Brno, Czechia, 2021.

4. **Z. Tu**, J. Zhang, N. Ma, J. Barker, "A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing". In Clarity workshop, 2021.

5. **Z. Tu**, J. Deadman, N. Ma, J. Barker, "Auditory-Based Data Augmentation for End-to-End Automatic Speech Recognition". In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022.

6. **Z. Tu**, N. Ma, J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners". In *Interspeech*, Incheon, Korea, 2022.

7. **Z. Tu**, N. Ma, J. Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction". In *Interspeech*, Incheon, Korea, 2022.

8. M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, **Z. Tu**, "The 2nd Clarity Enhancement Challenge for hearing aid speech intelligibility enhancement: Overview and Outcomes". *Submitted to* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.

1.7 Thesis Outline

This thesis is structured in the following way:

- Chapter 2 firstly reviews the progress in recent years on hearing aid speech enhancement. Then the review covers the development of both intrusive and non-intrusive

speech intelligibility prediction approaches, with a specific focus on those that can model different hearing abilities.

- Chapter 3 investigates the differentiable optimisation of hearing aid fittings with an intelligibility objective function for both quiet and noisy speech.
- Chapter 4 presents the proposed DNN-based hearing aid enhancement system for binaural intelligibility enhancement in noisy environments and studies its performance with both objective and subjective evaluation results.
- Chapter 5 presents the proposed ASR-based intelligibility predictor leveraging DNN hidden representations, and shows its advantages over a number of widely used metrics for both normal hearing and hearing impaired listeners.
- Chapter 6 presents the proposed non-intrusive uncertainty-based ASR intelligibility predictor, and studies its performances for both normal hearing and hearing impaired listeners.
- Chapter 7 concludes this thesis, provides answers to the research questions, discusses limitations of the work presents potential future directions.

Chapter 2

Background and Related Work

Hearing impairment, which can significantly decrease quality of life and lead to mental illness (Loughrey et al., 2018), has yet to be adequately resolved. Intelligibility improvement has been a major focus of hearing aid studies, as it can help hearing impaired listeners improve life quality by reducing the impediment to daily communication. For the purpose of developing better hearing aid algorithms, this thesis has been motivated to explore novel approaches for not only speech intelligibility enhancement, but also speech intelligibility prediction.

In a usual everyday environment, a speech signal can suffer from both external and internal degradation before being perceived by a hearing impaired listener, as shown in Figure 2.1. The internal degradation is caused by hearing impairment, and the external degradation can be caused by environmental noise, reverberation, electronic transmission, etc. As a consequence of the joint degradation, hearing impaired listeners fail to understand speech even with the usage of hearing aids in noisy environments. Being able to hear the speech but not able to understand it is a major complaint towards current hearing aids. Therefore, an ideal hearing aid is expected to tackle both external and internal degradation to improve the hearing impaired listeners' intelligibility.

In Section 2.1, this chapter first reviews the approaches to tackling internal degradation, i.e. hearing loss compensation algorithms. From early linear sound amplification algorithms to modern wide dynamic range compression and frequency lowering, hearing loss compensation research has been focusing on not only listening comfort but also intelligibility. After that, Section 2.2 covers techniques for tackling external degradation, i.e., speech denoising approaches. Beamformers and deep neural network (DNN) based speech denoising methods are the main focus, as they have shown the ability to effectively suppress external noises and have been top choices for modern assistive hearing devices.

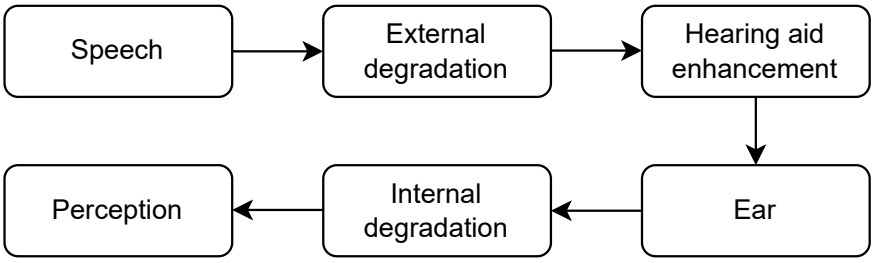


Fig. 2.1 Flow chart showing speech to the perception of a hearing impaired listener.

Intelligibility modelling is crucial for the development of hearing aid algorithms. When evaluating these algorithms, an accurate intelligibility predictor can save a large amount of time and expenses. In Section 2.3, intrusive intelligibility prediction methods are reviewed, including hearing impairment intelligibility prediction. These intrusive approaches make predictions for a degraded signal with additional information, which is usually the corresponding clean reference signal. Despite the success of intrusive approaches, the reference signals are usually not provided in realistic scenes. Therefore, non-intrusive intelligibility prediction approaches which require only the degraded signals are needed and covered in Section 2.4.

2.1 Hearing loss compensation

The research on fittings for hearing loss compensation has been an interest since over 80 years ago. The selected gain and frequency response was usually linear in the last century, that is, to provide constant gains varing with frequencies, but independent of sound levels. With the emergence of the wide dynamic range compression technology, more attention has been paid to nonlinear fittings, which accommodate the incoming sound energy levels at various frequencies. Additionally, frequency compression provides another perspective for hearing loss compensation by squeezing sounds in inaudible high frequencies into a smaller range of audible lower frequencies. In this section, the current major techniques for hearing loss compensation will be reviewed.

2.1.1 Linear amplification

The one-half gain rule for hearing aid fittings was first proposed by Lybarger (1963) and formed the basis for many prescribing formulas. It recommends providing the gain as half of the hearing thresholds, e.g., for a 50 dB hearing loss, a 25 dB gain needs to be provided. This rule can be easily computed and was found effective for the speech reception threshold except

for mild hearing impairments. The one-half gain rule was later examined and calibrated, e.g. by Berger et al. (1980) and Byrne and Tonisson (1976). The broad rationale of later derivation for fitting formulas is usually motivated by the idea of amplifying all frequency bands of speech to the most comfortable level or to be equally loud at a comfortable level.

Later, Byrne and Dillon (1986) revised the former procedure by Byrne and Murray (1986) and proposed the new National Acoustic Laboratories procedure (NAL-R) for gain and frequency response selection. The underlying motivation is to maximise speech intelligibility by maximising as much of a speech signal to be audible while still keeping the volume comfortable. It followed the idea of equalising loudness across all speech frequency bands to maximise the audibility and validated the procedure by Byrne and Murray (1986), which failed to do so. The rationale is that if one or a limited number of frequency bands dominates the overall loudness, the remaining frequency bands will probably be too soft or even unnoticeable and degrading speech intelligibility. Therefore, the NAL-R prescriptive fitting is based on the loudness equalisation, while also taking the one-half gain rule into consideration to determine the average gain. The NAL-R provided insertion gains at different frequencies which are computed as:

$$G_f = 0.05AHL + 0.31HL - B, \tag{2.1}$$

where HL is the hearing loss at the given frequency, AHL is the average hearing loss at [500, 1000, 2000] Hz, and the bias B is [-17, 8, -3, 1, 1, -1, -2. -2, -2] at [250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000] Hz, respectively. The biases are measured to achieve a flat audiogram, i.e., the relative gains vary across frequencies so that the corresponding loudness is equal. In the study by Byrne et al. (1990), NAL-R was found to need to be further calibrated for profound hearing impairment. The gain rule needs to follow the two-thirds rule when the high frequency hearing losses are profound, and NAL-RP was proposed based on this.

2.1.2 Nonlinear amplification

As shown in Figure 2.2, linear amplification limits its application as it can not handle speech at a large range of different levels. This is due to the due to the fact that the reduction of the dynamic range of hearing impairment is not linear, i.e., very loud speech is usually also uncomfortably loud for hearing impaired listeners, while normal conversational speech can be too soft for them to understand. To accommodate the need for adjusting insertion gain for different sound pressure levels, wide dynamic range compression has been widely applied in hearing aids since the 1990s.

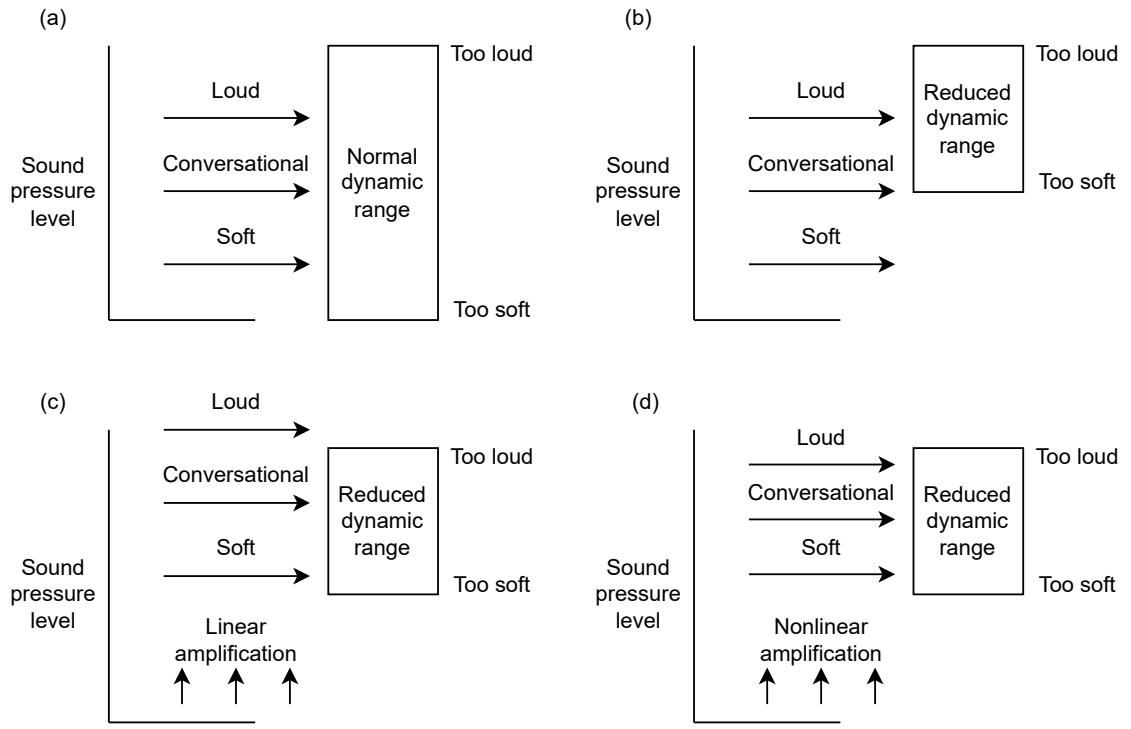


Fig. 2.2 Speech loudness perception of the normal and reduced dynamic range, adapted from (Kuk, 1996). The arrows represent the loudness range of input speech, and the box areas represent the loudness dynamic range of listeners. Sub-figure (a) and (b) show the normal dynamic range and the reduced dynamic range perceiving loud, conversational and soft speech, respectively. Sub-figure (c) presents the linear amplification effect for the reduced dynamic range. The loud speech can be uncomfortably too loud for impaired hearing, despite the soft speech can be perceived. Sub-figure (d) shows the nonlinear dynamic range compression amplification that provides insertion gain adaptively, so that speech at various sound pressure levels can be perceived comfortably.

NAL-NL1 (Byrne et al., 2001) was proposed to meet the need for wide dynamic range compression based on the similar rationale of NAL-RP. However, it does not strictly stick to the rule that loudness needs to be equal across all frequency bands, because this is likely to degrade the speech quality to an unacceptable level. Therefore, NAL-NL1 was derived following the principle that the speech should be amplified to a normal loudness or to a lower level while maximising intelligibility. The loudness normalisation is applied except for high loudness levels, for which the amplification targets are set to a lower level as it leads to higher intelligibility. NAL-NL2 (Keidser et al., 2011) made a further extension with more experimental data while keeping the same aim as NAL-NL1. More insertion gain is provided for low and high frequencies and less gain for middle frequencies, and hearing impaired listeners can gain better intelligibility without increasing loudness when using the NAL-NL2 presecription than the NAL-NL1.

Multiple additional nonlinear prescriptive fittings have been developed and widely used. For example, the CAMEQ, CAMEQ-2HF (Moore et al., 1999a, 2010) were motivated by the rationale that provides good audibility over a wide range of levels while still maintaining listening comfort. Their fitting procedures are based on loudness and quality judgments. The latest Desired Sensation Level method (Scollie et al., 2005) was derived to make sure the audibility of conversational speech is as much as possible when avoiding loudness discomfort. It also accommodates different requirements for speech in quiet and noisy environments. It is worth noting that there is a loudness difference among the NAL-NL2, DSL, and CAMEQ-2FH, that is, the overall gain of the DSL is the largest and the NAL-NL is the smallest among the three. Also, the NAL-NL2 also reduces high frequency gain when high hearing thresholds are present because of the evidence of less efficiency of using these ranges under such thresholds (Hogan and Turner, 1998).

For the nonlinear amplification by wide dynamic range compression, attack and release times are also important and were studied in (Alexander and Masterson, 2015; Gatehouse et al., 2006a,b). Attack time represents the time delay between the intensity of the input signal exceeding the threshold and the compressor compressing the intensity to the target value, while the release time is the opposite. Generally, a short attack time is required to prevent sudden intensity rise which could bring loudness discomfort, and the release time is always longer than the attack time. Meanwhile, wide dynamic range compression can sometimes decrease the intelligibility of speech signals, e.g., it can cause distortion of the signal envelopes and introduce modulation sidebands.

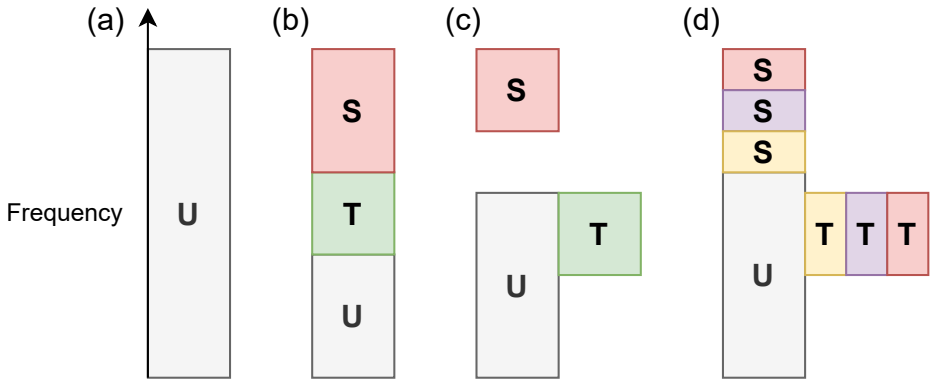


Fig. 2.3 Presentations of three frequency lowering techniques. Spectrum of an unprocessed signal (a) and that processed by three frequency lowering techniques: (b) frequency compression, (c) frequency transposition, (d) frequency composition. U, S, and T represent unprocessed, source, and target frequency bands, respectively.

2.1.3 Frequency lowering

Most hearing impaired listeners often suffer from high-frequency hearing loss due to aging and other effects. Frequency lowering has been a recent focus of hearing aids research. It is the technique of shifting a range of input frequencies to lower frequencies so that listeners with weak high frequency audibility could be provided with low frequency cues. Frequency transposition (Alexander et al., 2014; Parent et al., 1997), frequency composition (Kuriger and Lesimple, 2012; Salorio-Corbetto et al., 2017), and frequency compression (Ellis and Munro, 2015; Glista et al., 2009; Hopkins et al., 2014; Simpson et al., 2005) are three most widely utilised techniques concerned with frequency lowering.

Brief illustrations of common frequency lowering methodologies are shown in Figure 2.3, and the main differences among these techniques are how high frequency cues are dealt with. Frequency compression compresses high frequency bands of the source signals into lower bands while not overlapping with unprocessed original bands. For frequency transposition, source and target signals share the same bandwidth and the high frequency bands are directly transposed into lower bands. As for frequency composition, the source signal in the higher frequency domain is divided into subbands, which are compressed separately and aggregated into the lower frequency domain. A number of evaluations were done in the past decade and proved that these techniques can help hearing loss patients restore part of the information in quiet environments (Glista and Scollie, 2018). However, in noisy and complex scenes, these techniques failed to be functional. Meanwhile, frequency lowering techniques will also cause inherent distortions including reducing the spacing between harmonics, altering spectral peak levels, and modifying spectral shapes (McDermott, 2011).

2.2 Speech denoising for hearing aids

For listeners with normal hearing, the auditory system is capable of extracting target speech in noisy environments. This ability is usually severely decreased for hearing impaired listeners because of the raised hearing thresholds, the nonlinear auditory coding distortion, etc. Therefore, it is often insufficient for hearing aids to improve the intelligibility of speech in noise with only hearing loss compensation. Simply amplifying the incoming sounds to the auditory does not help much for target speech extraction. Therefore, a denoising module is desired for hearing aids to suppress unwanted sounds for the purpose of intelligibility improvement for hearing impaired listeners.

There are several types of interfering sounds that can heavily damage speech intelligibility, especially for hearing impaired listeners: (1) environmental noises that mask the crucial information of target speech for recognition, e.g, domestic noise such as washing machine, traffic noise, aircraft engine noise; (2) interfering voices with spectrum similar to that of speech, leading to increasing difficulty for target speech recognition; (3) substantial room reverberation produced by the reflection of hard surfaces such as walls, floors, and ceilings. It has been a long interest for hearing aids to take advantage of noise suppression techniques, including adaptive filtering (Vary and Martin, 2006) and spectral subtraction (Bentler and Chiou, 2006; Boll, 1979). More recently, beamforming, also referred to as a spatial filtering technique, stood out for noise suppression as it can take advantage of microphone arrays to extract sound from the target direction (Benesty et al., 2008). Additionally, with the rise of deep learning, DNN-based methods have shown great potentials for hearing aid speech denoising. The last two groups of noise suppression techniques are reviewed in the context of hearing aid processing.

2.2.1 Beamformers

Essentially, beamforming leverages the directional clues within the signals received by multiple microphones to enhance the sound from a target source and suppress sounds from other directions. Generally, the more microphones are used and the more widely spread the microphones are, the better beamforming performs. Meanwhile, hearing aids are usually fitted binaural and thus can be combined to form a beamforming array. Each device is supposed to output one processed signal for each ear. The binaural outputs depend on both how the directional cues are used and how the binaural cues are preserved. There have been a number of works proposing different beamformers for hearing aids, such as Best et al. (2017); Doclo et al. (2010, 2015); Moore et al. (2021). In this section, the classical

minimum variance distortionless response (MVDR) beamformer (Capon, 1969) is reviewed as an example.

A multi-channel noisy speech $y \in \mathbb{R}^{M \times t}$ consisting of a target signal x and interfering noise n can be transformed into the time-frequency domain and expressed as:

$$\mathbf{Y}(f,t) = \mathbf{d}(f)\mathbf{S}(f,t) + \mathbf{N}(f,t), \tag{2.2}$$

where $\mathbf{Y}, \mathbf{S}, \mathbf{N} \in \mathbb{R}^{M \times F \times T}$ are the time-frequency representations of the noisy speech, target speech and noise, respectively. M, F, T represent the number of microphones, frequency bins, and time indices. Meanwhile, $\mathbf{d}$ is the steering vector representing the direct path impulse responses between the target speech and microphones.

A beamformer essentially targets obtaining the weights $\mathbf{w}$ to sum the signals from microphones $\mathbf{Y} \in \mathbb{R}^{M \times F \times T}$ into a predicted target signal $\hat{\mathbf{S}} \in \mathbb{R}^{F \times T}$, which can be computed as:

$$\hat{\mathbf{S}} = \mathbf{w}^{\mathrm{H}}\mathbf{Y}, \tag{2.3}$$

where $\mathbf{w} \in \mathbb{R}^{M \times F}$ and $(.)^{\mathrm{H}}$ is the Hermitian Conjugate of a complex matrix. When applying beamforming to an utterance, the beamformer is regarded as a time-invariant beamformer if the weights stay unchanged for all frames. Otherwise, it is regarded as time-variant beamformer and can be adaptive to the change of an utterance.

The MVDR beamformer aims at minimising the power of the output signal while keeping the target speech distortionless, and the weights are formed as:

$$\mathbf{w}_{\mathrm{MVDR}}(f) = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}(f)\Phi_{\mathbf{YY}}(f)\mathbf{w}(f), \text{ s.t. } \mathbf{w}(f)^{\mathrm{H}}\mathbf{d}(f) = 1, \tag{2.4}$$

where $\Phi_{\mathbf{YY}}(f)$ is the covariance matrix of the noisy time-frequency representation $\mathbf{Y}$ at frequency f. One solution to the MVDR beamformer can be based on reference channel selection:

$$\mathbf{w}_{\mathrm{MVDR}}(f) = \frac{\Phi_{\mathbf{NN}}^{-\mathbf{1}}(f)\Phi_{\mathbf{SS}}(f)}{\mathrm{Trace}\left(\Phi_{\mathbf{NN}}^{-\mathbf{1}}(f)\Phi_{\mathbf{SS}}(f)\right)}\mathbf{u}, \tag{2.5}$$

where $\Phi_{\mathbf{NN}}$ and $\Phi_{\mathbf{SS}}$ are the covariance matrices of the noise and target speech, respectively. Also, $\mathbf{u}$ is the reference channel one-hot vector, i.e., the index of the reference microphone is 1 and the rest are 0s.

The covariance matrix $\Phi_{\mathbf{SS}}$ can be achieved with the estimated time-frequency mask $\mathbf{M}_s(t,f)$ of the target speech:

$$\Phi_{\mathbf{SS}}(f) = \frac{\sum_{t=1}^{T} \hat{\mathbf{S}}(t,f)\hat{\mathbf{S}}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathbf{M}_s(t,f)\mathbf{M}_s(t,f)}, \tag{2.6}$$

where the estimated $\hat{\mathbf{S}}$ is computed as:

$$\hat{\mathbf{S}}(t,f) = \mathbf{M}_s(t,f)\mathbf{Y}(t,f). \tag{2.7}$$

Similarly, the covariance matrix $\Phi_{\mathbf{SS}}$ can also be estimated with the noise mask $\mathbf{M}_n(t,f)$. This solution converts the problem of estimating the weight $\mathbf{w}$ into estimating the time-frequency masks of the time-frequency representations of target speech and noise. The estimation of the masks can be accomplished with statistic approaches, such as with complex Gaussian mixture model (Higuchi et al., 2017) and with DNNs (Erdogan et al., 2016). There are multiple additional beamformers based on different intuitions and different solutions. The beamformers are preferable for hearing aid applications as they are relatively computationally cheaper and introduce minimum distortions. However, the linear combination of signals from multiple microphones may not be able to suppress noises effectively, especially when the target speech is in close proximity to the noise sources.

2.2.2 DNN-based approaches

DNNs have been explored for noise suppression with data-driven optimisation since they show the powerful modelling ability on large databases, such as the methods proposed in Lu et al. (2013); Xu et al. (2013, 2014). In general, DNNs turn speech denoising into regression tasks, and they are optimised to map noisy speech to clean speech via back-propagation. It is believed that the design and choice of DNN architectures, optimisation objective functions, and speech representations can all be important to the performance of DNN-based approaches.

Early approaches focus more on single-channel noise suppression. Many of them train the networks with reconstruction objectives to predict the desired clean time-frequency representations given those of noisy speech. For example, Lu et al. (2013) leveraged an auto-encoder to predict the clean mel frequency power spectrum, which can be further converted to the waveform with the phase information from the noisy speech. Xu et al. (2014) proposed to train a basic DNN to restore the clean speech with the representation of log power spectral features, through which the waveform speech can be synthesised by inverse short-term Fourier transform. More recently, Défossez et al. (2020) proposed to

use a U-Net for waveform speech denoising with an optimisation objective operated in the time domain. Apart from that, a growing body of work has been utilising Generative Adversarial Nets (GANs) (Goodfellow et al., 2014), which are known to be able to improve the fidelity and perception of network outputs. Pascual et al. (2017) firstly proposed to take advantage of a GAN for waveform speech denoising, and combined the reconstruction with the GAN loss as the optimisation objective. Similarly, Michelsanti and Tan (2017); Soni et al. (2018) trained GANs for noise suppression but with time-frequency representations. Also, a number of works proposed to use speech evaluation metrics as the optimisation objectives for denoising DNNs. Martin-Donas et al. (2018); Zhao et al. (2018) implemented differentiable approximations to these measures so that they can propagate the prediction errors to optimise the networks. Zhang et al. (2018) used another DNN to approximate these evaluation measures and leveraged this network as the optimisation objective.

Speech separation can be regarded as a special task of speech denoising. For the separation task, more than one target speech signal is expected to be extracted, that is, the input signals contain overlapping speech from multiple sources and the separation model is supposed to generate each clean speech signal separately. One of the common challenges in training speaker-independent multi-talker speech separation models is the label permutation problem: it is difficult to match the predicted signal and the ground truth target signals to compute the correct loss for optimisation. Early DNN-based speech separation approach (Hershey et al., 2016) avoids this problem by leveraging clustering algorithms to minimise the distances among the time-frequency bins from the same source while to maximise the distances of those from different sources. The permutation invariant training technique (Yu et al., 2017) was proposed to tackle this problem by regarding the pair with minimum error among all potential prediction and label pairs and widely used in the later DNN-based speech separation methods. Similar to the speech denoising approaches aforementioned, the choices of network architecture, and optimisation objectives of speech separation models have been constantly studied and investigated. The time-domain speech separation approaches, which directly are usually optimised with SNR-based objectives and directly process waveform speech (Luo and Mesgarani, 2018, 2019), stood out in recent studies. There are also a number of works focusing only on the speech by one target talker, such as Wang et al. (2019); Žmolíková et al. (2019). These approaches generally leverage the speaker embedding from the target talker to extract the target speech.

Previously introduced DNN-based speech denoising approaches focus on single-channel speech, while there usually are microphone arrays deployed in modern hearing aids. Therefore, DNN-based multi-channel speech denoising approaches that take better advantage of spatial information are of more interest for the purpose of hearing aids. A number of

DNN-based multi-channel speech denoising approaches leverage DNNs to predict the time-frequency masks of target speech and apply the masks to beamformers as introduced in the previous section (Erdogan et al., 2016; Heymann et al., 2016). Essentially, the DNN-based beamformers still apply a linear combination to multi-channel signals, thus the performances are limited to the nature of beamformers. Another group of approaches extend single-channel models by combining the spectral and spatial information, such as Chakrabarty et al. (2018); Wang et al. (2018). Similarly, time-domain DNN-based approaches have also achieved success in multi-channel speech separation by directly encoding all channels into the networks (Gu et al., 2019; Zhang et al., 2020).

Despite the success of DNN-based speech denoising, they have been rarely applied in modern hearing aids. One of the underlying reasons is that DNNs often come with high latency, which is not acceptable for hearing aid applications. A number of studies conducted by Stone and Moore (1999, 2002, 2003, 2005) suggest that the disturbance increase as the latency increases. A latency as low as 20 to 30 ms can be disturbing for listeners with mild to moderate hearing loss. Meanwhile, Dillon et al. (2003) found that 10 ms latency can degrade sound quality on commercial hearing aids. A more recent study by Stone et al. (2008) suggests that the latency may need be as low as 5 to 6 ms for open-canal hearing aids, which have been progressively more popular in the market. As DNNs are generally heavily parameterised, they are computationally expensive and more difficult to satisfy the low latency requirement than beamformers. It is relatively easier for time-domain DNN-based approaches than frequency-domain approaches because the window size of the Fourier transformation is conventionally quite long. Despite that, more attention has been paid to reducing the ideal latency of DNN-based speech denoising (Wang et al., 2022).

2.3 Intrusive speech intelligibility prediction

As mentioned in Chapter 1, it is common to leverage additional information apart from the degraded speech signal itself for intelligibility prediction, and these approaches are regarded *intrusive*. A majority of intrusive approaches predict intelligibility by comparing some (psycho)acoustic representations of the reference and degraded speech signal. The additional information used in these approaches is usually the reference speech or the additive noise. Another group of intrusive approaches make intelligibility prediction with ASR models and compares the ASR outcomes with the speech transcriptions, which are the additional information. In this section, both the development of (psycho)acoustic representation-based and ASR-based intelligibility prediction approaches will be introduced.

Although a large amount of work has been done for speech intelligibility prediction, they usually have an implicit assumption that the listener's hearing ability is not degraded. Therefore, these approaches can be difficult for the purpose of evaluating hearing aid enhancement, i.e. the listener's hearing ability is imperfect. Fortunately, attention has also been drawn to the field of intelligibility prediction by hearing impaired observers. These approaches usually take advantage of existing knowledge of the human auditory periphery to build impaired hearing models and use these models to simulate the acoustic representations of the signal that hearing impaired listeners perceive and use these to perform intelligibility prediction. In this section, a number of approaches for hearing impairment intelligibility prediction will also be reviewed.

2.3.1 Acoustic representation-based approaches

The articulation index (AI) is one of the first objective speech intelligibility measures proposed by the Bell Telephone Laboratories (French and Steinberg, 1947). It was then improved and elaborated by Kryter (1962) and widely used for the evaluation of speech communication systems. The AI is computed based on the SNRs within multiple frequency bands. These SNRs are limited and subject to an auditory masking effect and then combined with corresponding perceptually motivated weight coefficients (Kryter, 1962). The AI was later extended and standardised to the speech intelligibility index (SII) (ANSI, 1997). These approaches target calculating the available average amount of speech information, and they use the long-term averaged speech spectrum as inputs, therefore, they can only perform well for simple linear degradation, e.g., stationary additive noise. Rhebergen and Versfeld (2005) made a further extension by computing the SII for each small time frame within a speech signal to take the modulation domain into consideration, and combined all SII values as the predicted intelligibility. This approach could perform well for speech in more fluctuating (i.e., non-stationary) noises in terms of the speech recognition thresholds (SRTs). Furthermore, Kates and Arehart (2005) proposed coherence speech intelligibility (CSII), which replaces the SNR in the SII with the signal-to-distortion ratio (SDR) in each frequency band. The SDRs are computed with coherence function (Carter et al., 1973), which is a measure of correlation in the frequency domain. The CSII can be applied to speech nonlinearly degraded by peak-clipping and centre-clipping, while it considers the analysis of only wideband signals, not the narrowband sub-components.

Since the AI was developed for simple linear degradation, Steeneken and Houtgast (1980) proposed the speech transmission index (STI) to predict the intelligibility of speech degraded by reverberation and some nonlinear degradation like clipping. For this purpose, a noise signal with a speech-shaped long-term averaged spectrum is modulated at a range

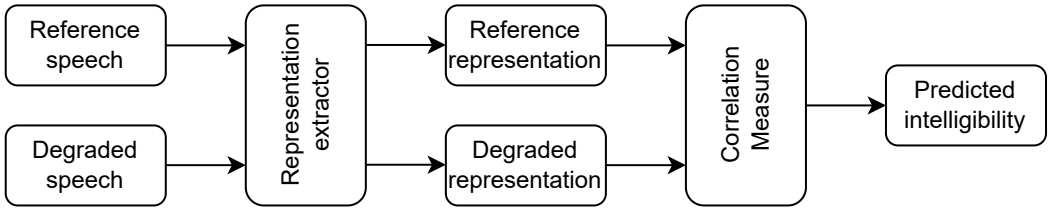


Fig. 2.4 General approach of intrusive intelligibility prediction.

of frequencies and then processed by the transmission system. Each modulated signal and its corresponding output response are used to compute an apparent SNR, and all the SNRs are combined with a group of psycho-acoustically derived weighting to achieve the overall STI value. After the proposition of the STI, researchers developed variations to use speech as the probe signals for more sophisticated degradation, such as the works of Hohmann and Kollmeier (1995); Payton and Braida (1999); Payton et al. (2002), etc. However, these speech-based STI methods were proved to be not able to adequately predict the intelligibility of nonlinearly processed speech (Hohmann and Kollmeier, 1995; Payton et al., 2002; van Buuren et al., 1999). Later, a normalised covariance measure (NCM) based STI variant showed its potential for nonlinear operations, including envelope thresholding and spectral subtraction (Goldsworthy and Greenberg, 2004). To compute the NCM-based STI, the temporal envelope normalised covariance, i.e., normalised cross correlation (NCC), between the reference and degraded speech at different frequencies are extracted. These covariances are then converted to the apparent SNR for the overall predicted intelligibility. Apart from that, STI and its many variants require the assumption that the degradation is stationary, therefore they fail to predict the intelligibility of speech with non-stationary distortions. The quasi-stationary STI is then proposed (Schwerin and Paliwal, 2014) to process the modulation envelope in short-time segments, and only requires the assumption of quasi-stationarity so that it can tackle non-stationary degradation.

The aforementioned approaches briefly reviewed the early progression of speech intelligibility prediction. Until then, the general approach was to make the correlation-based comparison between some extracted acoustic representations of the reference and degraded speech, as shown in Figure 2.4. However, these approaches are still less appropriate for degraded speech processed by some speech enhancement models, which can result in nonlinear and non-stationary distortions.

In recent years, more attention has been attracted to finding more appropriate acoustic representations for intelligibility prediction. The length of the analysis segment window can be one crucial factor when extracting representations. Some approaches estimate correlation values based on the complete speech signal, which tends to be several seconds, e.g., Goldswor-

thy and Greenberg (2004); Kates and Arehart (2005). Such a long analysis window leads to a restriction that the degradation needs to be stationary, and a few high amplitude regions could dominate the overall estimation. On the contrary, there are also some approaches using a very short analysis window less than 50 ms, e.g., Christiansen et al. (2010); Rhebergen and Versfeld (2005). This short window leads to a poor modulation frequency resolution and excludes low temporal modulations which contribute heavily to intelligibility. The results from Drullman et al. (1994); van den Brink (1964) showed that the analysis window length around 333-500 ms could be more suitable for intelligibility prediction. Motivated by this, Taal et al. (2011) proposed a short-time objective intelligibility measure (STOI), which has been arguably the most widely used intelligibility prediction approach. To compute the STOI value, the reference and degraded speech are firstly decomposed into time-frequency bins with a group of one-third octave bands. Then the short-time temporal envelope segments of the degraded speech are normalised and clipped. Finally, the degraded segments are compared with those of the reference speech to achieve the overall predicted intelligibility. The STOI showed a high correlation with the intelligibility of listening results for speech processed by some enhancement model, including the ideal time frequency segregation (Brungart et al., 2006) and two single-channel noise-reduction algorithms (Ephraim and Malah, 1984; Erkelens et al., 2007). Furthermore, Jensen and Taal (2016) proposed the extended-STOI (ESTOI) to improve the performance on speech with temporally modulated noise maskers.

Apart from the correlation-based approaches, there are several different groups of approaches for intelligibility prediction, and their acoustic representations are carefully designed. The speech-based envelope power-spectrum model (sEPSM) assumes that intelligibility can be predicted by the SNRs in the envelope domain (Jørgensen and Dau, 2011). It takes advantage of a group of gammatone filters (Moore and Glasberg, 1983), which can describe the shape of the impulse response on the basilar membrane, for envelope extraction. The sEPSM was further extended using a dynamic compressive gammachirp filterbank (Irino and Patterson, 2006) by Yamamoto et al. (2019). Another group of approaches estimate intelligibility by measuring the mutual information between the acoustic representations of reference and degraded speech. The K-nearest neighbour mutual information intelligibility measure (Taghia and Martin, 2013) uses the same acoustic representation as STOI. The speech intelligibility in bits (Van Kuyk et al., 2017), which estimates the amount of information shared between a talker and a listener in bits per second, uses an auditory model to extract representations. Moreover, the glimpsing model proposed by Cooke (2006) has also been investigated for intelligibility prediction. It leverages the "glimpse", which is defined as the proportion of time frequency regions where the SNR is higher than a predefined threshold.

The recent variant of the glimpsing model (Edraki et al., 2021) decomposes speech into spectro-temporal modulation subspace with the Gabor filterbank (Schädler et al., 2012).

2.3.2 ASR-based approaches

The acoustic representation-based intelligibility prediction approaches always make strong assumptions that the correlation between the representations of reference and degraded speech, or the SNR at (modulation) frequency bands is closely related to intelligibility. However, these assumptions can fail due to, e.g., different types of background noise, and processing of enhancement models. Employing ASR models for intelligibility modelling has the potential to overcome this, as it requires only a minimum set of assumptions, that is, how much a speech signal can be understood by an ASR model is related to the intelligibility achievable by humans. In addition, the acoustic representation-based approaches usually require the reference speech or the noise, while ASR-based approaches usually need only the transcription of the degraded speech. Therefore, there has been a growing interest in researchers using ASR models to predict intelligibility.

The aforementioned glimpsing model (Cooke, 2006) is one of the earliest works to show that the similarity of speech in noise recognition patterns between ASR models and humans. The quantitative results show that the consonant recognition identification performance of a missing data ASR model with the glimpses as input is similar to that of humans for speech in speech-shaped noise modulated with the envelope of multiple talker babble (Cooke, 2006). Later, Barker and Cooke (2007) also showed that the ASR can also approximate human recognition performance in terms of speaker intelligibility. However, the glimpse approaches can usually predict intelligibility for only speech with additive noise.

Later, Schädler et al. (2015) found that the SRTs of human listeners for speech in multiple noise conditions can be well predicted by ASR models. They trained and tested an ASR model with the German matrix test material, and the experimental results show that the SRTs of the ASR model for speech in multiple types of stationary noises and a fluctuating noise are much more correlated than that of the SII to the human SRTs. Furthermore, Spille et al. (2018b) found that the gap between ASR and human recognition can be further closed by a DNN-based ASR model, as DNNs have brought substantial progress to ASR in recent years.

The ASR-based intelligibility prediction has attracted increasing interest recently. More works have been conducted for hearing impairment intelligibility prediction, and non-intrusive intelligibility prediction and they will be introduced in the following sections.

2.3.3 Hearing impairment intelligibility prediction

The aforementioned intelligibility prediction approaches share an assumption that a listener's hearing ability is not degraded. Consequently, they can make poor intelligibility predictions when a listener is hearing impaired, as the internal degradation to the speech is not taken into their consideration. For the purpose of accurately predicting the intelligibility of hearing impaired listeners, a hearing loss model that models the listener's hearing ability needs to be included in the intelligibility predictor. Furthermore, binaural cues, which are usually used to localise the target speech source, are crucial for listeners to recognise speech in spatially separate noise, and studies also found that human is better than ASR models at taking advantage of binaural cues (Spille et al., 2018b). Therefore, the binaural cues can be also an important factor to predict how much a listener understands speech. In this section, a number of intelligibility predictors incorporating hearing loss models, with also their application to binaural signals, will be reviewed.

The hearing aid speech perception index version 1 (HASPIv1) (Kates and Arehart, 2014a) is one of the most widely used intelligibility predictors for hearing impaired listeners and incorporates a comprehensive auditory model (Kates, 2013). The auditory model simulates a number of hearing perception degradation due to hearing impairment. Kates and Arehart (2021) later revised the HASPI and proposed HASPI version 2, which will be reviewed in this section. In the following part of this thesis, the acronym 'HASPI' will be used to represent HASPI version 2 for convenience. The diagram of the auditory model within the HASPI is shown in Figure 2.5. An input speech is firstly resampled to 24 kHz, then processed by a middle ear model which consists of a two-pole infinite impulse response (IIR) high-pass filter with a cut-off frequency at 350 Hz and a one-pole IIR low-pass filter with a cut-off frequency at 5 kHz. A group of 32 fourth-order gammatone IIR filters (Cooke, 1993; Patterson et al., 1995) covering the range from 80 to 8000 Hz are used for auditory analysis. The bandwidths of the gammatone filterbank are broadened in responses to not only hearing losses (Moore et al., 1999b) but also signal intensity (Baker and Rosen, 2002), i.e. the root mean square (RMS) average. The dynamic-range compression due to OHC damage is modelled by the control filterbank (Ruggero et al., 1997). The control filterbank also consists of 32 gammatone IIR filters covering from 80 to 8 kHz, while the bandwidths are set to maximum, i.e., corresponding to the maximum hearing loss (Zhang et al., 2001). In addition, the OHC compression includes an 800-Hz low pass filter to provide a small time delay. The OHC compressed envelope is then multiplied by the envelope retrieved from the analysis filterbank. After that, the envelope is processed by simulated IHC compression. At last, the IHC compressed envelope is introduced to a 2 ms rapid adaptation. The envelopes of the reference and the degraded speech extracted from the HASPI auditory model are

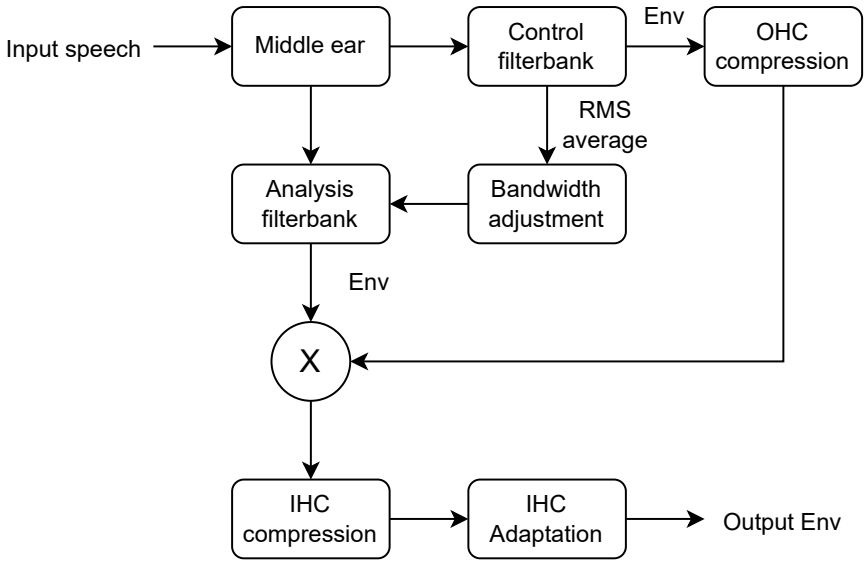


Fig. 2.5 HASPI auditory model for envelope extraction.

used for further modulation analysis. A set of five basis functions are used to acquire the short-time mel-frequency cepstral coefficients (Mitra et al., 2012), which are then filtered by ten modulation filters. The NCC between the reference and degraded filtered sequences are computed and averaged across the five basis functions to produce ten averaged modulation values. Eventually, these ten modulation outputs are mapped to the intelligibility score using an ensemble of ten neural networks. To accommodate the need of predicting the intelligibility of binaural signals, a better-ear strategy can be applied, i.e. regarding the larger score between the predicted left ear and the right ear ones as the eventual intelligibility score.

Apart from the HASPI, the Cambridge MSBG hearing loss model (Baer and Moore, 1993, 1994; Moore and Glasberg, 1993; Stone and Moore, 1999), a well-recognised hearing loss simulator, can be used as the front-end for hearing impairment intelligibility prediction. The MSBG model takes an input speech and simulates how a hearing impaired listener *perceives* the speech given the audiogram. The simulated signal can then be regarded as the degraded speech and used for intelligibility prediction with the normal approaches introduced in the previous sections. The diagram of the MSBG hearing loss model is shown in Figure 2.6. Given an input speech, a source to cochlea transformation filter is applied to simulate the acoustic changes during sound propagation from the free field to the eardrum (Shaw, 1974). The spectral smearing is then applied to simulate the reduced frequency selectivity caused by hearing impairment. This is done by applying a bandwidth broadened auditory filterbank (Moore and Glasberg, 1983) in the frequency domain. After that, the loudness recruitment simulates the phenomenon that the response to the speech of an impaired auditory

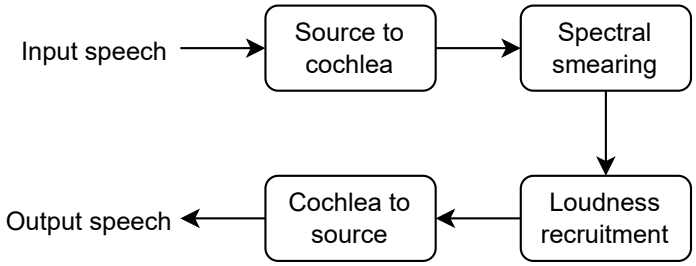


Fig. 2.6 Diagram of the MSBG hearing loss model.

is smaller than a normal one, while the responses to high-level sounds are similar. A group of gammatone filters retrieve the envelopes at different frequency bands, and then the envelopes are nonlinearly compressed according to the hearing abilities. The compressed envelopes are then used as the gain to adjust the amplitudes of the speech at different frequency bands. At the last, a cochlea to source transformation filter is applied to the recruited speech. For the purpose of intelligibility prediction, an additional predictor is desired to process the simulated hearing impaired degraded speech. The modified binaural STOI (MBSTOI) (Andersen et al., 2018a) is an improved version of the deterministic binaural STOI (Andersen et al., 2016) and can take advantage of binaural cues through an equalization-cancellation stage (Durlach, 1972) to predict binaural intelligibility. The combination of the MSBG hearing loss model and MBSTOI can therefore predict binaural intelligibility for hearing impaired listeners, e.g. Graetzer et al. (2021).

Additionally, the ASR-based intelligibility prediction approaches have been also used for hearing impairment intelligibility prediction. Kollmeier et al. (2016) extended the method proposed in Schädler et al. (2015) by introducing a hearing impairment effect simulation front-end to the ASR model. The front-end consists of a simulation of the elevated hearing threshold by setting the speech level to the audiogram level if it is below this value, and an additional suprathreshold distortion by adding Gaussian white noise with an individualised fitted standard deviation. The results of the German Matrix test in stationary and fluctuating noises show that the predicted SRTs of hearing impaired listeners by the ASR are correlated well with hearing impaired listeners. Similarly, Fontan et al. (2017) proposed to take advantage of a simplified MSBG model as the ASR front-end. The proposed system was evaluated with simulated hearing impaired listeners (i.e. normal hearing listeners listening to speech with simulated hearing impairment degradation) on a small vocabulary of French speech material. Experimental results show that though the ASR recognition results are generally worse than humans, they are correlated to human performance well.

2.4 Non-intrusive speech intelligibility prediction

Non-intrusive speech intelligibility prediction is always desired for its application in realistic scenarios, where a clean reference signal is difficult to access. As intrusive approaches usually require strictly aligned and clean reference signals, they are usually used to evaluate simulated scenarios. On the contrary, a degraded signal can be submitted alone to a reference-free non-intrusive intelligibility predictor. Therefore, the non-intrusive approaches are more likely to be used for real-life applications, including hearing aids, voice calls or cochlea implants.

Early non-intrusive approaches rely on the prior knowledge of acoustic features that are correlated with intelligibility. The application of these approaches can be limited, e.g., the speech to reverberation modulation energy ratio (SRMR) (Falk et al., 2010) and average modulation-spectrum area (ModA) (Chen et al., 2013) target only reverberant or dereverberated speech. Another group of approaches can be regarded as variants of intrusive approaches, especially STOI. Specifically, these approaches estimate the corresponding reference signal or features from a degraded speech signal and use the estimated reference signal or features to predict intelligibility in an intrusive way. In addition, data-driven non-intrusive intelligibility prediction has been drawing increasing attention in recent years. These approaches take advantage of machine learning models to learn the mapping from degraded signals to intelligibility labels. Lastly, ASR models have been also explored for non-intrusive intelligibility prediction. In this section, the four categories of the aforementioned approaches will be reviewed.

2.4.1 Acoustic representation-based approaches

Non-intrusive intelligibility prediction approaches based on acoustic prior knowledge are usually motivated by the observed correlation between acoustic features and intelligibility. SRMR (Falk et al., 2010) is a classic approach motivated by the finding that the ratio of low to high modulation frequency energy is related to reverberant or dereverberated speech intelligibility. In detail, a group of gammatone filters are used to process a degraded speech to obtain the temporal envelopes. The envelopes are then processed by 7 or 8 modulation filters, which are chosen on a per-signal basis. The SRMR is given by the summation of the first 4 modulation energy divided by the summation of the other modulation energy. The ModA (Chen et al., 2013) follows a similar idea by computing the area of the modulation spectrum and shows an advantage in predicting the intelligibility of reverberant speech perceived by cochlea implant listeners.

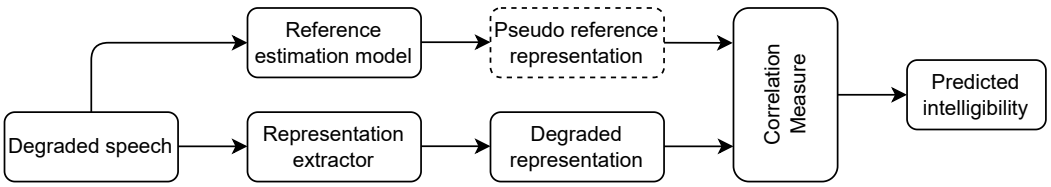


Fig. 2.7 Non-intrusive intelligibility prediction with estimated reference.

The across-band envelope correlation metric (ABECm) is also proposed for non-intrusive intelligibility prediction (Chen, 2016b). It is inspired by the phenomenon that the across-band envelope correlation carries important information for human speech perception. Therefore, the average correlation of adjacent bands is used as an intelligibility predictor. Additionally, the reduced speech dynamic range measure (rDRm) is found to be correlated with speech intelligibility (Chen, 2016a). Specifically, a degraded speech signal is divided into short non-overlapping segments, whose RMS levels are divided by the overall RMS level and regarded as relative-RMS-levels. The speech dynamic range is computed as the difference between the maximum and minimum relative-RMS-levels. Both ABECm and rDRm are experimented with sentences in multiple SNR levels with various noise maskers and show competitive results for intelligibility prediction.

2.4.2 Pseudo reference-based approaches

When the corresponding reference clean speech is not available for intrusive intelligibility prediction, an intuitive idea is to create a *pseudo* reference. With the pseudo reference, it is possible to make a non-intrusive prediction using intrusive approaches. As described in Figure 2.7, the approaches construct a reference estimation model to generate pseudo reference representation, which is then used for correlation measures like intrusive approaches. In some special cases, the pseudo reference speech signal itself is estimated for intelligibility prediction(Sørensen et al., 2016, 2017b).

The non-intrusive STOI (NI-STOI) proposed by Andersen et al. (2017) is one of the early works to construct a model for reference representation estimation. Specifically, it introduces a model to learn the principle components of modulations of a very long clean speech signal, and then project degraded speech modulation onto the learnt principle components to estimate the pseudo reference. The degraded short-time segmentation and estimated reference segmentation are used for correlation measures following the STOI approach to make intelligibility predictions. The experimental results show that NI-STOI can outperform SRMR for simulated speech in noises, but perform significantly worse than STOI, especially when speech is degraded by cafe noises. Similarly, Sørensen et al. (2017a) proposed the

non-intrusive codebook-based STOI (NIC-STOI) to take advantage of the codebook-based approach (Kavalekalam et al., 2016) to estimate clean envelope spectrum, and NIC-STOI shows close correlation with STOI. In addition, Karbasi et al. (2016) proposed to use a twin hidden Markov model to synthesise 1/3 octave band representation of pseudo reference, and the experiment shows very competitive results to STOI.

There are two major limitations to the pseudo reference based non-intrusive approaches. One is difficulty in accurate estimation, i.e., when the reference estimation model fails to make a reasonable estimation, the performance of the predictor can suffer from a significant drop. The other limitation is that its performance depends on the employed intrusive approach. If the intrusive predictor, i.e. usually STOI, does not perform well in some scenarios, it is still not possible to make an accurate prediction even if the pseudo reference can be estimated perfectly.

2.4.3 Data-driven approaches

Data-driven approaches have been increasingly popular for non-intrusive intelligibility prediction, thanks to the rapid development of large-scale machine learning techniques in speech processing. A common framework of these approaches is to learn a mapping from a degraded speech signal or its representations to the corresponding intelligibility score, which can be human listening results or scores from intrusive predictors like STOI. The mapping models are optimised with a large amount of training data, i.e. degraded speech and intelligibility pairs, and can be generalised to different evaluation scenarios.

The approach proposed by Sharma et al. (2010) is one of the early works of data-driven non-intrusive intelligibility prediction and uses a Gaussian mixture model to learn the mapping from degraded speech features to SII scores. Later, Sharma et al. (2016) proposed to employ a classification and regression tree to learn the mapping from short-time features of degraded speech to STOI scores. The experimental result shows the predicted STOI can be correlated well with ground truth STOI across a large range of SNRs for simulated noisy speech. In recent years, there has been a growing interest to take advantage of DNN models for intelligibility prediction. Andersen et al. (2018b) proposed to take advantage of a convolutional neural network, which takes short-time representations used in STOI as inputs and makes predictions on human recognition results. The performance was shown to be slightly better than STOI and approaching ESTOI. Similarly, Zezario et al. (2020) proposed the STOI-Net, which consists of a convolutional neural network and a long short-term memory with an attention mechanism, to predict STOI scores. Later, Zezario et al. (2022) proposed the MBI-Net to predict speech intelligibility from multi-channel input signals.

Data-driven approaches heavily rely on the quality of the training data used to optimise the models. If intrusive intelligibility scores estimated by methods like STOI are used for training objective, the performance of the models are then capped by the intrusive approach itself. Otherwise, a large number of human recognition results are required to optimise the model, which can be very expensive and time-consuming.

2.4.4 ASR-based approaches

Most ASR-based intelligibility prediction requires transcripts to measure the recognition results of ASR models, which are then used to correlate with those of human. Therefore, these approaches are usually considered intrusive. The approaches proposed in Holube and Kollmeier (1996) and Jürgens and Brand (2009) overcome this by looking at the dynamic time warping (DTW) ASR. DTW ASR models make predictions based on the measured distances between test degraded words and a number of pre-prepared template word recordings. The experiments also show that the predictions are more accurate if the template and test recordings are identical (Jürgens and Brand, 2009), which is similar to intrusive predictions.

Meanwhile, there are a number of works that leverage ASR-derived measures to enable non-intrusive prediction. Martinez et al. (2022) proposed to leverage the mean temporal distance to capture the temporal smearing effect (Hermansky et al., 2013) in the phoneme posteriogram generated by an ASR model. The mean temporal distances are then mapped to word error rate (WER) so that the transcripts are not needed during the evaluation. The estimated WER are then used to measure the SRTs for German Matrix test material. The experimental results show that the predicted SRTs are well correlated with human listeners. However, the generalisation ability requires more evaluation, as the training and test noises are similar in this work. Later, Roßbach et al. (2022) proposed to use a similar model to make intelligibility predictions for hearing impaired listeners at the utterance level. Surprisingly, even though the model is trained with a noisy German speech database, it can perform quite well for English speech in the evaluation. Additionally, Karbasi et al. (2022) investigated microscopic, i.e., word level, non-intrusive prediction with ASR models. In detail, a number of ASR-derived measures, including dispersion, entropy, log-likelihood ratio, etc., are used to map to the recognition correctness of each word within a matrix speech corpus. This method requires a number of human intelligibility labels to optimise the mapping model, which is a simple feed-forward network.

Chapter 3

Optimising Hearing Aid Fitting with the DHASP framework

3.1 Introduction

An appropriate amplification fitting tuned for the listener's hearing disability is critical for the good performance of hearing aids, which are expected to improve audibility and hopefully intelligibility. Typically, the amplification of the hearing aid at various frequencies closely matches the listener's audiogram, which is measured using pure tone, with a standardised mapping. Early hearing aid fitting prescriptions, including the National Acoustic Laboratories Revised (NAL-R) formula (Byrne and Dillon, 1986), aim to maximise speech intelligibility for a specified loudness level. With the introduction of commonly used wide dynamic range compression, more recent prescriptions, including NAL-NL1, NAL-NL2 (Byrne et al., 2001; Keidser et al., 2011) and CAMEQ, CAMEQ-2HF (Moore et al., 1999a, 2010), enable adaptive amplification with respect to the incoming sound levels. Generally, hearing aid fittings are designed to amplify incoming sounds in a way that the amplified sound can be perceived and understood comfortably.

The developments of most prescriptive fittings are based on data collected in subjective listening experiments, which are usually expensive and time-consuming. In this chapter, an alternative approach to finding the optimal fitting is explored. Inspired by recent advances in deep neural networks for speech processing, this chapter proposes a differentiable hearing aid speech processing (DHASP) framework in which a hearing aid processor with trainable parameters can be optimised via back-propagation. Using a differentiable approximation to an existing intelligibility model, the hearing aid is automatically tuned to maximise the predicted intelligibility of the speech signal for a specific individual.

Furthermore, amplification is often not enough to restore intelligibility for hearing impaired listeners. The most common complaint of hearing aid users is that they struggle to understand speech in noisy situations (Brons et al., 2015; Lesica, 2018). This is despite the fact that hearing aids are able to provide sufficient amplification, and despite the fact that modern hearing aids often include noise suppression algorithms. Ultimately better source separation algorithms might solve this problem in the future, but it is still desirable to investigate whether speech intelligibility in noise can be improved by data-driven approaches to parameter-tuning in current hearing aid designs. In particular, the potential for replacing traditional hearing aid fitting formulae with scene-dependent fitting algorithms is looked at in this chapter.

The fitting formulae approach to hearing aid gain setting is remarkably successful and widely deployed in modern hearing aids given that a single formula is used to cover all listening conditions. However, the question naturally arises as to whether better results could be achieved by using noise-dependent fittings, and if so, how should these fittings be optimised. This is particularly relevant now that environment classification algorithms are available to automatically detect whether a user is, say, in a domestic living room, in a noisy cafe or standing by a busy road intersection, thus it can be beneficial to be able to switch gain settings for different environments. It should also be considered that hearing aids now apply increasingly sophisticated (but imperfect) noise-reduction algorithms (e.g., adaptive filtering (Vary and Martin, 2006), spectral subtraction (Bentler and Chiou, 2006; Boll, 1979), spatial filtering (Levitt, 2001)) that can alter the signal in ways that have not been considered in the design of modern hearing aid fitting formulae.

In addition, recent hearing aids are using environmental classification algorithms (Lamarche et al., 2010; Nordqvist and Leijon, 2004) to allow the characteristics of the noise suppression algorithms to be tuned separately for different noise types (Bentler and Chiou, 2006). This further complicates the requirements of fitting formulae. Logically, hearing aid gains should be optimised in consideration of the perception of the *processed* noise-reduced signal that the hearing aid delivers. This chapter thus further explores the possibilities of improving current hearing aid performance by developing fittings specific to different listening environments and noise-reduction processing, and extends the DHASP framework to speech in noise.

This chapter is organised as follows. Section 3.2 presents the overview of the DHASP framework. Section 3.3 and Section 3.4 show the methods, experimental setup and results for clean speech and noisy speech, respectively. The last section summarises the works in this chapter.

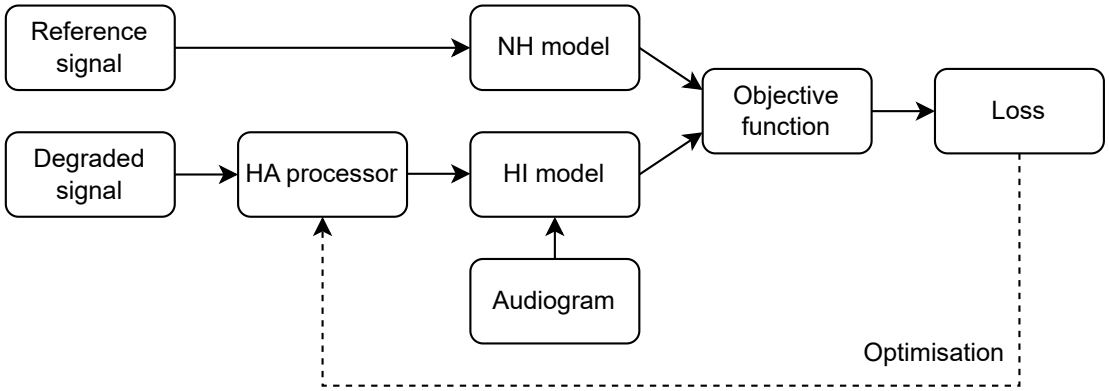


Fig. 3.1 Overall workflow of DHASP.

3.2 DHASP framework overview

The overall workflow of the proposed DHASP framework is shown in Figure 3.1. The degraded signal can represent a noisy speech signal processed with or without a noise suppression algorithm. To simulate the typical signal pathway of a hearing aid (HA) user, the degraded signal is enhanced by a HA processor before being processed by a hearing impaired (HI) model. Its difference from a reference signal processed by a normal hearing (NH) model is measured and used as the loss to optimise the HA processor with back-propagation. The NH and HI models are represented by a hearing loss model, whose characteristics are based on a listener's audiogram. All signals are presented at 65 dB as the sound pressure level (SPL) of normal conversation. A high-performance deep learning library PyTorch (Paszke et al., 2019) is used for the implementation to retrieve the gradients for the optimisation.

To study the performance of optimised amplification fittings, a finite impulse response (FIR) filter is used as the HA processor providing level-independent amplification. The amplification provided depends on six trainable parameters, which represent the insertion gains at [250, 500, 1000, 2000, 4000, 6000] Hz consistent with the frequencies used by a typical audiogram. The frequency response is then obtained with linear interpolation, and iFFT is applied to retrieve the impulse response. A Hann window is subsequently multiplied with the impulse response.

In Section 2.3.3, two approaches are introduced to model the intelligibility of hearing impaired listeners, that is using the Hearing Aid Speech Perception Index (HASPI) (Kates, 2013) and modelling the hearing ability loss with the MSBG hearing loss model (Baer and Moore, 1993, 1994; Moore and Glasberg, 1993; Stone and Moore, 1999). In the remaining of this chapter, the DHASP framework is validated with the approximations to these two models. Specifically, it is firstly validated for clean speech. The objective function is based

on a differentiable approximation to HASPI in this case. After that, the DHASP framework is used for the optimisation of hearing aid fitting for speech in noise. And the objective function is adopted from the MSBG hearing loss model.

3.3 Fitting optimisation for clean speech

In this section, the details of the differentiable approximation to the auditory model originally designed in the HASPI (Kates, 2013) are firstly introduced. The auditory model approximation is then used in the objective function to optimise the FIR filter which represents the hearing aid fitting. The performance of the optimised fittings is later evaluated with a clean speech corpus, and the results are presented at the end of this section.

3.3.1 Differentiable HASPI-based objective

The differentiable HASPI-based objective consists of the differentiable HASPI-based model for simplified hearing impairment simulation, and an objective function comparing the normal hearing and hearing impaired output. The workflow of the differentiable auditory model shown in Figure 3.2 is mainly adapted from the auditory processing used in HASPI. The model operates at 24 kHz and depends on the auditory thresholds given by the listener's audiogram at [250, 500, 1000, 2000, 4000, 6000] Hz. The auditory thresholds are set to zeros for the normal hearing model. Two groups of filterbanks, the dynamic-range compression, and a dB conversion process, are used to simulate the mechanisms in human audition considering the impact of hearing impairment. In contrast to the auditory model in the original HASPI (Kates, 2013), which uses infinite impulse response filters (IIR), the proposed DHASP framework employs FIR filters to avoid expensive recursive computation. The middle ear component and the inner-hair cell adaptation process in the HASPI model are not included for the same reason. The influence of the signal intensity on the analysis filter bank included in the HASPI model is not considered because of the difficulty in the differentiation implementation. All parameter settings used in the differentiable auditory model are the same as the model used in HASPI.

Analysis filterbank

The analysis filter bank consists of a total of $I = 32$ fourth-order FIR gammatone filters (Cooke, 1993). The i^{th} filter $h_a^{(i)}$ of the analysis filterbank is expressed as:

$$h_a^{(i)}(t) = A_a^{(i)} t^{\left(N^{(i)}-1\right)} e^{-2\pi b_a^{(i)} t} \cos\left(2\pi f_a^{(i)} t\right), \tag{3.1}$$

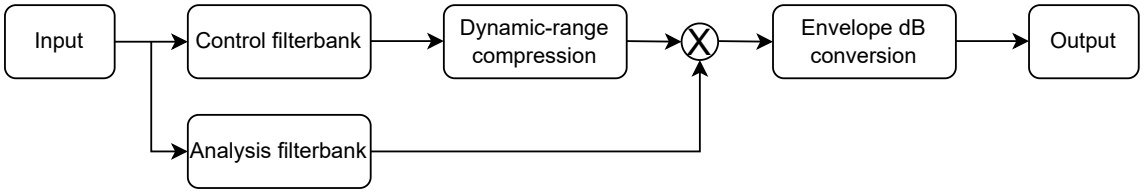


Fig. 3.2 Differentiable HASPI-based model.

where $A_a^{(i)}$ is the amplitude required to normalise the frequency response of the filter; $b_a^{(i)}$ and $f_a^{(i)}$ are the bandwidth and the centre frequency of the filter, respectively (Loweimi et al., 2019); $N(i)$ is the order of the filter which is set as 4 in the model. The centre frequencies f_a are in the Mel scale covering the range from 80 Hz to 8 kHz. The bandwidths b_a^{NH} are in the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1983) for the normal hearing model. To approximate the behaviour that the auditory filter bandwidths increase along with the hearing loss (Moore et al., 1999b), the bandwidths b_a^{HL} of the hearing loss model is expressed as:

$$b_a^{HL} = \left(1 + attn_o/50 + 2(attn_o/50)^6\right) b_a^{NH}, \tag{3.2}$$

where $attn_o$ is the hearing loss for outer-hair cells in dB, with a maximum attenuation of 50 dB (Kates, 2013).

Control filterbank

Another group of fourth-order FIR gammatone filters are used as control filterbank to simulate the two-tone suppression mechanism in the cochlea (Bruce et al., 2003; Heinz et al., 2001). The bandwidths of the control filters correspond to the maximum bandwidth allowed in the hearing loss model, i.e. 50 dB attenuation for outer-hair cell. The bandwidths of control filters are set wider so that they could reduce the gain of the parts of a signal, that stay outside the bandwidth of the analysis filters but still within the control filters (Kates, 2013). Each centre frequency of the control filter $f_c^{(i)}$ is shifted higher relative to the centre frequency of the corresponding analysis filter $f_a^{(i)}$ using a human frequency-position function to correspond to a fractional basal shift (Greenwood, 1990):

$$f_c^{(i)} = 165.4(10^{(1+s)\log_{10}\left(1+f_c^{(i)}/165.4\right)} - 1), \tag{3.3}$$

where s is the shift fraction which is set 0.02 in this model.

Dynamic-range compression

The dynamic-range compression is simulated following the control filtering. The input to the compression rule is each control signal envelope $E_c^{(i)}(n)$ in dB. The compression gain $G^{(i)}(n)$ in dB is computed as:

$$G^{(i)}(n) = -attn_o - (1 - 1/CR)(\theta_{low} - \hat{E}_c^{(i)}(n)), \tag{3.4}$$

where:

$$\hat{E}_c^{(i)}(n) = max(\theta_{low}, (min(E_c^{(i)}(n), \theta_{high})). \tag{3.5}$$

θ_{low} is the lower threshold set as $(attn_o + 30)$ dB sound pressure level, and θ_{high} is set as in the model. CR is the compression ratio which is 1.25 at 80 Hz and linearly increases to 3.5 at 8 kHz for the normal hearing model. This compression behaviour is consistent with the psychophysical estimates of dynamic-range compression in the human auditory system (Moore et al., 1999b). Increasing outer-hair cell damage leads to the reduction of compression ratio. The gain reduction G_{max_o} is set as 14 dB for the compression ratio of 1.25 at 80 Hz, and as 50 dB for the compression ratio 3.5 at 8 kHz. The outer-hair cell threshold is set as $1.25G_{max_o}$. If the hearing loss indicated by the audiogram is greater than the outer-hair threshold, $attn_o$ is set as G_{max_o} and inner-hair cell loss $attn_i$ is set as the remaining loss. On the contrary, $attn_o$ and $attn_i$ are set as 80% and 20% of the total loss, respectively. The compression gain $G^{(i)}(n)$ is then converted into the linear scale, and applied to the corresponding output of the analysis filtering.

Envelope dB conversion

The compressed analysis envelope is converted into dB at this stage. The inner-cell hair loss attenuation $attn_i$ is then added to the converted envelope.

Objective function

The reference envelope $E_r^{(i)}(n)$ and processed envelope $E_p^{(i)}(n)$ processed by the normal hearing and the hearing loss model, respectively, are smoothed using a 16 ms Hann window with 50% overlapping over the time period I. Given the smoothed envelopes $E_r^{(i)}(m)$ and $E_p^{(i)}(m)$, the objective function consists of a cepstral correlation measure function (Kates and Arehart, 2014a) and an energy control function. A set of half-cosine basis functions $b_j(i)$ are

used to compute the cepstral sequences:

$$C_r^{(j)}(m) = \sum_{i=1}^{I} b_j(i) E_r^{(i)}(m), \tag{3.6}$$

$$C_p^{(j)}(m) = \sum_{i=1}^{I} b_j(i) E_p^{(i)}(m), \tag{3.7}$$

where:

$$b_j(i) = cos[(j-1)\pi i/(I-1)]. \tag{3.8}$$

These basis functions are similar to the principal components for the short-time spectra of speech (Zahorian and Rothenberg, 1981) and have been used for consonant and vowel recognition (Nossair and Zahorian, 1991; Zahorian and Jagharghi, 1993). The normalised correlation is then expressed as:

$$R(j) = \frac{\sum_{m=0} C_r^{(j)}(m) C_p^{(j)}(m)}{\sqrt{\sum_{m=0} (C_r^{(j)}(m))^2} \sqrt{\sum_{m=0} (C_p^{(j)}(m))^2}}. \tag{3.9}$$

The final cepstral correlation is the average of $R(2)$ to $R(6)$.

To prevent the over-amplification of the trained hearing-aid processors, which brings discomfort to listeners, an energy control loss is introduced to constrain the processed envelope energy if it is higher than the corresponding reference envelope energy:

$$L_e^{(i)} = \sum_{m \in S} (E_p^{(i)}(m) - E_r^{(i)}(m)), \tag{3.10}$$

where:

$$S = \{m \mid E_p^{(i)}(m) - E_r^{(i)}(m) > 0\}. \tag{3.11}$$

Overall, the objective function used is expressed as:

$$L = -\frac{1}{5} \sum_{j=2}^{6} R(j) + \alpha \sum_i L_e^{(i)}, \tag{3.12}$$

where α is the energy loss weight, which is tuned empirically.

3.3.2 Experiments

Evaluation

HASPI is used to evaluate the performance of the proposed framework. HASPI is based on the auditory model proposed by Kates (2013), and it is designed to predict the speech intelligibility of hearing impaired listeners. The HASPI intelligibility score H is computed as a linear combination followed by a nonlinear scaling function:

$$H = \frac{1}{1 + e^{-(14.817C_C + 4.616C_B - 9.047)}} \tag{3.13}$$

Both basilar membrane vibration (BMV) correlation C_B, based on the temporal fine structures, the rapid oscillations close to the center frequency, and cepstral correlation C_C based on the envelopes, slower amplitude modulations, respectively, are taken into consideration.

NAL-R prescription is used as the baseline system. It prescribes a gain frequency curve given an audiogram. The hearing losses at [250, 500, 1000, 2000, 4000, 6000] Hz are used for the frequency response derivation to be consistent with the proposed framework. A FIR filter is then designed as the hearing aid processor given the frequency response curve.

Audiogram database

10 standard audiograms from Bisgaard et al. (2010), which cover a range of common audiograms in clinical practice, are used to evaluate the proposed framework and are shown in Figure 3.3 as the solid curves with crossing marks. N1 to N7 represent hearing impaired listeners with flat and moderately sloping audiograms, and S1 to S3 represent the steep-sloping group. The audiograms are ranked according to the hearing loss severity. As HASPI has a maximum hearing loss limit, the audiograms used in this work are capped at a deficit of 100 dB hearing loss.

Experimental setup

In the experiment of fitting optimisation for clean speech, DHASP is trained and evaluated on the TIMIT dataset (Garofolo et al., 1993). The training set consists of utterances from 462 speakers while utterances from another 50 speakers are used as the validation set. Utterances of the remaining 24 speakers are used as the final evaluation test set. In both training and evaluation, the input signal is normalised so that its root mean square (RMS) amplitude equals one and is regarded as 65 dB SPL to mimic everyday conversational speech. Utterance segments of 0.5 seconds long are randomly sampled as the input signals during training. The processors are trained with a batch size of 128 for 4000 epochs using the Adam optimiser

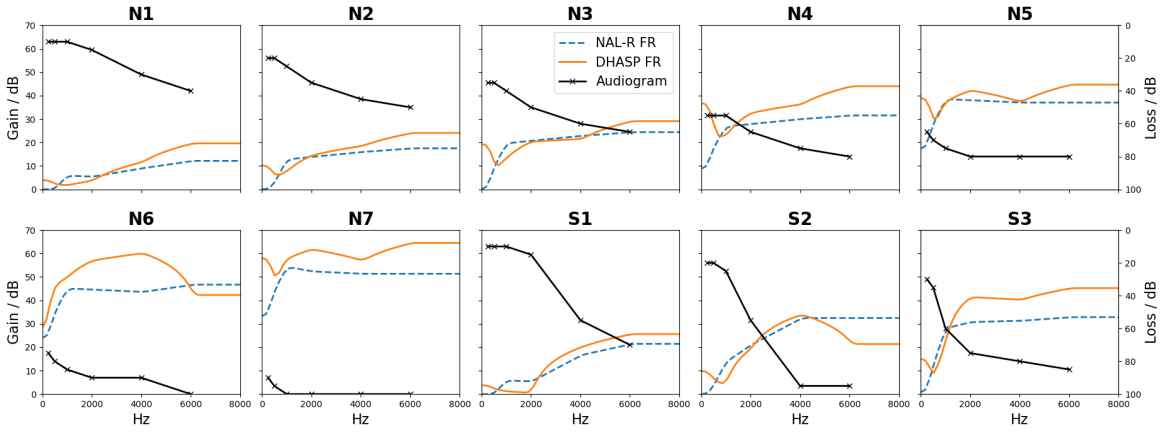


Fig. 3.3 Frequency responses of the DHASP optimised and the NAL-R fittings for standard audiograms for speech in quiet. Standard audiograms are presented as solid curves with cross markers. The hearing losses are capped at 100 dB. The dashed and solid curves represented the frequency responses of the NAL-R prescription filters and the trained DHASP filters, respectively.

(Kingma and Ba, 2014) and a learning rate of 0.001. Six trainable parameters which represent the frequency response gains of the processors at [250, 500, 1000, 2000, 4000, 6000] Hz are optimised. The parameters are all initialised to 1 dB for audiograms N1 to N6 and S1 to S3. For profound loss such as audiogram N7, the low gain initialisation leads to vanishing gradients. Therefore the parameters are initialised to 50 dB in the experiment. The energy loss coefficient α is set to 5e-5.

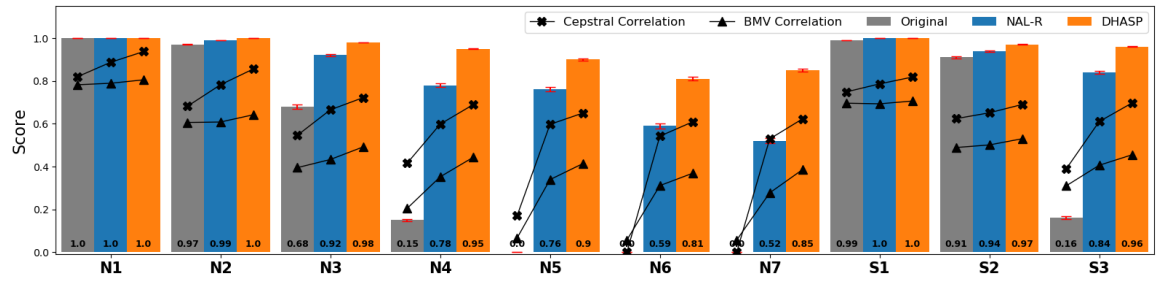


Fig. 3.4 HASPI intelligibility scores of the original, NAL-R processed, and DHASP processed signals for speech in quiet. The error bars indicate the standard error of the mean utterance intelligibility scores in the test dataset. The curves with the cross and triangle marks show the corresponding cepstral correlation scores and the BMV correlation scores.

3.3.3 Results

The frequency responses of the optimised filters by DHASP and the NAL-R prescription filters are shown in Figure 3.4 as the solid and dashed curves, respectively. Across all the standard audiograms the amplification provided by the optimised filters broadly follows the frequency response patterns of the NAL-R filters. In general, the optimised filters amplify the input signals more in the frequency region below 500 Hz. For audiograms with mild and moderate high frequency loss (N1-N5), the proposed filters have higher gains in the high frequency area. On the contrary, the proposed filters amplify less in the high frequency when the loss is severe as shown in N6 and S2. As the information in the high frequency is almost not recoverable due to the profound loss, the amplification in that area would not make a significant difference. DHASP ensures stable convergence for the training of all audiograms, while the convergence time increases along with the severity of the hearing loss.

Figure 3.3 shows the HASPI scores, including intelligibility scores H, cepstral correlation C_C and simulated BMV correlation C_B, of the unprocessed original signals, NAL-R processed signals, and DHASP-processed signals. Both processed signals had higher intelligibility scores than the unprocessed signals. The filters optimised by the DHASP framework achieved higher HASPI scores than the NAL-R prescription filters, with improvements significant across all the audiogram conditions [paired t-test, $p < .005$]. With increased hearing loss severity, the advantages of the proposed optimised processors are more significant compared to the NAL-R prescription. The variation of the intelligibility scores across all the utterances in the test dataset indicates that DHASP can achieve better performance with good consistency. It is not surprising that the cepstral scores of the optimised filters are higher than the NAL-R ones because the objective focuses on the envelope correlation. However, the optimised filters also consistently achieve higher BMV correlation scores.

3.4 Fitting optimisation for noisy speech

This section focuses on the fitting optimisation for speech in different noisy environments and the processed noisy speech by a common noise suppression algorithm. The differentiable approximation to the MSBG model, and the design of the optimisation objective are firstly explained in detail. The experimental setup is then presented, including the evaluation metrics and the database. The experimental results are presented at the end.

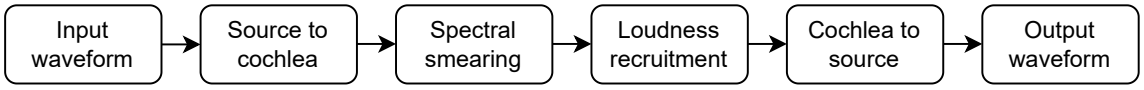


Fig. 3.5 Differentiable MSBG-based model.

3.4.1 Differentiable MSBG model-based objective

The MSBG model can be considered as a hearing impairment simulator, which consists of the simulations of acoustic transformation between sound source and cochlea, spectral smearing, and loudness recruitment. The structure of the model is shown in Figure 3.5. The hearing loss model proposed in this work is a differentiable approximation to the MSBG model, and the differences are in filter implementation and envelope retrieval. All infinite impulse response (IIR) filters in the MSBG model are approximated with FIR filters, and Hilbert transformation is used to extract the envelopes, so that the computation can be performed in parallel for fast optimisation using GPUs.

Source to cochlea transformation

The transformation of the sound pressure level from a sound source to the cochlea is derived from the combination of a free field and a middle ear transfer function. The free field transfer function (Shaw, 1974) approximates the acoustic changes during sound propagation from the free field to the eardrum. The middle ear transfer function (Killion, 1978) simulates the acoustic alterations of sound in the middle- and inner-ear before arriving at the cochlea. The overall transformation is implemented using an FIR filter, whose frequency response is the combination of the free field and the middle ear frequency-gain tables.

Spectral smearing

Spectral smearing (Baer and Moore, 1993) is used in the hearing loss model to simulate reduced frequency selectivity, which is one of the major deficits in the sound analysis ability of cochlear hearing loss. Experimental results showed that this technique leads to little effect on the intelligibility of speech in quiet, but a large effect on speech in noise (Baer and Moore, 1993) or interfering speech (Baer and Moore, 1994). This is generally consistent with the phenomenon that impaired frequency selectivity contributes largely to the difficulty of understanding speech in noise for listeners with cochlea hearing loss.

Input waveform signals are first processed by STFT with Hamming windows. Smearing is then performed to the power spectrogram, and the phase remains unchanged for iSTFT after smearing. Given the input spectrogram X and the output spectrogram Y, the spectrum

smearing function is expressed as:

$$Y = A_N^{-1} A_W X, \tag{3.14}$$

where A_N and A_W are the matrices representing the normal and the widened auditory filterbanks, respectively. For each auditory filter within the filterbank, the form is given by:

$$W(g) = (1+pg)\exp(-pg), \tag{3.15}$$

where $W(g)$ is the intensity weighting function describing the filter shape in the frequency domain, g is the frequency difference from the centre frequency f_c divided by f_c, and p is the parameter determining the sharpness of the auditory filter (Moore and Glasberg, 1983). The value of p is computed as:

$$p = \frac{4f_c}{r \times \mathrm{ERB}}, \tag{3.16}$$

where ERB is the equivalent rectangular bandwidth (Glasberg and Moore, 1990) calculated as $24.7 \times (0.00437 f_c + 1)$, and the widening factor r differs for the lower and upper sides of the filter, denoted as r_l and r_u, respectively. The values of r_l and r_u are dependent on the degree of hearing loss. Each auditory filter is at last calibrated by dividing $24.7 \times \mathrm{ERB}(r_l + r_u)/2$ to remove an upward tilt in the excitation pattern, which is caused by the increase of the bandwidth as the f_c grows.

Loudness recruitment

It is observed that the response of a damaged cochlea to low-level sounds is much smaller than a normal one, while the response to high-level sounds is roughly the same as normal (Moore and Glasberg, 1993). This is simulated by a recruitment mechanism as stated below.

A group of gammatone filters are firstly used to extract the fine structures $x(n)$ of the smeared waveform signal. The i^{th} filter $h^{(i)}$ of filterbank is expressed as:

$$h^{(i)}(t) = A^{(i)} t^{\left(N^{(i)}-1\right)} e^{-2\pi b^{(i)} t} \cos\left(2\pi f^{(i)} t\right), \tag{3.17}$$

where $f^{(i)}$ is the centre frequency, $b^{(i)}$ is the bandwidth computed as $1.019 \times \mathrm{ERB}$, N is the order of the gammatone filter (4 in this study), and $A^{(i)}$ is the amplitude to normalise the filters. The number of filters I differs according to the hearing abilities. The bandwidths of the filters are broadened two or three times for moderate or moderate to severe hearing loss, respectively. The outputs of the filters are aligned in the time domain to ensure the peaks for all channels are coincident with a pulse input. This alignment will make the mixture of

the outputs of the filters generally sound almost identical to the input signal. The envelope $E(n)$ of each channel is retrieved with Hilbert transformation followed by a group of low pass filters for smoothing. The waveform output signal $y(n)$ is then recruited as:

$$y(n) = \sum_{i=1}^{I} \left(\frac{E^{(i)}(t)}{E_\theta} \right)^{\left(\frac{\theta}{\theta - \mathrm{HL}^{(i)}} - 1 \right)} x^{(i)}(n), \tag{3.18}$$

where HL is the audiometric hearing loss in dB, θ is the maximal loudness threshold which is set 105 dB, and E_θ is the corresponding envelope magnitude.

Cochlea-to-source transformation

The recruited signal is lastly processed by a cochlea-to-source transfer FIR filter, whose frequency response is the additive inverse in dB of the frequency response of the source-to-cochlea transformation filter.

Objective function

Given the simulated reference signal $y_r(n)$, i.e. the clean normal hearing signal, and the processed hearing impaired signal $y_p(n)$, STFT is firstly applied to retrieve the corresponding spectrograms $Y_r(m,k)$ and $Y_p(m,k)$. The objective function consists of a spectrogram reconstruction loss L_{spec} and a sound pressure level loss L_{spl}. L_{spec} is expressed as:

$$L_{spec} = 20 \log_{10} \left(\frac{1}{mk} \sum_{m,k} \left| Y_p(m,k) - Y_r(m,k) \right| \right), \tag{3.19}$$

and L_{spl} is computed as:

$$L_{spl} = 20 \log_{10} \left(\sqrt{\frac{1}{n} \sum_{n} y_p(n)^2} - \sqrt{\frac{1}{n} \sum_{n} y_r(n)^2} \right). \tag{3.20}$$

The overall objective function is expressed as:

$$L = \begin{cases} L_{spec} + \alpha L_{spl}, & \text{if } L_{spl} \geq 0 \\ L_{spec}, & \text{otherwise} \end{cases}, \tag{3.21}$$

where α is a weighting coefficient. L_{spl} is used to prevent over-amplification, which could lead to listening discomfort for the listeners, and thus it is not applied if negative.

3.4.2 Experiments

Evaluation

HASPI is used for the evaluation of the proposed optimised fittings within the DHASP framework in terms of intelligibility improvement for hearing impaired listeners. In addition, the hearing aid speech quality index (HASQI) (Kates and Arehart, 2014b), which is developed based on the same physiological auditory model used in HASPI, is used to evaluate speech quality. Also, segmental frequency weighted signal-to-noise (FWSNR) ratio is also used for evaluation, and the implementation is from (Loizou, 2013). The FWSNRs are measured after the amplified signals processed by the MSBG model.

The open-source NAL-R prescription is used as the baseline prescription, which is widely recognised and more importantly tested in subjective experiments. To be consistent with the HA processor, the derivation of the NAL-R frequency response is based on the hearing losses at [250, 500, 1000, 2000, 4000, 6000] Hz. The Wiener filtering algorithm, popular for noise suppression in hearing aids, is used in this work as the noise suppression front end. Fittings optimised on clean data are regarded as the general fittings (G) that stay invariant for all environments. Meanwhile, the custom fittings include the fittings optimised on noisy data (Cn) and the fittings optimised on the Wiener filtering enhanced noisy data (Cw).

Database

Three standard audiograms (Bisgaard et al., 2010) that represent different hearing loss categories are used in this study: N1 (mild hearing loss), N2 (moderate hearing loss), and N4 (moderate to severe hearing loss). Their hearing losses at different frequencies are shown in Table 3.1.

Table 3.1 Hearing losses of the audiograms used in this study.

	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	6 kHz
N1	10 dB	10 dB	10 dB	15 dB	30 dB	40 dB
N2	20 dB	20 dB	25 dB	35 dB	45 dB	50 dB
N4	55 dB	55 dB	55 dB	65 dB	75 dB	80 dB

The noisy speech corpus introduced by Valentini-Botinhao et al. (2016) is used, which was mixed using speech utterances from the Voice Bank Corpus (Veaux et al., 2013), and noises from the recordings of the first channel of the Demand database (Thiemann et al., 2013). A set of 56 speakers is used for training, and another set of 28 speakers is divided into the validation and test sets. There are around 400 utterances from each speaker. 10 types of noises are mixed with the utterances, at the SNRs of 0, 5, 10, and 15 dB according to

ITU-T P.56 standard (ITU-T, 1993). All signals are filtered with a high pass filter whose cut-off frequency is 80 Hz to eliminate non-speech interference. Three categories of noises are selected in the experiments: traffic, kitchen, and babble. The traffic noise and kitchen noise are comparatively more stationary, and mainly distributed in low and high frequencies, respectively. The babble noise is less stationary, and its spectrum overlaps clean speech.

Experimental setup

The Wiener filter implementation is based on the method proposed by Plapous et al. (2006). The parameter setting in the differentiable hearing loss model is consistent with the MSBG model. In the spectral smearing, $[r_l, r_u]$ are set as [4.0, 2.0], [2.4, 1.6], and [1.6, 1.1] for the moderate to severe, moderate, and mild hearing loss, respectively. In the loudness recruitment, the gammatone filterbank consists of 36, 28, or 19 filters respectively for the three types of hearing loss. The HA processors are trained with a batch size of 128 for 500 epochs using the Adam optimiser with a learning rate of $1e$-2. The parameters are initialised as the NAL-R fitting. The weighting coefficient in Eq. 3.21 is set to 5.

3.4.3 Results

Figure 3.6 shows the frequency responses of the optimised general (G) and custom (Cn, Cw) fittings along with the baseline NAL-R prescription for the three hearing loss categories in various noise conditions. First, it can be observed that in general, the frequency responses of optimised fittings have higher gains in low and high frequencies and lower gains around 1 kHz than the NAR-R prescription. This is broadly consistent with the results of subjective hearing experiments reported in Mackersie et al. (2020) and the previous section. The subjective experiments for hearing-aid self-fitting showed that hearing impaired listeners prefer higher gain in high frequency, and lower gain around 1 kHz for speech in noise compared to the NAL-NL2 fitting, which provides even more gain in low and high frequencies and less gain in middle frequency than NAL-R.

Second, compared to the frequency responses of the processors optimised using noisy data (Cn), those optimised using Wiener filtering denoised data (Cw) are more similar to those optimised using clean data (G). This is expected as Wiener filtering is able to suppress the noise to some extent. Among the three noise types, the kitchen noise energy is mainly distributed in high frequencies, and thus the processors optimised in kitchen noise provide more gain in low frequency and less gain in high frequency. On the contrary, the processors optimised in traffic and babble noises, whose energy is mostly in low frequencies, show more gain in high frequency and less in low frequency. It can also be observed that as the

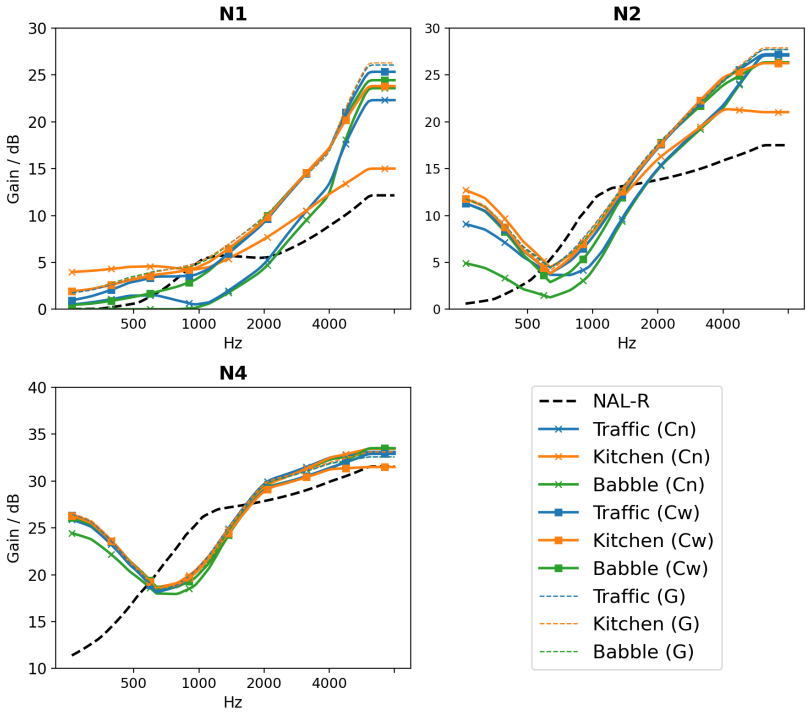


Fig. 3.6 Frequency responses of NAL-R fitting, custom fittings optimised with noisy data (Cn), custom fittings optimised with Wiener filtering enhanced noisy data (Cw), and the general fittings (G) for different hearing losses.

Table 3.2 Evaluation scores of various fittings applied to noisy speech before and after the enhancement of Wiener filtering. N: NAL-R prescriptive fittings, G: optimised general fittings, +W: using enhanced noisy speech by Wiener filtering, Cn: custom fittings optimised on noisy data, Cw+W: custom fittings optimised on Wiener filtering enhanced noisy data with Wiener filtering. The single best score in each group is indicated in bold.

		N	N+W	G	G+W	Cn	Cw+W
Traffic							
N1	HASPI	0.92	0.93	0.92	0.95	0.93	0.95
	HASQI	0.25	0.23	0.26	0.27	**0.28**	0.26
	FWSNR	5.40	6.08	5.82	6.40	**7.59**	6.22
N2	HASPI	0.87	0.81	0.88	0.91	0.89	0.91
	HASQI	0.18	0.13	0.20	0.19	**0.21**	0.19
	FWSNR	5.63	6.00	7.08	6.31	**7.61**	6.30
N4	HASPI	0.21	0.16	0.52	0.40	0.52	0.39
	HASQI	0.03	0.02	0.08	0.05	0.08	0.05
	FWSNR	0.91	2.06	5.12	4.44	**5.24**	4.40
Kitchen							
N1	HASPI	0.99	0.98	0.99	0.99	0.99	0.99
	HASQI	0.42	0.30	0.42	0.35	**0.44**	0.35
	FWSNR	9.54	8.77	7.42	8.65	**10.71**	8.76
N2	HASPI	0.97	0.91	0.98	0.98	0.98	0.98
	HASQI	0.23	0.17	0.30	0.24	**0.30**	0.24
	FWSNR	**9.64**	8.54	7.84	7.77	8.39	7.65
N4	HASPI	0.23	0.20	0.61	0.53	0.61	0.51
	HASQI	0.03	0.03	0.08	0.07	0.08	0.07
	FWSNR	1.86	2.21	6.55	5.71	**6.63**	5.65
Babble							
N1	HASPI	0.68	0.65	0.66	0.67	0.68	0.67
	HASQI	0.15	0.13	0.15	0.14	**0.16**	0.14
	FWSNR	4.26	4.21	4.41	4.36	**6.00**	4.23
N2	HASPI	0.57	0.50	0.59	0.59	**0.60**	0.59
	HASQI	0.11	0.08	0.12	0.10	0.12	0.10
	FWSNR	4.35	4.07	5.38	4.35	**5.48**	4.24
N4	HASPI	0.13	0.09	**0.31**	0.24	0.30	0.23
	HASQI	0.02	0.02	**0.06**	0.04	0.05	0.04
	FWSNR	0.37	1.01	3.52	2.81	**3.66**	2.80

hearing loss severity increases from N1 to N4, the differences among processors optimised with different settings are smaller.

Table 3.2 presents the HASPI, HASQI, and FWSNR scores of the various fittings evaluated in this study. The optimised fittings specific to different listening environments and noise-reduction processing outperform NAL-R in most of the evaluation scores. The improvement over NAL-R is in general larger for more severe hearing loss than that for mild hearing loss, but the benefit can be seen for all three listeners' audiograms and across different noise conditions. This shows there are potential advantages in listening-condition specific fittings that can be learned using the proposed data-driven approach.

Comparing the scores between the optimised custom fittings and the general prescriptive fittings, it is clear that the custom fittings produced higher FWSNRs than the general fittings in all cases. The custom fittings also achieved overall higher or approximately equal HASPI and HASQI scores than the general fittings for mild and moderate hearing losses. For moderate to severe hearing loss in the babble noise condition, the general fitting scores are marginally better than the custom fitting scores. This could be due to the fact that the custom fittings provided overall higher insertion gain, and HASPI and HASQI are sensitive to signal presentation level when the hearing threshold is high, i.e., when the hearing loss is more severe.

It can also be seen that Wiener filtering does not lead to better performance for hearing loss based on the objective evaluation. Our direct measurement without the hearing loss model suggests that Wiener filtering can improve the SNR of noisy speech on average by 6.18, 8.62, and 5.06 dB for traffic, kitchen and babble noise, respectively. However, the HASPI scores suggest that only listeners with mild and moderate hearing losses can gain intelligibility benefits from Wiener filtering in the environment with traffic noise, where the noise is relatively stationary and distributed in low frequencies. For more severe hearing loss, or in the kitchen and babble noise, Wiener filtering did not improve the HASPI score. The SNR improvement by Wiener filtering also did not translate into improvement to the FWSNR in all the tested conditions. While it improves the FWSNR for NAL-R fittings in traffic noise, there is no clear performance pattern for other noisy environments. Overall, the objective experiments suggest the potential advantage of the custom hearing aid fitting optimisation, while challenging the intelligibility benefit of Wiener filtering for environmental noise suppression.

3.5 Conclusions

This chapter presents the DHASP framework for hearing aid fitting optimisation with hearing impaired models. The framework is fully differentiable, therefore can optimise the fittings using a differentiable objective function with the back-propagation algorithm. FIR filters are used to carry the amplification frequency responses with respect to different listeners' audiograms, characteristics of the noise environments, and whether noise suppression algorithms are being applied. The objective functions used in the DHASP are differentiable approximations to widely-used auditory models which take hearing impairment into account. The differentiable HASPI-based and MSBG model-based objective functions are used for speech in quiet and noise, respectively. According to objective evaluation results, the hearing aid amplification fittings optimised by the DHASP framework outperform the NAL-R prescription processors given a range of standard audiograms.

It has been argued that a data-driven hearing aid fitting algorithm can be more flexible than current prescribed fitting formulae. The objective evaluation results suggest that the optimisation-based approach has the potential to outperform prescribed fittings, and that noise-dependent optimisation is particularly promising, with the greatest benefits for mild and moderate hearing losses. In addition, it is observed that the fittings optimised by two different auditory-based models show similar patterns compared to the NAL-R prescriptive fittings from the mild to the moderate to severe hearing losses, i.e., more amplification in the low and high frequencies, and less amplification in the middle frequencies around 1000 Hz.

With the introduction of differentiable optimisation, DHASP has the potential to help the further fine-tuning of the hearing aid fitting as well. Moreover, this framework can also be used for the optimisation of more powerful models like deep neural networks due to the differentiable characteristic. Therefore, it has the potential to help tackle various more complex challenges, such as speech denoising and separation, for hearing impaired listeners. In the next chapter, the DHASP framework incorporates a DNN for speech denoising to enhance speech.

Chapter 4

Incorporating DNN-based Denoising into the DHASP framework

4.1 Introduction

Hearing impairment is usually associated with the decreased sensitivity of sound loudness, i.e., weak sounds can no longer trigger auditory neural activities and strong sounds trigger less (Lesica, 2018). However, the incoming sound is not supposed to be lifted by the same amount as the reduced hearing threshold, i.e., there are fewer dB between the quietest sound that can be heard and the loudest sound that can be tolerated by a hearing impaired listener. This is due to the decrease in the dynamic range, which is like having fewer bits available in an audio signal. As a result, speech fidelity is going to be lost and thus difficult to be understood, especially for those with severe hearing impairment.

The previous chapter introduces the usage of the DHASP framework to optimise hearing aid amplification fitting. However, this is not sufficient to help hearing impaired listeners understand speech in noise. Decreased sensitivity is an over-simplified understanding of hearing impairment, which also reduces the frequency selectivity of auditory, causes temporal smearing, and brings profound distortion to neural activity patterns. The detailed frequency or temporal cues, and the undistorted neural activity patterns are important for sound source localisation and separation. Consequently, hearing impaired listeners are more likely to fail to understand speech in noisy scenes, i.e., when suffering from additional external distortions such as environmental noise and reverberation.

For that reason, a denoising module that suppresses external distortion is equally important as an amplification module for an ideal hearing aid. In fact, noise suppression has become a popular feature of modern hearing aids. Even so, research has shown that many noise

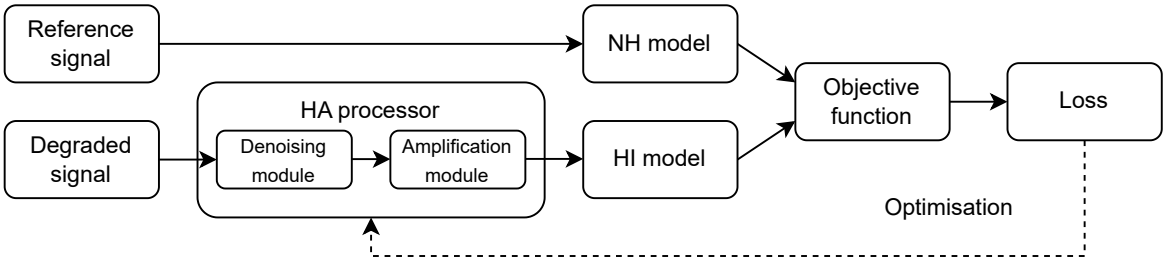


Fig. 4.1 Overall workflow of DHASP including a denoising module and an amplification module.

reduction features provide only limited benefit on intelligibility improvement (Magnusson et al., 2013), especially single channel noise suppression algorithms. In some situations, noise suppression can improve the quality of speech but harms intelligibility (Brons et al., 2014). Therefore, a denoising module that can significantly improve the intelligibility of hearing impaired listeners for speech in noise is wanted for an ideal hearing aid.

Recently, multi-channel noise suppression systems have shown progressive improvement in terms of intelligibility enhancement, as introduced in Section 2.2. These systems can take good advantage of spatial cues to extract target speech. DNN-based multi-channel systems have shown particularly significant improvement. Motivated by that, a DNN-based multi-channel denoising module is included in the HA processor for noise suppression within the DHASP framework in this chapter, as shown in Figure 4.1. It is worth noting that hearing aids require real-time processing with a latency requirement below 10 ms, and most DNN models take utterance-level inputs thus leading to high latency. Therefore, the proposed HA processor is implemented to meet the requirement of an ideal low latency, i.e., the processor is causal, and the output from the processor at the current sample does not use any information from input samples more than 5 ms into the future. The proposed HA processor is validated with a large-scale database from the first round Clarity Enhancement Challenge (CEC1) (Graetzer et al., 2021).

4.2 Method

As presented in the HA processor block in Figure 4.1, a denoising module M_D and an amplification module M_A need to be optimised for noise suppression and hearing loss compensation. As a result, the approach proposed in this chapter is designed to optimise the two modules in two separate optimisation stages. The overall workflow of the approach is shown in Figure 4.2. In the first stage, M_D is optimised with a signal-to-noise ratio (SNR) loss for noise and reverberation suppression.

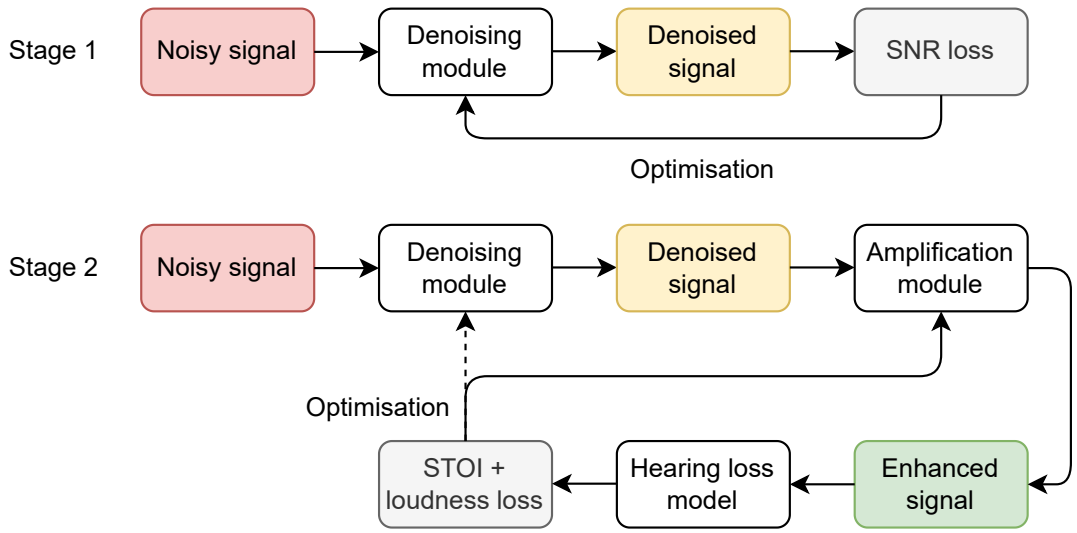


Fig. 4.2 Overall workflow of the two-stage optimisation for the denoising and the amplification modules. In the second stage, the denoising module can be jointly optimised together with the amplification module.

In the second stage, a differentiable hearing loss model M_{HL} is incorporated and M_A is optimised with an objective function consisting of an STOI loss (Taal et al., 2011) and a loudness loss (Steinmetz and Reiss, 2021) for the compensation of hearing impairment. This is similar to the approach introduced in Section 3.4, i.e., the optimised amplification is dependent on not only the hearing impairment but also the denoising effect. Additionally, M_D can be jointly optimised in the second stage. All components are implemented with PyTorch (Paszke et al., 2019), and the back-propagation algorithm is used to compute gradients for the optimisation. M_D, M_A and M_{HL} are described in this section.

4.2.1 Denoising module

The denoising module M_D aims to suppress disturbances caused by both noise and speech interferers. Conv-TasNet (Luo and Mesgarani, 2019) is an end-to-end convolutional time-domain audio separation network and has shown its success for single-channel speech separation and denoising tasks. In order to exploit the spatial information provided by multi-channel signals in the Clarity Challenge, the multi-channel (MC) Conv-TasNet is used in this work as M_D. The MC-Conv-TasNet has been proven effective for a joint denoising, dereverberation and separation task in terms of SNR and ASR recognition improvement (Zhang et al., 2020). In this work, it is further validated for intelligibility improvement in a more realistic environment setting with a larger SNR range.

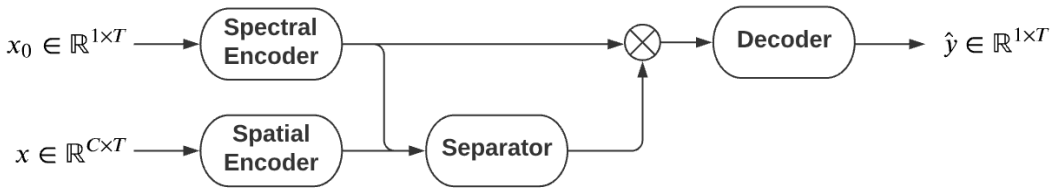


Fig. 4.3 Structure of MC-Conv-TasNet

The structure of MC-Conv-TasNet is shown in Figure 4.3. It incorporates a spectral encoder, a spatial encoder, a separator and a decoder. Given a multi-channel noisy signal $x \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the number of signal samples, the spectral encoder takes one channel as the input and maps segments of this channel $x_0 \in \mathbb{R}^{1\times T}$ to high-dimensional features with a 1-D convolutional layer. Meanwhile, the spatial encoder extracts the spatial information from x with a 2-D convolutional layer. Outputs of both spectral and spatial encoders are utilised by the separator, which then computes a mask for the target features. The separator consists of multiple 1-D convolutional blocks, which include multiple 1-D convolutional layers, PReLU activations, normalisation layers, and residual connections. Finally, the decoder reconstructs a single channel output $\hat{y} \in \mathbb{R}^{1\times T}$ with the estimated features provided by the separator.

Different from (Luo and Mesgarani, 2019; Zhang et al., 2020), SNR rather than scale-invariant SNR (SI-SNR) is used as the objective, so that the signal level stays consistent as it is critical for the down-streaming amplification. The SNR loss $\mathrm{L_D}(y,\hat{y})$ is expressed as:

$$\begin{aligned}\mathrm{L_D}(y,\hat{y}) &= -10\log_{10}\frac{\|y\|^2}{\|y-\hat{y}\|^2+\tau\|y\|^2} \\ &= 10\log_{10}\left(\|y-\hat{y}\|^2+\tau\|y\|^2\right)-10\log_{10}\|y\|^2,\end{aligned} \tag{4.1}$$

where $\hat{y}$ and y are the estimated and reference signals, respectively, and $\tau = 10^{-\mathrm{SNR_{max}}/10}$ is a soft threshold preventing examples that are well denoised dominating the gradients within a training batch (Wisdom et al., 2020). $\mathrm{SNR_{max}}$ is set to 30 dB according to Wisdom et al. (2020).

4.2.2 Amplification module

The amplification module M_A aims to implement individualised enhancement to the denoised signals to maximise the intelligibility for the hearing impaired listeners. In this work, both a Conv-TasNet and a finite impulse response (FIR) filter are compared as candidates to be used as the amplification module. The structure of the amplification Conv-TasNet is roughly consistent with the denoising MC-Conv-TasNet. The difference is that the amplification

Conv-TasNet does not deploy a spatial encoder, as it takes the single channel output from the denoising module as the input. The amplification FIR is the same as the processor in the DHASP described in the previous chapter. The amplification module takes the denoised signal $\hat{y} \in \mathbb{R}^{1\times T}$ as the input and produces the amplified signal $\hat{z} \in \mathbb{R}^{1\times T}$.

STOI is used in the objective function as the target is to achieve maximal intelligibility. A loudness constraint term is also included, otherwise, the signal could be over-amplified as STOI is almost regardless of signal level. Specifically, STOI predicts intelligibility by computing the cross correlation between the acoustic representations of the processed signal and the reference signal. When amplifying the processed signal, more information can be leaked after the processing of the hearing loss simulator. Therefore, a loudness constraint is needed, and the objective function is expressed as:

$$\mathrm{L_A}(y,\hat{z}) = -\mathrm{STOI}(y, M_{HL}(\hat{z})) + \alpha\|\Gamma(y) - \Gamma(M_{HL}(\hat{z}))\|^2, \tag{4.2}$$

where α is a weighting coefficient, Γ is the loudness computing formula according to ITU-R BS.1770-4 (Radiocommunication Sector of ITU, 2011), and M_{HL} represents the hearing loss the model which will be introduced in the next section.

4.2.3 Hearing loss model

The hearing loss model M_{HL} used in this work is a differentiable approximation to the MSBG model (Baer and Moore, 1993, 1994; Moore and Glasberg, 1993; Stone and Moore, 1999) released in the challenge, and the detailed approximation implementation explained in the previous chapter. Different from the MSBG model, the differentiable hearing loss model takes advantage of FIR filters and Hilbert transformation for fast parallel computing. The model takes the audiogram of a listener as input, and simulates free field, middle- and inner-ear transformation, spectral smearing, and loudness recruitment. For more details, see Section 3.4.

4.3 Experimental setup

This section presents the detailed experimental settings to evaluate the proposed HA processor. The database used in the CEC1 challenge (Graetzer et al., 2021) which provides simulated domestic noisy environments is used in this work, and the results are retrieved from both objective and subjective evaluations. The baseline systems include the CEC1 challenge baseline with also a number of CEC1 participants. In addition, the detailed configuration of the proposed HA processor is also introduced in this section.

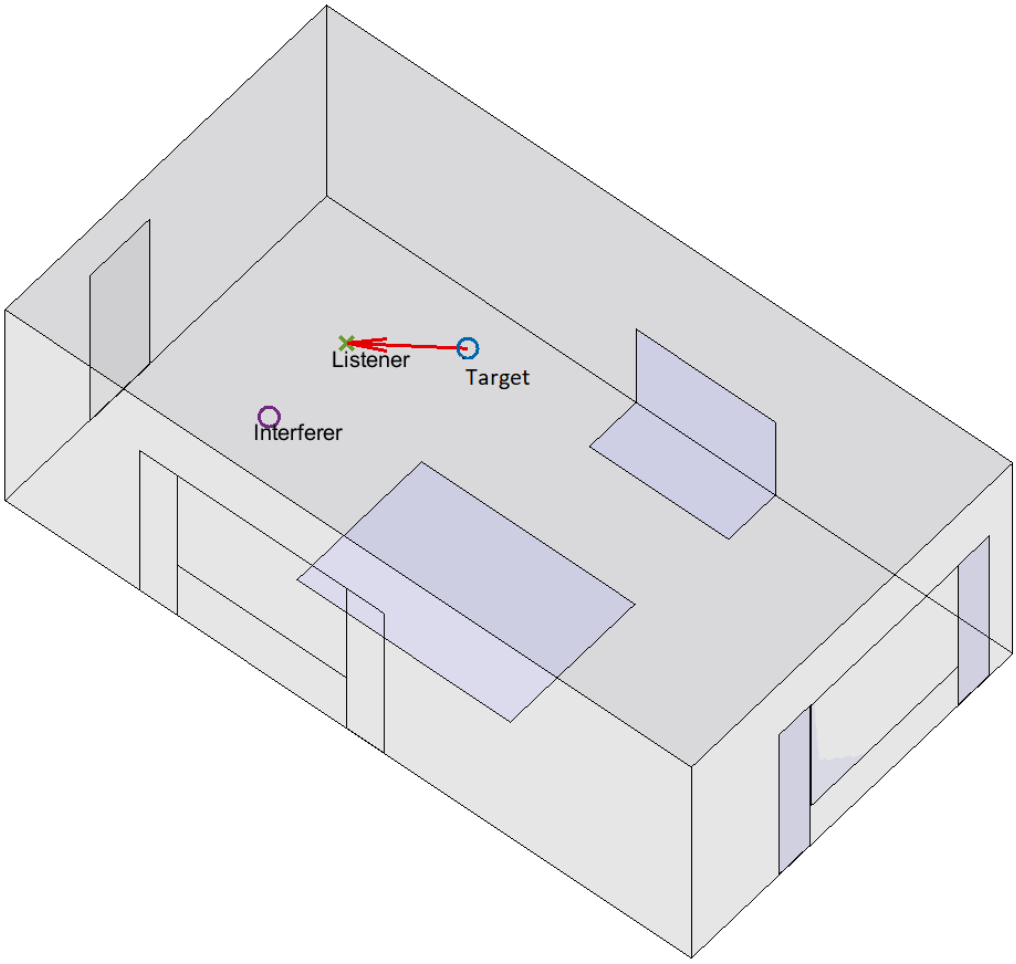


Fig. 4.4 An example scene in CEC1 database.

4.3.1 Overview of CEC1 Database

The CEC1 database provides a set of 10,000 simulated domestic scenes for hearing aid speech processing. Each scene is a simulated room in a cuboid shape where there are a target speaker, an interfering source, and a hearing impaired listener, as shown in Figure 4.4[1]. The simulated rooms are built with room impulse responses generated by the Real-time framework of the Auralization of interactive Virtual ENvironments (RAVEN) (Schröder and Vorländer, 2011). The target speech material contains British English sentences read by 40 readers, and each utterance consists of 7 to 10 words. For the interfering sources, half of the scenes use a speech from the Open-source Multi-speaker Corpora of the English Accents in the British Isles (Demirsahin et al., 2020) and the other half use domestic noises from the Freesound (Font et al., 2013). In order to simulate the use of hearing aids, a group of

[1] Adopted from https://claritychallenge.org/docs/cec1/cec1_scenario

head-related impulse responses are drawn from the OlHead-HRTF database (Denk et al., 2018).

In each scene, the speech at the listener is firstly generated by convolving the target speech with the binaural room impulse response, which is created in RAVEN and the head-related impulse response from OlHead-HRTF. The speech is then mixed with a speech or noise interferer with a specified speech-weighted SNR in the frequency domain. The SNRs for the speech interferers are from 0 to 12 dB, and the SNRs for the noise interferers are from -6 to 6 dB. The SNR is calculated with respect to a specific setup. In detail, both speech and noise before SNR calculation are convolved by a speech-weighted filter, and the overall SNR is defined at the better ear. Additionally, the target speech always begins two seconds after the start of the interferer.

Among the 10,000 simulated scenes provided in CEC1, 6,000 of which are used as the training set (*train*), 2,500 are treated as the development set (*dev*), and 1,500 are used for the final evaluation set (*eval*). Utterances from 24 speakers are selected for *train*, 10 for *dev*, and 6 for *eval*. Each scene incorporates a six-channel noisy signal, which consists of the front, mid, and rear microphone inputs for both the left and right ear, and a dual-channel clean anechoic signal from the left and right ear positions. The sampling rate of the signals is 44.1 kHz.

Bilateral pure-tone audiograms are used to characterise listeners' hearing abilities by recording the hearing thresholds at [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz. 100 generated audiograms are provided in *train* and *dev*, and another 50 audiograms from real listeners for the *eval*.

4.3.2 System setup

Denoising module

The network configuration of M_D is described in this section. In general, the parameter settings follow those used by Zhang et al. (2020). The signals are downsampled to 22.05 kHz to be operated by the network. 256 and 128 filters are used in the spectral and spatial encoders, respectively. The length of the encoder filters is 20 samples, and thus the latency is less than 1 ms as the network is configured causal. 256 and 512 channels are used in the bottleneck 1×1 convolutional block and the convolutional blocks, respectively. The kernel size in the convolutional blocks is 3. 6 convolutional blocks with dilation factors of 1, 2, 4, ..., 32 are repeated 4 times within the separator.

All six channels of noisy signals are used as the input of M_D, and one channel of anechoic signals is used as the reference (dependent on the left or right ear) for training. M_D is trained

for 200 epochs on 2-second long segments. Adam optimiser (Kingma and Ba, 2014) is used for training with the initial learning rate of $1e$-3. Gradient clipping with a maximum L2-norm of 5 is applied. The convolution layers and layer normalisation in M_D are implemented causally. An NVIDIA Tesla V100 SXM2 GPU is used for training M_D, and two modules are trained in total for the left and right ear. In this work, M_D is not jointly optimised in the second stage, otherwise it can cause strong artefacts.

Amplification module

Both the Conv-TasNet and an FIR filter are selected to be optimised as the amplification module, noted as M_A^C and M_A^F, respectively. As hearing losses cause complicated non-linear degradation, M_A^C is expected to provide such an amplification that can be a better fit to this degradation. In contrast, M_A^F is optimised to provide a simple and linear amplification which processes signals with constraints, i.e., avoids distortion or artefacts. The configuration of M_A^C is consistent with M_A, except for the number of separator convolutional blocks being two. The implementation of M_A^F is detailed described in the previous chapter, and the length of the FIR filter is 882. The latency of M_A^F used for evaluation is more than 5 ms, while we further reduced the tap size of the FIR filter to 220 and the difference is minimal.

The single-channel output of M_D is used as the input, and M_A produces a single-channel amplified signal for hearing loss compensation to each ear. The amplified signals are hard clipped from -1 to 1 after amplification to meet the CEC1 rule, and then upsampled to 44.1 kHz for the processing of M_{HL}. M_A^C is trained for 50 epochs with the initial learning rate of $1e$-3 and M_A^F is trained for 20 epochs with the learning rate of $5e$-2.

4.3.3 Evaluation

The evaluation consists of an objective evaluation and a subjective evaluation, both conducted by the Clarity challenge organisers. In the objective evaluation, each scene within the *eval* set is evaluated with three audiograms. The combination of the MSBG hearing loss simulator and MBSTOI is used as the evaluation metric. MBSTOI is an improved version of binaural STOI, which is arguably the most widely used intelligibility evaluation metric. The detailed description of MBSTOI can be seen in Section 2.3.3. An enhanced speech signal is first processed by the MSBG model given the corresponding audiogram. The processed signal and the corresponding clean reference signal are used to compute the MBSTOI score.

In the subjective evaluation, each scene within the *eval* set is evaluated by one hearing impaired listener. For scenes with noise interferer, the listener is asked to follow the instruction: “In the speech in noise test, you will hear a sentence and a loud distracting noise (e.g.,

Table 4.1 Overview of the systems submitted to CEC1.

System	Beamforming	DNN-based denoising	Amplification
Žmolíková and Cernock (2021)	MVDR	Conv-TasNet	DNN
Tammen et al. (2021) (a)	wBLCMP	DNN post-filter	MBDRC
Tammen et al. (2021) (b)	wBLCMP		MBDRC
Yang et al. (2021)	RLS adaptive	MC-Conv-TasNet	Linear equaliser
Moore et al. (2021)	MVDR		CAMFIT + AGC
Chen et al. (2021)		DCCRN	Dynamic equaliser
Kendrick (2021)		U-Net	Linear
Gajecki and Nogueira (2021)		Binaural Conv-Tasnet	

a washing machine). You need to repeat what the talker is saying." For scenes with speech interferer, as the target speech always starts later than the interferer speech, the instruction for the listener is: “In the two-talker test, you will hear two talkers speaking at the same time. One talker will start later than the other. You must repeat what this second talker is saying." The repeated speech by the listener is then transcribed by an ASR to retrieve the recognition results, which are then compared with the reference prompt to calculate the percentage of the correctly recognised words. Also, the listener responses and the corresponding prompt texts need aligning before the correct words can be counted.

4.3.4 Baselines

CEC1 provides a baseline hearing aid implementation, which consists of the CAMFIT algorithm (Moore et al., 1999a) and a configuration of the OpenMHA (Kayser et al., 2021) for a behind-the-ear model. The CAMFIT provides the compression ratios for a multiband compression system where the centre frequencies are at [177, 297, 500, 841, 1414, 2378, 4000, 6727] Hz. The OpenMHA configuration involves the multiband dynamic range compression (MBDRC) plugin for hearing loss compensation and directional processing to improve the SNR levels.

The approach described in this chapter has also been evaluated against the various other systems that were submitted to CEC1. In general, all these systems can be broadly described in terms of beamforming, DNN-based denoising and amplification. A brief overview of these systems is presented in Table 4.1.

Žmolíková and Cernock (2021) proposed a system consisting of three parts: a beamforming module, a post-enhancement DNN, and a listener-adjustment DNN. The beamforming

module uses a minimum variance distortionless response beamformer to take advantage of spatial cues of a multi-channel signal, with the time-frequency mask estimated by a complex gaussian mixture model. The Conv-TasNet is used for post-enhancement with a multi-task optimisation objective consisting of STOI, SNR and PMSQE (Martin-Donas et al., 2018). The listener-adjustment DNN is an auxiliary network taking audiograms into consideration and outputs the amplified signal for hearing loss compensation.

Similarly, Tammen et al. (2021) proposed a system consisting of (1) a weighted binaural linearly constrained minimum power (wBLCMP) beamformer targeting at minimising output power when ensuring the desired speech component undistorted, (2) a DNN-based minimum variance distortionless response post-filter for further interferer suppression, (3) a multiband compression for hearing loss compensation which is the same as the one in OpenMHA. The full system is denoted as Tammen et al. (2021) (a); the system without the post-filter is denoted as Tammen et al. (2021) (b).

Yang et al. (2021) also proposed to combine DNN and beamformer for intelligibility improvement. A DNN-based single-channel speech enhancement is firstly trained with a multi-resolution spectral loss (Wisdom et al., 2019). The DNN denoised single channel signals are then used for recursive least squares (RLS) adaptive beamforming. In the beamformer's time-frequency space, the coefficients are amplified for the compensation of hearing loss. It is worth noting that training of the system involves an on-the-fly data augmentation by generating new scenes combining target speech and interfering sources from different existing scenes in the original training dataset.

Moore et al. (2021) proposed a system without using DNN for noise suppression. A binaural minimum variance distortionless response beamformer is used to improve SNRs. Meanwhile, a broadband automatic gain control is used following linear hearing loss compensation.

Moreover, Chen et al. (2021), Kendrick (2021) and Gajecki and Nogueira (2021) proposed to use DNN-based systems for noise suppression. Chen et al. (2021) leveraged a deep complex convolution recurrent network (DCCRN) for denoising followed by a weighted prediction error filtering for dereverberation. Kendrick (2021) employed a convolutional U-Net for denoising and a subsystem consisting of an amplification filter bank processor, a compressor and soft clipping. Gajecki and Nogueira (2021) proposed to introduce attention layers to two Conv-TasNet for binaural denoising.

Table 4.2 **Objective** evaluation results. M_D: MC-Conv-TasNet based denoising module; M_A^C: Conv-TasNet based amplification module; M_A^F: FIR based amplification module.

Method	MBSTOI	
	Noise interferer	Speech interferer
Žmolíková and Cernock (2021)	0.678	0.715
Moore et al. (2021)	0.653	0.676
Kendrick (2021)	0.639	0.701
Yang et al. (2021)	0.632	0.670
Tammen et al. (2021) (a)	0.611	0.636
Tammen et al. (2021) (b)	0.607	0.634
Chen et al. (2021)	0.524	0.521
Gajecki and Nogueira (2021)	0.481	0.549
Baseline	0.282	0.335
$M_D + M_A^C$	0.672	0.704
$M_D + M_A^F$	**0.693**	**0.741**

4.4 Results

4.4.1 Objective results

The results of MBSTOI objective evaluation are shown in Table 4.2. Both the denoising effectiveness and the benefit of hearing loss compensation decide the MBSTOI scores of the MSBG model processed signals. As the baseline system does not include a functional noise suppression module, there is a large gap between its objective scores and those of all CEC1 participants. Systems proposed by Chen et al. (2021) and Gajecki and Nogueira (2021) manage to suppress interferers to some degree with DNNs to improve objective evaluation scores significantly compared to the baseline. Meanwhile, systems consisting of both beamformers and DNNs, i.e. those proposed by Tammen et al. (2021); Yang et al. (2021); Žmolíková and Cernock (2021), can achieve better MBSTOI scores. It is worth noting that pure beamforming system (Moore et al., 2021) obtain very competitive results, though DNNs are usually considered to be able to bring more significant noise suppression. The system proposed by Moore et al. (2021) can achieve very high MBSTOI scores thanks to its denoising DNN and amplification subsystem.

The proposed systems can reach the top objective scores. It is worth noting that M_A^F performs overall the best in terms of both noise and speech interferers. As the FIR filter has such a simple structure that could have better generalisation ability, it performs better compared to deep neural network based M_A^C.

Table 4.3 **Subjective** evaluation results. M_D: MC-Conv-TasNet based denoising module; M_A^F: FIR based amplification module.

Method	Correctness (per cent)	
	Noise interferer	Speech interferer
Tammen et al. (2021) (b)	**86.726**	**86.885**
Yang et al. (2021)	85.532	4.444
Tammen et al. (2021) (a)	84.914	83.929
Moore et al. (2021)	83.613	82.895
Žmolíková and Cernock (2021)	75.424	81.498
Kendrick (2021)	72.222	77.778
Chen et al. (2021)	60.593	44.681
Baseline	33.202	51.152
$M_D + M_A^F$	80.426	82.432

Most CEC1 participating systems consist of a denoising module for interferer suppression and an amplification module for hearing loss compensation. The system in Žmolíková and Cernock (2021) and this work show that the amplification modules that are optimised with approximated MSBG models help reach the top objective evaluations.

Also, the proposed system with M_A^F reaches the highest MBSTOI scores, and the system proposed by Žmolíková and Cernock (2021) with a DNN-based amplification module achieves the second highest scores. Meanwhile, the proposed system with M_A^C amplification which is much more heavily parameterised reaches the third highest MBSTOI scores. This suggests that the over-parameterisation of the amplification module, i.e. the optimised amplification module with a large number of parameters, can do harm to the performance on the evaluation set.

The variance of the objective evaluation results achieved by the systems with only DNNs for interferer suppression is larger than those with beamformers. This suggests that the quality of the denoising module can be highly dependent on the DNN structures, training techniques, etc. Meanwhile, beamformers can benefit from the CEC1 scenarios in which the first two seconds of the signal contain only interferers, and the listeners, target sources and interfering sources all are at fixed locations.

4.4.2 Subjective results

The subjective evaluation results are shown in Table 4.3. In general, systems proposed by participants can significantly outperform the CEC1 baseline. Approaches with beamformers can achieve top subjective results, while those using only DNN for denoising bring less

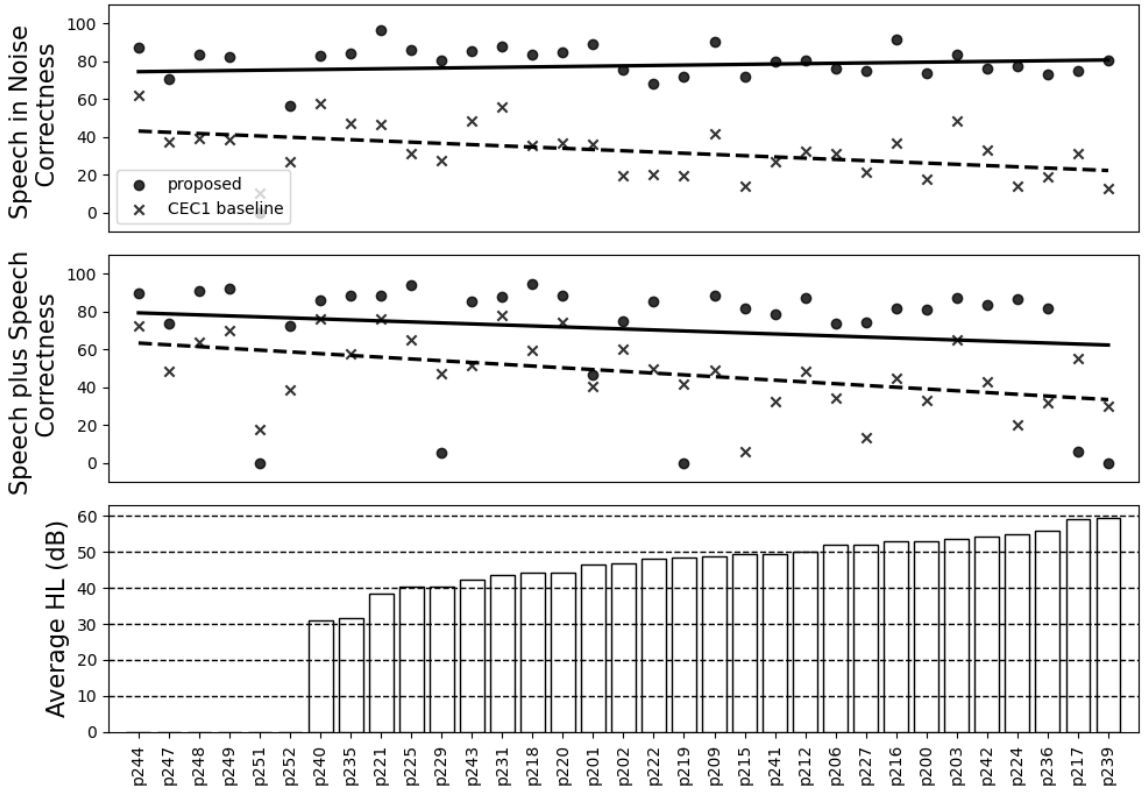


Fig. 4.5 Subjective evaluation correctness of the proposed approach and the CEC1 baseline, and average hearing thresholds of each listener. Subjective evaluation correctness of the proposed approach and the CEC1 baseline, and average hearing thresholds of each listener. The top panel presents the correctness of scenes with speech in noise interferers. The middle panel presents the correctness of scenes with speech in speech interferers. The bottom figure presents the average hearing loss (HL) across different frequencies and both ears of each listener. In the top and middle panels, the regression lines of both the proposed and CEC1 baseline correctness are presented.

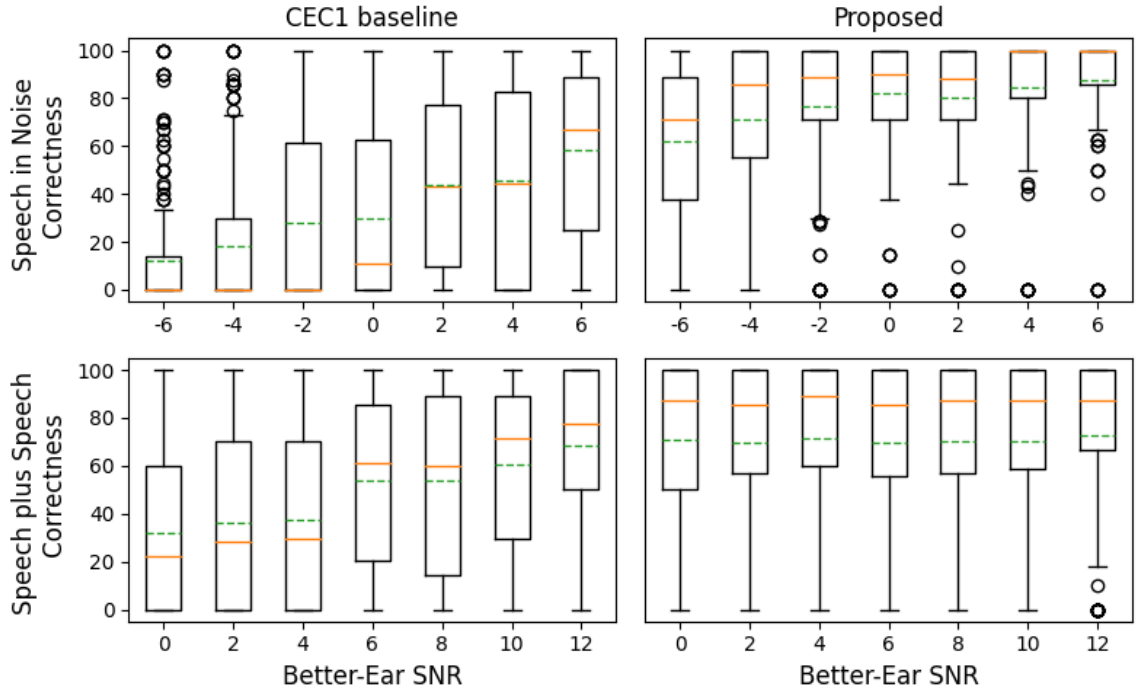


Fig. 4.6 Box plots of subjective evaluation correctness against better-ear SNRs of the unprocessed scenes. The top rows presents the recognition correctness of scenes with speech in noise interferers, and the bottom rows show those of scenes with speech plus speech interferers. The median and mean correctness are shown in orange solid line and green dashed line, respectively. The box represents the range from the lower to the upper quartile of the correctness scores, and the whiskers extend to 1.5 times the interquartile range.

intelligibility improvement. The ranking in terms of scenes with noise interferers is quite consistent with that of scenes with speech interferers. The system proposed by Yang et al. (2021) is an outlier achieving very low subjective scores for scenes with speech interferers. The reason is that the recognition correctness of scenes with speech interferers can be heavily biased, as the listeners are asked to repeat what the second talkers say, i.e., the target speech, while the system of Yang et al. (2021) can completely eliminate the interfering speech resulting in no second talkers appearing.

The overall rankings indicate the benefit of beamformers for subjective intelligibility improvement. One major advantage of beamforming is the better utilisation of spatial clues for noise suppression. The limits of the CEC1 scenes, i.e., the fixed first two seconds of interferers and source and listener positions, make the task easier for beamformers. In addition, DNNs usually bring more distortion, especially for low SNR signals, despite they can suppress more noises. Therefore, the intelligibility of DNN-processed signals may be more degraded compared to distortionless beamformers. This is also reflected by the comparison between the two systems proposed in Tammen et al. (2021): the full system with DNN post-filter gains a slight advantage in terms of MBSTOI scores but performs slightly worse in the subjective evaluation.

Figure 4.5 presents the listening recognition correctness of the CEC1 baseline and the proposed system $M_D + M_A^F$, and average hearing thresholds of each listener. The results of the first six listeners, i.e., from p244 to p252, are conducted by normal hearing listeners and not computed for the overall results shown in Table 4.3. For the scenes with noise interferers, the proposed system with a DNN-based noise suppression module can bring a significant improvement compared to the CEC1 baseline for all hearing impaired listeners. For the scenes with speech interferers, only a limited number of hearing impaired listeners can not gain benefit from the proposed system, which could be caused the listening test instruction confusion, as the correctness of these listeners (p217, p219, p229, p239) is close to zero.

For the CEC1 baseline, it can be observed that the correctness decreases with the growth of average hearing losses. This phenomenon is more significant for scenes with speech interferers. Meanwhile, the correctness decreasing trend is not obvious for the proposed system. This suggests that the proposed system can help gain more intelligibility improvement for more severe hearing impaired listeners.

The subjective recognition performance along SNRs is shown in Figure 4.6. The SNRs are those of raw signals, i.e., those without any processing. For both scenes with noise and speech interferers, the proposed system can gain significant improvements, especially for those with low SNRs. It can also be observed that the scenes with noise interferers in low SNRs are not as intelligible as those with high SNRs with the enhancement of the proposed

systems. However, recognition correctness distributions of scenes with speech interferers are similar across all SNRs.

4.5 Conclusions

In this chapter, a DNN-based denoising module is included in the DHASP framework for noise suppression. The amplification FIR filter is then optimised with the denoised signals with the objective consisting of differentiable approximations to the MSBG hearing loss model and STOI. The proposed hearing aid processor is validated in the CEC1. It achieves the top performance in terms of objective evaluation, while not being as competitive for subjective evaluation.

Meanwhile, beamformer-based systems stand out in the subjective evaluation thanks to two major reasons: (1) beamformers introduce less external processing distortion compared to DNNs; (2) the CEC1 scenes are static and a two-second non-target speech period is provided in the beginning of a scene signal, thus it is easier for beamformers to take advantage of spatial cues. Beamformers do not achieve high objective intelligibility scores as they can preserve interferers, which does not necessarily degrade subjective intelligibility. The reason for this difference is that it is typical for most intelligibility predictors, especially intrusive ones, to consider the signal as 'one source' and measure the amount of distortion by comparing the whole signal to the reference. However, this is only true when the listener hears a signal as a single source and the difference appears as distortion artefacts. If the cues for segregation are preserved for noisy speech, the listener can attend selectively to the target speech, i.e., the interferer may cause some masking but this will be interpreted as missing information rather than a mismatch. There is nothing much in MBSTOI and many intelligibility measurements that really capture this. Therefore, there is a significant gap between the objective and subjective evaluation results, and thus more accurate intelligibility predictors are wanted for better development of speech enhancement algorithms, especially for hearing aids.

Chapter 5

Intrusive Intelligibility Prediction with ASR Hidden Representations

5.1 Introduction

Accurate objective speech intelligibility prediction plays an important role in the development of hearing aids, because subjective listening experiments can be time-consuming and expensive (Falk et al., 2015). Most approaches make predictions by comparing the acoustic representations of reference and degraded speech signals. In Chapter 2, it was noted that appropriate representations are crucial for accurate prediction. Although much progress has been made in accurate intelligibility prediction, many proposed approaches fail for speech processed by some enhancement models, particularly those that cause non-linear and non-stationary distortions. One potential reason is that the hand-picked acoustic representations of these approaches are usually not explicitly correlated with recognition.

The speech recognition performance of recent DNN-based ASR systems is approaching that of humans, and more importantly, they have also shown similar patterns in speech recognition performances, e.g. (Fontan et al., 2017; Schädler et al., 2015). Therefore, it has been of interest to use DNN-based ASR for intelligibility prediction. Compared to the acoustic representations proposed in the aforementioned intelligibility prediction algorithms, hidden representations of DNN-based ASR are optimised to directly correlate with recognition. In this chapter, these ASR hidden representations are used to measure the similarity between a pair of degraded and reference speech signals, and the measured similarity is regarded as an intelligibility predictor.

This chapter is organised as follows. Section 5.2 presents the extraction of ASR hidden representations and the similarity measurement. Section 5.3 describes the experimental setup

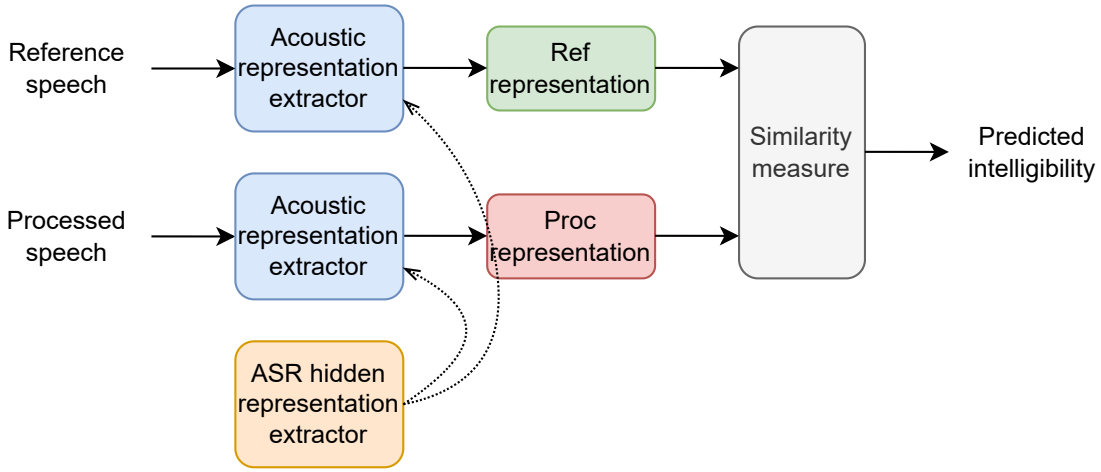


Fig. 5.1 A general framework for intrusive intelligibility prediction. The proposed approach in this chapter uses an ASR model as the representation extractor.

including databases, model configuration, and evaluation metrics. The results and analyses of two datasets are presented in Section 5.4 and Section 5.5. The last section summarises the work in this chapter.

5.2 Similarities between ASR hidden representations

This section will describe how to leverage hidden representations from an ASR model for intelligibility prediction. The majority of intrusive intelligibility prediction approaches fall into a similar framework shown in Figure 5.1, such as STOI and HASPI. An acoustic representation extractor is used to extract the representations of the processed speech signal $\hat{x}$ and its corresponding reference speech x. The similarity between the reference representations H and processed representations $\hat{H}$ is measured and used to correlate to the intelligibility.

In this chapter, an ASR model is proposed to be used as the representation extractor for intelligibility prediction. The ASR model based on transformer architectures (Vaswani et al., 2017) has achieved great success recently and is used for hidden representation extraction. As the hierarchy of such powerful transformer-based ASR models usually consists of multiple levels, and the representations at different levels are differently expressive, the performances of hidden representations at different levels within the ASR are also investigated. Since the ASR model used in this work takes a single-channel speech signal as the input, the hidden representation can thus only represent this single channel. In order to extend the proposed approach to binaural signals, a better-ear policy is applied in the similarity computation, i.e. regarding the larger score between the predicted left-ear and the right-ear similarities as

the eventual intelligibility score. In this section, the ASR model will first be described in detail. Then the different hidden representations investigated in this work will be introduced. Finally, the similarity computation will be explained.

5.2.1 DNN-based ASR model

Figure 5.2 shows the architecture of the transformer-based ASR model used in this work. It consists of a convolutional neural network (CNN) based PreNet (Han et al., 2020), a transformer-based encoder, and a transformer-based decoder. The PreNet is a stack of convolutional layers for a better understanding of global context. Both the encoder and decoder are composed of a number of transformer blocks. Each encoder transformer block consists of a multi-head self-attention sub-layer and a position-wise fully connected feed-forward sub-layer (Vaswani et al., 2017). A residual connection and layer normalisation (Ba et al., 2016) are applied to both sub-layers. In the multi-head attention sub-layer, the input features are firstly mapped to query Q, key K with embedding length d_k, and value V with embedding length d_v, and the attention mechanism is computed as:

$$Attention\,(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{5.1}$$

The projection and attention mechanisms are run in parallel multiple times, and the concatenation of attention outputs is then multiplied by a linear projection matrix. Compared to the encoder transformer block, an extra multi-head attention sub-layer is inserted to perform the attention mechanism on the encoder output features. In addition, a positional mask is used in the decoder multi-head self-attention sub-layer to enforce that only the known previous decoded outputs are dependent.

The ASR model is optimised with the joint CTC-attention mechanism (Kim et al., 2017), i.e., a combination of Connectionist Temporal Classification (CTC) (Graves et al., 2006) and attention-based sequence-to-sequence (seq2seq) (Chorowski et al., 2015). The CTC leverages repeatable intermediate label representation and a special blank label for ASR decoding, and the loss function can be expressed as:

$$L_{CTC} = -\log\left(\sum_{\pi\in\beta^{-1}(l)}\prod_{m=1}^{M}P(z^m_{\pi_m})\right), \tag{5.2}$$

where β is a function that removes repeated intermediate and blank labels, π_m is the intermediate and blank label sequence, $P(z^m_{\pi_m})$ is the probability of π_m at time m, and l is the target label sequence. The seq2seq loss function is the sum of divergences between the ground

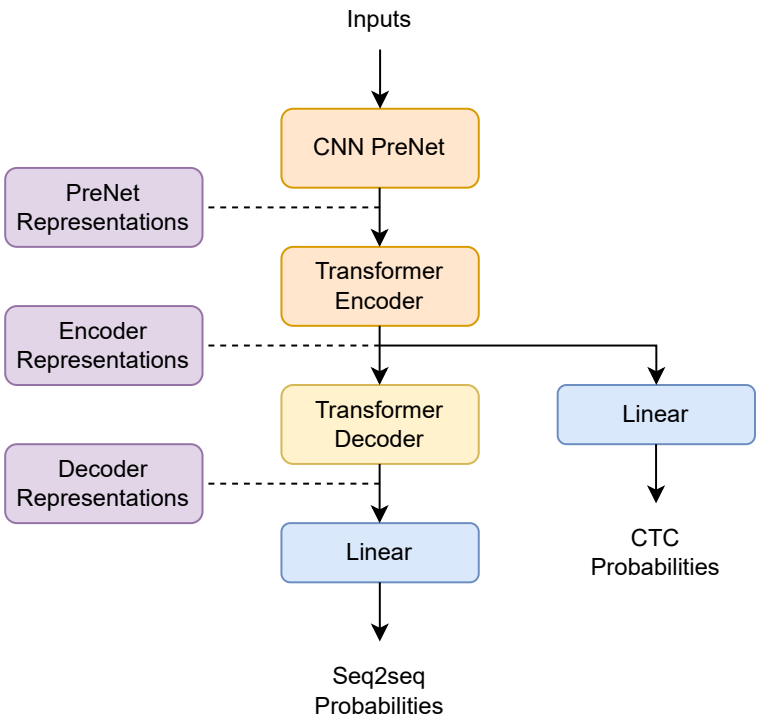


Fig. 5.2 ASR architecture and hidden representations at three different levels.

truth label z_u and predicted token $\hat{z}_u$ at u-th position in the transcript sequence:

$$L_{seq2seq} = \sum_u P(z_u)(\log P(z_u) - \log P(\hat{z}_u)). \tag{5.3}$$

The overall loss function for ASR optimisation is:

$$L = \lambda L_{CTC} + (1-\lambda) L_{seq2seq}, \tag{5.4}$$

where λ is a predefined weighting coefficient.

5.2.2 Hidden representations

The transformer-based ASR model is hierarchical and consists of multiple DNN-based blocks. The knowledge from different levels of the ASR model can be different, and thus can produce different effects for intelligibility prediction. This study investigates three hidden representations, as shown in Figure 5.2, including outputs of the CNN PreNet $H^{pre} \in R^{T^{pre} \times d^{pre}}$, outputs of the transformer encoder $H^{enc} \in R^{T^{enc} \times d^{enc}}$, and outputs of the transformer decoder $H^{dec} \in R^{T^{dec} \times d^{dec}}$. The PreNet representations H^{pre} are viewed as

low-level acoustic features. Meanwhile, the encoder representations h^{enc} can be viewed as high-level acoustic representations, as ASR models using CTC decoding do not learn a language model, and CTC output intermediate labels are independent from each other. In contrast, the seq2seq decoder is usually considered as an internal language model (Meng et al., 2021). Therefore, the decoder representations H^{dec} are viewed as hidden representations with learnt language knowledge. Due to the structure of the transformer-based ASR model, the length of PreNet and encoder representations are determined based on the input signals. In contrast, the decoder representations have indeterminate lengths, which indicates that the lengths of the processed and the reference representations can be different.

5.2.3 Similarity computation

The cosine similarity is used in this work to measure the similarity between two hidden representations, as it is naturally well-scaled to the range from 0 to 1. Given a pair of hidden representations at a single time step, $h, \hat{h} \in R^d$, the cosine similarity is computed as:

$$\rho = \cos(h, \hat{h}) = \frac{h \cdot \hat{h}}{\|h\| \|\hat{h}\|}, \tag{5.5}$$

where $\|\cdot\|$ is the $L2$ norm. For PreNet and encoder representations, the reference and processed representations of each time step are matched. The similarity at each time step ρ_t for the binaural reference and processed representations is computed from the pair of representations at this time step, i.e., $\rho_t = \cos(h_t, \hat{h}_t)$. The overall similarity between the reference and processed representations is computed as:

$$sim(H, \hat{H}) = \frac{1}{T} \sum_{t=1}^{T} \rho_t. \tag{5.6}$$

For decoder representations, the representations of the reference and processed signals could have variable time steps, i.e., T, $\hat{T}$could be different. Therefore, for each pair of sequences of decoder representations $\{H, \hat{H}\}$, the fast dynamic time warping algorithm proposed in (Salvador and Chan, 2007) is applied to find the warped path. The overall similarity is computed as the similarity of the warped pair:

$$sim(H, \hat{H}) = sim(H_w, \hat{H}_w) = \frac{1}{T_w} \sum_{t=1}^{T_w} \cos(H_w(t), \hat{H}_w(t)), \tag{5.7}$$

where H_w and $\hat{H}_w$ are the warped representations, T_w is the total time steps after warping.

The similarity computation for single-channel signals are introduced above. For the binaural signals, the better-ear policy is applied, i.e., the maximal similarity among the reference and processed single channel pairs $\{h^l, \hat{h}^l\}$, $\{h^l, \hat{h}^r\}$, $\{h^r, \hat{h}^l\}$, $\{h^r, \hat{h}^r\}$ at each time step is selected as the overall similarity at this time step. For PreNet and encoder representations, the overall similarity between the binaural reference and processed representations is therefore computed as:

$$sim(H^{bi}, \hat{H}^{bi}) = \frac{1}{T} \sum_{t=1}^{T} \max \left\{ \rho_t^{ll}, \rho_t^{lr}, \rho_t^{rl}, \rho_t^{rr} \right\}. \tag{5.8}$$

Similarly for the decoder representations, all four representations of the reference and processed signals could have variable time steps, i.e., T^l, T^r, $\hat{T}^l$, $\hat{T}^r$ could be different. And the overall binaural similarity is then computed as:

$$sim(H^{bi}, \hat{H}^{bi}) = \max \left\{ sim(H_w^l, \hat{H}_w^l), sim(H_w^l, \hat{H}_w^r), sim(H_w^r, \hat{H}_w^l), sim(H_w^r, \hat{H}_w^r) \right\}. \tag{5.9}$$

5.3 Experimental setup

5.3.1 Databases

The experiments are conducted on two very different databases: the Noisy Grid corpus (Barker and Cooke, 2007) and the first round Clarity Prediction Challenge (CPC1) corpus (Barker et al., 2022). Both of them contain a large number of degraded speech signals and their corresponding references, together with the recognition results of human listeners. The listeners in the Noisy Grid corpus are normal hearing, whereas those in the CPC1 are hearing impaired. Utterances in the Noisy Grid corpus are single-channel, and strictly controlled in terms of speech material and noise levels. In addition, no speech enhancement is applied to speech signals in the Noisy Grid corpus. In contrast, utterances in the CPC1 corpus are binaural, and generated to simulate everyday domestic scenes. Furthermore, various speech enhancement algorithms are applied to these utterances.

5.3.2 ASR configuration

The SpeechBrain (Ravanelli et al., 2021) LibriSpeech transformer ASR recipe is used in this work. 80-channel log mel-filterbank coefficients are used as inputs with a 25 ms window with a stride of 10 ms. The sampling rate of the input signals is 16k Hz covering the mel-filterbank features from 0 to 8k Hz. The PreNet consists of three 2D convolutional layers, and the encoder and the decoder consist of 12 and 6 transformer blocks, respectively. The weighting

coefficient λ is set 0.3 for training, and 0.4 for decoding. The dimensions at the one-time step for PreNet, encoder, and decoder hidden representations are 10 240, 768, and 768, respectively.

The ASR model used for intelligibility prediction is from the SpeechBrain released model, which is trained on the 960-hour LibriSpeech database (Panayotov et al., 2015). Therefore, the model is of the knowledge for well-performed clean speech recognition. Furthermore, the ASR model is finetuned on the experimental databases to incorporate the knowledge of degraded speech. Unless stated otherwise, the ASR model is finetuned for ten epochs on the training set. For the noisy Grid corpus, the inference of the ASR model is strictly constrained within the Grid grammar, as was also the case for the listening experiments.

5.3.3 Evaluation

Three performance evaluation measures, including root mean square error (RMSE), normalised cross-correlation coefficient (NCC), and Kendall's Tau coefficient (KT), are exploited as the evaluation metrics to measure the correlation between the predicted intelligibility and the listener word correctness scores (WCS). The WCS is computed as the words that are correctly recognised divided by the total number of words in an utterance, which is regarded as a proxy for intelligibility. As the first two metrics RMSE and NCC could be invalid for non-linear correlations, a logistic function $f(x) = 1/[1+\exp(ax+b)]$ is applied to the predicted intelligibility to examine the monotonic relation, following the conventions of previous works, including Andersen et al. (2018a); Taal et al. (2011). Each database consists of a training set, a development set, and an evaluation set. The ASR model is trained with the data in the training set. The parameters a and b of the logistic function are optimised on the development set. And the fitted predictions are evaluated on the evaluation set.

5.4 Monaural speech in SSN with normal hearing listeners

5.4.1 Corpus description

The Noisy Grid corpus is an extension to the original Grid corpus (Cooke et al., 2006) with added speech-shaped noise (SSN) at 11 different SNR levels from -14 dB to 6 dB, plus one at 40 dB. Each Grid utterance consists of six words following the structure of "command-color-preposition-letter-digit-adverb", and the words are randomly selected within a limited vocabulary of $[4,4,4,25,10,4]$ words for each sentence location, respectively. The listeners are asked to identify "color", "letter", and "digit" in the listening tests, therefore the WCS for each utterance can only be $[0,1/3,2/3,1]$. In order to make the distribution of WCS

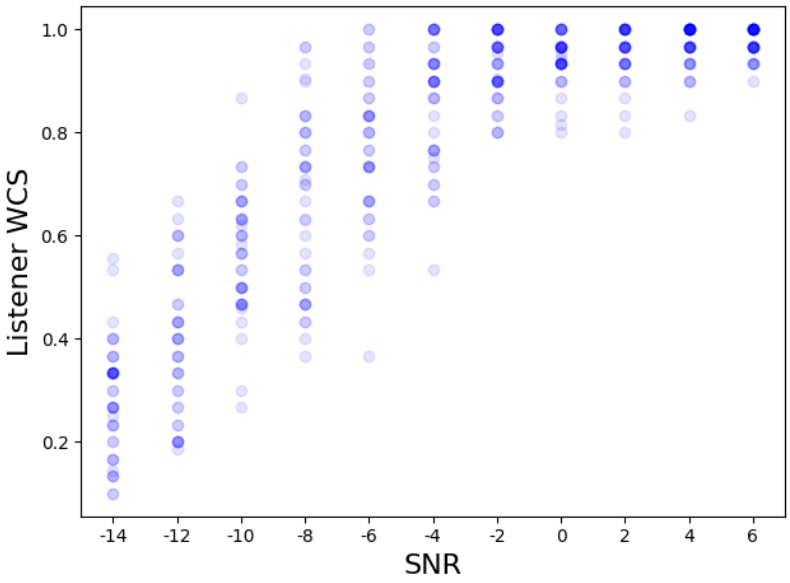


Fig. 5.3 Scatter plot of the listener word correctness scores (WCS) distribution in the evaluation set at different SNR levels. The opaqueness is correlated with the density.

relatively more continuous, the reported WCS is averaged over ten utterances at the same SNR level. The added SSN is created by shaping Gaussian noise so that its long-term average spectrum is the same as that of the average of speech signals within the clean Grid corpus. The human listening results are conducted by 20 normal-hearing listeners. The original database comprises utterances spoken by 34 speakers, and all the listening test results are reported in Barker and Cooke (2007). In this work, the database is divided into the training set for ASR optimisation consisting of the utterances from 22 speakers, the development set consisting of the utterances from 6 speakers, and the evaluation set consisting of utterances from 6 speakers. Figure 5.3 shows the listener WCS distribution at different SNR levels.

5.4.2 Baselines

A number of intrusive intelligibility predictors are used as the baselines in the experiment. These baselines are all widely-used and well-recognised, and predict intelligibility by comparing a degraded speech signal and its corresponding reference signal. In addition, the ASR recognition WCS is also used as one of the baselines. As in the proposed approach, the intelligibility scores predicted by the baselines are fitted by a logistic function, and the fitting parameters are optimised on the development set.

Coherence speech intelligibility index

The coherence speech intelligibility index (CSII) (Kates and Arehart, 2005) is a widely-used variant of the speech intelligibility index (SII). The CSII replaces the signal-to-noise ratio (SNR) of each frequency band with the signal-to-distortion ratio (SDR). The SDR is estimated with the coherence function, which is the correlation in the frequency domain. Three values representing different amplitude levels of CSII, i.e. low-CSII, mid-CSII, and high-CSII, are computed and linearly combined to achieve the overall prediction using the coefficients in Kates and Arehart (2005).

Normalised covariance measure

The normalised covariance measure (NCM) (Goldsworthy and Greenberg, 2004) is an improved version of the speech transmission index (STI) (Steeneken and Houtgast, 1980). To measure the NCM, a group of band-pass filters is applied to both the degraded and the reference signals to extract the temporal envelopes. The normalised covariance between the degraded and reference envelope at each band is then measured and converted to an apparent SNR. These SNR values are clipped, and at last combined subject to a band-wise weighting function.

Short-time objective intelligibility measure

The short-time objective intelligibility (STOI) measure (Taal et al., 2011) takes advantage of short-time temporal envelope segments with a duration of 386 ms, which is suggested as the optimal duration for intelligibility prediction (Drullman et al., 1994; van den Brink, 1964). The segments extracted from the degraded signals are normalised and clipped so that the SDR is higher than 15 dB. The predicted intelligibility is computed as the mean of all the NCC between the reference and degraded segments across both time and frequency bands.

Extended short-time objective intelligibility measure

The extended short-time objective intelligibility (ESTOI) measure was proposed in Jensen and Taal (2016) to improve the performance of STOI in the situation where *modulated* noise sources are present. The ESTOI computes the spectral correlation between the degraded and reference signals, instead of the correlation of the envelope segments. In addition, the clipping in STOI is removed.

Table 5.1 Evaluation results on the Noisy Grid corpus in terms of RMSE, NCC, and KT. The down arrow indicates the smaller the better, and the up arrows indicate otherwise.

	RMSE ↓	NCC ↑	KT ↑
CSII	0.100	0.928	0.766
NCM	**0.083**	**0.950**	**0.801**
STOI	0.146	0.850	0.671
ESTOI	0.103	0.926	0.761
SIIB	0.131	0.877	0.691
HASPI	0.197	0.716	0.526
ASR WCS	0.139	0.854	0.697
PreNet representations	0.129	0.905	0.726
Encoder representations	0.129	0.915	0.747
Decoder representations	0.115	0.923	0.761

Speech intelligibility in bits

The speech intelligibility in bits (SIIB) (Van Kuyk et al., 2017) predicts intelligibility based on information theory. A mutual information estimator is used to estimate the mutual information in bits between the representations of the degraded and reference speech signals. The representations are extracted by an auditory model that simulates both time and frequency masking (Rhebergen et al., 2006; Slaney et al., 1993). The SIIB used in this work is the modified version using a Gaussian channel proposed in Van Kuyk et al. (2018). As it is suggested that the duration of input signals to SIIB should be larger than 20 seconds, the Grid signals are all repeated 20 times and concatenated.

Hearing aid speech perception index

The hearing aid speech perception index (HASPI) version 2 (Kates and Arehart, 2021) incorporates an elaborate auditory model, that can simulate hearing impairment, to extract estimated envelopes of the degraded and reference signals. The cepstral coefficient correlations at a number of modulation rates are then computed and averaged over a group of basis functions. The results are then fed into an ensemble of neural networks to generate the predicted intelligibility. As the listeners in the Noisy Grid corpus listening experiments are normal hearing, the hearing thresholds input to HASPI are set as zeros.

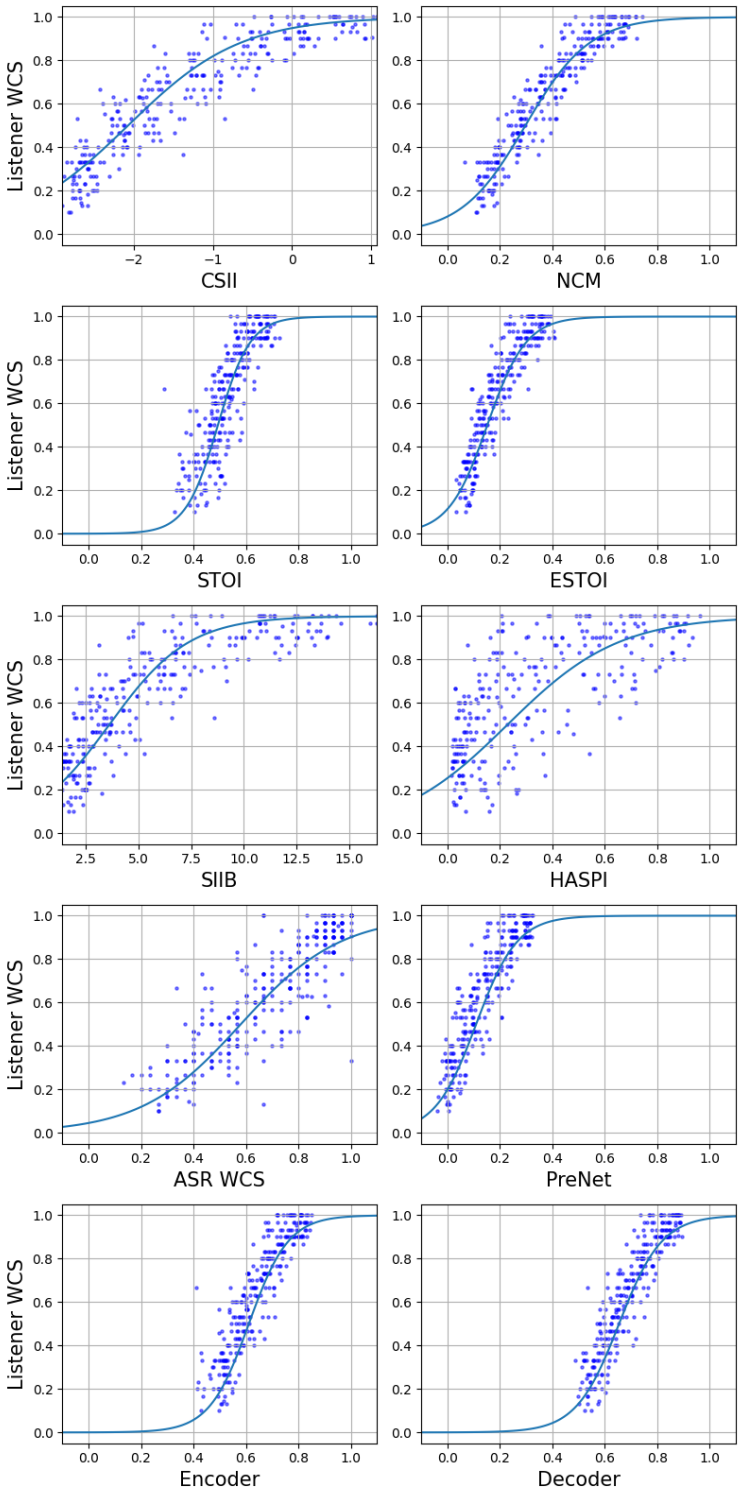


Fig. 5.4 Scatter plots of all intelligibility predictions on the Grid corpus evaluation set, along with the logistic fitting functions.

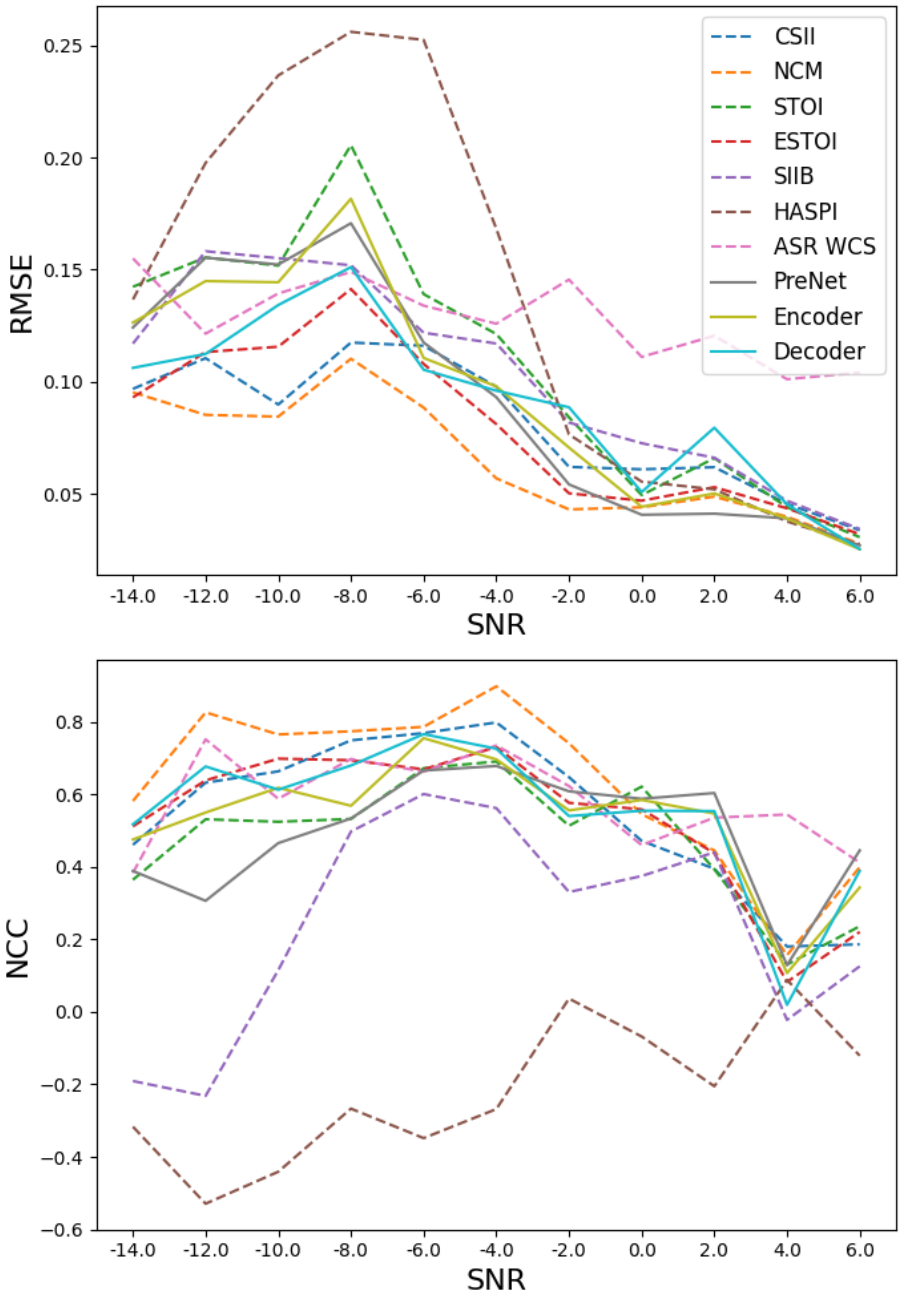


Fig. 5.5 RMSE and NCC of the intelligibility predictors at different SNRs. The dashed lines represent the baseline approaches, and the solid lines represent the proposed ASR hidden representation-based approaches.

5.4.3 Results

It is observed that over 90% of utterances, whose SNRs are equal to or higher than 0 dB, have perfect WCS in the listening tests. In order to even the distribution of the database, only the results of utterances whose SNRs are lower than 0 dB are reported. The overall evaluation results of the baselines and proposed ASR hidden representation-based intelligibility predictors are listed in Table 5.1. Additionally, the scatter plots of the predictions against the listeners' recognition results are shown in Figure 5.4. Other than that, the RMSE and NCC of the predictors at different SNRs are shown in Figure 5.5.

NCM performs the best in terms of all three evaluation metrics, followed by CSII. Essentially, both these two approaches measure the SNRs at a number of frequency bands and combined the SNRs subject to a frequency-weighted function. For speech degraded by stationary noise, as is the case of Noisy Grid corpus, NCM and CSII with relatively simple mechanisms can perform exceptionally well, thanks to the frequency-weighted functions, which reflect the importance of different frequency bands to human speech perception.

ESTOI, which measures the spectral normalised correlation, can also make accurate predictions, as it is proposed to improve the performance over STOI when speech is degraded by modulated noise sources. Meanwhile, the mutual information-based SIIB, which estimates how much information is shared between the clean reference signal produced by a speaker and the degraded signal received by a listener, can also achieve competitive results. However, HASPI performs poorly even though it incorporates a detailed auditory model. One reason could be that the prediction of HASPI heavily relies on an ensemble of neural networks, which are specifically optimised towards other intelligibility databases listed in Kates and Arehart (2021) and do not generalise well to the Noisy Grid corpus.

Although the performance of ASR WCS is fairly competitive, it is outperformed by the proposed ASR hidden representation-based approaches. This suggests that the similarities of hidden representations can be a better predictor for speech intelligibility compared to ASR recognition performance. By comparing the performance of different levels of representations, it can be observed that high-level representations can support more accurate intelligibility predictions. In addition, as shown in the scatter plots in Figure 5.4, the similarities between the reference and degraded signals of higher-level representations are stronger. This indicates that the ASR model can naturally extract the features that are closely related to recognition with the growth of network depth.

It can be observed in Figure 5.5 that the RMSE values of most intelligibility predictors at SNR levels equal to or higher than 0 dB are minimal, because both the predicted intelligibility and human recognition results are almost perfect. It is also clear that there is a peak for almost every predictor at the -8 dB SNR in terms of RMSE, because the listener WCS at

-8 dB is most widely distributed as shown in Figure 5.3. In addition, there is a clear trend that the RMSE values decline from -8 db to 0 dB, where the listener WCS clearly increases and is distributed approximately in the range of 0.4 to 1.0. Interestingly, the RMSE of ASR WCS is less affected by SNR levels. Despite the fact that the ASR WCS is similar to the human WCS in low SNR ranges, it still makes recognition mistakes in high SNR ranges where listeners can achieve almost perfect recognition performance.

For NCC at different SNR levels also shown in Figure 5.5, most approaches have slight improvement from -14 dB to -2 dB. The sudden drops from 0 dB could be due to the ceiling effect, i.e. the most listener WCS is 1 and the predictions can be noisily distributed slightly smaller than 1. Apart from the poorly performing HASPI, ASR WCS is still an outlier in terms of its NCC score. Its NCC is relatively flat across SNR levels for the same reason that its RMSE scores are poorer in the high SNR range.

5.5 Processed binaural speech in domestic noise with hearing impaired listeners

5.5.1 Corpus description

CPC1 (Barker et al., 2022) provides a large number of processed binaural speech signals by machine learning hearing-aid systems and the corresponding responses from hearing impaired listeners. Each signal represents a simulated mixture of a target speech and an interfering noise within a simulated cuboid-shaped living room, enhanced by a hearing-aid system given the audiogram (i.e., a pure-tone measure of hearing thresholds at different frequencies) of a listener. Both the binaural processed signals and the corresponding anechoic reference signals are provided. The ground truth intelligibility is presented as the listener WCS. A total of 6 speakers, 10 hearing aid systems and 27 listeners are included.

The CPC1 includes two tracks: (1) *closed-set*, that is the listeners and systems in the evaluation set are overlapped with those in the training data; (2) *open-set*, that is the systems or listeners in the evaluation set are not included in the training data. For both tracks, the scenes in the training data are split into 70% and 30% as a training set and a development set, and the results on the evaluation data are reported. The experimental details can be found in the CPC1 overview paper (Barker et al., 2022).

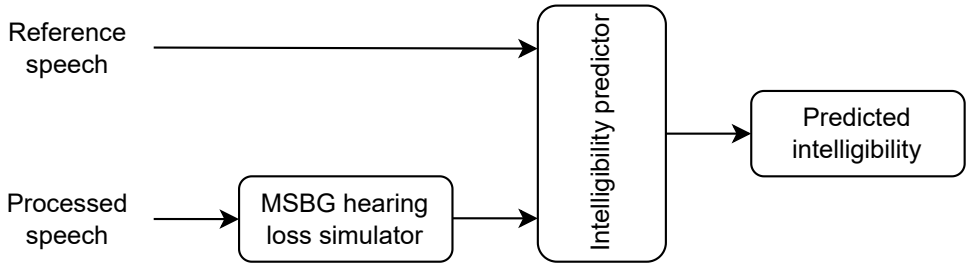


Fig. 5.6 Intelligibility prediction for hearing impaired listeners with the MSBG hearing loss simulator.

5.5.2 Baselines

The baseline approaches in the Noisy Grid corpus are also used as the baselines for CPC1. However, most of these approaches cannot take hearing abilities into consideration and can process only monaural signals. In order to simulate the hearing impairment of the listeners, the Cambridge MSBG hearing loss simulator (Baer and Moore, 1993, 1994; Moore and Glasberg, 1993; Stone and Moore, 1999) is applied to the processed speech signals given listeners' audiograms, as shown in Figure 5.6. For the purpose of leveraging binaural information, a simple but effective better ear (BE) policy is used. Similarly to the binaural setup introduced in Section 5.2.3, the maximal predicted intelligibility of the four pairs of processed and reference signals, i.e., left processed and left reference, left processed and right reference, right processed and left reference, right processed and right reference signal pairs, is regarded as the overall prediction. Therefore, BE-CSII, BE-NCM, BE-STOI, BE-ESTOI, and BE-SIIB can be used applied to CPC1 signals as the baselines. The combination of MSBG simulation and BE is also applied to ASR recognition WCS, so that it can be used as another baseline approach.

The aforementioned baselines leverage the MSBG model to simulate hearing impairment. Meanwhile, HASPI itself incorporates a well-designed auditory model that simulates hearing impairment. Therefore, the BE-HASPI can make intelligibility predictions for CPC1 binaural speech given different audiograms. In addition, the CPC1 introduces its official baseline as the combination of the MSBG hearing loss model and modified binaural STOI (MBSTOI) (Andersen et al., 2018a). MBSTOI is the modified version of the deterministic binaural STOI (Andersen et al., 2016) and can take advantage of binaural cues through an equalization-cancellation stage (Durlach, 1972).

Table 5.2 Evaluation results on both CPC1 *closed-set* and *open-set* in terms of RMSE, NCC, and KT.

	RMSE ↓	NCC ↑	KT ↑
Closed-set			
BE-CSII	0.287	0.615	0.412
BE-NCM	0.289	0.607	0.388
BE-STOI	0.273	0.662	0.421
BE-ESTOI	0.253	0.719	0.446
BE-SIIB	0.303	0.566	0.364
MBSTOI	0.285	0.621	0.398
BE-HASPI	0.254	0.717	0.445
ASR WCS	0.250	0.729	**0.523**
PreNet representations	0.347	0.299	0.182
Encoder representations	0.237	0.758	0.487
Decoder representations	**0.231**	**0.773**	0.498
Open-set			
BE-CSII	0.404	0.493	0.358
BE-NCM	0.308	0.580	0.409
BE-STOI	0.371	0.559	0.418
BE-ESTOI	0.294	0.640	0.466
BE-SIIB	0.336	0.521	0.387
MBSTOI	0.365	0.529	0.391
BE-HASPI	0.267	0.676	0.469
ASR WCS	0.250	0.723	**0.534**
PreNet representations	0.356	0.254	0.136
Encoder representations	0.241	0.751	**0.534**
Decoder representations	**0.235**	**0.763**	0.530

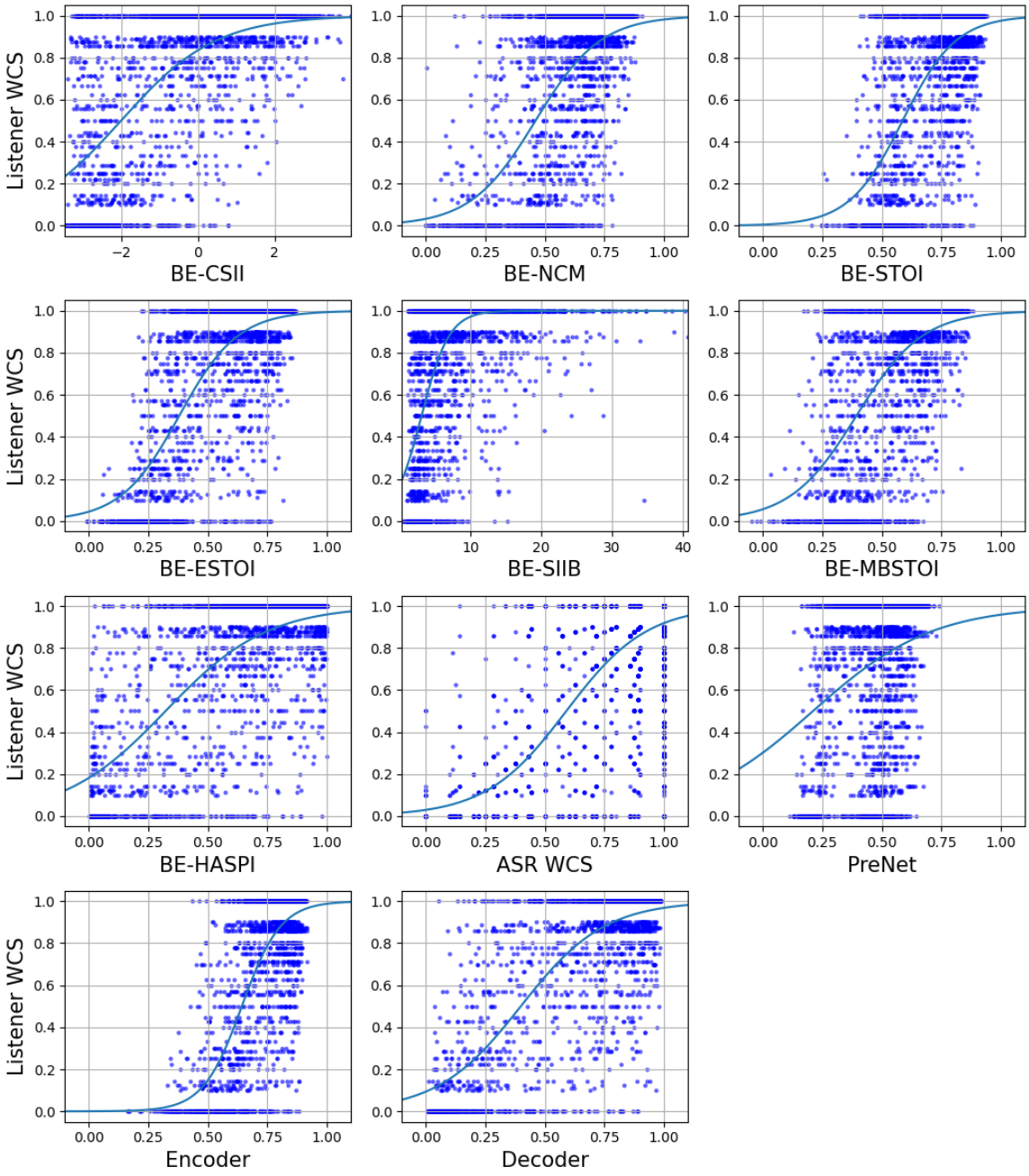


Fig. 5.7 Scatter plots of all intelligibility predictions on the CPC1 *closed-set* evaluation set, along with the logistic fitting functions.

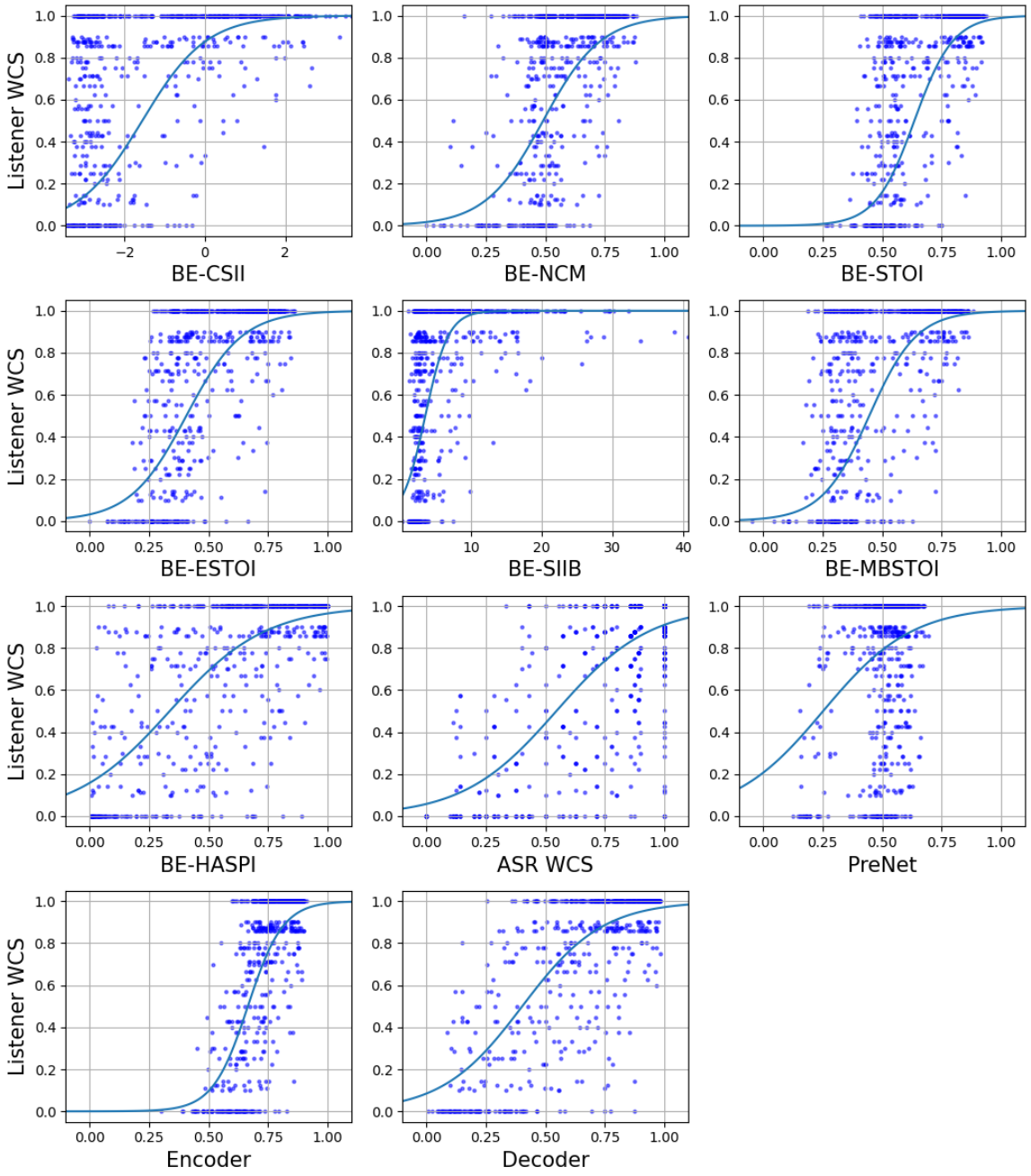


Fig. 5.8 Scatter plots of all intelligibility predictions on the CPC1 *open-set* evaluation set, along with the logistic fitting functions.

5.5.3 Results

ASR training

The training of the default ASR model starts from the pretrained model on the LibriSpeech[1] (LS). Therefore, it has a detailed knowledge of clean speech. Furthermore, it is optimised with the LS *train-clean-100* set combined with noises from the training set in the first round Clarity Enhancement Challenge (Graetzer et al., 2021) (CLS) for ten epochs. Finally, the ASR model is optimised on the CPC1 training set for another ten epochs. In addition, the MSBG hearing loss model is used to process the signals when training and testing on CPC1.

Overall results

Table 5.2 presents the performance of various baselines and the proposed ASR hidden representation-based approaches. Figure 5.7 and Figure 5.8 show the listener WCS against the predicted intelligibility by the baselines and proposed approaches with their corresponding logistic mapping functions for the closed-set and open-set. Generally, the predicted intelligibility is less accurate for the open-set, because neither the listener nor the speech enhancement systems in the evaluation set are seen in the training set. Meanwhile, the rankings for different approaches are similar for the closed- and open-set.

In contrast to the Grid corpus results, the performances of BE-NCM and BE-CSII are not as competitive in the CPC1 database. Especially, BE-CSII performs the worst for the open-set. The two classic approaches NCM and CSII are simple yet effective for speech in stationary noises, but can fail to make accurate intelligibility predictions for speech in complex environments and being processed by enhancement systems. This is not surprising as they are designed for stationary degradation and a limited number of types of non-linear processing.

On the contrary, BE-ESTOI shows relatively satisfactory results for the CPC1 database. As ESTOI performs also well in the noisy Grid corpus as shown in Figure 5.1, it shows its consistency in intelligibility prediction for very different speech signals. However, both BE-STOI and MBSTOI fail to reach as competitive performances as ESTOI, and their evaluation results on both closed- and open-sets are similar. This indicates the equalization-cancellation stage in MBSTOI might not bring many benefits. The major difference between ESTOI and STOI is that ESTOI computes the spectrally normalised correlation rather than the envelope segment correlation. Therefore, the correlation between the reference and degraded signals in the frequency domain can be more crucial to intelligibility prediction than in the time domain.

[1] huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

It can also be observed that BE-HASPI can make relatively accurate intelligibility predictions in the CPC1 database. Given that it performs very poorly in the noisy Grid corpus, the contribution to the good CPC1 performance could be due to the incorporated elaborate hearing loss model.

Both the results of the closed-set and open-set indicate that the proposed ASR high-level hidden representations-based approach could outperform the baselines at intelligibility prediction in terms of RMSE and NCC. The ASR WCS predictions are advantageous with regard to KT because WCS is discrete, i.e., in which case *tied* pairs are more likely to appear. Interestingly, the low-level representations from the PreNet can achieve even better results than ASR WCS in the noisy Grid corpus, but perform poorly for CPC1. Between the two high-level hidden representations, the decoder ones including language model knowledge are better than the encoder ones which represent high-level acoustic features in terms of RMSE and NCC, while the KT scores are close. It can also be observed in the scatter plots that the similarities of the decoder representation are more spread than those of the encoder representation, i.e., the similarity between a severely degraded speech signal and its reference signal can be very low.

Data mismatch

For the purpose of further investigating the influence of data mismatch (i.e., different distribution of training and evaluation data) on ASR models for intelligibility prediction, four different ASR models with different training data knowledge (LS, LS+CLS, LS+CPC1, LS+CLS+CPC1) are probed. The MSBG model is used for all models as preprocessing for hearing loss simulation. ASR models trained on CLS can be considered to have knowledge of noisy speech, and those trained on CPC1 can be considered to have knowledge of hearing-aid processed noisy speech. The correlations between the predicted intelligibility with decoder representations and the ground truth WCS are shown in Table 5.3. The results show that the ASR models trained with CPC1 training data (LS+CPC1, LS+CLS+CPC1) could make optimal predictions, while the latter one is slightly better in terms of RMSE and NCC because it has knowledge of noisy speech. Meanwhile, the ASR models with no knowledge of CPC1 data (LS, LS+CLS) could also achieve competitive results. It is worth noting that the ASR model trained only on clean LS signals could still outperform the baseline system.

Hearing loss simulation

The influence of the MSBG hearing loss model is also investigated. The intelligibility prediction results of ASR models trained on LS+CLS+CPC1 with and without the MSBG

Table 5.3 Evaluation results on the *closed-set* of decoder representations from different ASR models.

MSBG	Training data	RMSE ↓	NCC ↑	KT ↑
with	LS	0.264	0.692	0.449
	LS+CLS	0.243	0.746	0.464
	LS+CPC1	0.233	0.768	**0.503**
	LS+CLS+CPC1	**0.231**	**0.773**	0.498
w/o	LS+CLS+CPC1	0.234	0.767	0.476

model for hearing loss simulation are also shown in Table 5.3. The results indicate that the MSBG hearing loss model can offer a slight advantage in hearing impaired intelligibility prediction for the ASR hidden representations.

Listener- and system-wise correlation

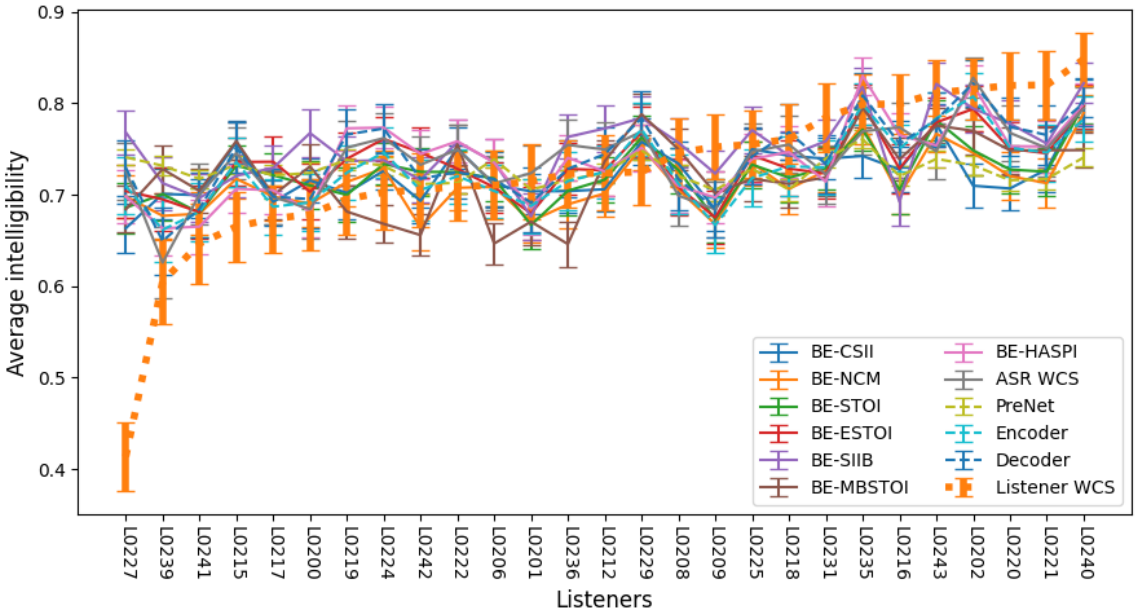
The results of the listening experiments provided by CPC1 can be noisy, because of the not strictly constrained speech materials, the large size vocabulary, etc. Therefore, both the listener WCS and the predicted intelligibility scores are averaged on listeners and hearing-aid systems for more conclusive analysis. The average listener WCS, the predicted intelligibility from the baselines, and the proposed ASR hidden representation similarity from the ASR trained on LS+CLS+CPC1 with the MSBG hearing loss model on different listeners and hearing-aid systems with their corresponding error bars are shown in Figure 5.9a and Figure 5.9b. The listener- and system-wise evaluation results on the *closed-set* are measured and shown in Table 5.4. It can be observed that BE-HASPI can perform exceptionally well for both listener- and system-wise intelligibility prediction. This is opposite to its performance on the noisy Grid corpus, which provides the WCS from normal hearing listeners. Therefore, it is sensible to believe that the hearing loss simulation incorporated by HASPI is pretty reliable. Meanwhile, the proposed decoder representation performs the best for system-wise intelligibility prediction in terms of RMSE.

5.6 Conclusions

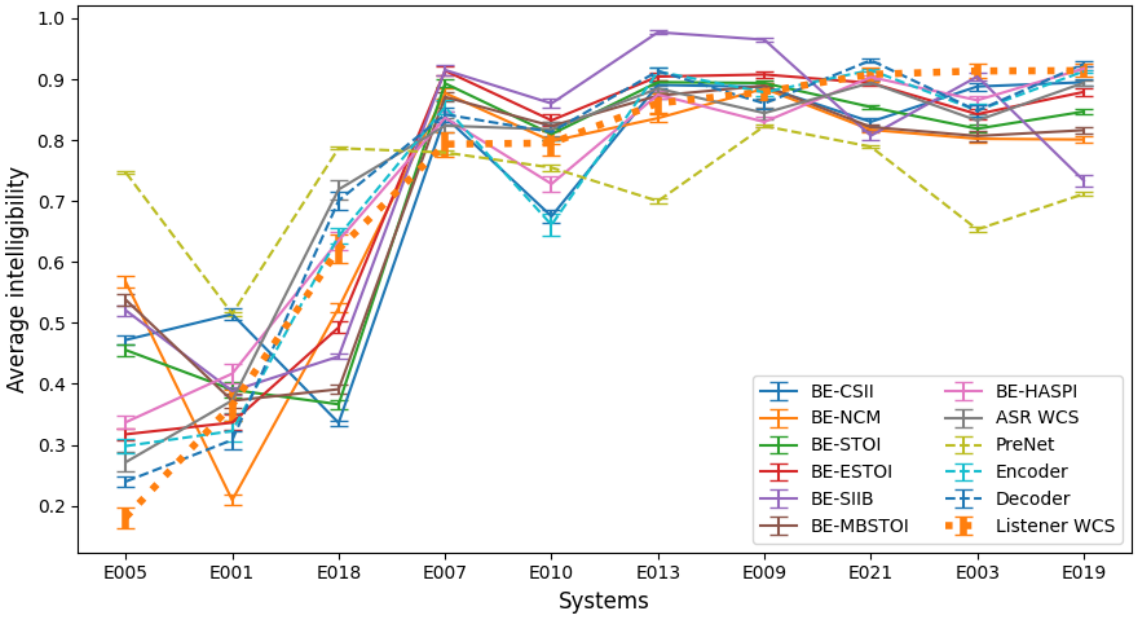
In this chapter, ASR hidden representations have been investigated for intelligibility prediction. In detail, the similarity between hidden representations of a degraded speech signal and its corresponding reference signal extracted by a DNN-based ASR model is regarded as an intelligibility indicator. This proposed approach has shown its efficacy in the experi-

Table 5.4 Listener- and system-wise evaluation results on the *closed-set* of predicted intelligibility.

	RMSE ↓	NCC ↑	KT ↑
Listener-wise			
BE-CSII	0.072	0.575	0.356
BE-NCM	0.076	0.490	0.373
BE-STOI	0.073	0.541	0.373
BE-ESTOI	0.073	0.529	0.362
BE-SIIB	0.084	0.355	0.385
MBSTOI	0.078	0.412	0.299
BE-HASPI	**0.070**	**0.593**	0.425
ASR WCS	0.074	0.515	0.430
PreNet representations	0.087	-0.043	-0.002
Encoder representations	0.071	0.571	**0.510**
Decoder representations	0.073	0.540	0.459
System-wise			
BE-CSII	0.146	0.801	0.644
BE-NCM	0.150	0.792	0.289
BE-STOI	0.130	0.846	0.289
BE-ESTOI	0.078	0.948	0.378
BE-SIIB	0.151	0.794	0.244
MBSTOI	0.147	0.798	0.244
BE-HASPI	0.062	**0.984**	**0.778**
ASR WCS	0.053	0.983	0.733
PreNet representations	0.229	0.342	0.022
Encoder representations	0.067	0.962	0.689
Decoder representations	**0.049**	0.982	0.733



(a) Listener-wise



(b) System-wise

Fig. 5.9 Listener- and system-wise average intelligibility with standard errors on the *closed-set*.

ments on the noisy Grid corpus, which contains the responses of normal hearing listeners on speech with additive stationary noises, and the CPC1 database, which contains the responses of hearing impaired listeners on simulated domestic noisy speech processed by complex enhancement algorithms.

The experimental results show that classic intelligibility prediction approaches NCM and CSII can achieve the most accurate prediction for speech in stationary SSN, but they fail on simulated domestic noisy speech with complex non-linear processing. Meanwhile, ESTOI, the improved version of arguably the most widely used STOI, can achieve consistently high performance on both databases. In addition, HASPI, which incorporates an elaborate auditory model, is found that can make accurate intelligibility predictions for hearing impaired listeners.

On the other hand, the proposed ASR hidden representation-based intelligibility prediction can achieve competitive performance on the noisy Grid corpus, and the best on the CPC1 database. In addition, the experimental results show it can be better than using ASR recognition results for intelligibility prediction. Detailed analysis shows that the high-level hidden representations, which also contain the language model knowledge, can achieve the best performance.

Chapter 6

Non-intrusive Intelligibility Prediction with Unsupervised ASR Uncertainty

6.1 Introduction

Accurate intelligibility prediction has always been of great interest for its importance in developing speech enhancement-related applications. Intelligibility prediction is a naturally *unobtrusive* task, i.e., when humans are judging whether a signal is intelligible or not, they are not doing so with respect to some external reference. Meanwhile, when predicting the intelligibility of a degraded speech signal, additional inputs, such as a corresponding reference clean speech, and transcription of the degraded speech, are typically required as seen in the previous chapter. However, these intrusive prediction approaches can be difficult to be applied in real-world scenarios. In these scenarios, it can be expensive or impractical to obtain additional inputs. Therefore, *non-intrusive* intelligibility prediction, which only requires the degraded speech itself, has been a growing research topic.

Conventional non-intrusive intelligibility prediction takes advantage of acoustic features, such as the speech to reverberation modulation energy ratio (SRMR) (Falk et al., 2010) and the across-band envelope correlation metric (ABECm) (Chen, 2016b). These acoustic features of degraded speech are observed to be correlated to intelligibility in certain scenarios and thus be used for prediction. Different approaches have used different acoustic features, for example, SRMR is based on reverberation characteristics, while ABECm uses the average envelope correlation of adjacent bands. One major disadvantage of these methods is that the application is usually limited to a certain scenario, e.g., SRMR is designed for reverberant and dereverberated speech, and part of the ABECm foundation is based on the assumption that these correlations are important for recognising noise-vocoded speech. Another group of non-

intrusive approaches try to generate *pseudo* clean reference signals, and can be considered as variants of intrusive prediction methods, such as Andersen et al. (2017), Sørensen et al. (2017a), and Karbasi et al. (2016). A clean feature estimation model is usually constructed and used to produce an estimated reference for computing STOI-like scores. Therefore, clean signals are usually required to optimise the estimation model. Recently, a number of data-driven methods are proposed, such as Andersen et al. (2018b); Sharma et al. (2016); Zezario et al. (2020). These methods train a classification and regression tree or neural networks to predict intelligibility from features of noisy signals, therefore requiring a number of expensive human listening results or scores from intrusive predictors like STOI. These approaches are limited by the quality of training data and the intrusive prediction results.

Apart from the aforementioned approaches, another promising candidate for non-intrusive intelligibility prediction is to take advantage of ASR models, such as Holube and Kollmeier (1996); Jürgens and Brand (2009); Karbasi et al. (2022); Spille et al. (2018a), given that they can perform similarly to human speech recognition in terms of recognition patterns in certain situations (Barker and Cooke, 2007; Fontan et al., 2017; Schädler et al., 2015). It is natural that the recognition results of an ASR model can be used as an intelligibility predictor. However, these approaches are not entirely non-intrusive as some forms of reference are still used. For example, transcription is needed to calculate the recognition correctness, and temporal alignment is needed if the task is phoneme recognition. Furthermore, the recognition results are not necessarily a good prediction of speech intelligibility, e.g., an ASR model can sometimes make a correct *guess* even if the intelligibility of the speech is low. Therefore, non-intrusive ASR-based intelligibility models are desired.

Early non-intrusive ASR-based approaches turn to dynamic time warping ASR so that no transcripts are needed (Holube and Kollmeier, 1996; Jürgens and Brand, 2009). Specifically, representations of a test degraded speech signal extracted by a designed perception model are compared with those of a number of prior template speeches. The minimal distance between the test representations and template representations is used to correlate with intelligibility. It is also found that the approach performs best when the test and prior speech material are identical, which is then similar to intrusive prediction. Recently, there is growing attention to leveraging ASR-derived measures for non-intrusive prediction. Roßbach et al. (2022) exploited mean temporal distance to capture the temporal smearing effect (Hermansky et al., 2013) in the phoneme posteriogram generated by an ASR model, and map the distance to intelligibility. Karbasi et al. (2022) investigated a number of word-level posterior-related measures for microscopic intelligibility prediction. These approaches optimised ASR-derived features and optimised a mapping model with training data pairs, i.e., degraded speech

features and human intelligibility scores, and can generalise the prediction ability to the degraded speech in the evaluation.

Uncertainty of speech recognition is similar to the definition of speech intelligibility, which can be characterised by the probability of correct word recognition (Allen, 1995). Meanwhile, the ASR uncertainty is associated with the probability of models making correct predictions (Kalgaonkar et al., 2015). The uncertainty estimation is crucial for ASR application as it can help improve robustness in critical tasks, i.e., it is extremely important to assess reliability or probability of correctness for decisions made by ASR models (Jiang, 2005). A large number of recent ASR uncertainty estimation methods construct and optimise an estimator on top of the original ASR model to predict the uncertainty of a given utterance (Kalgaonkar et al., 2015; Ragni et al., 2018; Swarup et al., 2019). Recently, a word-level ASR uncertainty estimation method is also proposed by Oneaţă et al. (2021). Most approaches train the uncertainty estimation model with supervision, that is the ratio of recognition mistakes made by an ASR model is needed. Meanwhile, a sequence-level uncertainty estimation method for auto-regressive structured prediction tasks is proposed by Malinin and Gales (2021).

Motivated by the connection between ASR uncertainty and speech intelligibility, this chapter investigates how to estimate the uncertainty of an ASR model and correlate it to intelligibility. The uncertainty of a model is typically estimated with supervision, i.e., to train a model with the uncertainty labels, which are the human intelligibility scores in this work. However, the supervised approaches share the same problem with data-driven intelligibility prediction, which is the difficulty to obtain expensive human listening recognition labels. Therefore, this chapter proposes an unsupervised ASR uncertainty estimation method, which does not require intelligibility labels to make predictions. In addition, most ASR-based approaches focus on word- or phoneme-level prediction, and are limited to matrix tests, i.e., limited vocabulary and grammar in the speech material. It is desirable to propose an ASR-based approach that can perform well for sequence-level everyday speech. In this chapter, an unsupervised sequence-level uncertainty estimation is formulated, and the estimated uncertainty is used to correlate speech intelligibility. It is explored how accurately the ASR uncertainty is able to model intelligibility, compared to a range of intelligibility predictors, including not only non-intrusive approaches but also widely used intrusive approaches. The experimental materials cover matrix speech with additive noises for normal hearing listeners, and processed simulated speech in domestic noises for hearing impaired listeners.

This chapter is organised as follows. Section 6.2 presents the derivation of unsupervised uncertainty estimation. Section 6.3 very briefly introduces the experimental setup which is similar to that used in the previous chapter, allowing a better comparison of intrusive

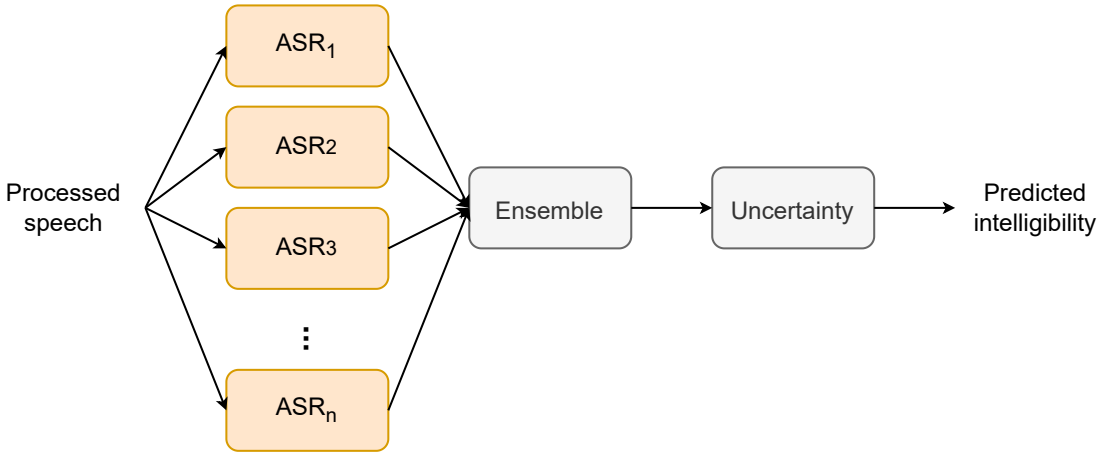


Fig. 6.1 An ensemble of ASR models is used to estimate the uncertainty of a processed speech, which is then used for intelligibility prediction.

vs non-intrusive approaches, using the Noisy Grid and CPC1 databases. The results and analyses are presented in Section 6.4 and Section 6.5. The last section summarises the work in this chapter.

6.2 Unsupervised ASR uncertainty estimation

In this section, it is described how two sequence-level ASR uncertainty measures, *confidence* and *entropy*, are formulated using an ensemble method following the derivation in Malinin and Gales (2021). The ensemble of models can be interpreted from a Bayesian perspective, i.e., regarding model parameters θ as random variables and using a prior $\mathrm{p}(\theta)$ to compute the posterior $\mathrm{p}(\theta|D)$ with a given dataset D. As Bayesian inference is usually intractable for models like deep neural networks, it is possible to take advantage of an approximation $\mathrm{q}(\theta)$ to $\mathrm{p}(\theta|D)$ with a family of models with different parameters (Hoffmann and Elster, 2021). Monte-Carlo Dropout (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) are two major approaches to generating ensembles, and the latter approach is exploited in this work. A brief overview of the proposed approach is shown in Figure 6.1. A group of ASR models are ensembled to estimate the uncertainty given a processed speech, and the estimated uncertainty is used to correlate with speech intelligibility.

6.2.1 Sequence-level uncertainty estimation

Given the ASR training dataset containing variable-length sequences of input acoustic features $\{x_1,\ldots,x_N\} = x \in X$, and the corresponding transcript targets $\{y_1,\ldots,y_L\} = y \in Y$,

an ensemble of M ASR models $\{\mathrm{P}(y|x;\theta^{(m)})\}$ can be trained to achieve the approximated posterior $\mathrm{q}(\theta)$. The sequence-level predictive posterior $\mathrm{P}(y|x,\theta)$ can be computed as the expectation of the ensemble:

$$\mathrm{P}(y|x,\theta) = \mathbb{E}_{\mathrm{q}(\theta)}[\mathrm{P}(y|x,\theta)] \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{P}\left(y|x,\theta^{(m)}\right), \tag{6.1}$$

where $\theta^{(m)} \sim \mathrm{q}(\theta) \approx \mathrm{p}(\theta|D)$. The sequence-level entropy $\mathrm{H}(y|x,\theta)$ can be expressed as:

$$\mathrm{H}(y|x,\theta) = \mathbb{E}_{\mathrm{P}(y|x,\theta)}[-\ln\mathrm{P}(y|x,\theta)] = -\sum_{y\in Y}\mathrm{P}(y|x,\theta)\ln\mathrm{P}(y|x,\theta). \tag{6.2}$$

It is usually not possible to compute the posterior $\mathrm{P}(y|x,\theta)$ as Y is an infinite set with variable-length transcript sequences. However, an autoregressive ASR model could factorise the posterior into a product of conditionals:

$$\mathrm{P}(y|x,\theta) = \prod_{l=1}^{L}\mathrm{P}(y_l|y_{<l},x;\theta), y_l \in \{\omega_1,\ldots,\omega_K\}, \tag{6.3}$$

where ω represents the byte-pair encoding (BPE) token, and K is the size of BPE vocabulary. The BPE tokeniser is one of the popular subword-based tokenisation approaches, which target compressing a very large vocabulary size and avoiding character-based tokenisation that leads to very long sequences. The BPE tokenisation is designed to ensure that the most commonly occurred words in the vocabulary are represented as a single token, while the rarely appeared words are divided into more subword tokens. It is now widely used in ASR tasks.

Confidence is usually considered as the maximum predicted probability, and the sequence-level confidence C_S in this work is regarded as a combination of token-level confidence. In order to make a fair comparison of sequences with variable lengths, a length normalisation rate is used (Cover, 1999), and C_S is computed as:

$$C_S = \exp\left[\frac{1}{L}\ln\sum_{l=1}^{L}\max\frac{1}{M}\sum_{m=1}^{M}\mathrm{P}(y_l|y_{<l},x;\theta^{(m)})\right]. \tag{6.4}$$

Entropy computation is usually challenging as the expectations of y are practically intractable, i.e., there are K^L possible candidates for a L-length sequence y_L, and a forward-pass inference needs to be conducted for each hypothesis y. Meanwhile, beam-search in ASR inference stage is able to provide high-quality hypotheses and can therefore be considered as a form of importance-sampling that yields hypotheses from high-probability space. By using

B top hypotheses within a beam, the approximated sequence-level entropy H_S with simple Monte-Carlo estimation can be computed as:

$$\begin{aligned} H_S &= -\sum_{b=1}^{B} \frac{\pi_b}{L^{(b)}} \ln \mathrm{P}(y^{(b)}|x,\theta), \\ \pi_b &= \frac{\exp \frac{1}{T} \ln \mathrm{P}(y^{(b)}|x,\theta)}{\sum_k^B \exp \frac{1}{T} \ln \mathrm{P}(y^{(k)}|x,\theta)}, \end{aligned} \tag{6.5}$$

where a calibration temperature T can be introduced to adjust the distribution of hypotheses, and:

$$\ln \mathrm{P}(y^{(b)}|x,\theta) = \sum_{l^{(b)}=1}^{L^{(b)}} \ln \frac{1}{M} \sum_{m=1}^{M} \mathrm{P}(y_l^{(b)}|y_{<l}^{(b)},x;\theta^{(m)}). \tag{6.6}$$

As higher entropy indicates more uncertainty, negative entropy is used to form a measure that is correlated with intelligibility in this work.

6.2.2 Token-level ASR posterior

The ASR models used in this work are based on the transformer architecture (Vaswani et al., 2017), which has shown impressive results recently. The model consists of a convolutional neural network-based front-end, a transformer-based encoder, and a transformer-based decoder. A mechanism combining the Connectionist Temporal Classification (CTC) and attention-based sequence to sequence (seq2seq) is used for the optimisation Kim et al. (2017). When estimating the uncertainty, the predictive posterior for each token is expressed as:

$$\mathrm{P}(y_l|y_{<l},x;\theta^{(m)}) = \lambda \mathrm{P}_{CTC}(y_l|y_{<l},x;\theta^{(m)}) + (1-\lambda)\mathrm{P}_{seq2seq}(y_l|y_{<l},x;\theta^{(m)}), \tag{6.7}$$

where λ is a weighting coefficient.

6.3 Experimental setup

The experimental setup is roughly the same as described in Section 5.3. The two very different databases Noisy Grid corpus (Barker and Cooke, 2007) and CPC1 (Barker et al., 2022) corpus are used to evaluate the proposed ASR uncertainty-based non-intrusive approach. As introduced before, the Noisy Grid corpus provides monaural utterances in speech-shaped noises and the recognition results by normal hearing listeners. Meanwhile, the CPC1 corpus provides a large number of binaural speech examples in simulated domestic environments then processed by several speech enhancement models, and the responses from hearing

impaired listeners. Both databases are divided into a training set, a development set, and an evaluation set.

The ASR configuration and training are also identical to the ones used in the previous chapter, i.e., the ASR architecture follows the SpeechBrain LibriSpeech transformer ASR recipe (Ravanelli et al., 2021), and the ASR model is from the released model which is pretrained on the 960-hour LibriSpeech dataset (Panayotov et al., 2015). For the purpose of constructing an ensemble of ASR models, 6 models are finetuned on the training set of the experimental databases. As a result, an ensemble of these 6 finetuned ASR models are used for uncertainty estimation. Different random seeds are used for the training of these 6 models, i.e., the training process for example the order of loading data batches is different for each model. In addition, the weighting coefficient λ is set to 0.4 with respect to the settings in the recipe, and the temperature T and B the beam size are set 1 and 10, respectively. The hyperparameter settings for uncertainty estimation are discussed for the CPC1 experiments.

As for evaluation metrics, RMSE, NCC, and KT are all included together with a logistic fitting function $f(x) = 1/[1+\exp(ax+b)]$, as RMSE and NCC are usually used for linear correlation measure thus invalid for non-linear correlations. The fitting function is applied to the uncertainty to predict the word correctness score of the listener responses, and the parameters are tuned on the development set.

6.4 Monaural speech in SSN noise with normal hearing listeners

6.4.1 Baselines

The performance of the novel approaches will be compared to a number of baselines. In the previous chapter, NCM achieves the best intrusive performance in the Noisy Grid corpus, thus selected as a baseline approach. Meanwhile, STOI and ESTOI are arguably the most widely used intrusive approaches. In addition, one of the most classic non-intrusive predictor SRMR is also used. The performances of the ASR WCS obtained from the ensemble of ASR models, and the intrusive ASR hidden representation (HR) based proposed in the previous are also reported for comparison.

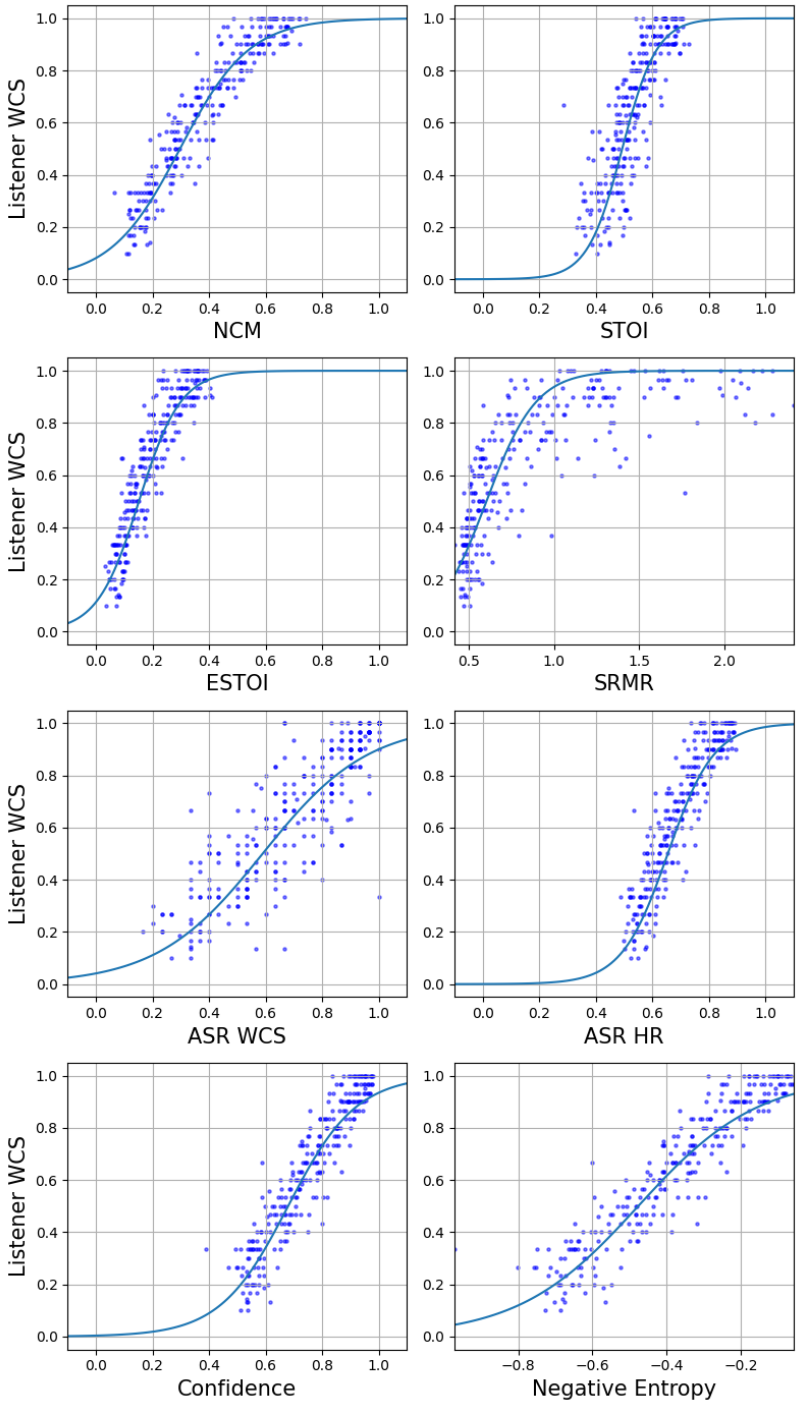


Fig. 6.2 Scatter plots of all intelligibility predictions on the Grid corpus evaluation set, along with the logistic fitting functions.

Table 6.1 Evaluation results on the Noisy Grid corpus in terms of RMSE, NCC, and KT.

	RMSE ↓	NCC ↑	KT ↑
NCM	0.083	0.950	0.801
STOI	0.146	0.850	0.671
ESTOI	0.103	0.926	0.761
SRMR	0.150	0.851	0.661
ASR WCS	0.144	0.844	0.695
ASR HR	0.115	0.923	0.761
ASR Uncertainty C_S	0.093	0.937	0.790
ASR Uncertainty $-H_S$	0.094	0.936	0.791

6.4.2 Results

Overall results

Table 6.1 lists the evaluation results on the Noisy Grid test set. The results show that the non-intrusive SRMR can achieve competitive performance close to the widely used STOI, despite it not requiring a reference signal. Even so, it does not perform as well as other intrusive approaches, including ESTOI, NCM, ASR WCS and ASR ASR. Meanwhile, the ASR uncertainty based approaches can make more accurate predictions than most intrusive approaches including ASR HR and ESTOI. Although there is a minor performance gap between the ASR ASR uncertainty based approaches and NCM, the results are surprising given that they are estimated with only the degraded speech signals. It can be observed that the negative entropy performs overall slightly better than the confidence, but the difference is very minimal.

Figure 6.2 shows the predicted intelligibility scores of NCM, STOI, ESTOI, SRMR, ASR WCS, ASR HR, and the confidence and negative entropy from the ensemble of ASR models along with their logistic mapping functions. It can be observed that SRMR generally fits well for low intelligibility speech, while in the high listener WCS region, the data points are dispersed. As SRMR is computed as the ratio between low and high frequency modulation, this indicates that a low ratio is not a necessary condition to low intelligibility, i.e., some speech can still be quite intelligible when the ratio is low. The ASR uncertainty measures, on the contrary, fit very well for both low and highly intelligible speech signals.

Data mismatch

The influence of data mismatch, i.e., the distribution gap between the ASR training data and evaluation data, is also investigated. Four ensembles consisting of ASR models trained with different training data are used: (1) ASR models trained only on clean LibriSpeech (LS);

Table 6.2 Evaluation results on ASR ensembles trained by different training databases.

Training data	Measure	RMSE ↓	NCC ↑	KT ↑
LS	C_S	0.172	0.762	0.572
	$-H_S$	0.166	0.788	0.595
	WCS	0.206	0.607	0.440
CGrid	C_S	0.224	0.521	0.329
	$-H_S$	0.235	0.444	0.302
	ASR WCS	0.148	0.825	0.650
DGrid	C_S	0.098	0.925	0.767
	$-H_S$	0.099	0.924	0.768
	ASR WCS	0.115	0.901	0.754
NGrid	C_S	**0.093**	**0.937**	0.790
	$-H_S$	0.094	0.936	**0.791**
	ASR WCS	0.144	0.844	0.695

(2) LibriSpeech pretrained models finetuned on Clean Grid corpus (CGrid); (3) LibriSpeech pretrained models finetuned on clean Grid corpus mixed with DEMAND noise (Thiemann et al., 2013) (DGrid); (4) LibriSpeech pretrained models finetuned on Noisy Grid speech.

The results of Noisy Grid test set for each ensemble of ASR models are shown in Table 6.2. It shows that a stronger prior knowledge of the test data, i.e., closer training and evaluation data distribution, leads to a higher correlation between ASR uncertainty and speech intelligibility based on the results of CGrid, DGrid, and NGrid. However, it can be observed that when the ASR models have no knowledge of the noisy signals, the confidence and negative entropy of LS finetuned ensemble could outperform the CGrid ensemble. It is also worth noting that ASR models optimised on DGrid, i.e., different types of noises from the Noisy Grid test set, could also produce competitive results.

6.5 Processed binaural speech in domestic noise with hearing impaired listeners

6.5.1 Baselines

The proposed approach is also compared against a number of existing intrusive and non-intrusive baselines for the CPC1 database. In the previous chapter, ASR HR shows the overall best performance and is included here as one of the intrusive baselines. ASR WCS is also used for comparison. Other than those, two well-performing intrusive approaches BE-ESTOI and BE-HASPI are considered, as BE-ESTOI shows accurate prediction ability and BE-HASPI incorporates an elaborate hearing loss model. In addition, MBSTOI, which

is the baseline system of CPC1, is used as another baseline system as it can take better advantage of binaural cues. These approaches leverage the MSBG model to simulate the effect of hearing hearing as a front-end loss except for BE-HASPI.

For the non-intrusive baselines, SRMR is included as it is one of the most representative conventional approaches based on prior knowledge of acoustic features. Similarly to BE-ESTOI and BE-HASPI, the better-ear policy is applied to accommodate binaural information. In addition, three non-intrusive approaches in the CPC1 are taken into consideration. Roßbach et al. (2022) proposed to leverage the mean temporal distance of triphone posteriogram generated by an ASR to predict intelligibility, which is similar to the work proposed in Martinez et al. (2022). It is worth noting that the training material for the ASR is a large simulated noisy German speech corpus. The other two approaches (Close et al., 2022; Zezario et al., 2022) are data-driven. In detail, Zezario et al. (2022) proposed to use a DNN to learn the mapping from cross-domain features to listener WCS. The cross-domain features include STFT processed spectrogram, learnable filter bank extracted features (Ravanelli and Bengio, 2018), and latent representations extracted by a large-scale self-supervised model WavLM (Chen et al., 2022). The WavLM is also trainable during the training of the DNN. Similarly, Close et al. (2022) proposed to train a prediction DNN to fit the normalised spectrogram to listener WCS.

6.5.2 Results

Overall results

The overall results of the CPC1 *closed-* and *open-set* are shown in Table 6.3. For the *closed-set*, the ASR uncertainty measures can achieve the performance very close to the best performing intrusive approach ASR HR, and surpass all the other intrusive predictors in terms of RMSE and NCC. ASR WCS performs the best with regard to KT as WCS are discrete, i.e., *tied* pairs are more likely to appear.

Meanwhile, the ASR uncertainty measures still stand out among non-intrusive approaches. They largely outperform the conventional acoustic feature-based BE-SRMR. The ASR-based approach proposed in Roßbach et al. (2022) also does not perform as well as the ASR uncertainty. As the training material is German and the predictor is based on the triphone posteriogram, the intelligibility predictor does not have access to language-specific information. This is consistent with the finding in the previous chapter, that an intelligibility predictor can be more accurate if the language information is taken into consideration, i.e. the decoder hidden representation is better than the high-level encoder hidden representation for intelligibility estimation. The data-driven approach proposed in Zezario et al. (2022)

Table 6.3 Evaluation between the listening results WCS and predicted intelligibility measures on CPC1 evaluation set.

		RMSE ↓	NCC ↑	KT ↑
Closed-set				
Intrusive	BE-ESTOI	0.253	0.719	0.446
	BE-HASPI	0.254	0.717	0.445
	MBSTOI	0.285	0.621	0.398
	ASR WCS	0.249	0.731	0.526
	ASR HR	**0.231**	0.773	0.498
Non-intrusive	BE-SRMR	0.354	0.244	0.152
	Close et al. (2022)	0.334	0.43	-
	Roßbach et al. (2022)	0.259	0.70	-
	Zezario et al. (2022)	**0.231**	**0.78**	-
	ASR Uncertainty C_S	0.234	0.767	0.497
	ASR Uncertainty $-H_S$	0.233	0.768	**0.499**
Open-set				
Intrusive	BE-ESTOI	0.294	0.640	0.466
	BE-HASPI	0.267	0.676	0.469
	MBSTOI	0.365	0.529	0.391
	ASR WCS	0.253	0.717	**0.530**
	ASR HR	**0.235**	**0.763**	**0.530**
Non-intrusive	BE-SRMR	0.358	0.213	0.116
	Roßbach et al. (2022)	0.321	0.54	-
	Zezario et al. (2022)	0.244	0.75	-
	ASR Uncertainty C_S	0.248	0.729	0.512
	ASR Uncertainty $-H_S$	0.246	0.734	0.512

Table 6.4 Listener- and system-wise evaluation results on the *closed-set* of predicted intelligibility.

		RMSE ↓	NCC ↑	KT ↑
Listener-wise				
Intrusive	BE-HASPI	**0.070**	**0.593**	0.425
	ASR WCS	0.074	0.515	0.430
	ASR HR	0.073	0.540	**0.459**
Non-intrusive	BE-SRMR	0.082	0.274	0.322
	ASR Uncertainty C_S	0.074	0.526	**0.459**
	ASR Uncertainty $-H_S$	0.074	0.495	0.425
System-wise				
Intrusive	BE-HASPI	0.062	**0.984**	**0.778**
	ASR WCS	0.053	0.983	0.733
	ASR HR	**0.049**	0.982	0.733
Non-intrusive	BE-SRMR	0.219	0.463	0.111
	ASR Uncertainty C_S	0.052	0.979	0.733
	ASR Uncertainty $-H_S$	0.054	0.975	0.733

can achieve the best performance in the *closed-set* thanks to the knowledge provided by the large-scale self-supervised pretrained model WavLM. Other than that, it does require the listener WCS label to train the model.

The results on the *open-set* are similar those on the *closed-set*. It is worth noting that the top performing non-intrusive approaches have a larger performance drop compared to the ASR HR. This indicates that the performance of non-intrusive approaches is more sensitive to the mismatch gap between the training and evaluation data.

Unlike the results in the Noisy Grid database, the intelligibility prediction accuracy of the confidence is slightly lower than negative entropy. As the CPC1 speech material is much more complicated, i.e., the speech material has a larger vocabulary, the degradation is simulated with domestic noises and room impulse responses, and non-linear processing by various DNN-based speech enhancement, the entropy can reflect the uncertainty better than confidence.

Listener- and system-wise correlation

As the CPC1 listener recognition results can be noisy, the average intelligibility prediction accuracy over different listeners and systems is investigated. The listener- and system-wise correlation results are shown in Table 6.4. As shown in the previous chapter, BE-HASPI can

Table 6.5 Evaluation results on the *closed-set* of different ensembles of ASR models trained on different databases, and with or without using MSBG hearing loss simulation.

MSBG	Training data	Measure	RMSE ↓	NCC ↑	KT ↑
with	LS	ASR WCS	0.269	0.674	0.455
		ASR HR	0.264	0.692	0.449
		ASR Uncertainty C_S	0.283	0.630	0.426
		ASR Uncertainty $-H_S$	0.278	0.646	0.423
	LS+CLS	ASR WCS	0.244	0.742	0.503
		ASR HR	0.243	0.746	0.464
		ASR Uncertainty C_S	0.250	0.731	0.444
		ASR Uncertainty $-H_S$	0.245	0.738	0.446
	LS+CPC1	ASR WCS	0.248	0.735	0.527
		ASR HR	0.233	0.768	0.503
		ASR Uncertainty C_S	0.236	0.764	0.504
		ASR Uncertainty $-H_S$	0.233	0.768	0.505
	LS+CLS+CPC1	ASR WCS	0.249	0.731	0.526
		ASR HR	0.231	0.773	0.498
		ASR Uncertainty C_S	0.234	0.767	0.497
		ASR Uncertainty $-H_S$	0.233	0.768	0.499
w/o	LS+CLS+CPC1	ASR WCS	0.249	0.730	0.525
		ASR HR	0.234	0.767	0.476
		ASR Uncertainty C_S	0.241	0.751	0.472
		ASR Uncertainty $-H_S$	0.239	0.754	0.477

achieve the most accurate listener-wise prediction in terms of RMSE and NCC, thanks to its elaborate hearing loss model. Meanwhile, the ASR confidence achieves slightly worse results than the intrusive ASR HR-based approach, but better than ASR WCS and BE-SRMR. Though the ASR negative entropy prediction is the same as confidence with regard to RMSE, the NCC and KT evaluations are not as good.

For the system-wise evaluation results, it can be observed that all ASR-based approaches perform well concerning RMSE, and approaching BE-HASPI concerning NCC and KT. The ASR confidence is slightly better than the negative entropy, and both of them can achieve similar results to the intrusive ASR HR-based prediction.

Data mismatch

Motivated by the performance gap between the CPC1 *closed-* and *open-set*, the influence of the mismatch between the training and evaluation data is explored. Four ensembles of ASR models: (1) trained by only clean LibriSpeech speech (LS), (2) LS pretrained then

finetuned with noisy LibriSpeech mixed with Clarity noise (LS+CLS), (3) LS pretrained then finetuned with CPC1 databases (LS+CPC1), (4) LS+CLS trained then finetuned with CPC1 databases (LS+CLS+CPC1), are used for uncertainty estimation in the *closed-set*. All four ensembles are strong models of clean speech recognition as they are all pretrained with LibriSpeech. CLS represents the data distribution of noisy speech, and CPC1 represents the data distribution of noisy speech processed by various hearing aid algorithms. Table 6.5 presents the results of the four ASR-based approaches.

For ASR HR and uncertainty based approaches, it is clear that the performance improves when the training data is more similar to the evaluation data, i.e., the ASR ensembles trained with CPC1 are better than the ensemble trained with only CLS, and the CLS trained ensemble performs better than only LS trained ensemble. However, this is not the case for ASR WCS-based prediction, whose best performance is achieved when the ensemble of ASR models is trained with LS+CLS.

In general, the intrusive ASR HR-based approach could make more accurate predictions than the non-intrusive ASR uncertainty-based ones when the training data is the same. In spite of that, the performance gap is quite small when CPC1 is used for training. This indicates that despite the non-intrusive uncertainty approaches being not as generalisable as intrusive HR, they can still make very accurate intelligibility predictions. It is also worth mentioning that when the training data is significantly different from the evaluation data, the ASR uncertainty measures perform worse than ASR. This pattern can also be observed in the Noisy Grid results, i.e. when the ensemble of ASR models is trained with CGrid or LS.

Hearing loss simulation

The influence of the MSBG hearing loss model is also investigated and the results are shown in Table 6.2. The performance of the ensemble of ASR models trained on LS+CLS+CPC1 with and without the MSBG model for hearing loss simulation is presented. The uncertainty measures can gain more advantages with the MSBG model, compared to the HR-based prediction. This could be due to the ceiling effect, as the CPC1 listening results are quite noisy and the ASR HR approach has already achieved very top performance, thus does not have enough room for further improvement.

Hyperparameters for uncertainty estimation

The influence of three major hyperparameters for uncertainty estimation is investigated with CPC1 database, including the ensemble size, temperature, and TopK beam size. The results are shown in Table 6.6. Ideally, the estimated uncertainty should be more accurate with the

Table 6.6 The effect of tuning uncertainty estimation hyperparameters on system performance as measured by RMSE, NCC and KT. Results are shown separately for the confidence-based (C_s) and entropy-based ($-H_s$) uncertainty estimates.

	Hyperparameter	Measure	RMSE ↓	NCC ↑	KT ↑
Ensemble size	1	C_S	0.233	0.769	0.497
		$-H_S$	0.233	0.769	0.499
	3	C_S	0.233	0.768	0.498
		$-H_S$	0.233	0.769	0.500
	6	C_S	0.234	0.767	0.497
		$-H_S$	0.233	0.768	0.499
Temperature	0.1	C_S	0.238	0.756	0.491
		$-H_S$	0.237	0.758	0.492
	0.5	C_S	0.235	0.764	0.495
		$-H_S$	0.235	0.763	0.496
	1	C_S	0.234	0.767	0.497
		$-H_S$	0.233	0.768	0.499
	1.5	C_S	0.236	0.764	0.485
		$-H_S$	0.235	0.764	0.490
	2	C_S	0.236	0.763	0.482
		$-H_S$	0.236	0.761	0.486
Beam size	5	$-H_S$	0.233	0.768	0.499
	10	$-H_S$	0.233	0.768	0.499
	15	$-H_S$	0.233	0.768	0.499
	20	$-H_S$	0.233	0.768	0.499
	50	$-H_S$	0.233	0.769	0.500

increase in ensemble size. However, the intelligibility prediction does not gain benefit from using only one ASR model to using six models as an ensemble. All the ASR models are finetuned for only ten epochs from the same LS trained model, and the CPC1 database is relatively small for training a modern end-to-end ASR model. These lead to minor variability among different ASR models. Therefore, increasing the ensemble size does not contribute to the CPC1 experiments.

Temperature is a parameter that changes the probability distribution of the softmax function. The lower the temperature is, the model is more confident in its classification. On the contrary, the softmax probability distribution is flatter when the temperature is high. Five temperature values from 0.1 to 2 are tried, and the results show that the most accurate intelligibility prediction is made when the temperature is 1.

By increasing the number of top candidates in the beam search, i.e. beam size, the entropy estimation is supposed to be more accurate, because more samples are obtained. The beam size is set from 5 to 50 in the experiments. The experimental results show that when the beam size is set 50, there indeed is an improvement in terms of NCC and KT, but the difference is marginal. This could be due to the same reason as the ensemble size, i.e., the uncertainty estimation does not vary much when the models are quite similar.

In conclusion, the hyperparameters for uncertainty estimation including ensemble size, temperature, and TopK beam size do not make a significant difference in the intelligibility estimation. The potential reason is that the variance among the ASR models in the ensemble is not significant. As the CPC1 database is relatively small, the randomisation of the finetuning process can not lead to significant differences among the ASR models.

6.6 Conclusions

In this chapter, an ASR uncertainty-based intelligibility prediction approach has been proposed. Specifically, an ensemble of ASR models has been leveraged to infer the recognition uncertainty, which shows a high correlation with human intelligibility. The proposed approach has three major advantages: (1) the approach is non-intrusive, thus does not require a clean reference speech signal for prediction; (2) the uncertainty is estimated without supervision, therefore no listener recognition results are needed to train the model if considering only monotonicity; (3) the predicted intelligibility is utterance-level, and the language model information can be well utilised.

The experimental results have shown that the non-intrusive ASR uncertainty-based intelligibility prediction approach can make accurate predictions in both Noisy Grid and CPC1 databases. In the Noisy Grid experiments, which contain simulated speech linearly

degraded by SSN, the uncertainty prediction achieves the second best results among a number of intrusive and non-intrusive approaches. In the CPC1 experiments, which contain simulated domestic noisy speech processed by complex DNN-based enhancement and the responses by hearing impaired listeners, the performance of uncertainty prediction is only slightly behind the best performing intrusive HR-based approach, and a DNN-based data-driven approach, which requires the knowledge from both large-scale pretrained model and a large number of listener intelligibility labels. Between the two uncertainty measures, confidence and negative entropy, the difference is minimal. Confidence performs slightly better in the Noisy Grid database, while negative entropy performs better in the CPC1.

Further analysis shows that the ASR uncertainty is sensitive to data mismatch between the ASR training data and intelligibility evaluation data. When the distributions of the training and evaluation data are close, the non-intrusive uncertainty-based prediction exhibited a performance very similar to the intrusive hidden representation based approach. Otherwise, it can even be even worse than ASR word correctness scores. Despite that, ASR uncertainty estimated by mismatched training data can still achieve competitive results, and better than many other leading intrusive and non-intrusive approaches.

Chapter 7

Conclusions

This thesis has been focusing on data-driven approaches for speech intelligibility enhancement and prediction for hearing aids. For the purpose of improving the intelligibility of noisy speech for hearing impaired listeners, a differentiable hearing aid speech processing framework was proposed. This framework could optimise hearing aid fittings together with a DNN-based denoising model using an intelligibility-based objective function. The first half of this thesis presented this framework for data-driven speech intelligibility enhancement and was organised to address the following research questions:

- How well can data-driven optimised hearing aid fittings perform in terms of intelligibility improvement for speech in noisy and noise-free environments?
- Can the hearing aid fittings optimised for different noisy environments provide benefits over general fittings?
- How well can hearing aid speech enhancement models with a DNN-based denoising module perform in noisy environments?

An accurate speech intelligibility predictor can be crucial for the development of hearing aid enhancement algorithms, because it can reduce the requirement for expensive subjective evaluation with listening experiments. The second half of this thesis focused on intelligibility prediction with ASR models, and provided an intrusive and a non-intrusive ASR-based approach to address the following research questions:

- How well can the features extracted by ASR models perform in terms of robust intelligibility prediction?
- How can ASR models predict intelligibility non-intrusively, i.e., without using extra information like reference signals or transcription?

The final chapter concludes this thesis by first reviewing the contributions with respect to the above research questions, and then discussing some of the limitations of the work before presenting potential future research directions.

7.1 Contributions

7.1.1 Speech intelligibility enhancement for hearing aids

Motivated by the recent success of data-driven approaches, this thesis explored their application to the optimisation of hearing aid fittings by constructing the DHASP framework. Specifically, the framework consists of an FIR filter representing a frequency gain amplification table, and an optimisation objective incorporating a differentiable approximation to an auditory model. The auditory model takes hearing abilities into consideration and models the intelligibility of speech in noise judged by hearing impaired listeners. With such an optimisation objective, the hearing aid amplification fittings are trained to maximise speech intelligibility given a speech database.

In Chapter 3, the DHASP framework was first optimised with a clean speech database, TIMIT (Garofolo et al., 1993), at conversational levels. In the experiments, the optimisation objective was a combination of an approximation to HASPI, which models the intelligibility of hearing impaired listeners, and an energy constraint, which prevents over-amplification. Compared to the widely used and recognised classic NAL-R fittings, the optimised fittings provided an intelligibility improvement when evaluated using objective measures. Also, it was observed that the optimised fittings tend to provide more amplification in the low and high frequencies, while less amplification in the middle frequencies, i.e., around 1 kHz, similar to the pattern shown in Gonçalves Braz et al. (2022).

Having considered the case of clean speech (i.e., noise-free conditions), the DHASP framework was extended to handle speech in various noisy environments, and a differentiable approximation to the MSBG hearing loss model was used as the optimisation objective. Given one specific audiogram, the amplification fittings are optimised in the context of the noise conditions, i.e. so as to be customised to different noisy environments, for example, traffic, babble, and kitchen noises. In addition, the effects of a conventional hearing aid denoising approach, Wiener filtering, were explored, i.e., the optimised fittings were also customised to whether the denoising function is turned on. Objective evaluation results showed that the data-driven optimised fittings that are customised to different noisy environments and the usage of the denoising feature can outperform both generally optimised fittings and NAL-R. Additionally, it was found that Wiener filtering does not necessarily improve speech

intelligibility. Furthermore, it was also found that the fittings optimised with the MSBG hearing loss model provided more amplification at low and high frequencies, while less gain in middle frequencies around 1 kHz compared to NAL-R, which is similar to those of the HASPI optimised fittings.

Taken together, these experimental results showed that the DHASP framework could perform well in terms of intelligibility improvement, and the customised fittings could gain benefit over general fittings. However, it should be noted that these conclusions are based on objective evaluation and have not been validated with hearing impaired listeners.

In Chapter 4, the DHASP framework was further extended to more complex scenarios with a DNN-based denoising module and evaluated with both objective and subjective evaluations. The CEC1 database (Graetzer et al., 2021) provides a large number of simulated domestic scenes, each of which consists of a target talker, a hearing impaired listener, an interfering source and room acoustics. To tackle the intelligibility degradation caused by the interfering sources and reverberation, a DNN-based denoising module was trained to extract the target speech in the DHASP framework. The denoised speech signals were then used to optimise the amplification fittings for hearing loss compensation with an objective of the approximated MSBG hearing loss model and approximated STOI. Both objective and subjective evaluation results showed that the data-driven speech intelligibility enhancement model with a DNN-based denoising module could provide a significant improvement over a conventional hearing aid model. Furthermore, it was also observed that methods combining DNNs and beamformers can enhance speech with minimal distortion. These processed speech signals, though not achieving the best objective intelligibility scores, are most intelligible to hearing impaired listeners among those processed by many other approaches.

7.1.2 Speech intelligibility prediction

It was observed in the findings of Chapter 4 that there is often significant disagreement between the objective and subjective evaluation results, i.e., systems with lower objective scores can achieve the top subjective performance. This indicates there is a need for more accurate speech intelligibility predictors for the development of hearing aid speech enhancement algorithms. Motivated by the recent progress of DNN-based ASR models, which can perform similarly to humans in some recognition tasks, this thesis has proposed to exploit them for intelligibility prediction in this thesis.

Most current widely used intrusive intelligibility predictors follow a scheme that measures the similarity between the extracted acoustic representations of a processed speech signal and its corresponding clean reference signal. The accuracy of the prediction is heavily dependent on the quality of the extracted representations. In Chapter 5, a DNN-based ASR model was

used as a representation extractor for the intrusive intelligibility prediction. The motivation is that DNNs are naturally trained to be good representation extractors, and thus ASR-based DNNs are expected to extract the representations that are crucial for the recognition of speech. The proposed approach was validated with two very different databases. The Noisy Grid corpus (Barker and Cooke, 2007) provides a large number of matrix test speech mixed with speech-shaped noise and their corresponding recognition performances by normal hearing listeners. In contrast, the CPC1 (Barker et al., 2022) database provides a large number of enhanced speech signals that simulate noisy domestic environments and their corresponding recognition performances by hearing impaired listeners. The representations extracted by the ASR model achieved very competitive prediction accuracy for the Grid corpus, and significantly outperformed a large number of popular intelligibility predictors. Experimental results also showed that high-level representations of the DNN-based ASR which contains language information can achieve the most accurate intelligibility prediction in the CPC1 database.

Intrusive approaches are usually difficult to apply in realistic environments, in which the reference signals are not available. As a result, non-intrusive intelligibility predictors are more desirable. In Chapter 6, an ASR uncertainty-based non-intrusive approach was proposed. Specifically, the sequence-level recognition uncertainty of a given processed speech signal was estimated with an ensemble of ASR models, and the uncertainty was then used to correlate to speech intelligibility. The uncertainty estimation is unsupervised, therefore does not require uncertainty labels, i.e., intelligibility labels from human listening experiments in this context. The proposed ASR uncertainty-based non-intrusive approach was also validated with the Noisy Grid corpus and the CPC1 database. It achieved the best performance for the Noisy Grid corpus, and very competitive prediction accuracy for the CPC1 database among a number of leading intrusive and non-intrusive approaches. The results suggested that the uncertainty estimated from an ensemble of ASR models can be naturally very correlated to human speech recognition results.

For both the proposed intrusive and non-intrusive approaches in this thesis, the quality of the ASR models is crucial. The DNN-based ASR models used in this work were pre-trained with the LibriSpeech database (Panayotov et al., 2015) which contains 960 hours of utterances, and therefore possess strong knowledge of clean speech recognition. The pre-trained models were then finetuned with the noisy or processed speech signals to reduce the mismatch between the ASR training data and the evaluation data. Experiments showed that the smaller the mismatch is, the more accurate is the intelligibility prediction that can be achieved with the ASR models. Furthermore, it was also found that the ASR representation-based intrusive

approach is more robust than the ASR uncertainty-based non-intrusive approach when the training and evaluation data mismatch is significant.

7.2 Limitations and future research

This thesis explored the application of data-driven approaches to hearing aid speech intelligibility enhancement, including amplification and denoising. However, this is still far from adequately solving the problem of hearing impairment. First of all, intelligibility is only one attribute of speech, while there could be a number of other attributes that are important for the satisfaction of hearing aid usage. Improving intelligibility can sometimes lead to the degradation of quality, e.g., intelligibility enhancement can distort the target speech to keep the cues for recognition, which may lead to listening dissatisfaction. Similarly, intelligibility is not necessarily correlated to listening effort, i.e., a processed speech signal can be quite intelligible but still require a lot of effort to understand for a hearing impaired listener. This can also lead to a decline in life quality. Also, high intelligibility does not necessarily lead to high comprehension, i.e., one can misinterpret a sentence while understanding most words. Secondly, the noisy speech intelligibility enhancement lacks consideration of environmental awareness, which is crucial for the everyday usage of hearing aids. Complete suppression of interfering sources may remove important non-speech sounds, such as fire alarms.

Additionally, the proposed approaches for speech intelligibility enhancement still need to be further refined for potential applications. Wide dynamic range compression is a standard for modern hearing aids that allows adaptive amplification. However, the data-driven optimised fittings proposed in this thesis are linear and may only be suitable for speech at conversational levels. Further work should be conducted to extend the data-driven optimised fittings to be compatible with the wide dynamic range compression (Gonçalves Braz et al., 2022). Despite the success of DNN-based noise suppression models, they have rarely been deployed for real-time applications. The major reason is that the DNN models are usually over-parameterised and computationally expensive. This leads to long processing times and high power consumption, which makes it particularly difficult to deploy the DNN-based denoising models on modern hearing aids. For future research, DNN compression techniques such as knowledge distillation can be used to achieve a DNN with a much smaller number of parameters while keeping a similar noise suppression performance (Tan and Wang, 2021).

Regarding speech intelligibility prediction, a major challenge for the proposed approaches is generalisation, i.e., the ability to predict the intelligibility of a speech signal that is very different from the signals used for training the underpinning ASR models. The difference can be caused by different noisy environments, being processed by different enhancement

systems, etc. Therefore, a future direction for improving ASR-based intelligibility prediction is to generate a larger database for ASR training by introducing speech in more variant noisy scenes and processed by more diverse speech enhancement models. It is also worth noting that the ASR representation-based intelligibility prediction approach does not perform as well as the non-intrusive approach in the Noisy Grid corpus. The reason could be that the over-parameterisation of the DNNs leads to noisy representations for relatively simple tasks, i.e., part of the extracted representations are redundant and not related to the recognition tasks. Therefore, it is worthwhile to conduct further research on the ASR representations and how much each of them is correlated to speech intelligibility.

References

Alexander, J. M., Kopun, J. G., and Stelmachowicz, P. G. (2014). Effects of frequency compression and frequency transposition on fricative and affricate perception in listeners with normal hearing and mild to moderate hearing loss. *Ear and Hearing*, 35(5):519.

Alexander, J. M. and Masterson, K. (2015). Effects of WDRC release time and number of channels on output SNR and speech recognition. *Ear and Hearing*, 36(2):e35.

Allen, J. B. (1995). How do humans process and recognize speech? In *Modern methods of speech processing*, pages 251–275. Springer.

Andersen, A. H., De Haan, J. M., Tan, Z.-H., and Jensen, J. (2016). A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4995–4999. IEEE.

Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2017). A non-intrusive short-time objective intelligibility measure. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089. IEEE.

Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2018a). Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions. *Speech Communication*, 102:1–13.

Andersen, A. H. et al. (2018b). Non-intrusive speech intelligibility prediction using convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1925–1939.

ANSI, A. (1997). S3. 5-1997, Methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19:90–119.

Arslan, E., Orzan, E., and Santarelli, R. (1999). Global problem of drug-induced hearing loss. *Annals of the New York Academy of Sciences*, 884(1):1–14.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baer, T. and Moore, B. C. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *The Journal of the Acoustical Society of America*, 94(3):1229–1241.

Baer, T. and Moore, B. C. (1994). Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *The Journal of the Acoustical Society of America*, 95(4):2277–2280.

Baker, R. J. and Rosen, S. (2002). Auditory filter nonlinearity in mild/moderate hearing impairment. *The Journal of the Acoustical Society of America*, 111(3):1330–1339.

Barker, J., Akeroyd, M., Cox, T. J., Culling, J. F., Firth, J., Graetzer, S., Griffiths, H., Harris, L., Naylor, G., Podwinska, Z., Porter, E., and Munoz, R. V. (2022). The 1st Clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction. In *Proc. Interspeech 2022*, pages 3508–3512.

Barker, J. and Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417.

Benesty, J., Chen, J., and Huang, Y. (2008). *Microphone array signal processing*, volume 1. Springer Science & Business Media.

Bentler, R. and Chiou, L.-K. (2006). Digital noise reduction: An overview. *Trends in Amplification*, 10(2):67–82.

Berger, K. W., Hagberg, E. N., and Rane, R. L. (1980). A reexamination of the one-half gain rule. *Ear and Hearing*, 1(4):223–225.

Best, V., Roverud, E., Mason, C. R., and Kidd Jr, G. (2017). Examination of a hybrid beamformer that preserves auditory spatial cues. *The Journal of the Acoustical Society of America*, 142(4):EL369–EL374.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in amplification*, 14(2):113–120.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

Brons, I., Houben, R., and Dreschler, W. A. (2014). Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. *Trends in hearing*, 18.

Brons, I., Houben, R., and Dreschler, W. A. (2015). Acoustical and perceptual comparison of noise reduction and compression in hearing aids. *Journal of Speech, Language, and Hearing Research*, 58(4):1363–1376.

Bruce, I. C., Sachs, M. B., and Young, E. D. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *The Journal of the Acoustical Society of America*, 113(1):369–388.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6):4007–4018.

Byrne, D. and Dillon, H. (1986). The national acoustic laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing*, 7(4):257–265.

Byrne, D., Dillon, H., Ching, T., Katsch, R., and Keidser, G. (2001). NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures. *Journal of the American academy of audiology*, 12(1).

Byrne, D. and Murray, N. (1986). Predictability of the required frequency response characteristic of a hearing aid from the pure-tone-audiogram. *Ear and Hearing*, 7(2):63–70.

Byrne, D., Parkinson, A., and Newall, P. (1990). Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired. *Ear and Hearing*, 11(1):40–49.

Byrne, D. and Tonisson, W. (1976). Selecting the gain of hearing aids for persons with sensorineural hearing impairments. *Scandinavian Audiology*, 5(2):51–59.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proc. of the IEEE*, 57(8):1408–1418.

Carter, G., Knapp, C., and Nuttall, A. (1973). Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. *IEEE Transactions on Audio and Electroacoustics*, 21(4):337–344.

Chakrabarty, S., Wang, D., and Habets, E. A. (2018). Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 476–480. IEEE.

Chen, F. (2016a). Modeling noise influence to speech intelligibility non-intrusively by reduced speech dynamic range. In *Proc. Interspeech 2016*, pages 1359–1362.

Chen, F. (2016b). Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation. *Biomedical Signal Processing and Control*, 24:109–113.

Chen, F., Hazrati, O., and Loizou, P. C. (2013). Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. *Biomedical Signal Processing and Control*, 8(3):311–314.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Chen, X., Shi, Y., Xiao, W., Wang, M., Wu, T., Shang, S., Meng, Q., and Zheng, N. (2021). A cascaded speech enhancement for hearing aids in noisy-reverberant conditions. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, 28.

Christiansen, C., Pedersen, M. S., and Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, 52(7-8):678–692.

Close, G., Hollands, S., Goetze, S., and Hain, T. (2022). Non-intrusive speech intelligibility metric prediction for hearing impaired individuals. In *Proc. Interspeech 2022*, pages 3483–3487.

Cooke, M. (1993). *Modelling auditory processing and organisation*, volume 7. Cambridge University Press.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Cooke, M. et al. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Cover, T. M. (1999). *Elements of Information Theory*. John Wiley & Sons.

Davis, A., McMahon, C. M., Pichora-Fuller, K. M., Russ, S., Lin, F., Olusanya, B. O., Chadha, S., and Tremblay, K. L. (2016). Aging and hearing health: the life-course approach. *The Gerontologist*, 56(Suppl_2):S256–S267.

Demirsahin, I., Kjartansson, O., Gutkin, A., and Rivera, C. (2020). Open-source multi-speaker corpora of the English accents in the British Isles. In *Proc. the Twelfth Language Resources and Evaluation Conference*, pages 6532–6541.

Denk, F., Ernst, S. M., Heeren, J., Ewert, S. D., and Kollmeier, B. (2018). The Oldenburg Hearing Device (OlHeaD) HRTF Database. *University of Oldenburg, Tech. Rep*.

Dillon, H., Keidser, G., O'Brien, A., and Silberstein, H. (2003). Sound quality comparisons of advanced hearing aids. *The Hearing Journal*, 56(4):30–32.

Doclo, S., Gannot, S., Moonen, M., Spriet, A., Haykin, S., and Liu, K. R. (2010). Acoustic beamforming for hearing aid applications. In *Handbook on array processing and sensor networks*, volume 9, pages 269–302. Wiley Hoboken, NJ, USA.

Doclo, S., Kellermann, W., Makino, S., and Nordholm, S. E. (2015). Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine*, 32(2):18–30.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5):2670–2680.

Durlach, N. I. (1972). Binaural signal detection-equalization and cancellation theory.

Défossez, A., Synnaeve, G., and Adi, Y. (2020). Real time speech enhancement in the waveform domain. In *Proc. Interspeech 2020*, pages 3291–3295.

Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (2021). A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction. In *Proc. Interspeech 2021*, pages 206–210.

Ellis, R. J. and Munro, K. J. (2015). Benefit from, and acclimatization to, frequency compression hearing aids in experienced adult hearing-aid users. *International Journal of Audiology*, 54(1):37–47.

Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations*.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121.

Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., and Roux, J. L. (2016). Improved mvdr beamforming using single-channel mask prediction networks. In *Proc. Interspeech 2016*, pages 1981–1985.

Erkelens, J. S., Hendriks, R. C., Heusdens, R., and Jensen, J. (2007). Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1741–1752.

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., and Scollie, S. (2015). Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine*, 32(2):114–124.

Falk, T. H., Zheng, C., and Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proc. of the 21st ACM International Conference on Multimedia*, pages 411–412.

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., and Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, 60(9):2394–2405.

Fowler, E. P. (1936). A method for the early detection of otosclerosis: A study of sounds well above threshold. *Archives of otolaryngology*, 24(6):731–741.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119.

Gajecki, T. and Nogueira, W. (2021). Binaural speech enhancement based on deep attention layers. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *STIN*, 93:27403.

Gatehouse, S., Naylor, G., and Elberling, C. (2006a). Linear and nonlinear hearing aid fittings–1. patterns of benefit: adaptación de auxiliares auditivos lineales y no lineales–1. patrones de beneficio. *International Journal of Audiology*, 45(3):130–152.

Gatehouse, S., Naylor, G., and Elberling, C. (2006b). Linear and nonlinear hearing aid fittings–2. patterns of candidature: Adaptación de auxiliares auditivos lineales y no lineales–2. patrones de selección de candidatos. *International Journal of Audiology*, 45(3):153–171.

Gelderblom, F. B., Tronstad, T. V., and Viggen, E. M. (2017). Subjective intelligibility of deep neural network-based speech enhancement. In *Proc. Interspeech 2017*, pages 1968–1972.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138.

Glista, D. and Scollie, S. (2018). The use of frequency lowering technology in the treatment of severe-to-profound hearing loss: a review of the literature and candidacy considerations for clinical application. In *Seminars in hearing*, volume 39, pages 377–389. Thieme Medical Publishers.

Glista, D., Scollie, S., Bagatto, M., Seewald, R., Parsa, V., and Johnson, A. (2009). Evaluation of nonlinear frequency compression: Clinical outcomes. *International Journal of Audiology*, 48(9):632–644.

Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679–3689.

Gonçalves Braz, L., Fontan, L., Pinquier, J., Stone, M. A., and Füllgrabe, C. (2022). Opra-rs: a hearing-aid fitting method based on automatic speech recognition and random search. *Frontiers in Neuroscience*, 16:57.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Gourévitch, B., Edeline, J.-M., Occelli, F., and Eggermont, J. J. (2014). Is the din really harmless? long-term effects of non-traumatic noise on the adult auditory system. *Nature Reviews Neuroscience*, 15(7):483–491.

Graetzer, S., Barker, J., Cox, T. J., Akeroyd, M., Culling, J. F., Naylor, G., Porter, E., and Muñoz, R. V. (2021). Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In *Proc. Interspeech 2021*, pages 686–690.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of the 23rd international conference on Machine learning*, pages 369–376.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species——29 years later. *The Journal of the Acoustical Society of America*, 87(6):2592–2605.

Gu, R., Wu, J., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. (2019). End-to-end multi-channel speech separation. *arXiv preprint arXiv:1905.06286*.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., and Wu, Y. (2020). ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. In *INTERSPEECH*, pages 3610–3614.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (2001). Auditory nerve model for predicting performance limits of normal and impaired listeners. *Acoustics Research Letters Online*, 2(3):91–96.

Hermansky, H., Variani, E., and Peddinti, V. (2013). Mean temporal distance: Predicting ASR error from temporal properties of speech signal. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7423–7426. IEEE.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200. IEEE.

Higuchi, T., Ito, N., Araki, S., Yoshioka, T., Delcroix, M., and Nakatani, T. (2017). Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):780–793.

Hoffmann, L. and Elster, C. (2021). Deep ensembles from a Bayesian perspective. *arXiv preprint arXiv:2105.13283*.

Hogan, C. A. and Turner, C. W. (1998). High-frequency audibility: Benefits for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 104(1):432–441.

Hohmann, V. and Kollmeier, B. (1995). The effect of multichannel dynamic compression on speech intelligibility. *The Journal of the Acoustical Society of America*, 97(2):1191–1195.

Holube, I. and Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3):1703–1716.

Hopkins, K., Khanom, M., Dickinson, A.-M., and Munro, K. J. (2014). Benefit from non-linear frequency compression hearing aids in a clinical setting: The effects of duration of experience and severity of high-frequency hearing loss. *International Journal of Audiology*, 53(4):219–228.

Humes, L. E., Kewley-Port, D., Fogerty, D., and Kinney, D. (2010). Measures of hearing threshold and temporal processing across the adult lifespan. *Hearing research*, 264(1-2):30–40.

Irino, T. and Patterson, R. D. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE Transactions on Audio, Speech, and Language processing*, 14(6):2222–2232.

ITU-T, P. (1993). Objective measurement of active speech level. *ITU-T Recommendation*.

Jensen, J. and Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487.

Jürgens, T. and Brand, T. (2009). Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5):2635–2648.

Kalgaonkar, K., Liu, C., Gong, Y., and Yao, K. (2015). Estimating confidence scores on ASR results using recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4999–5003. IEEE.

Karbasi, M. et al. (2016). Twin-HMM-based non-intrusive speech intelligibility prediction. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 624–628. IEEE.

Karbasi, M., Zeiler, S., and Kolossa, D. (2022). Microscopic and blind prediction of speech intelligibility: Theory and practice. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2141–2155.

Kates, J. (2013). An auditory model for intelligibility and quality predictions. In *Proc. of Meetings on Acoustics (ICA2013)*, volume 19, page 050184. Acoustical Society of America.

Kates, J. M. (2008). *Digital hearing aids*. Plural publishing.

Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, 117(4):2224–2237.

Kates, J. M. and Arehart, K. H. (2014a). The hearing-aid speech perception index (HASPI). *Speech Communication*, 65:75–93.

Kates, J. M. and Arehart, K. H. (2014b). The hearing-aid speech quality index (HASQI) version 2. *Journal of the Audio Engineering Society*, 62(3):99–117.

Kates, J. M. and Arehart, K. H. (2021). The hearing-aid speech perception index (HASPI) version 2. *Speech Communication*, 131:35–46.

Kavalekalam, M. S., Christensen, M. G., Gran, F., and Boldt, J. B. (2016). Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 191–195. IEEE.

Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., and Hohmann, V. (2021). Open community platform for hearing aid algorithm research: open master hearing aid (openMHA). *arXiv preprint arXiv:2103.02313*.

Keidser, G., Dillon, H., Flax, M., Ching, T., and Brewer, S. (2011). The NAL-NL2 prescription procedure. *Audiology research*, 1(1).

Kendrick, P. (2021). Hearing aid speech enhancement using U-Net convolutional neural networks. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Killion, M. C. (1978). Revised estimate of minimum audible pressure: Where is the "missing 6 dB"? *The Journal of the Acoustical Society of America*, 63(5):1501–1508.

Kim, S., Hori, T., and Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kochkin, S. (2010). Marketrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *The Hearing Journal*, 63(1):19–20.

Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., and Brand, T. (2016). Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by Plomp with a quantitative processing model. *Trends in hearing*, 20:2331216516655795.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697.

Kuk, F. K. (1996). Theoretical and practical considerations in compression hearing aids. *Trends in Amplification*, 1(1):5–39.

Kuriger, M. and Lesimple, C. (2012). Frequency composition™: A new approach to frequency-lowering.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Lamarche, L., Giguère, C., Gueaieb, W., Aboulnasr, T., and Othman, H. (2010). Adaptive environment classification system for hearing aids. *The Journal of the Acoustical Society of America*, 127(5):3124–3135.

Lesica, N. A. (2018). Why do hearing aids fail to restore normal auditory perception? *Trends in Neurosciences*, 41(4):174–185.

Levitt, H. (2001). Noise reduction in hearing aids: A review. *Journal of Rehabilitation Research and Development*, 38(1):111–122.

Liberman, M. C. and Kujawa, S. G. (2017). Cochlear synaptopathy in acquired sensorineural hearing loss: Manifestations and mechanisms. *Hearing research*, 349:138–147.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., et al. (2017). Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734.

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. CRC press.

Loughrey, D. G., Kelly, M. E., Kelley, G. A., Brennan, S., and Lawlor, B. A. (2018). Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: a systematic review and meta-analysis. *JAMA otolaryngology–head & neck surgery*, 144(2):115–126.

Loweimi, E., Bell, P., and Renals, S. (2019). On learning interpretable CNNs with parametric modulated kernel-based filters. In *Proc. Interspeech 2019*, pages 3480–3484.

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440.

Luo, Y. and Mesgarani, N. (2018). TasNet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

Luo, Y. and Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266.

Lybarger, S. (1963). Simplified fitting system for hearing aids. *Canonsburg, Pennsylvania: Radioear Corp*.

Mackersie, C. L., Boothroyd, A., and Garudadri, H. (2020). Hearing aid self-adjustment: Effects of formal speech-perception test and noise. *Trends in Hearing*, 24.

Magnusson, L., Claesson, A., Persson, M., and Tengstrand, T. (2013). Speech recognition in noise using bilateral open-fit hearing aids: The limited benefit of directional microphones and noise reduction. *International Journal of Audiology*, 52(1):29–36.

Malinin, A. and Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Marrone, N., Mason, C. R., and Kidd Jr, G. (2008). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *The Journal of the Acoustical Society of America*, 124(5):3064–3075.

Martin-Donas, J. M., Gomez, A. M., Gonzalez, J. A., and Peinado, A. M. (2018). A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal Processing Letters*, 25(11):1680–1684.

Martinez, A. M. C., Spille, C., Roßbach, J., Kollmeier, B., and Meyer, B. T. (2022). Prediction of speech intelligibility with dnn-based performance measures. *Computer Speech & Language*, 74:101329.

Mateer, E. J., Huang, C., Shehu, N. Y., and Paessler, S. (2018). Lassa fever–induced sensorineural hearing loss: A neglected public health and social burden. *PLoS neglected tropical diseases*, 12(2):e0006187.

Mathers, C. D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442.

McCormack, A. and Fortnum, H. (2013). Why do people fitted with hearing aids not wear them? *International Journal of Audiology*, 52(5):360–368.

McDermott, H. J. (2011). A technical comparison of digital frequency-lowering algorithms available in two current hearing aids. *PLoS One*, 6(7).

Meng, Z., Parthasarathy, S., Sun, E., Gaur, Y., Kanda, N., Lu, L., Chen, X., Zhao, R., Li, J., and Gong, Y. (2021). Internal language model estimation for domain-adaptive end-to-end speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 243–250. IEEE.

Michelsanti, D. and Tan, Z.-H. (2017). Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *Proc. Interspeech 2017*, pages 2008–2012.

Mitra, V., Franco, H., Graciarena, M., and Mandal, A. (2012). Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4120. IEEE.

Moore, A. H., Hafezi, S., Vos, R., Brookes, M., Naylor, P. A., Huckvale, M., Rosen, S., Green, T., and Hilkhuysen, G. (2021). A binaural mvdr beamformer for the 2021 Clarity enhancement challenge: ELO-SPHERES consortium system description. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Moore, B., Glasberg, B., and Stone, M. (1999a). Use of a loudness model for hearing aid fitting: III. a general method for deriving initial fittings for hearing aids with multi-channel compression. *British Journal of Audiology*, 33(4):241–258.

Moore, B. C. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The journal of the acoustical society of America*, 74(3):750–753.

Moore, B. C. and Glasberg, B. R. (1993). Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech. *The Journal of the Acoustical Society of America*, 94(4):2050–2062.

Moore, B. C., Glasberg, B. R., and Stone, M. A. (2010). Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF. *International Journal of Audiology*, 49(3):216–227.

Moore, B. C., Vickers, D. A., Plack, C. J., and Oxenham, A. J. (1999b). Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism. *The Journal of the Acoustical Society of America*, 106(5):2761–2778.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165.

Nordqvist, P. and Leijon, A. (2004). An efficient robust sound classification algorithm for hearing aids. *The Journal of the Acoustical Society of America*, 115(6):3033–3041.

Nossair, Z. B. and Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *The Journal of the Acoustical Society of America*, 89(6):2978–2991.

Oneaţă, D., Caranica, A., Stan, A., and Cucu, H. (2021). An evaluation of word-level confidence estimation for end-to-end automatic speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 258–265. IEEE.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Parent, T. C., Chmiel, R., and Jerger, J. (1997). Comparison of performance with frequency transposition hearing aids and conventional hearing aids. *Journal-American Academy of Audiology*, 8:355–365.

Pascual, S., Bonafonte, A., and Serrà, J. (2017). SEGAN: Speech enhancement generative adversarial network. In *Proc. Interspeech 2017*, pages 3642–3646.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037.

Patterson, R. D., Allerhand, M. H., and Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894.

Payton, K. L. and Braida, L. D. (1999). A method to determine the speech transmission index from speech waveforms. *The Journal of the Acoustical Society of America*, 106(6):3637–3648.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P., and Goldsworthy, R. (2002). Computing the STI using speech as a probe stimulus. *Past, present and future of the speech transmission index*, pages 125–138.

Plapous, C., Marro, C., and Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2098–2108.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Radiocommunication Sector of ITU (2011). Recommendation ITU-R BS.1770-4. Algorithms to measure audio programme loudness and true-peak audio level.

Ragni, A., Li, Q., Gales, M. J., and Wang, Y. (2018). Confidence estimation and deletion prediction using bidirectional recurrent neural networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 204–211. IEEE.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Recio-Spinoso, A. and Cooper, N. P. (2013). Masking of sounds by a background noise–cochlear mechanical correlates. *The Journal of Physiology*, 591(10):2705–2721.

Rhebergen, K. S. and Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America*, 120(6):3988–3997.

Roßbach, J., Huber, R., Röttges, S., Hauth, C. F., Biberger, T., Brand, T., Meyer, B. T., and Rennies, J. (2022). Speech intelligibility prediction for hearing-impaired listeners with the LEAP model. In *Proc. Interspeech 2022*, pages 3498–3502.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America*, 101(4):2151–2163.

Sachs, M. B., Voigt, H. F., and Young, E. D. (1983). Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology*, 50(1):27–45.

Sachs, M. B. and Young, E. D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *The Journal of the Acoustical Society of America*, 68(3):858–875.

Salorio-Corbetto, M., Baer, T., and Moore, B. C. (2017). Evaluation of a frequency-lowering algorithm for adults with high-frequency hearing loss. *Trends in hearing*, 21.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131(5):4134–4151.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, 54(sup2):100–107.

Schröder, D. and Vorländer, M. (2011). Raven: A real-time framework for the auralization of interactive virtual environments. In *Forum acusticum*, pages 1541–1546. Aalborg Denmark.

Schwerin, B. and Paliwal, K. (2014). An improved speech transmission index for intelligibility prediction. *Speech Communication*, 65:9–19.

Scollie, S., Seewald, R., Cornelisse, L., Moodie, S., Bagatto, M., Laurnagaray, D., Beaulac, S., and Pumford, J. (2005). The desired sensation level multistage input/output algorithm. *Trends in amplification*, 9(4):159–197.

Sharma, D. et al. (2016). A data-driven non-intrusive measure of speech quality and intelligibility. *Speech Communication*, 80:84–94.

Sharma, D., Hilkhuysen, G., Gaubitch, N. D., Naylor, P. A., Brookes, M., and Huckvale, M. (2010). Data driven method for non-intrusive speech intelligibility estimation. In *2010 18th European Signal Processing Conference*, pages 1899–1903. IEEE.

Shaw, E. A. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6):1848–1861.

Simpson, A., Hersbach, A. A., and McDermott, H. J. (2005). Improvements in speech perception with an experimental nonlinear frequency compression hearing device. *International Journal of Audiology*, 44(5):281–292.

Slaney, M. et al. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35(8).

Sliwinska-Kowalska, M., Davis, A., et al. (2012). Noise-induced hearing loss. *Noise and Health*, 14(61):274.

Soni, M. H., Shah, N., and Patil, H. A. (2018). Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5039–5043. IEEE.

Sørensen, C., Boldt, J. B., Gran, F., and Christensen, M. G. (2016). Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1358–1362. IEEE.

Sørensen, C. et al. (2017a). Non-intrusive intelligibility prediction using a codebook-based approach. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 216–220. IEEE.

Sørensen, C., Xenaki, A., Boldt, J. B., and Christensen, M. G. (2017b). Pitch-based non-intrusive objective intelligibility prediction. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 386–390. IEEE.

Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (2018a). Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, 48:51–66.

Spille, C., Kollmeier, B., and Meyer, B. T. (2018b). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, 52:123–140.

Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Steinmetz, C. J. and Reiss, J. (2021). pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society.

Stone, M. A. and Moore, B. C. (1999). Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses. *Ear and Hearing*, 20(3):182–192.

Stone, M. A. and Moore, B. C. (2002). Tolerable hearing aid delays. II. Estimation of limits imposed during speech production. *Ear and Hearing*, 23(4):325–338.

Stone, M. A. and Moore, B. C. (2003). Tolerable hearing aid delays. III. Effects on speech production and perception of across-frequency variation in delay. *Ear and Hearing*, 24(2):175–183.

Stone, M. A. and Moore, B. C. (2005). Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects. *Ear and Hearing*, 26(2):225–235.

Stone, M. A., Moore, B. C., Meisenbacher, K., and Derleth, R. P. (2008). Tolerable hearing aid delays. V. Estimation of limits for open canal fittings. *Ear and Hearing*, 29(4):601–617.

Swarup, P., Maas, R., Garimella, S., Mallidi, S. H., and Hoffmeister, B. (2019). Improving ASR confidence scores for alexa using acoustic and hypothesis embeddings. In *Interspeech*, pages 2175–2179.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Taghia, J. and Martin, R. (2013). Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):6–16.

Tammen, M., Gode, H., Kayser, H., Nustede, E. J., Westhausen, N. L., Anemüller, J., and Doclo, S. (2021). Combining binaural LCMP beamforming and deep multi-frame filtering for joint dereverberation and interferer reduction in the Clarity-2021 challenge. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Tan, K. and Wang, D. (2021). Towards model compression for deep learning based speech enhancement. *IEEE/ACM transactions on audio, speech, and language processing*, 29:1785–1794.

Thiemann, J., Ito, N., and Vincent, E. (2013). Demand: a collection of multi-channel recordings of acoustic noise in diverse environments. In *Proc. Meetings Acoust*, pages 1–6.

Tremblay, K. and Miller, C. (2014). How neuroscience relates to hearing aid amplification. *International Journal of Otolaryngology*, 2014.

Valentini-Botinhao, C., Wang, X., Takaki, S., and Yamagishi, J. (2016). Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152.

van Buuren, R. A., Festen, J. M., and Houtgast, T. (1999). Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality. *The Journal of the Acoustical Society of America*, 105(5):2903–2913.

van den Brink, G. (1964). Detection of tone pulse of various durations in noise of various bandwidths. *The Journal of the Acoustical Society of America*, 36(6):1206–1211.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2017). An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*, 25(1):115–119.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2018). An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2153–2166.

Vary, P. and Martin, R. (2006). *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*, pages 5998–6008.

Veaux, C., Yamagishi, J., and King, S. (2013). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE.

Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J. R., Saurous, R. A., Weiss, R. J., Jia, Y., and Moreno, I. L. (2019). VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proc. Interspeech 2019*, pages 2728–2732.

Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (2018). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wang, Z.-Q., Wichern, G., Watanabe, S., and Le Roux, J. (2022). STFT-domain neural speech enhancement with very low algorithmic latency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Willems, P. J. (2000). Genetic causes of hearing loss. *New England Journal of Medicine*, 342(15):1101–1109.

Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., and Saurous, R. A. (2019). Differentiable consistency constraints for improved deep speech enhancement. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 900–904. IEEE.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., and Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3846–3857.

World Health Organization (2021). *World Report on Hearing*. World Health Organization.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2013). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19.

Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K., and Nakatani, T. (2017). Predicting speech intelligibility using a Gammachirp envelope distortion index based on the signal-to-distortion ratio. In *Proc. Interspeech 2017*, pages 2949–2953.

Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K., and Nakatani, T. (2019). Speech intelligibility prediction with the dynamic compressive gammachirp filterbank and modulation power spectrum. *Acoustical Science and Technology*, 40(2):84–92.

Yang, S. J., Wisdom, S., Gnegy, C., Lyon, R. F., and Savla, S. (2021). Listening with googlears: Low-latency neural multiframe beamforming and equalization for hearing aids. In *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.

Zahorian, S. A. and Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America*, 94(4):1966–1982.

Zahorian, S. A. and Rothenberg, M. (1981). Principal-components analysis for low-redundancy encoding of speech spectra. *The Journal of the Acoustical Society of America*, 69(3):832–845.

Zezario, R. E., Chen, F., Fuh, C.-S., Wang, H.-M., and Tsao, Y. (2022). MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids. In *Proc. Interspeech 2022*, pages 3944–3948.

Zezario, R. E. et al. (2020). Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model. In *APSIPA ASC*, pages 482–486. IEEE.

Zhang, H., Zhang, X., and Gao, G. (2018). Training supervised speech separation system to improve stoi and pesq directly. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5374–5378. IEEE.

Zhang, J., Zorilă, C., Doddipatla, R., and Barker, J. (2020). On end-to-end multi-channel time domain speech separation in reverberant environments. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6389–6393. IEEE.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, 109(2):648–670.

Zhao, Y., Xu, B., Giri, R., and Zhang, T. (2018). Perceptually guided speech enhancement using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5074–5078. IEEE.

Žmolíková, K. and Cernock, J. (2021). BUT system for the first Clarity enhancement challenge. *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*.

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., and Černockỳ, J. (2019). SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814.