Mutual information based speech intelligibility prediction and its application to hearing aid fitting

by

Ganlong Wang

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Electrical Engineering.

Victoria University of Wellington
2022

Abstract

Speech enhancement was, is, and will be the key technology for digital speech transmission. When developing speech enhancement algorithms, the intelligibility or the quality of the processed speech needs to be evaluated. Intelligibility is more fundamental than quality. The evaluation of speech intelligibility can be carried out through subjective listening tests and objective metrics. Carrying out subjective listening tests is time-consuming and costly. Using the objective metrics is more efficient. Thus, there has been an increasing research in speech intelligibility prediction.

Speech intelligibility can be measured in terms of the information received by a listener. This thesis aims at developing a mutual information-based speech intelligibility predictor (SIP) and using the mutual information-based predictor to assist the fitting process in hearing instruments. To achieve this goal, this thesis carried out three studies.

First, it studied the modeling of the transmitted message. For mutual information-based SIPs, the most important thing is to determine the transmitted message. This thesis studied two approaches of modeling of the message: one is the continuous-valued sound, and the other is the discrete-valued linguistic message. Two corresponding SIPs were developed. By comparing their predicted intelligibility results with the psychometric curves, which are the subjective intelligibility scores, it shows that the modeling of discrete-valued message gives a better match to the psychometric curves.

Second, based on the modeling of the discrete-valued message, this thesis proposed a mutual information-based SIP. Since the discrete-valued message cannot be obtained from a speech signal, the proposed SIP calculates the mutual information between the clean speech and the received

speech, instead of calculating the mutual information between the message and the received speech. The proposed SIP considers frequency correlation for both the clean speech and the received speech. The evaluation results show that the proposed SIP performs better than the existing state-of-the-art mutual information-based SIPs.

Third, this thesis proposed an automatic fitting tool for the nonlinear frequency compression (NFC) operator, which is a frequency lowering operator used in hearing instruments. The automatic fitting tool adjusts the parameter in the NFC by maximizing the mutual information between the message and the frequency-lowered speech. To evaluate the automatic fitting tool, the parameter was also searched by listening tests. The results show that the parameter determined by the automatic fitting tool is consistent to the parameter determined by the listening tests.

Acknowledgments

First and foremost, I thank Bastiaan Kleijn for his patience, support, and supervision throughout my studies. His ways of thinking inspired me to seek the truth and changed my attitude towards the research. His seriousness in research and hard work ethics have influenced me a lot and will be beneficial to all my life. I would like to deeply express my gratitude to him.

I would also like to thank Victoria University of Wellington and my colleagues for their support during the Covid-19 period. Their support and help make me calm down and feel secure. Without them I cannot get out of this difficult situation quickly.

I also thank Mengzhu Yan and Ge Lai for their encouragement and accompany. I enjoyed the time when we went to the Victoria Recreation Center together. This helps me balance the work and life a bit.

Finally, I thank my family for their continual love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Hearing instruments are intended to improve speech intelligibility for hearing-impaired people. When developing an enhancement algorithm, its performance is evaluated through a listening test, which is time-consuming. Moreover, audiologists need to fit a hearing aid for each individual user due to their different audiograms and cognitive characteristics [4]. A less experienced audiologist cannot maximize the intelligibility improvement for hearing aid users. For these reasons, hearing aid manufactures are seeking a more efficient way to measure speech intelligibility for their products.

Nowadays, there is a growing consumer market for hearing aids. According to the statistical data from World Health Organization in the year of 2015, there are 360 million people worldwide, which takes up over 5% of the world's population, having disabling hearing loss [5]. Among them, 328 million are adults and 32 million are children. Among the group of elderly people, hearing loss is more prevalent, because approximately one-third of people over 65 years old suffer disabling hearing loss. With the help from hearing aids, hearing-impaired people can benefit in their daily life. However, the current production of hearing aids meets less than 10%

of the global need.

In the development of an enhancement algorithms, speech intelligibility of the enhanced speech needs to be measured. Currently, there are different subjective measures for speech intelligibility, such as nonsense syllable test, word and sentence tests [6–8]. These subjective tests provide a reliable way to assess speech intelligibility, but they are time-consuming and require access to trained listeners. For example, Diagnostic Rhyme Test (DRT) is a widely used word test for evaluating speech intelligibility. In this test, a listener is presented to a word, and is asked to select the presented one from a pair of words. This process is repeated for 192 different pairs of words [7].

An efficient speech intelligibility measure is also required for the fitting of hearing aids. After a hearing aid has been manufactured with new enhancement algorithms, the hearing aid needs to be fitted for individual hearing-impaired person. For example, frequency lowering techniques are used by hearing aids for the listeners with severe hearing loss [9]. Nonlinear frequency compression (NFC) is one of the frequency lowering techniques and its parameters need to be fitted [10]. Speech intelligibility prediction establishes the starting point for the fitting, as it produces a mean value for all hearing-impaired listeners having the same audiogram. The fitting effect is largely dependent on the experience of audiologists. An efficient speech intelligibility measure can be expected to provide assistance during the fitting process.

In summary, speech intelligibility can be used to guide the development of intelligibility enhancement algorithms and the fitting of hearing aids. However, speech intelligibility obtained by the subjective tests is a time-consuming process. To obtain speech intelligibility more efficiently, objective intelligibility measure is necessary.

1.2 Limitation of the Existing Speech Intelligibility Predictors

Since listening test is costly, there have been increasing research in objective speech intelligibility predictors (SIPs). As speech intelligibility is a measure of how well the meaning of a spoken message is transmitted to a listener [11], it can be quantified in terms of information. In [12, 13], SIIB (speech intelligibility in bits) and SIIB$^{\text{Gauss}}$ (SIIB for Gaussian communication channel) estimate speech intelligibility by calculating the mutual information between a message signal and a received signal. The *message* is defined as the content of a speech signal and it has nothing to do with the other information, such as talker, environment, emotion, etc. The message is embedded in speech and carries all the information that contributes to speech intelligibility. SIIB and SIIB$^{\text{Gauss}}$ model the transmission from the message and the clean speech as a production channel, which is an additive noise channel.

SIIB and SIIB$^{\text{Gauss}}$ achieved good intelligibility prediction results [13]. They consider correlation in frequency for both the clean speech and the received speech. As SIIB and SIIB$^{\text{Gauss}}$ calculate mutual information as the sum of the mutual information for each band, the bands should be independent. However, the bands of the original log-auditory spectrogram are correlated for the clean speech and the received speech. To remove the correlation across frequency bands, Karhunen-Loève transform (KLT) is applied to both the clean speech and the received speech.

There are three problems in SIIB and SIIB$^{\text{Gauss}}$. First, they assume a constant correlation coefficient between the transformed message and the transformed clean speech across the channels. However, it is unlikely to be true, because the variances of the transformed message across the transformed bands are not approximately constant. In other words, the correlation coefficient should gradually decrease, as the index of the transformed band increases.

Second, SIIB and SIIB$^{\text{Gauss}}$ treat the transformed bands as uncorrelated in frequency for the transformed received speech. This is not true, as the KLT matrix is derived based on the covariance matrix of the clean speech, instead of the covariance matrix of the received speech.

Third, SIIB and SIIB$^{\text{Gauss}}$ require the numbers of bands in the clean speech and the received speech to be identical, since they calculate mutual information for each transformed band individually. For the frequency-lowered speech, the high frequency components are lowered at low frequency. Since the bandwidth of the frequency-lowered speech is smaller than the bandwidth of the original clean speech, SIIB and SIIB$^{\text{Gauss}}$ cannot be directly applied for speech intelligibility prediction.

This thesis aims at solving these problems. The research goal will be outlined in the next section.

1.3 Research Goals

The overall goal of this thesis is to develop a mutual information-based SIP and use it to assist the fitting of hearing instruments. In order to fulfill the overall goal, three objectives have been established to guide this research.

1. Evaluate two assumptions of the probability distribution of the message. The first one assumes the message has a continuous-valued distribution, and the second one assumes the message has a discrete-valued distribution.

2. Propose a better mutual information-based SIP based on the correct modeling of the message.

3. Propose an automatic fitting tool for the nonlinear frequency compression operator, which is one of the frequency lowering operators. This research goal is not part of solving the three problems discussed in Section 1.2.

1.4 Organization of the Thesis

This thesis consists of seven chapters. The remainder of the thesis is organized as follows. Chapter 2 presents the literature review. Chapters 3 to 6 present the main contributions of the thesis. Chapter 7 concludes the work and discusses the possible future works. A brief overview of each chapter is as follows.

Chapter 2: Literature Review

This chapter introduces the current SIPs based on different categories, which include SNR, correlation coefficient, neural network, and mutual information. The transmission of the message is studied for the perspective of communication theory. Finally, it introduces the background of the frequency lowering operators in hearing instruments.

Chapter 3: Speech Transmission Model based on Continuous-valued Message

This chapter studies the validity of the modeling of the continuous-valued message. The transmitted message, the clean speech, and the received speech are assumed Gaussian distributed. The transmitted message is modeled as continuous-valued sound. The constant correlation coefficients between the transformed message and the transformed clean speech are replaced by the correlation coefficients that were calculated via CHAINS data set. The SIP is based on the continuous-valued message. Speech intelligibility was estimated for four types of noisy speech signals. The experimental results show that the predicted intelligibility did not match well to the psychometric curves, which are subjective intelligibility results.

Chapter 4: Speech Transmission Model based on Discrete-valued Message

This chapter studies the validity of the modeling of the discrete-valued message. The transmitted message is modeled as discrete-valued linguis-

tic unit. Given a speech signal, the discrete-valued message cannot be obtained. However, the mutual information rate between the transmitted message and the clean speech can still be estimated from the linguistic study. Mutual information between the transmitted message and the received speech is calculated for each band, and then summed up across the frequency bands. The evaluation results show that the predicted intelligibility based on the discrete-valued message matches to the psychometric curves better than the predicted intelligibility based on the continuous-valued message.

Chapter 5: Mutual Information based Speech Intelligibility Predictor

This chapter proposes a mutual information-based SIP for normal-hearing listeners based on the findings in the Chapters 3 and 4. As the discrete-valued message cannot be generated from a clean speech signal, the SIP calculates the mutual information between the clean speech and the received speech. The proposed SIP performs better than the state-of-the-art mutual information-based SIP.

Chapter 6: Mutual Information based Frequency Lowering Fitting

This chapter proposes an automatic tool, which is a mutual information-based SIP, for the fitting of the NFC operator. The NFC operator is one of the frequency lowering operators, and has two parameters: one is the cut-off frequency and the other is the compression ratio [10]. Since the NFC operator does not change the signal below the cutoff frequency, the mutual information between the clean speech and the frequency-lowered speech is infinite. To solve this problem, the modeling of the continuous-valued message is used. As the NFC operator changes the spectrum of speech, the frequency-lowered speech is not familiar to the listeners. This requires acclimatization procedure before carrying out speech recognition test. To improve the efficiency, we propose a sound distinction test instead recognition test. The sound distinction test and the proposed SIP show similar

compression ratios that maximize the intelligibility of frequency-lowered speech.

Chapter 7: Conclusions and Future Work

This chapter summarizes the findings in Chapters 3 to 6 and present the future work.

Chapter 2

Literature Review

This chapter starts by introducing basic concepts about speech intelligibility, such as the definition and enhancement of speech intelligibility. Then, a review of current objective measures of speech intelligibility is given. It addition, the message transmission is investigated from the perspective of communication theory. At last, the background of frequency lowering is introduced.

2.1 Fundamentals of Speech Intelligibility

2.1.1 What is speech intelligibility?

The word *intelligibility* refers to how much a conveyed message can be interpreted correctly by a recipient. In speech communication, it is the extent to which a spoken message can be understood by a listener. Speech intelligibility can be measured quantitatively at different levels of abstraction, such as the acoustic signal waveform, the sequence of phonemes, and the sequence of spoken words [14]. For example, when speech is measured at the level of spoken words in a listening test, if 100 words are spoken by a speaker and 90 words are received correctly by a listener, then the speech intelligibility is 0.9.

2.1.2 What affects speech intelligibility?

Speech intelligibility can be affected by each module of a communication system, i.e., the speaker, the transmission channel, and the listener. In this section, we introduce how each module can affect the speech intelligibility.

When speech is produced, its intelligibility is affected by the speech material, the speaker, and the loudness. As the contextual information can be used by the listener, the speech sentence transmitted over a noisy channel has a higher intelligibility than the non-sense syllables or single words. In addition, the speaker's ability to speak clearly and linguistically correctly affects speech intelligibility [15, 16]. Speech intelligibility is also reduced if the speech level is too low or too high [17–19]. This phenomenon might be caused by the excessive vocal effort, which leads to the change in speech intelligibility [20, 21].

When speech is transmitted through the channel, its intelligibility is subject to the characteristics of channel, such as the environmental noise, reverberation time, and the spatial configuration of the speaker and the noise source [22–25]. For example, people often feel hard to understand each other when talking in a cocktail party, where the target speaker is masked by the surrounding irrelevant speakers. By changing the location of the listener with respect to the target speaker and the other irrelevant speakers, we can improve the transmission channel, thus increase the intelligibility of the received speech.

The cognitive characteristic and hearing ability of listeners also have profound influence on the speech intelligibility. Native listeners can better understand a noisy speech signal than non-native listeners. In addition, hearing-impaired people usually have higher hearing threshold at high frequency [26] than normal-hearing people. This makes it particularly difficult for them to understand speech, because the high frequency covers some phonemes, which are important to speech intelligibility. For example, the phoneme /s/ has most signal power located at high frequency, and plays an important role in grammar. Without the ability to perceive

high frequency information, /s/ is hardly recognized by listeners.

2.1.3 How to enhance speech intelligibility?

As speech intelligibility is affected by the speaker, the transmission channel, and the listener, it can be enhanced from these three aspects.

When people talk in a noisy environment, like in a cocktail party, the speaker might raise his or her voice level, change the pitch or speech rate to make it audible by the listener [22, 27]. The phenomenon of this involuntary change in vocal tract is known as Lombard effect [28–30].

Speech intelligibility can also be enhanced based on the characteristics of the transmission channel. In mobile communication, the listener receives both far-end speech from mobile phone and environmental noise. If the environmental noise is loud, speech intelligibility decreases. To increase speech intelligibility, the mobile phone can detect the environmental noise and reprocess the speech signal prior to play-out [23, 31, 32].

Speech intelligibility is able to be enhanced on the listener's side. In hearing aid industry, various signal processing algorithms have been developed, which aim at compensating the hearing loss of hearing-impaired people [33]. One approach is to amplify speech level, such as automatic gain control (AGC) and wide dynamic-range compression (WDRC), which automatically adjusts the gain based on the level of input signals [34–38].

2.2 Current Objective Measures of Speech Intelligibility

This section starts by introducing the purpose of speech intelligibility prediction. Then, different speech intelligibility predictors are described.

2.2.1 Speech intelligibility prediction

As speech intelligibility measures how well the meaning of a spoken message is transmitted, it can be used to guide the development of speech enhancement algorithms. Speech intelligibility can be evaluated by the subjective listening tests that require the recruitment of trained listeners. This is time-consuming and costly. To obtain speech intelligibility more efficiently, many speech intelligibility predictors have been developed.

SIPs can be classified into intrusive SIPs and non-intrusive SIPs. Intrusive SIPs require a reference signal (either clean speech or noise) and a degraded signal, while non-intrusive SIPs require only a degraded signal. The current intrusive SIPs perform better than the non-intrusive SIPs [39, 40]. Intrusive SIPs consist of model-driven metrics and data-driven metrics. The model-driven metrics use a certain relationship between the reference and the degraded speech to predict speech intelligibility. The data-driven metrics train the relationship, and predict speech intelligibility based on the trained relationship.

For model-driven metrics, the certain relationships include mutual information, correlation coefficient, coherence, and SNR. SIIB and $\text{SIIB}^{\text{Gauss}}$ calculate the mutual information between the message and the degraded signal [13, 41]. STOI and ESTOI predict speech intelligibility based on the correlation coefficient between the temporal envelopes of the clean and degraded speech signals in short-time segments [42, 43]. HASPI uses correlation coefficients based on the temporal envelopes and the narrow-band time domain signals produced by the auditory filter bank [44]. CSII uses the magnitude-squared coherence to estimate the powers of speech and distortion, which are then used for intelligibility prediction [45]. sEPSM and mr-sEPSM predict speech intelligibility based on the SNR of the signal envelopes after the processing by the auditory modulation filterbank [46, 47].

For data-driven metrics, the relationship can be represented by a neural network or a model whose parameters are determined through training.

In [48], a neural network using the auditory spectra of speech as inputs is proposed. This neural network consists of a convolutional layer and three fully connected layers. In [49], a neural network using the time-domain speech as inputs is proposed. This neural network is based on U-net [50], which consists of multiple layers of convolution and transpose convolution at encoder and decoder, respectively. Differently from neural network, wSTMI [51] proposes a linear model, whose weight parameters are optimized through training. wSTMI is based on STMI [52], which predicts speech intelligibility by calculating the correlation coefficient between the spectro-temporal modulation spectrograms of the clean and the degraded signals. wSTMI assigns trained weights for the correlation coefficients to predict speech intelligibility.

At the moment, there is no conclusion about which SIP performs best, as the performance strongly depends on the data sets. However, we can still obtain a view about which SIPs perform relatively well. In [41], where SIIB and $\text{SIIB}^{\text{Gauss}}$ are compared with ESTOI, HASPI, and the other nine SIPs, SIIB and $\text{SIIB}^{\text{Gauss}}$ perform best among 13 data sets. In [51], the latest developed wSTMI was compared with SIIB, $\text{SIIB}^{\text{Gauss}}$, ESTOI, HASPI, and the other nine SIPs. By using some new data sets, wSTMI performs best, with ESTOI ranked second, HASPI ranked third, and $\text{SIIB}^{\text{Gauss}}$ ranked fourth.

The theory of speech intelligibility prediction starts developing from the articulation index theory in 1940s [19, 53, 54]. The articulation index describes the relationship between the amount of audible speech (in terms of percentage) and the channel signal-to-noise ratio (SNR). After the articulation index theory, many speech intelligibility predictors have been proposed. These predictors are based on different categories that include the generic SNR [45,47,55,56], the correlation coefficient [44,51,57], neural network [48, 49], and mutual information [13, 41]. In the following section, we will give an overview of the state-of-the-art speech intelligibility predictors from each category.

2.2.2 Generic SNR based metric

The SNR-based metric estimates speech intelligibility by calculating the power of narrowband speech and noise signals. Instead of calculating SNR directly, some speech intelligibility predictors use the signal-to-noise and distortion ratio (SDR), or calculate the SNR indirectly. Thus, we use the word "generic" to include these speech intelligibility predictors.

The SNR can be calculated in the auditory domain and modulation domain. The auditory domain refers to the space produced by the narrowband signals after the processing of the auditory filter. The modulation domain refers to the space produced by the modulated signals of these narrowband signals. We will describe two SIPs related to the auditory domain and one SIP related to the modulation domain.

Speech intelligibility index

Speech intelligibility index (SII) [55] is the first standard intelligibility predictor that is evolved from the articulation index (AI) theory [19]. Speech intelligibility can be derived from SII based on a transfer function. The transfer function depends on the speech material and the proficiency of the talkers and the listeners.

We first have a quick review of the articulation index theory, since it is fundamental to the SII. When calculating the articulation index, the speech and the noise are decomposed into a set of narrowband signals, and they are assumed independent in frequency. Let P and P_j be the probabilities of recognition error for a fullband speech and its narrowband speech signal at band j, respectively. As the bands are independent, we have [19]

$$P = P_1 \cdot P_2 \cdots P_J, \tag{2.1}$$

where J is the total number of frequency bands. The relationship between the probability of recognition error and the audibility of speech can be expressed as [53]

$$P_j = 10^{-AI_j} \tag{2.2}$$

where AI_j is the articulation index for the narrowband speech at band j. Substitute (2.2) into (2.1), we have

$$\begin{aligned} P &= 10^{-\sum_{j=1}^{J} AI_j} \\ &= 10^{-AI}, \end{aligned} \tag{2.3}$$

where AI is the articulation index for the fullband speech. Thus, the articulation index of a speech signal can be calculated as

$$AI = \sum_{j=1}^{J} AI_j. \tag{2.4}$$

The calculation of AI was improved later by considering the spread of masking, the distortion caused by extremely high level voice and the band importance function. The AI changed its name to SII and was standardized in 1997 [55]. Fig. 2.1 shows the diagram for computing the SII. Here we introduce main steps of the calculation of SII. A detailed calculation can be found in [55]. The disturbance spectrum D_j is determined as the maximal value between the masking spectrum Z_j and the internal noise spectrum X_j:

$$D_j = \max(Z_j, X_j),$$

where Z_j depends on self-masking and the masking from lower frequency bands. The level distortion factor L_j is given by

$$L_j = \min(1 - \frac{S_j - U_j - 10}{160}, 1),$$

where S_j is the power spectrum of the clean speech, and U_j is the reference spectrum level of normal speech [55]. The SNR for band j is calculated as

$$SNR_j = \frac{S_j - D_j + 15}{30}.$$

The SNR_j value is limited within the range between 0 and 1. The final SII value is computed as

$$SII = \sum_{j=1}^{18} W_j \cdot L_j \cdot SNR_j,$$

where W_j is the band importance function.

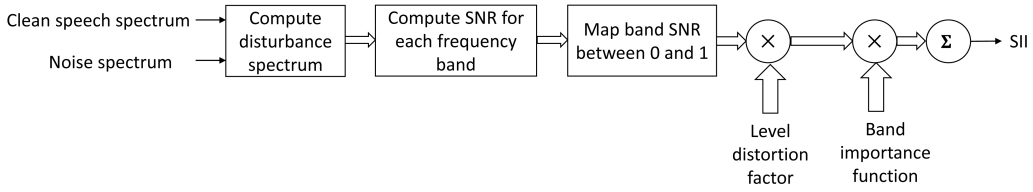


Figure 2.1: Diagram of the SII.

As the SII measure computes the band SNR based on the long-term spectrum, it can effectively predict intelligibility of speech affected by stationary noise, but cannot be applied for fluctuating noise maskers. Moreover, speech enhancement algorithms may disrupt the one-to-one band relationship between the clean speech and the degraded speech. For example, frequency compression operators [58, 59] that are used in hearing instruments generate a speech signal with reduced bandwidth. The intelligibility cannot be evaluated by the SII, because the SII is based on independent frequency bands.

Coherence-based speech intelligibility index

Instead of only using SNR, the coherence-based speech intelligibility index (CSII) is based on signal-to-noise and distortion ratio (SDR), which is computed from the magnitude-squared coherence (MSC) between the clean speech and the processed speech. Compared with the SII, CSII takes into account signal distortion in hearing aids, like peak-clipping [45], [60].

The MSC measures the strength of linear relationship between two signals. Assuming $S_{XX}(w)$ and $S_{YY}(w)$ are the autospectral densities of the input and the output of the system, and $S_{XY}(w)$ is the cross-spectral density, the MSC between the spectra of the input and the output can be com-

puted as

$$|\gamma(w)|^2 = \frac{|S_{XY}(w)|^2}{S_{XX}(w)S_{YY}(w)}. \tag{2.5}$$

In a linear system, the MSC between the input and the output is 1. In a non-linear system, it takes a value between 0 and 1. The SDR is computed as [61]

$$SDR(w) = \frac{|\gamma(w)|^2}{1 - |\gamma(w)|^2}. \tag{2.6}$$

Note that scaling of a speech signal does not reduce speech intelligibility, as long as the loudness of speech is still in the suitable range. Since the linear relationship between the original signal and the scaled signal does not change, the MSC is 1. If we use the SII in this scenario, the SII will predict that speech intelligibility is reduced.

When calculating CSII for normal-hearing listeners, the signal is divided into a sequence of overlapped segments. Based on the energy of the segment, all the segments are classified into three levels that are high-level, mid-level, and low-level. CSII is calculated for each level. The speech intelligibility is estimated by applying a logistic function to the weighted sum of the CSII of different levels. Fig. 2.2 shows the diagram of the computation of CSII for one level. In the computation of the CSII, the SDR is calculated for each auditory critical band from the Table I of [55]. The ro-ex filter in [62] is used to model the spectral shaping properties of the auditory filter. The band $SDR_{j_{dB}}$ is calculated as

$$SDR_{j_{dB}} = 10\log_{10}\frac{\sum_k G_j(w_k)|\gamma(w_k)|^2 S_{YY}(w_k)}{\sum_k G_j(w_k)(1 - |\gamma(w_k)|^2)S_{YY}(w_k)},$$

where $G_j(w_k)$ is the magnitude of the jth band's ro-ex filter at frequency w_k. As with the computation of the SII, the CSII is a weighted average of SDR over the whole critical bands.

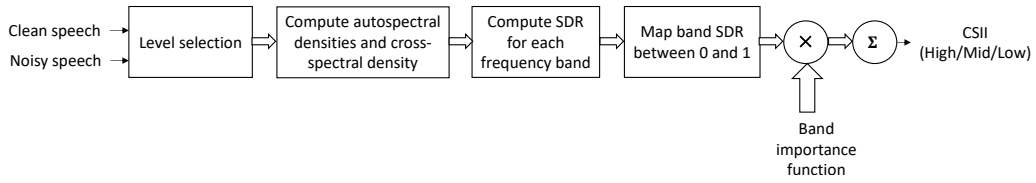


Figure 2.2: Diagram of the CSII.

For hearing-impaired listeners, the CSII considers the increased hearing threshold and speech power amplification. The hearing threshold is modeled as an internal noise. The new distortion level is the sum of the interval noise level and the original distortion level. The amplification increases speech gain, thus it partly restores the SDR.

Speech transmission index

The speech transmission index (STI) predicts speech intelligibility based on the modulation transfer function (MTF) in the modulation domain [56, 63–65]. It measures the quality of the transmission channel. The modulation transfer function (MTF) can be derived from the reverberation time of an auditorium. Thus, the STI can be used as a design tool for auditorium acoustics.

The STI does not calculate the SNR directly but derives an apparent SNR for each critical band from the MTF which measures the depth of the modulation envelope of the received signal. As a narrow-band (a critical band) speech can be viewed as a noise signal that is intensity-modulated by low-frequency signal, it can be written as

$$x(t) = \sqrt{1 + \cos(2\pi F t)}s(t),$$

where F is the modulation frequency and $s(t)$ is a narrow-band noise. A full-band speech signal is the sum of the narrow-band speech signals with different center frequencies. The modulation frequency ranges from

0.63 Hz to 12.7 Hz [66–69]. Speech signal has the frequency range from 100 Hz to 8 kHz, which means the center frequencies of narrow-band noise signals range from 100 Hz to 8 kHz. The intensity envelope of $x(t)$ is computed as

$$I_i(t) = I_{IN} \cdot (1 + \cos(2\pi F t)), \tag{2.7}$$

where $I_i(t)$ is the envelope of $x(t)$, I_{IN} is the mean of the input intensity envelope. Let $x(t)$ be transmitted through the system and the intensity envelope of the output signal is denoted by

$$I_o(t) = I_{OUT} \cdot (1 + m_F \cdot \cos(2\pi F t)), \tag{2.8}$$

where I_{OUT} is the mean of the output intensity envelope and m_F is called modulation index with the value between 0 and 1. The larger the modulation index, the less degradation the received signal suffers during the transmission. The MTF is obtained by computing the modulation indexes for different modulation frequencies.

The STI predicts speech intelligibility based on the apparent SNR which can be derived from the MTF. Assuming the noise in the transmission channel has a constant intensity envelope I_N, then the output intensity envelope can be expressed as

$$\begin{aligned} I_o(t) &= I_i(t) + I_N \\ &= I_{IN} \cdot (1 + \cos(2\pi F t)) + I_N \\ &= (I_{IN} + I_N) \cdot (1 + \frac{I_{IN}}{I_{IN} + I_N} \cos(2\pi F t)) \\ &= I_{OUT} \cdot (1 + m_F \cdot \cos(2\pi F t)). \end{aligned} \tag{2.9}$$

Then, the SNR at the modulation frequency F can be calculated by

$$SNR = \frac{m_F}{1 - m_F}.$$

Fig. 2.3 shows the diagram of computing the STI in practice. For the

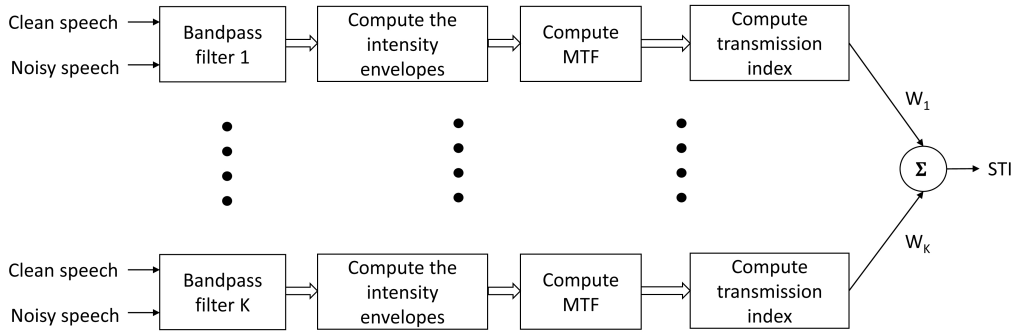


Figure 2.3: Diagram of the STI.

narrowband signals in each octave band, the MTF is computed as

$$MTF(f) = \alpha\sqrt{\frac{P_{yy}(f)}{P_{xx}(f)}},$$

where α is the ratio of the means of the intensity of the clean and noisy speech signals. $P_{xx}(f)$ and $P_{yy}(f)$ are the autospectral densities of the intensity envelopes of the clean speech and the noisy speech, respectively [70]. From the MTF, we can derive the average SNR (transmission index) in that octave band, and finally obtain the STI.

Compared with the SII, the STI computes the SNR indirectly from the MTF. The advantage of the STI is that it considers both reverberation and additive noise. As with the SII metric, the STI is based on the one-to-one mapping between the frequency bands of the clean speech and the degraded speech. Thus, it is unable to predict intelligibility of speech processed by frequency lowering operators.

2.2.3 Correlation coefficient based metric

The second type of SIPs estimate speech intelligibility based on the correlation coefficient between the clean speech and the degraded speech. The correlation coefficient can be calculated in the auditory domain and the

modulation domain. We will describe two SIPs (STOI and HASPI) that calculate the correlation coefficient in the auditory domain and one SIP (wSTMI) that calculates the correlation coefficient in the modulation domain.

Short-time objective intelligibility

The short-time objective intelligibility (STOI) metric estimates speech intelligibility by measuring the strength of linear relationship between the envelopes of the clean speech and the degraded speech in each one-third octave band [57, 71]. Human auditory perception depends on the firing rate, which is the amount of electrical pulses generated by auditory neurons within a unit time. As the envelopes determine the firing rate in the one-third octave bands over time, the strength of linear relationship between the envelopes reflects the strength of linear relationship between the firing rates of the clean speech and the degraded speech.

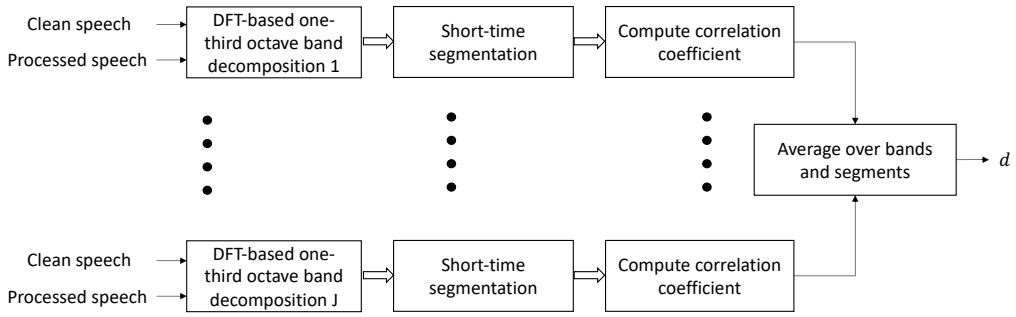


Figure 2.4: Diagram of the STOI.

Fig. 2.4 shows the diagram of computing STOI. Let $\hat{x}(k, m)$ denote the short-time Fourier transform (STFT) of the clean speech at the kth bin and the mth frame. Using STFT can capture speech features over time, as speech is short-term stationary. The temporal envelope of the clean speech

in each one-third octave band is calculated by

$$X_j(m) = \sqrt{\sum_{k \in K_j} |\hat{x}(k,m)|^2},$$

where j is the index of the one-third octave band and K_j is the set of bins belonging to the jth one-third octave band. Thirty consecutive frames with the total length of 384 ms comprise the short-time segmentation of the clean speech, which is denoted by

$$\mathbf{x}_{j,m} = [X_j(m-N+1), X_j(m-N+2), \cdots, X_j(m)]^T, \tag{2.10}$$

where $N = 30$. Similarly, we can compute the short-time segmentation of the processed speech $\mathbf{y}_{j,m}$. As the processed speech may contain severely degraded frames, which can excessively affect the final intelligibility score, the clipping procedure is introduced to upper bound this degradation by

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{\|\mathbf{x}_{j,m}\|}{\|\mathbf{y}_{j,m}\|}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20})\mathbf{x}_{j,m}(n)\right),$$

where $\|\cdot\|$ and n represent the ℓ_2 norm and the frame index among N frames, respectively. $\bar{\mathbf{y}}_{j,m}(n)$ denotes the normalized and clipped envelope of the processed speech. The value of β is chosen as -15 to indicate the lowest signal-to-distortion bound. Thus, the time frequency dependent intermediate intelligibility is computed by

$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}})^T(\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}})}{\|\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\|\|\bar{\mathbf{y}}_{j,m} - \mu_{\bar{\mathbf{y}}_{j,m}}\|}, \tag{2.11}$$

where $\mu_{(\cdot)}$ represents the sample average of the corresponding vector. Finally, the overall intelligibility score is computed as

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m},$$

where J and M represent the number of one-third octave bands and the total short-time segments of speech signals, respectively.

It should be noted that STOI does not consider the band correlation in frequency. To solve this problem, ESTOI was proposed [43]. The supervector (2.10) in STOI is changed to a spectrogram matrix in ESTOI. However, when calculating the intermediate intelligibility in ESTOI, which is similar to the intermediate intelligibility (2.11) in STOI, the dimensionality of the supervectors of the clean speech and the degraded speech should be identical. Since the dimensionality of the clean speech is larger than the dimensionality of the frequency-lowered speech, STOI and ESTOI cannot be used to predict the intelligibility of frequency-lowered speech.

Hearing-aid speech perception index

The Hearing-Aid Speech Perception Index (HASPI) version 1 [44] also uses the correlation coefficient to predict speech intelligibility. Unlike STOI and ESTOI that are based on the temporal envelope of the narrowband signal, HASPI uses both the envelope and the temporal fine structure (TFS).

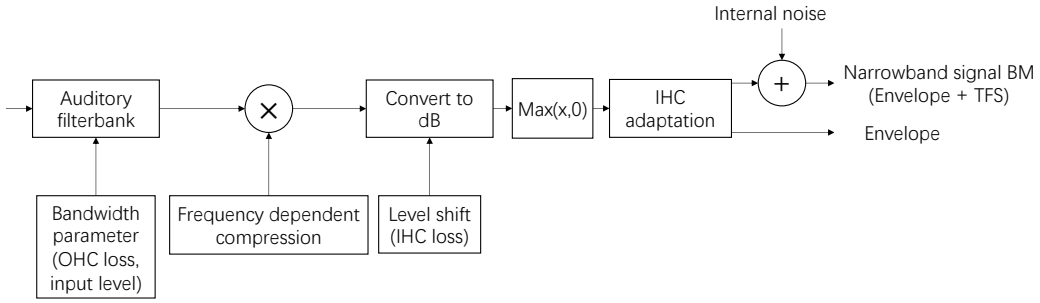


Figure 2.5: Peripheral auditory system in HASPI.

The envelope and the TFS for each frequency band are obtained through a complex peripheral auditory system which is shown in Fig. 2.5. The bandwidth of the auditory filterbank is controlled by the outer hair cell loss and the input signal level. The outputs of the auditory filterbank are the narrowband signal and its envelope. A frequency-dependent compression gain, which is also controlled by the outer hair cell loss and the

input signal level, applies to the output of the auditory filterbank. Then, the signals are converted from linear scale to dB scale. A linear level shift is applied to the dB signal due to inner hair cell loss. The Max operator performs as a half-wave rectifier. Inner hair cell adaptation simulates the forward masking in the auditory nerve. A constant low-level noise which represents the auditory threshold is added on the basilar membrane (BM) vibration signal.

The correlation coefficients are calculated for both the envelope and the narrowband signal, which are referred to as cepstral correlation and auditory coherence, respectively. In the calculation of the cepstral correlation, the cepstral signals are obtained by applying a transform on the envelopes of the clean speech and the degraded speech through five cosine basis functions. In the calculation of the auditory coherence, the correlation coefficients are calculated for low, mid, and high level energy of the segments. The speech intelligibility is estimated by applying a logistic function to the weighted sum of the cepstral correlation and the auditory coherence.

Weighted spectro-temporal modulation index

The weighted spectro-temporal modulation index (wSTMI) [51] predicts speech intelligibility based on the correlation coefficient calculated in the modulation domain. Let $X(f, n)$ denote the signal representation in the auditory domain for the clean speech, where f is the frequency index and n is the frame index. In wSTMI, 11 spectral and five temporal modulation filters are applied on $X(f, n)$ to generate 55 filter spectrograms $\tilde{X}(f, n; s_i, r_j)$, where s_i and r_j are spectral and temporal modulation frequencies, respectively. Similarly, the filtered spectrograms for the degraded speech are denoted by $\tilde{Y}(f, n; s_i, r_j)$.

The normalized cross-correlation for each filtered spectrogram and each

frequency is calculated as

$$d(f;s_i,r_j)=\frac{\langle\tilde{X}(f,n;s_i,r_j)-\mu_{\tilde{X}},\tilde{Y}(f,n;s_i,r_j)-\mu_{\tilde{Y}}\rangle}{||\tilde{X}(f,n;s_i,r_j)-\mu_{\tilde{X}}||\,||\tilde{Y}(f,n;s_i,r_j)-\mu_{\tilde{Y}}||}, \tag{2.12}$$

where $||\cdot||$ denotes ℓ_2 norm, and

$$\mu_{\tilde{X}}=\frac{1}{N}\sum_{n=1}\tilde{X}(f,n;s_i,r_j), \tag{2.13}$$

where N is the total number of time frames. An intermediate intelligibility $\rho(s_i,r_j)$ is defined as

$$\rho(s_i,r_j)=\frac{1}{F}\sum_{f}d(f;s_i,r_j), \tag{2.14}$$

where F is the total number of frequency bins. wSTMI estimates speech intelligibility as

$$\text{wSTMI}=\sum_{i=1}^{S}\sum_{j=1}^{R}w(s_i,r_j)\rho(s_i,r_j)+b, \tag{2.15}$$

where $w(s_i,r_j)$ and b are parameters optimized by minimizing the root mean square error between the wSTMI scores and the subjective intelligibility scores. Compared with the other SIPs, wSTMI achieves the best intelligibility prediction results. There are two factors that make wSTMI achieve the best performance. First, it considers modulation domain, while most SIPs do not. Second, the parameter optimization is based on data-driven, which is able to simulate the complex signal processing mechanism in the brain.

2.2.4 Neural network based metric

The neural network based metrics estimate speech intelligibility by training a neural network. There are different neural networks that can be used for intelligibility prediction. In [48], the neural network has three fully connected layers and uses the auditory spectrogram as the input. In [49], the

neural network is based on U-Net [50] and uses the time-domain speech signals as the input. The SIP in [49] performs better than the SIP in [48]. It also performs better than the classical SIPs for the seen data sets. However, it does not perform as well as the classical SIPs for unseen data sets.

2.2.5 Mutual information based metric

The mutual information-based measures predict speech intelligibility from the information theoretical point of view [12, 13, 41]. As discussed in Section 2.2.2, speech intelligibility was first estimated through the AI (articulation index). In [54], the relationship between the AI and the Shannon channel capacity was studied. It has shown that the AI is an approximation of the channel capacity.

As mutual information measures dependence beyond the conventional second-order statistics (e.g., correlation coefficient), it is able to quantify more complicated relationship between two variables. For one-dimensional variables that have Gaussian distribution, the mutual information is calculated as

$$I = -\frac{1}{2}\log_2(1-\rho^2), \tag{2.16}$$

where ρ is the correlation coefficient between the two variables. Thus, the mutual information metric is the same as the correlation metric.

Various mutual information-based metrics have been developed. There are two main differences among them. First, which two variables are used for calculating mutual information? Second, how is the mutual information calculated? In [72–74], the two variables are clean speech and degraded speech. In [12, 13, 41], the message embedded in clean speech is considered as the original variable, instead of the clean speech itself. The calculation of mutual information depends on the probability distribution of two variables. For specific distributions, the mutual information can be directly calculated [12, 41, 72]. For unspecific distributions, the mutual information can be calculated through the non-parametric approach

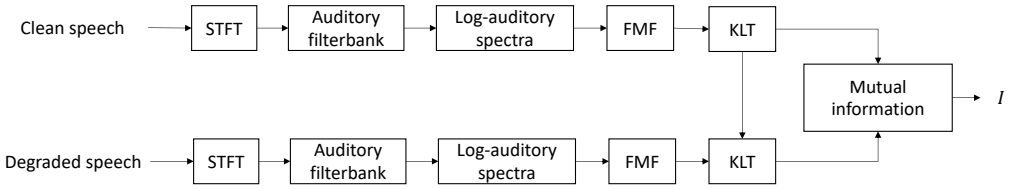


Figure 2.6: Diagram of SIIB and SIIB$^{\text{Gauss}}$.

(k-nearest neighbor) [13] or the parametric approach (Gaussian Mixture Model) [73].

SIIB and SIIB$^{\text{Gauss}}$ [13, 41] are two state-of-the-art mutual information-based metrics. Fig. 2.6 summarizes their calculation procedures. As with many other speech intelligibility metrics, human auditory system (modules of STFT, Auditory filterbank, and Log-auditory spectra) is used to derive the representations of the clean speech and the degraded speech. Forward masking function (FMF) [75] simulates the temporal masking in the auditory system. The module of KLT is used to remove the correlation in the frequency for the clean and the degraded speech. The KLT matrix is derived from the clean speech and is also used for the degraded speech. At last stage, the mutual information is calculated for each transformed band and them summed up over all the bands.

2.3 Message Transmission and Communication Theory

As speech consists of the message and the other information (e.g., talker information and environmental information), it can be viewed as a code of message. In this section, we study the message transmission from the perspective of communication theory.

2.3.1 Information rate of message and channel capacity

First, let us do a rough mathematical computation for the maximal information rate between the message and the clean speech. Take the language of English as an example, the fastest rate at which people produce and perceive speech signals is about 20 to 30 phonemes per second [76]. As English consists of 42 phonemes, the maximal entropy per phoneme is $-\log_2 \frac{1}{42} \approx 5.4$ bit/phoneme, if these phonemes have a uniform distribution. However, the realistic information rate per phoneme is lower, because of the dependence between adjacent phonemes and their non-uniform distribution. Thus, an upper bound for the maximal information rate of the transmitted message carried by an English speech signal is about 108 (20×5.4) to 162 (30×5.4) bit/s.

Next, we compare the upper bound of the maximal information rate of a message with the channel capacity of the voice communication. A conventional voice communication system has a bandwidth about 3 kHz. Assuming a SNR of 15 dB for the voice channel, according to Shannon's theorem [77], the channel capacity is

$$\begin{aligned} C &= 3000 \cdot \log_2(1 + 10^{1.5}) \\ &\approx 15\,\text{kbit/s}, \end{aligned}$$

which is the maximal rate at which information can be transmitted over this channel with an arbitrarily small probability of error. Comparing the information rate of message with the channel capacity of the voice communication, we can see the channel capacity is extremely higher than the information rate of the message. This suggests that the message embedded in a speech signal is sparse.

2.3.2 Linguistic channel code

When two people talk in a noisy environment or read out the same sentence, the message may still be understood. This shows speech production

is actually a coding process, which shares the same idea of channel coding in communication theory.

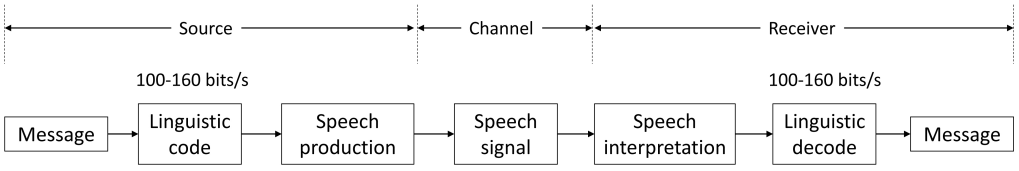


Figure 2.7: Message transmission over voice channel.

Fig. 2.7 illustrates the message transmission, which includes three parts: source, channel, and receiver. The channel capacity of human ear is much larger than the information rate of message. The source part shows how a message is encoded into a speech signal, and the receiver part illustrates the reversed process. Now let us study how a message is encoded into a speech signal. Since speech signals contain both the message and the other information (e.g., speaker information and environmental information), the dimensionality of speech should be larger than the dimensionality of the message. As shown in Fig 2.7, this dimension extension is realized by the linguistic code, which maps a message to a linguistic codeword. According to the information theory, a one-to-one function does not change the entropy of the input and output variables, which means the entropy of the linguistic codewords is identical to the entropy of the message [78]. However, as linguistic codewords have a higher dimensionality, the minimum distance of linguistic codewords becomes larger, which makes it possible for error correction at the receiver side. In addition, as the linguistic codewords do not span the whole space, the remaining space can be used to transmit the nonlinguistic information.

From the above analysis, we can see the linguistic code shares the same idea with the channel code in communication theory (technological channel code), which also maps a message into a high-dimensional space by in-

troducing redundant bits. The difference between linguistic channel code and technological channel code is the constraint of length on the codeword. To make the information rate achieve the channel capacity, in communication theory the codeword is extremely long, which means a large sequence of information bits are encoded into one codeword. However, in linguistic code one message is mapped into one codeword, which suggests that the codeword length of linguistic channel code is shorter than the technological channel code. The constraint of length on the linguistic codeword probably comes from the inability of human brain to decode a long codeword.

2.3.3 Linguistic decode for hearing-impaired listeners

For hearing-impaired people, the hearing loss makes them unable to receive the message correctly, which means a decreased bit rate of received message from the information theoretical point of view. This is not caused by the linguistic decode of hearing-impaired people, but the speech interpretation process, which is unable to provide the correct linguistic codewords.

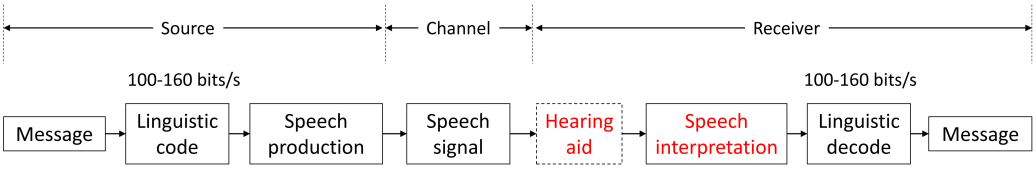


Figure 2.8: Message transmission over voice channel for hearing-impaired people.

The impaired part in message transmission for hearing-impaired people is marked in red color in Fig 2.8. If a hearing-impaired person has hearing loss at high frequency, he or she cannot construct the corresponding linguistic codewords. However, in Section 2.3.2 we know that the lin-

guistic codewords do not span the whole space of speech signal, which means the linguistic codewords can be represented by a lower dimensional space, i.e., a narrower frequency range. Since human can understand all the speech signal, the frequency resolution of human auditory system is higher than the frequency resolution of speech. As long as the frequency resolution of frequency-lowered speech is lower than the frequency resolution of the human auditory system, the frequency-lowered speech can be learned by the listeners. In other words, as long as the residual hearing can provide enough space for the new linguistic codewords, it is possible to transmit the whole linguistic information reliably.

With the help of hearing aid, the intelligibility of received speech can be enhanced. Based on the audiogram of a hearing-impaired person, the hearing aid can decide the dimensionality of the new space, which represents the new frequency range for the linguistic information. Ideally, if the hearing-impaired people can extract the linguistic codewords from the input speech signal and there exists a one-to-one mapping function for the standard and new linguistic codewords, the hearing-impaired people can still receive the whole linguistic information.

2.4 Frequency Lowering in Hearing Instruments

In this section, we give a background of frequency lowering techniques that are used in current hearing instruments, and their corresponding implementation.

2.4.1 Frequency lowering techniques

Hearing impairment usually starts at high frequency. Some people can have severe hearing loss at high frequency, such that they completely lose the audibility for high frequency. The region, where audibility is completely lost, is referred to as *dead region*.

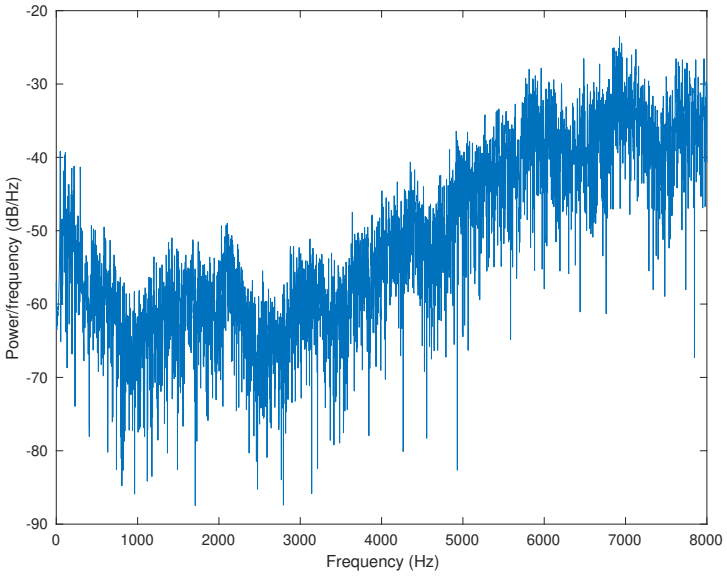


Figure 2.9: Periodogram of sound /s/.

Losing high-frequency information can damage speech intelligibility. For example, the sound /s/ has its main components at high frequency. It plays an important grammatical role, such as pluralization (cake and cakes), possession (Peter's cake), contraction (it and it's) and third person singular (he eats cake), and it has most of the signal power located at high frequency. From the periodogram of /s/, as shown in Fig. 2.9, we can see that a lot of power is located above 3 kHz. Without receiving the high-frequency information, /s/ is hardly recognized by listeners. For hearing-impaired children, obtaining the information at high frequency is particularly important, because developing a language system needs to learn and imitate the sounds that they can hear.

Frequency lowering aims to recover high-frequency information by moving signal at high frequency to low frequency, such that the signal becomes audible at low frequency. Based on whether the bandwidth of frequency-lowered components changes or not, frequency lowering techniques can be classified into *frequency transposition* and *frequency compression*, which are illustrated in Fig. 2.10. For frequency transposition, the

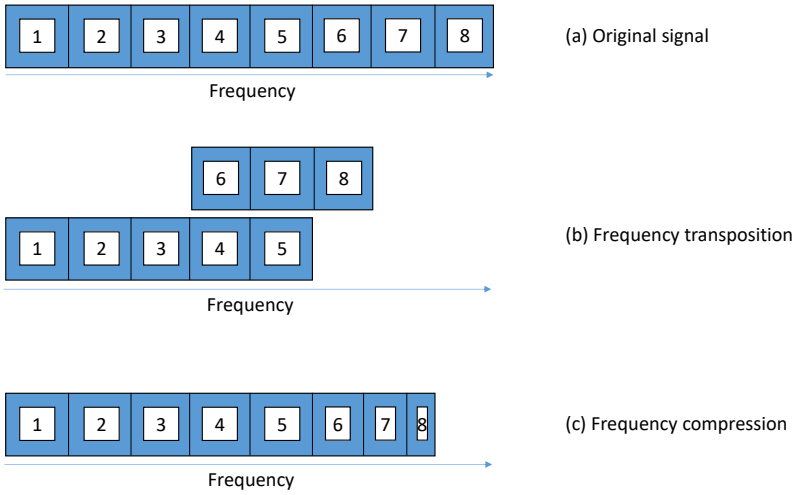


Figure 2.10: Illustration of frequency lowering. Figures are modified from [1].

bandwidth of high-frequency components remains the same. The high-frequency components are extracted and added directly to low frequency. For frequency compression, the bandwidth of high-frequency components is squeezed. There is no overlap of low-frequency and high-frequency components.

2.4.2 Implementation of frequency transposition

Since speech is short-time stationary, the power spectrum of speech varies over time. When implementing frequency lowering, we use the short-time Fourier transform (STFT) to obtain the magnitude and phase information at a time instant. Both magnitude information and phase information contribute to speech intelligibility [79–81].

As shown in Fig. 2.10, frequency lowering techniques used by the current commercial hearing instruments can be classified into frequency transposition and frequency compression. In frequency transposition, the band-

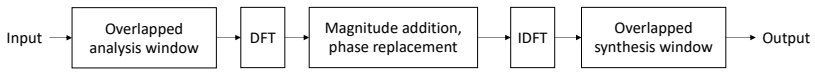


Figure 2.11: Diagram of frequency transposition.

width of frequency-lowered components does not change. Thus, the bandwidth of destination area is the same as the bandwidth of original area. Frequency lowering can be performed by directly shifting the magnitude of STFT to low frequency [82,83]. The phase information at low frequency is replaced by the phase information at high frequency. The frequency-lowered speech is synthesized by using the overlap-add method [84]. A general diagram of frequency transposition is shown in Fig. 2.11.

2.4.3 Implementation of non-linear frequency compression

In frequency compression, the bandwidth of the original area is squeezed. The signal at high frequency cannot be shifted directly to low frequency through STFT. Instead, the signal at high frequency is represented by a set of oscillators and is transferred to low frequency by changing the frequency of oscillators [58]. A perfect reconstruction of the original signal can be obtained by the oscillators, if their magnitude and frequency are selected properly.

The perfect reconstruction can be seen from the filter bank perspective, which is shown in Fig. 2.12. A band-limited signal is defined by

$$x : \mathbb{Z} \to \mathbb{R}, \quad \texttt{with}\, x \in \texttt{BL}^1[-\frac{1}{2}\omega_0, \frac{1}{2}\omega_0], \tag{2.17}$$

where $\omega_0 \in [0, 2\pi)$. The subband signal obtained from a band-limited signal is also band-limited. At the kth subband, the subband signal $\bar{x}_k(n)$ can

[1] BL denotes a subspace of bandlimited sequences.

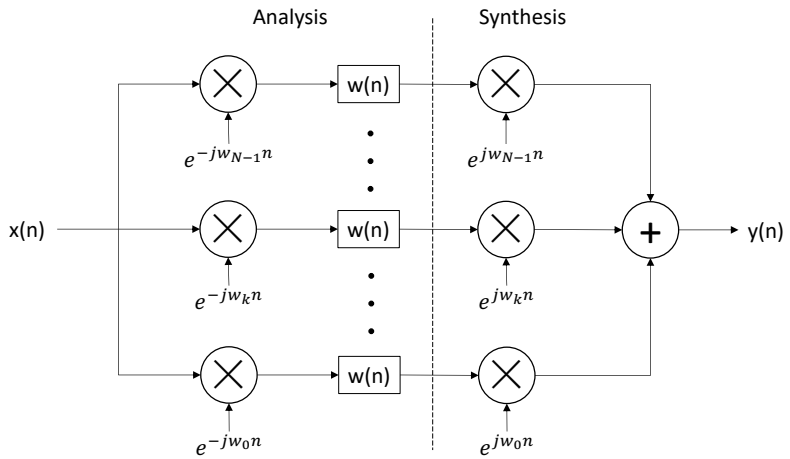


Figure 2.12: Diagram of N-channel filter bank. Figure is modified from [2].

be obtained by spectral shifting and low-pass filtering $x(n)$:

$$\begin{aligned} \bar{x}_k(n) &= w(n) * (x(n)\,\mathrm{e}^{-j\omega_k n}) \\ &= \sum_{m=-\infty}^{\infty} w(n-m)\,x(m)\,\mathrm{e}^{-j\omega_k m}, \end{aligned} \tag{2.18}$$

where $*$ is the convolution operator, $\omega_k = \frac{2\pi}{N}k$, $w(n)$ is the low-pass filter realized by a smooth window function, N is the number of subbands. A perfect reconstruction of $x(n)$ can be realized by summing up the demodulated subband signals followed by a normalization:

$$y(n) = \sum_{k=0}^{N-1} \bar{x}_k(n)\,\mathrm{e}^{jw_k n}. \tag{2.19}$$

Substitute (2.18) into (2.19), we have

$$\begin{aligned} y(n) &= \sum_{k=0}^{N-1} \sum_{m=-\infty}^{\infty} w(n-m)\,x(m)\,\mathrm{e}^{-j\omega_k m}\,\mathrm{e}^{jw_k n} \\ &= \sum_{m=-\infty}^{\infty} w(n-m)\,x(m) \sum_{k=0}^{N-1} \mathrm{e}^{j\omega_k(n-m)}. \end{aligned} \tag{2.20}$$

Since

$$\sum_{k=0}^{N-1} \mathrm{e}^{j\omega_k(n-m)} = \begin{cases} N, & m = n - r\,N, r \in \mathbb{Z} \\ 0, & \text{otherwise,} \end{cases} \tag{2.21}$$

(2.20) can be written as

$$y(n) = N \sum_{r\in\mathbb{Z}} w(r\,N)x(n - r\,N). \tag{2.22}$$

If the window $w(n)$ has the window length $N_w \leq N$,

$$w(r\,N) = 0, \quad r = \pm 1, \pm 2, \cdots \tag{2.23}$$

Thus, a perfect reconstruction can be obtained by scaling $y(n)$:

$$x(n) = \frac{1}{N\,w(0)}\,y(n). \tag{2.24}$$

If $N_w > N$, the window needs to be appropriately set up such that (2.23) is satisfied. In practice, this window can be set as the product of a normal window function and the sinc function. The reason why a perfect reconstruction of $x(n)$ can be achieved by using fewer FFT bins is that the frequency response of the base band window function is overlapped with the frequency response of the modulated window function, and the sum of these frequency responses is a constant.

Based on (2.19), we can generate a set of oscillators to realize a perfect construction of the input signal. The amplitude and frequency of oscillator at kth FFT bin are determined by the subband signal $\bar{x}_k(n)$. The complex-valued subband signal $\bar{x}_k(n)$ can be represented in the polar coordinate as

$$\bar{x}_k(n) = |\bar{x}_k(n)|\mathrm{e}^{j\angle\bar{x}_k(n)}. \tag{2.25}$$

We only consider non-negative frequency bands. Each subband is represented by a single oscillator. The frequency of the oscillator can be estimated by

$$v_k(n) = \Delta\theta_k(n) + \omega_k, \tag{2.26}$$

where $\Delta\theta_k(n)$ is the principal value (between $\pm\pi$) (between $\pm\frac{p_w}{2}$) of the phase difference

$$\angle\bar{x}_k(n) - \angle\bar{x}_k(n-1), \tag{2.27}$$

and $p_w \in [0, 2\pi)$ is the bandwidth of the low-pass filter $w(n)$. The phase of the oscillator can be calculated as the accumulation of the frequency

$$\begin{aligned} \phi_k(n) &= \sum_{m=0}^{n} v_k(m) \\ &= w_k\, n + \angle\bar{x}_k(n), \end{aligned} \tag{2.28}$$

where $v_k(0)$ is the initial phase of the subband signal. For the oscillators representing DC component and Nyquist frequency, their magnitude is $|\bar{x}_k(n)|$. For the other oscillators, their magnitude is $2\,|\bar{x}_k(n)|$. By summing the outputs of the oscillators, we have

$$\begin{aligned} y(n) &= 2\sum_{k=1}^{\frac{N}{2}-1} |\bar{x}_k(n)|\cos(\phi_k(n)) + |\bar{x}_0(n)|\cos(\phi_0(n)) + |\bar{x}_{\frac{N}{2}}(n)|\cos(\phi_{\frac{N}{2}}(n)) \\ &= \sum_{k=0}^{N-1} |\bar{x}_k(n)|\, \mathrm{e}^{j\phi_k(n)} \\ &= \sum_{k=0}^{N-1} |\bar{x}_k(n)|\, \mathrm{e}^{j(w_k\, n + \angle\bar{x}_k(n))} \\ &= \sum_{k=0}^{N-1} \bar{x}_k(n)\, \mathrm{e}^{jw_k\, n}. \end{aligned} \tag{2.29}$$

From eqs. (2.19) and (2.24), we can prove that (2.29) generates a perfect reconstruction of the original input signal.

To use (2.29) to obtain a perfect reconstruction of the input signal, the subband signal $\bar{x}_k(n)$ needs to be calculated at every instant. Since the main lobe bandwidth of the frequency response of $w(n)$ is much lower than 2π, $\bar{x}_k(n)$ can be approximately treated as a band-limited signal. Thus,

we can downsample $\bar{x}_k(n)$ to reduce the computation complexity. The decimation rate needs to be properly selected, such that the downsampling process does not cause aliasing, which means the bandwidth of the downsampled subband signal does not exceed π. Assuming that the subband signal is downsampled by an integer D, (2.26) can be rewritten as

$$v_k(m) = \frac{\Delta\theta'_k(n)}{D} + \omega_k, \quad m = n - D + 1, n - D + 2, \cdots, n \tag{2.30}$$

where $\Delta\theta'_k(n)$ is the principal value (between $\pm\pi$) of the phase difference:

$$\angle\bar{x}_k(n) - \angle\bar{x}_k(n - D). \tag{2.31}$$

Note that the frequency estimated in (2.30) is an approximation of the true frequency, due to the averaging of the phase difference. However, the phase at the integer times of D, which is the accumulation of the estimated frequency, is exactly the same phase of the original subband signal. Thus, the synthesized signal $y(n)$ at the integer times of D is exactly the same as the original input signal $x(n)$.

Chapter 3

Speech Transmission Model based on Continuous-valued Message

In this chapter, we study the validity of a speech transmission model based on continuous-valued message. The message refers to the information that is related to speech intelligibility, and it is independent of talker and acoustic environment. The speech transmission model can be used by speech intelligibility predictors [13, 41], which calculate the amount of information about the message in a received speech signal. The speech transmission model comprises a *production channel* and an *acoustic channel*. For the production channel, the input is the message and the output is clean speech. For the acoustic channel, the input is clean speech and the output is degraded speech.

A good speech intelligibility predictor depends on the proper modeling of speech transmission. In [13, 41], the speech intelligibility predictors SIIB and $\text{SIIB}^{\text{Gauss}}$ have not addressed two problems yet. First, the correlation coefficients between the transformed message and the transformed clean speech are assumed to be equal across channels. This is an oversimplified assumption, as we will show that correlation coefficients vary

across channels. Second, the channels in the transformed degraded speech are assumed to be independent. Since the KLT matrix is obtained from the clean speech rather than the degraded speech, applying this KLT matrix directly on the degraded speech hardly makes the channels in the transformed degraded speech independent. In this chapter, we address these two problems, and compare the improved model with SIIB and SIIB$^{\text{Gauss}}$.

3.1 Transmission Model

The original transmission model was proposed in [12, 13]. We denote the message, the clean speech, and the degraded speech by multidimensional continuous-valued random variables M, X, and Y, respectively. The message is considered as continuous-valued sound. The transmission model is represented by a Markov chain:

$$M \rightarrow X \rightarrow Y, \tag{3.1}$$

where the production channel is represented by $M \rightarrow X$, and the acoustic channel is represented by $X \rightarrow Y$. The transmission over the production channel is modeled as [85]

$$X_k = M_k + P_k, \tag{3.2}$$

where k is the channel index, M_k is the message, P_k is the production noise, X_k is the clean speech. The production noise models the inter- and intra-talker variability. Since the acoustic channel can have different forms, such as additive noise, reverberation, etc., we cannot model the acoustic channel by an additive model as we did for the production channel. Fig. 3.1 illustrates the diagram of the speech transmission model for the continuous-valued message.

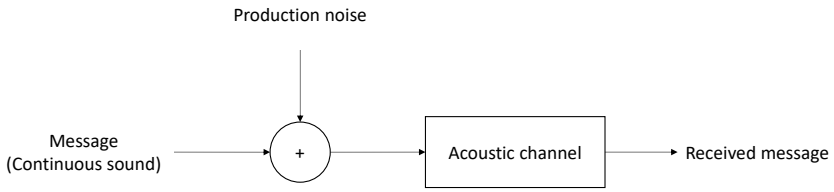


Figure 3.1: Speech transmission model for continuous-valued message.

Given the above Markov chain, we need to determine the exact form of the variables in the chain. Instead of the common STFT spectra, we choose log-auditory spectra for these variables. The auditory spectra take account of the frequency resolution of human ear. We apply a logarithm on the auditory spectra for two reasons. First, a logarithm can separate the vocal tract signal and the excitation signal. This is based on the fact that the convolution of the vocal tract signal and the excitation signal in the time domain is equivalent to the multiplication of their frequency responses. The frequency responses consist of magnitude spectra and phase spectra that both contribute to speech intelligibility [80, 86, 87]. It is unknown whether these two spectra contribute in a complementary or independent fashion [80]. In computational auditory scene analysis, it has been demonstrated that the ideal binary mask based on the magnitude spectra can improve speech intelligibility [88–90]. Thus, in the current study, we only consider magnitude spectra. The second reason for using the logarithm is that it simulates the human loudness perception of sound. As speech intelligibility depends on human hearing, taking the human hearing perception into account benefits the modeling of the transmission of the message.

The modeling of the message varies in different communities. In the engineering community [13, 41], the message is considered as continuous-valued sound. However, in the linguistic community [91, 92], the message is considered as discrete-valued linguistic message, which includes phonemes, syllables, words, etc. The linguistic messages are by nature

discrete-valued. Due to the success of the speech intelligibility predictor [41] that uses the modeling of continuous-valued message, we first improve this transmission model. In the next chapter, we study the case, where the message is considered as discrete-valued linguistic message.

3.2 Gaussian Distributed Pseudo Message

In our model, we hypothesize that speech intelligibility can be estimated by the mutual information between the message and the degraded speech. In the modeling of continuous-valued message, the message is found out through ensemble average of the same speech message recorded by different talkers. Given only one speech signal, we cannot find out the underlying message. However, if we know the statistical properties of the message, we can generate a pseudo message $\hat{M}$, such that $I(M;Y) = I(\hat{M};Y)$. Thus, we can use $I(\hat{M};Y)$ to estimate speech intelligibility. For the sake of simplicity, in the following sections we assume the message M, the clean speech X, and the degraded speech Y have multidimensional Gaussian distribution.

3.2.1 Mutual information for pseudo message

The mutual information between the message and the received speech is calculated as

$$I(M;Y) = h(M) + h(Y) - h(M,Y), \tag{3.3}$$

where $h(\cdot)$ denotes the differential entropy of a random variable. To have $I(M;Y) = I(\hat{M};Y)$, we need to ensure

$$h(\hat{M}) = h(M) \tag{3.4}$$

$$h(\hat{M},Y) = h(M;Y). \tag{3.5}$$

Let us first figure out the condition to satisfy (3.4), then we will determine the condition to satisfy (3.5). Since

$$h(M) = \int_M p_M(m)dm, \tag{3.6}$$

we need to have

$$p_{\hat{M}}(m) = p_M(m) \tag{3.7}$$

to satisfy (3.4). The message is continuous-valued sound and is represented in the log-auditory spectrogram domain. For a simplified model, we assume the message has a multivariate normal distribution, i.e.,

$$M \sim \mathcal{N}(\mu_M, \Sigma_M), \tag{3.8}$$

where

$$\begin{aligned} \mu_M &= \mathrm{E}[M] \\ \Sigma_{M_{i,j}} &= \mathrm{Cov}[M_i, M_j]. \end{aligned} \tag{3.9}$$

Thus, in order to satisfy (3.4), the pseudo message $\hat{M}$ should also be multivariate normally distributed and have the same mean vector and covariance matrix as the original message M.

Next, we find out the condition to satisfy (3.5). The differential entropy $h(M, Y)$ is calculated as

$$\begin{aligned} h(M,Y) &= \int_Y \int_M p_{M,Y}(m,y)dmdy \\ &= \int_Y \int_M \int_X p_{M,Y,X}(m,y,x)dxdmdy \\ &= \int_Y \int_M \int_X p_{M,X}(m,x)p_{Y|M,X}(y|m,x)dxdmdy. \end{aligned} \tag{3.10}$$

Due to the Markov chain (3.1), we have

$$p_{Y|M,X}(y|m,x) = p_{Y|X}(y|x). \tag{3.11}$$

Substitute (3.11) into (3.10), we have

$$h(M,Y) = \int_Y \int_M \int_X p_{M,X}(m,x)p_{Y|X}(y|x)dxdmdy. \tag{3.12}$$

Thus, in order to satisfy (3.5), we need to have

$$p_{\hat{M},X}(m,x) = p_{M,X}(m,x). \tag{3.13}$$

Let us denote $Z = [M;X]$ and $\hat{Z} = [\hat{M};X]$. For a simplified model, we again assume Z has a multivariate Gaussian distribution, i.e.,

$$Z \sim \mathcal{N}(\mu_Z, \Sigma_Z), \tag{3.14}$$

where

$$\begin{aligned} \mu_Z &= \mathrm{E}[Z] \\ &= \begin{bmatrix} \mu_M \\ \mu_X \end{bmatrix}, \end{aligned} \tag{3.15}$$

and

$$\Sigma_Z = \begin{bmatrix} \Sigma_M & \Sigma_{MX} \\ \Sigma_{XM} & \Sigma_X \end{bmatrix}. \tag{3.16}$$

In order to satisfy (3.13), $\hat{Z}$ should have Gaussian distribution, and

$$\mu_{\hat{Z}} = \mu_Z \tag{3.17}$$

$$\Sigma_{\hat{Z}} = \Sigma_Z. \tag{3.18}$$

Thus, we have

$$\mu_{\hat{M}} = \mu_M \tag{3.19}$$

$$\Sigma_{\hat{M}} = \Sigma_M \tag{3.20}$$

$$\Sigma_{\hat{M}X} = \Sigma_{MX}. \tag{3.21}$$

We conclude that (3.19) and (3.20) are the necessary conditions to satisfy (3.4). To satisfy (3.5), the additional condition (3.21) also needs to be met. In conclusion, if the generated pseudo message can satisfy eqs. (3.19) to (3.21), we have

$$\begin{aligned} I(M;Y) &= I(\hat{M};Y) \\ &= h(\hat{M}) + h(Y) - h(\hat{M},Y). \end{aligned} \tag{3.22}$$

3.2.2 Message with frequency independent bands

When generating the pseudo message, eqs. (3.20) and (3.21) need to be satisfied. As the covariance matrices in eqs. (3.20) and (3.21) do not have to be diagonal, the correlation in frequency needs to be considered. We first consider a simpler case, where the frequency bands in a message are independent. Then, the covariance matrices in eqs. (3.20) and (3.21) become diagonal, and can be written as

$$\begin{aligned} \sigma_{\hat{M}_k} &= \sigma_{M_k} \\ \mathrm{Cov}(\hat{M}_k, X_k) &= \mathrm{Cov}(M_k, X_k). \end{aligned} \tag{3.23}$$

We denote $\hat{P}_k$ as the pseudo production noise. X_k can be written as

$$X_k = \hat{M}_k + \hat{P}_k. \tag{3.24}$$

Since the production noise represents the talker variability, it is reasonable to assume

$$\mu_{P_k} = 0, \tag{3.25}$$

and the message and the production noise are independent. These assumptions also apply to the pseudo message and the pseudo production noise. Thus, (3.23) can be written as

$$\sigma_{\hat{M}_k} = \sigma_{M_k} \tag{3.26}$$

$$\mathrm{Cov}(\hat{M}_k, \hat{P}_k) = 0. \tag{3.27}$$

As the pseudo message has the same variance and mean of the true message, we can generate the pseudo message by scaling the true message and adding a noise signal. To satisfy eqs. (3.19), (3.26) and (3.27), we generate the pseudo message and the pseudo production noise as

$$\hat{M}_k = a_k(X_k - \mu_{X_k}) + b_k N_k + \mu_{X_k} \tag{3.28}$$

$$\hat{P}_k = (1 - a_k)X_k + a_k\mu_{X_k} - b_k N_k - \mu_{X_k}, \tag{3.29}$$

where a_k and b_k need to be solved. $N_k \sim \mathcal{N}(0,1)$ is a white noise signal. Substitute (3.28) and (3.29) into eqs. (3.26) and (3.27), we have

$$a_k = \rho^2_{M_k X_k} \tag{3.30}$$

$$b_k = \rho_{M_k X_k} \sqrt{1 - \rho^2_{M_k X_k}} \sigma_{X_k}, \tag{3.31}$$

where $\rho_{M_k X_k}$ is the correlation coefficient between the message and the clean speech.

3.2.3 Message with frequency correlated bands

For a multi-dimensional message with frequency components that are independent, the pseudo message only needs to satisfy eqs. (3.19), (3.26) and (3.27). The off-diagonal components in eqs. (3.26) and (3.27) are zeros. For frequency correlated multi-dimensional message, the pseudo message also needs to satisfy the conditions on these off-diagonal components.

Speech signals have non-zero off-diagonal components in the covariance matrix. Thus, (3.28) cannot be directly used to generate a pseudo message. To show this, we take a two-dimensional signal as the example, where the frequency bands in the message are correlated, but the frequency bands in the production noise are independent. By using (3.28), the pseudo message is generated as

$$\begin{aligned} \hat{M}_1 &= a_1(X_1 - \mu_{X_1}) + b_1 N_1 + \mu_{X_1} \\ \hat{M}_2 &= a_2(X_2 - \mu_{X_2}) + b_2 N_2 + \mu_{X_2}. \end{aligned} \tag{3.32}$$

Now we show that this pseudo message does not satisfy eqs. (3.20) and (3.21). In (3.20), the off-diagonal component

$$\begin{aligned} \mathrm{Cov}(\hat{M}_1, \hat{M}_2) &= \mathrm{E}[(\hat{M}_1 - \mu_{X_1})(\hat{M}_2 - \mu_{X_2})] \\ &= \mathrm{E}[(a_1(X_1 - \mu_{X_1}) + b_1 N_1)(a_2(X_2 - \mu_{X_2}) + b_2 N_2)] \\ &= a_1 a_2 \mathrm{E}[(M_1 - \mu_{M_1})(M_2 - \mu_{M_2})] \\ &\neq \mathrm{Cov}(M_1, M_2). \end{aligned} \tag{3.33}$$

In (3.21), the off-diagonal component

$$\begin{aligned}\mathrm{Cov}(\hat{M}_1, X_2) &= \mathrm{E}[(\hat{M}_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \mathrm{E}[(a_1(X_1 - \mu_{X_1}) + b_1 N_1)(X_2 - \mu_{X_2})] \\ &= a_1 \mathrm{Cov}(X_1, X_2),\end{aligned} \tag{3.34}$$

and

$$\begin{aligned}\mathrm{Cov}(M_1, X_2) &= \mathrm{E}[(M_1 - \mu_{M_1})(X_2 - \mu_{X_2})] \\ &= \mathrm{E}[(X_1 - P_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \mathrm{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \mathrm{Cov}(X_1, X_2).\end{aligned} \tag{3.35}$$

Thus,

$$\mathrm{Cov}(\hat{M}_1, X_2) \neq \mathrm{Cov}(M_1, X_2). \tag{3.36}$$

To generate the pseudo message for a speech signal, we can remove the frequency correlation in the speech before applying (3.28). Recall that the transmission over production channel is modeled as

$$X = M + P, \tag{3.37}$$

where M and X consist of bands that are correlated in frequency. Since M and P are independent, we have

$$\Sigma_X = \Sigma_M + \Sigma_P. \tag{3.38}$$

In the next section, we will study the statistical characteristics of these covariance matrices. At the moment, let us assume the production noise is a white noise. Then, Σ_P is a scaled identity matrix cI, where c is a constant. We have

$$\begin{aligned}\Sigma_M + \Sigma_P &= U \Lambda_M U^T + cIUU^T \\ &= U(\Lambda_M + cI)U^T,\end{aligned} \tag{3.39}$$

where the column vectors in U are the eigenvectors of Σ_M. We put (3.39) into (3.38), and it forms an eigendecomposition of Σ_X:

$$\Sigma_X = U(\Lambda_M + cI)U^T. \tag{3.40}$$

Thus, Σ_M and Σ_X have the same eigenvectors. By applying the transform matrix U^T on both sides of (3.37), the bands are not correlated in frequency for the transformed speech $U^T X$ and the transformed message $U^T M$. The additive model still holds. Since the bands are uncorrelated in frequency, we can use (3.28) to calculate the pseudo message.

3.3 Statistical Characteristics of Speech

When generating a pseudo message, we need to consider if the frequency bands in the original message are correlated. For independent frequency bands, (3.28) can be used directly. For correlated frequency bands, a KLT needs to be applied on the clean speech before using (3.28) to generate the pseudo message. In this section, we study the statistical characteristics of speech, and derive the correlation coefficient between the transformed message and the transformed speech. The derived correlation coefficients are used to generate the pseudo message.

3.3.1 Covariance matrices

The frequency correlation of the message can be illustrated by its covariance matrix. To generate the message, we used the CHAINS data set [93], which contains 33 sentences and each sentence was recorded by 36 talkers.

Let us denote the short-time Fourier transform (STFT) of clean speech by $X_s(j,t)$, where s refers to STFT, j is the FFT bin index, and t is the time index. The sampling frequency is 16 kHz. Hann windows with 50% overlap are used and the window length is 25 ms. Thirty gammatone filters are equally ranged on the ERB-rate scale, which ranges from 100 Hz to 6500 Hz. The log-auditory spectra $X_k(t)$ is calculated as

$$X_k(t) = 10 \log_{10} \left(\sum_j W_k^2(j) \left| X_s(j,t) \right|^2 \right), \tag{3.41}$$

where $W_k(j)$ is the magnitude response of kth gammatone filter. The speech recorded by the nth talker is denoted by $X_k^n(t)$. As with [12], the unbiased message for talker n is estimated as

$$M_k^n(t) = \frac{1}{N-1} \sum_{\substack{r=1 \\ r \neq n}}^{N} X_k^r(t). \tag{3.42}$$

The unbiased production noise for talker n is estimated as

$$\begin{aligned} P_k^n(t) &= X_k^n(t) - M_k^n(t) \\ &= X_k^n(t) - \frac{1}{N-1} \sum_{\substack{r=1 \\ r \neq n}}^{N} X_k^r(t). \end{aligned} \tag{3.43}$$

For each talker, we have derived the message and the production noise. Let us denote the message from talker n by

$$M^n = \begin{bmatrix} M_1^n(t) \\ M_2^n(t) \\ \vdots \\ M_k^n(t) \\ \vdots \\ M_K^n(t). \end{bmatrix} \tag{3.44}$$

We generate the message by concatenating the messages from each talker

$$M = [M^1, M^2, \cdots, M^n, \cdots, M^N]. \tag{3.45}$$

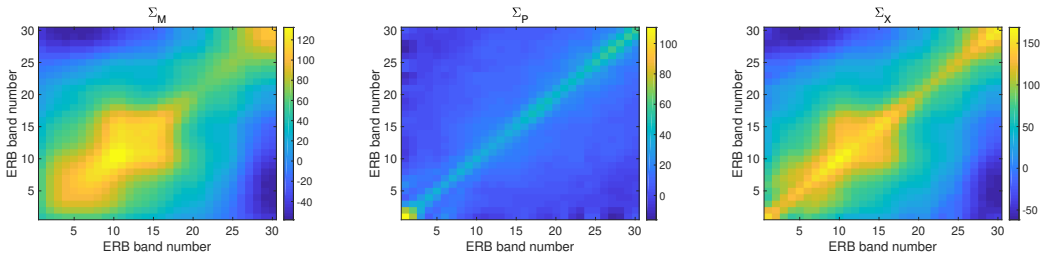
Similarly, we generate the production noise as

$$P = [P^1, P^2, \cdots, P^n, \cdots, P^N], \tag{3.46}$$

and the clean speech as

$$X = [X^1, X^2, \cdots, X^n, \cdots, X^N]. \tag{3.47}$$

The covariance matrices for the message, the production noise, and the clean speech are shown in Fig 3.2. Fig. 3.2a shows the frequency bands of the message are correlated. Fig. 3.2b shows the frequency bands of the production noise are also correlated. However, compared with the message in Fig. 3.2a, the production noise can be approximately viewed as a white noise. Due to the additive model, the frequency bands of the clean speech are correlated, which is shown in Fig. 3.2c.



(a) Message. (b) Production noise. (c) Clean speech.

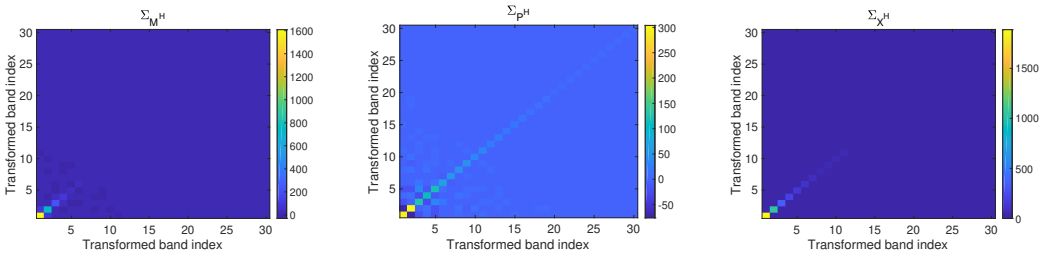
Figure 3.2: Covariance matrices of the original signals.

3.3.2 Pseudo message for speech

As discussed in Section 3.2.3, when generating the pseudo message, the bands of the clean speech should be uncorrelated in frequency. Since speech has frequency correlated components, we need to apply KLT on the clean speech before generating the pseudo message. In this section, we first check the covariance matrices of the transformed message, the transformed production noise, and the transformed clean speech. Then, we derive the correlation coefficients between the transformed message and the transformed clean speech.

When the production noise is white, the KLT matrix for the clean speech is the same as the KLT matrix for the message. Fig. 3.3 shows the covariance matrices for the transformed message, the transformed production noise, and the transformed clean speech. Since the KLT matrix was derived based on the clean speech, the bands of the transformed clean speech

are uncorrelated, as is shown in Fig. 3.3c. Fig. 3.3b shows the bands of the transformed noise are approximately uncorrelated. The energy goes down with the band index. This indicates that the covariance matrix of the production noise shows some similarity with the covariance matrix of the clean speech. Thus, assuming the production noise as a white noise is an approximation. Although the production noise is not exactly white, Fig. 3.3a shows the KLT matrix derived from the clean speech can also effectively remove the frequency correlation in the message. Since we assume the message, the production noise, and the clean speech are multivariate normally distributed, after KLT, the bands of the transformed message, the transformed production noise, and the transformed speech are independent. Thus, we can use (3.28) to derive the pseudo message based on the transformed speech.



(a) Transformed message.

(b) Transformed production noise.

(c) Transformed clean speech.

Figure 3.3: Covariance matrices of the transformed signals.

When using (3.28), we need to know the correlation coefficients between the transformed message and the transformed speech. The transformed message is

$$M^H = U^T M, \tag{3.48}$$

and the transformed speech is

$$X^H = U^T X. \tag{3.49}$$

The correlation coefficients $\rho_{M_k^H X_k^H}$ are shown by the blue dots in Fig. 3.4. We fit these dots with an exponential function

$$\rho_k^{\text{fit}} = ae^{b\,k+c}, \tag{3.50}$$

where k is the index of the transformed band. Nonlinear least squares method was used to find the parameters a, b, c. The fitted curve is

$$\rho_k^{\text{fit}} = 0.661e^{-0.132k+0.497}. \tag{3.51}$$

The fitted correlation coefficients are shown by the red curve in Fig. 3.4 and are used to generate the pseudo message.

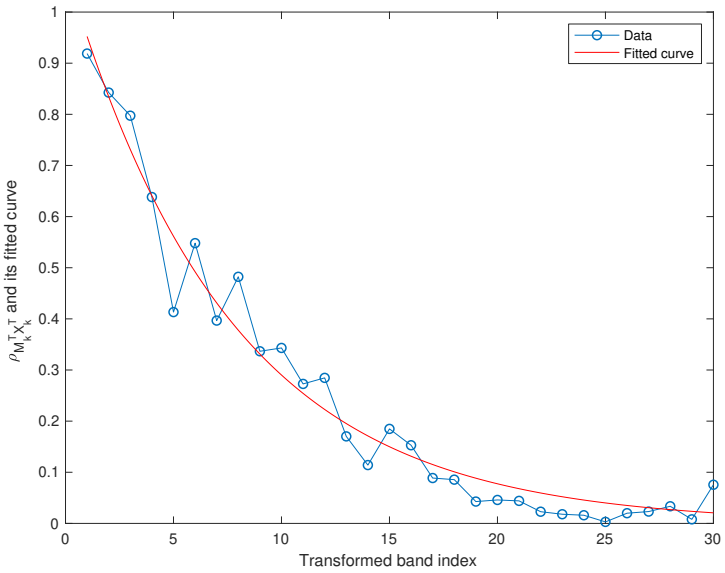


Figure 3.4: Correlation coefficients between the transformed message and the transformed speech.

3.4 Evaluation

In this section, we evaluate the validity of the transmission model based on the continuous-valued message. As speech intelligibility is quantified

by the information about the message in the received speech, the mutual information curve for a good modeling of the message should have a similar trend as the intelligibility curve. Thus, to evaluate the validity of the model, we compare the intelligibility curve with the mutual information curve under different conditions.

3.4.1 Data set

The psychometric curve derived from the Kjems AN data set [94] was used in the evaluation. In this data set, the clean speech is degraded by four additive noises, which are speech shaped noise, babble noise, bottling noise, and car noise. The clean speech sentences are from the Dantale II corpus [95]. Speech intelligibility was measured under different SNRs for each individual noise. The psychometric curve was fitted to the intelligibility data. With the psychometric curve, we can obtain the intelligibility for every SNRs. These intelligibility data points can be compared with the mutual information data points to see if these two curves have similar trend.

3.4.2 Estimation of the mutual information

We represent the signals in the form of a log-auditory spectrogram. The parameters of the log-auditory spectrogram were the same as described in Section 3.3.1. The clean speech and the received speech were calculated as (3.41). The mutual information is given by

$$\begin{aligned} I(M;Y) &= I(U^T M;Y) \\ &= I(\hat{M}^H;Y) \\ &= h(\hat{M}^H) + h(Y) - h(\hat{M}^H,Y), \end{aligned} \tag{3.52}$$

where U^T is the KLT matrix obtained from X, $\hat{M}^H$ is the pseudo transformed message for $U^T M$.

For a K-dimensional Gaussian variable X, we denote its covariance matrix by Σ_X and the unbiased estimate of the covariance matrix by S_X. Let L denote the number of samples. We have [96, p. 100]

$$\frac{|(L-1)S_X|}{|\Sigma_X|} \sim \prod_{k=1}^{K} \chi^2_{L-k}, \tag{3.53}$$

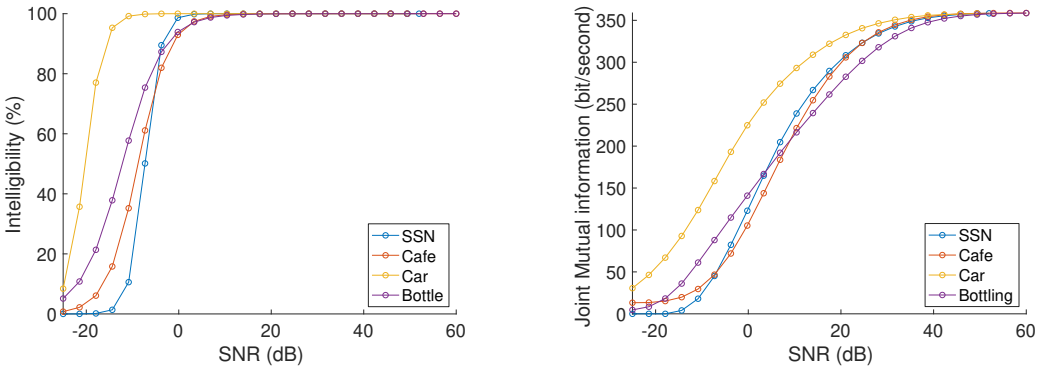
where χ^2_{L-k}, for $k = 1, \cdots, K$, denote independent central chi-square random variables with $L-k$ degrees of freedom. Since $|S_X|$ is a biased estimate of $|\Sigma_X|$, using $|S_X|$ directly leads to a biased estimate of $h(X)$. A bias-corrected estimate of the differential entropy can be calculated as [97, 98]

$$h(X) = \frac{1}{2\ln 2}\left(\ln\left((2\pi e)^K |\Sigma_X|\right) - K\ \ln\frac{2}{L-1} - \sum_{i=1}^{K} \Psi\left(\frac{L-i}{2}\right)\right), \tag{3.54}$$

where Ψ is the polygamma function.

3.4.3 Results and discussion

Fig. 3.5a shows the listening test results, which are illustrated by the psychometric curve. For car noise, the intelligibility from 0 to 100% corresponds to SNRs from -25 to -15 dB. For bottling noise, the intelligibility from 0 to 100% corresponds to SNRs from -20 to 5 dB. Thus, the SNR interval that makes the intelligibility saturate is between 10 and 25 dB. Fig. 3.5b shows the objective results, which are illustrated by the mutual information curve. For car noise, the SNR ranges from -25 to 20 dB. For bottling noise, the SNR ranges from -15 to 35 dB. Thus, the SNR interval that makes the intelligibility saturate is about 45 dB.



(a) Psychometric curve. (b) Mutual information.

Figure 3.5: Subjective and objective results.

The reason why the slope of the mutual information curve is more gradual than the slope of the psychometric curve is that we assume the message is continuous-valued. When the message has Gaussian distribution and the correlation coefficient ρ_{MX} of the production channel is large, the environmental channel SNR needs to be high to make $I(X;Y)$ reach $I(M;X)$. We illustrate this with a one-dimensional Gaussian variable. The transmission model follows the Markov chain in (3.1). The acoustic channel $X \rightarrow Y$ is modeled as an additive white Gaussian noise (AWGN) channel. The mutual information between M and Y is calculated as

$$I(M;Y) = -\frac{1}{2}\log_2(1-\rho_{MX}^2\rho_{XY}^2), \tag{3.55}$$

where ρ_{MX} is the correlation coefficient for the production channel,

$$\begin{aligned}\rho_{XY}^2 &= \frac{\sigma_X^2}{\sigma_X^2+\sigma_N^2} \\ &= \frac{\mathrm{SNR}_{XY}}{1+\mathrm{SNR}_{XY}}.\end{aligned} \tag{3.56}$$

We carried out two experimental results for $\rho_{MX} = 0.9$ and $\rho_{MX} = 0.3$, respectively. Fig. 3.6 illustrates their corresponding mutual information curves. The curve with larger ρ_{MX} shows wider SNR transition interval.

For $\rho_{MX} = 0.9$, the SNR interval is about 30 dB. For $\rho_{MX} = 0.3$, the SNR interval is about 20 dB.

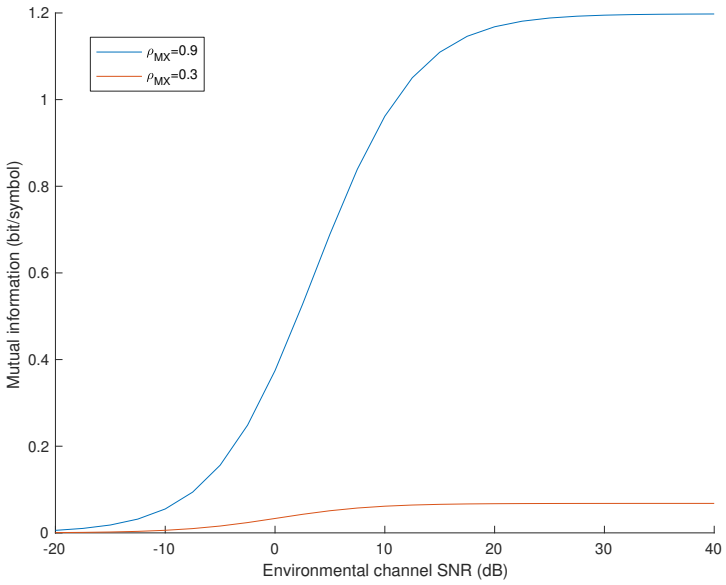


Figure 3.6: Impact of the production channel on the mutual information.

For real speech, Fig. 3.4 shows about one-third of the correlation coefficients are larger than 0.3. Thus, the continuous model requires a large SNR interval for $I(M;Y)$ to reach the saturation point. The inconsistency in the SNR interval between Fig. 3.5a and Fig. 3.5b indicates the modeling of the message should be improved. However, as the curves in the two figures have approximately consistent order, the mutual information metric based on the continuous-valued message shows the merit for objective intelligibility estimation.

3.5 Summary

In this chapter, we developed a message transmission model, where the message is assumed continuous-valued. This model can be used to quan-

tify the speech intelligibility by calculating the mutual information between the message and the received speech.

As the message cannot be obtained for one speech signal, we propose to generate a pseudo message, such that $I(\hat{M}, X) = I(M; X)$. When generating the pseudo message, the frequency bands of the clean speech should be uncorrelated for the proposed model. Thus, a KLT matrix is applied on the clean speech. We use the correlation coefficients between the transformed message and the transformed clean speech to generate the pseudo message.

To evaluate the validity of the transmission model based on continuous-valued message, we use this model to calculate the mutual information between the message and the noisy speech for four types of noise under different SNRs. The mutual information curves were compared with the psychometric curves. We found that the mutual information curves have a larger SNR interval than the psychometric curve. This larger interval is caused by the large $I(M; X)$, which is calculated based on the assumption that the message is continuous-valued. In the next chapter, we will study the inconsistency in the SNR interval between the mutual information curve and the psychometric curve.

Chapter 4

Speech Transmission Model based on Discrete-valued Message

In this chapter, we study the transmission model based on the assumption of discrete linguistic units. We first investigate the relationship between the subjective metric of speech intelligibility and the objective metric of mutual information. Then, we show the advantage of the discrete modeling of the linguistic units. Finally, we propose a speech intelligibility predictor that is based on the discrete modeling of linguistic units.

4.1 Background

Speech intelligibility measures comprehensibility of a speech signal under a given condition. The comprehensibility can be represented by the proportion of linguistic units that are correctly received by listeners. For example, word is a type of linguistic unit, and speech intelligibility is usually represented by the percentage of identified words. As linguistic units are morphemes, words, and sentences, which are discrete, it leads us to develop a model for discrete messages.

In the last chapter, continuous sound is regarded as the message , while in this chapter discrete linguistic unit is regarded as the message. Table. 4.1 shows some linguistic concepts that are used in our study. Linguistic units are meaningful language units. The smallest linguistic unit is morpheme, and the largest linguistic unit is sentence. It should be noted that a morpheme is a larger unit than a phoneme. For example, the word "unpredictable" consists of four morphemes ["un" "pre" "dict" "able"]. The morpheme "able" consists of three phonemes: /ə/, /b/, /l/.

Table 4.1: Linguistic items

Item	Description
Phone	Distinct speech sound, regardless of whether the exact sound is critical to the meanings of words.
Phoneme	Smallest unit of sound that is critical to the meanings of words. A phoneme may contain several different phones.
Syllable	A unit of pronunciation having one vowel sound, with or without surrounding consonants.
Morpheme	Smallest meaningful unit of a word.
Linguistic unit	Meaningful language unit. Morpheme, word, sentence.

Since speech intelligibility is related to the semantic information, we are interested in quantifying the mutual information between the transmitted linguistic units and the received linguistic units. In our study, we refer to the mutual information rate of the linguistic units as *message rate*. We can estimate the message rate based on the rate of morpheme, word, or sentence. In English, approximately 170,000 words are currently in use, as indicated by the Second Edition of the Oxford English Dictionary [99]. We make a crude assumption that the words are uniformly distributed, thus an upper bound on the entropy of the word is $\log_2 170000 \approx 17.4$ bit/word.

An average person speaks 120 - 150 words per minute. This gives an upper bound on the estimate of the message rate, which is 34.8 - 43.5 bit/second. In [100], the message rate was estimated based on discrete syllables. It estimated that the message rate is about 39 bit/second for all the studied languages. In [101], continuous speech sound is viewed as message, and it estimated that the message rate is around 100 bit/second. We can see that the message rate based on the discrete linguistic unit does not match to the message rate based on the continuous sound. In the following sections, we study the speech intelligibility prediction based on the discrete linguistic unit.

4.2 Discrete Modeling of Linguistic Unit

4.2.1 Message transmission model

The message transmission model for discrete linguistic unit is shown in Fig. 4.1. In the discrete modeling, the message represents morpheme, word, and sentence. Unlike the continuous model in Fig. 3.1, the discrete model has additional module of encoder, classification, and decoder. The discrete linguistic unit is encoded into a codeword, which has frequency and time correlation. The encoding procedure is a one-to-one mapping. The production noise models inter- and intra-talker variability. The modules of classification and decoder together maps the received noisy speech to discrete linguistic unit. The classification module quantizes a continuous sample into discrete sample. The dimensionality of the input and the output of the classification module does not change. The decoder is a one-to-one mapping.

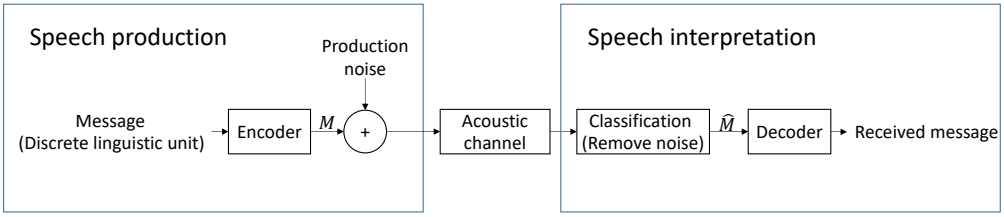


Figure 4.1: Message transmission model for discrete linguistic unit.

4.2.2 Advantage of the discrete modeling of the message

The advantage of the discrete modeling of message is that it can provide reliable communication. From Shannon's noisy-channel coding theorem, we know that if the rate of symbol is below channel capacity, there exists a code such that the symbols can be transmitted with arbitrarily small chance of error. As continuous sound has infinite number of symbols, reliable transmission cannot be realized. When we transmit information via speech, the information should be correctly received. Thus, the speech information can be passed reliably through multiple relays, i.e., from one person to a next person. This indicates that reliable speech communication requires a discrete message.

The reliable transmission of linguistic units is realized in a similar way as the repetition code, where symbols are repeated to increase the distance between any two codewords. Speech is frequency and time correlated, which means linguistic units are mapped into a large span of frequency and time. Both frequency and time are dimensions. When mapping a symbol into a higher dimensional space, redundancy is introduced in the codeword. Thus, they are more robust to noise. Fig. 4.2 illustrates an example, where repetition code can provide more reliable transmission. For the blue curve, the binary symbols are transmitted directly over the noisy channel. For the red curve, each binary symbol is repeated 10 times before transmission. At the same SNR, the red curve shows a higher transmitted

information per symbol than the blue curve.

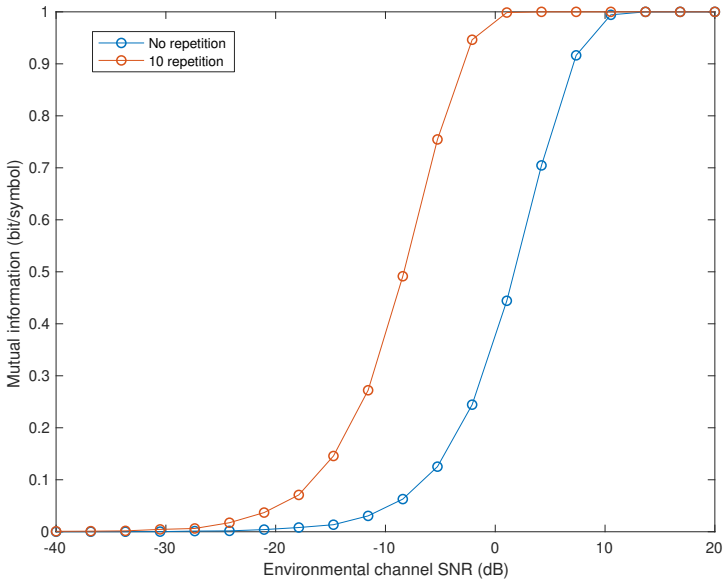


Figure 4.2: Higher dimensionality makes the transmission robust to noise.

4.3 Better Fit with Psychometric Curve

In Chapter 3, the mutual information is calculated between the continuous-valued message and the four types of noisy speech signals. When comparing the SNR transition widths of the psychometric curve and the mutual information curve in Fig. 3.5, we find that the mutual information curve has a wider SNR transition width than the psychometric curve. This is caused by the continuous modeling of the message. If we model the message as discrete linguistic units, the SNR transition width can be reduced. There are two reasons for this. First, the mutual information $I(M;X)$ for the continuous Gaussian modeling of message is always higher than the discrete modeling of message. Thus, it requires a wider SNR interval for $I(M;Y)$ to approach $I(M;X)$. M is the message, X is the clean speech, and Y is the received speech. Second, the length of the codeword for dis-

crete message can affect the steepness of the mutual information curve, while the length of the codeword for continuous message dose not. In this section, we will discuss this in more detail.

4.3.1 Continuous modeling and discrete modeling

When calculating the mutual information between the message and the clean speech, the continuous modeling of message gives a higher mutual information than the discrete modeling of message. For the speech production channel modeled by (3.2), the mutual information at channel k is calculated as

$$\begin{aligned} I(M_k; X_k) &= h(X_k) - h(X_k|M_k) \\ &= h(X_k) - h(M_k + P_k|M_k) \\ &= h(X_k) - h(P_k), \end{aligned} \tag{4.1}$$

where P_k is Gaussian noise. Given a certain variance of M_k, the continuous Gaussian modeling of M_k produces a Gaussian distribution of X_k. The modeling of discrete-valued M_k also produces a continuous variable X_k, but the distribution of X_k is not Gaussian. As the Gaussian distribution for a given variance has maximal $h(X_k)$, the continuous Gaussian modeling of the message always produces a larger $I(M_k; X_k)$ than the discrete modeling of message.

When the environmental channel SNR gradually increases, $I(M; Y)$ for the continuous modeling and $I(M; \hat{M})$ for the discrete modeling will approach $I(M; X)$. Since the continuous modeling has a larger $I(M; X)$, it requires $I(M; Y)$ to have a larger SNR to reach $I(M; X)$. To verify this, we simulated $I(M; Y)$ for the continuous Gaussian modeling and $I(M; \hat{M})$ for the discrete modeling, respectively. We take a one-dimensional signal as the example. The mutual information $I(M; Y)$ for the continuous modeling is calculated as

$$I(M; Y) = -\frac{1}{2} \log_2(1 - \rho_{MX}^2 \rho_{XY}^2), \tag{4.2}$$

where $\rho_{MX} = 0.99$ and ρ_{XY} is a set of values, which correspond to the environmental SNR from -30 dB to 30 dB.

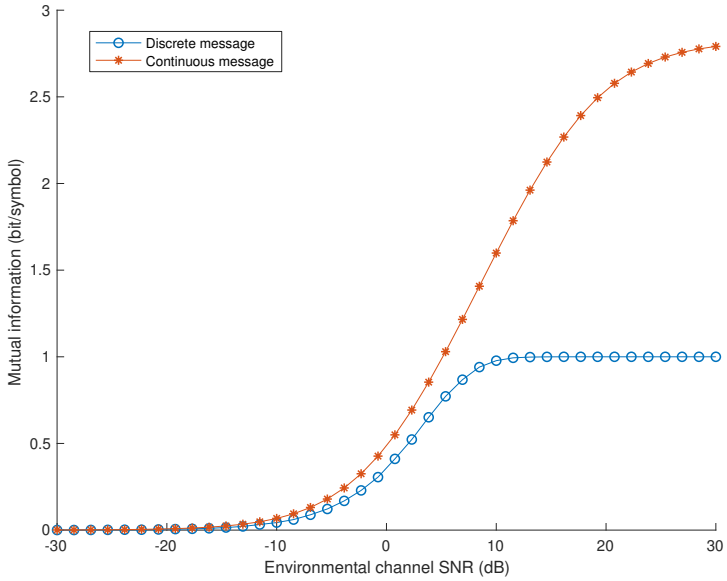


Figure 4.3: Mutual information for discrete modeling and continuous modeling of message.

For the discrete modeling, we generated a binary sequence from $\{-1, 1\}$ with approximately equal probability. Thus, the variance of the discrete signal was 1. The number of samples was 10^5. We generated a white noise as the production noise and added it to the discrete signal to simulate the clean speech. The variance of the production noise was derived, such that the correlation coefficient between the binary sequence and the clean speech was $\rho_{MX} = 0.99$. The environmental noise was generated as white noise with different SNRs from -30 dB to 30 dB. At the receiver side, the decoder selects a binary signal from $\{-1, 1\}$, which is closest to the received signal. The mutual information $I(M; \hat{M})$ is calculated as

$$I(M; \hat{M}) = H(M) + H(\hat{M}) - H(M, \hat{M}), \tag{4.3}$$

where $H(\cdot)$ denotes the entropy. Fig. 4.3 illustrates the mutual information curve for the continuous modeling and the discrete modeling, re-

spectively. For the discrete modeling of message, the mutual information $I(M;\hat{M})$ approaches 1 bit/symbol, which is equal to $I(M;X) = H(M)$. For the continuous modeling of message, the mutual information $I(M;Y)$ approaches $-\frac{1}{2}\log_2(1-0.99^2) = 2.8$ bit/symbol, which is equal to $I(M;X)$. In Fig. 4.3, we can see for the same variance of clean speech, the continuous modeling of message requires a larger SNR to reach $I(M;X)$ than the discrete modeling of message.

4.3.2 Length of the codeword

The second reason that the discrete modeling can provide a better fit of mutual information curve to the psychometric curve is that the discrete modeling can generate a steeper mutual information curve for a longer codeword. This is because for a given codeword with higher dimensions, its noisy samples are more likely to locate on the sphere of a ball. When the balls of each codeword just contact, adding a little bit more noise will make noisy samples of one codeword move to the space of another codeword. Thus, the probability of correct transmission suddenly drops, which means a quick drop of mutual information. In speech communication, we decode the current linguistic information by taking into account the preceding and the following linguistic information. For example, we decode a phoneme in the context of word, and decode a word in the context of a sentence. By using the context information, the length of the codeword is increased. Thus, the discrete modeling gives a sharper mutual information curve.

The mathematical proof is given in the following. We denote a codeword by

$$\boldsymbol{X} = [X_1, X_2, \cdots, X_L], \tag{4.4}$$

where L is the dimension of the codeword. We generate a set of code-

words, such that the power of each codeword is constant:

$$\begin{aligned} P_X &= X_1^2 + X_2^2 + \cdots + X_L^2 \\ &= LP_{X_i}, \end{aligned} \tag{4.5}$$

where P_{X_i} is the average power for each dimension. We also set the mean of the codewords for each dimension as constant:

$$\mu_{X_i} = \mathrm{E}[X_1] = \mathrm{E}[X_2] = \cdots = \mathrm{E}[X_L]. \tag{4.6}$$

Now we have two symbols $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ that are mapped into the L dimensional space. The Euclidean distance between two symbols is

$$\begin{aligned} D_{\boldsymbol{x}_a\boldsymbol{x}_b} &= ||\boldsymbol{x}_a - \boldsymbol{x}_b||_2 \\ &= \sqrt{(x_{a1} - x_{b1})^2 + \cdots + (x_{aL} - x_{bL})^2} \\ &= \sqrt{x_{a1}^2 + \cdots + x_{aL}^2 + x_{b1}^2 + \cdots + x_{bL}^2 - 2x_{a1}x_{b1} - \cdots - 2x_{aL}x_{bL}} \\ &\approx \sqrt{2P_X - 2L\mathrm{E}^2[X_i]} \\ &= \sqrt{2L(P_{X_i} - \mu_{X_i}^2)} \\ &= D, \end{aligned} \tag{4.7}$$

where D is a constant. Thus, the distance between any two codewords is approximately constant for very long codewords.

Given a codeword $\boldsymbol{x}_a$, we generate a noisy sample $\boldsymbol{y}$:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}_a + \boldsymbol{n} \\ &= \begin{bmatrix} x_{a1} \\ x_{a2} \\ \vdots \\ x_{aL} \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_L \end{bmatrix}, \end{aligned} \tag{4.8}$$

where $\boldsymbol{n}$ is a L-dimensional white noise. The distance between the codeword and the noisy sample is $D_{\boldsymbol{x}_a\boldsymbol{y}} = ||\boldsymbol{n}||_2$. The probability of correct recognition is

$$P_c = P(D_{\boldsymbol{x}_a\boldsymbol{y}} < D_{\boldsymbol{x}_i\boldsymbol{y}} | \boldsymbol{x}_a), \tag{4.9}$$

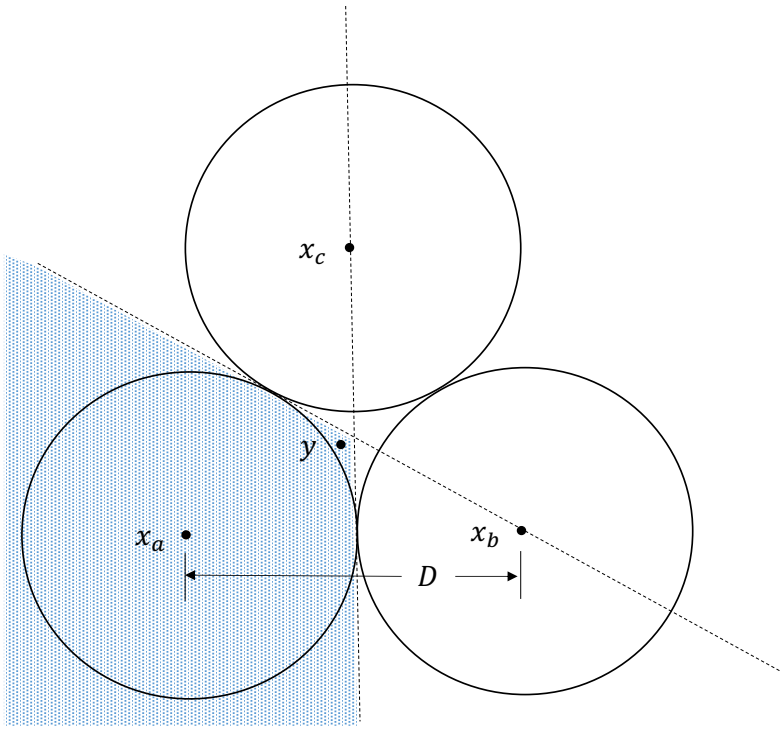


Figure 4.4: An example of the distributions of codewords and noisy samples. The shaded area illustrates the region of noisy samples belonging to x_a.

where $\boldsymbol{x}_i$ denotes any codeword except $\boldsymbol{x}_a$. Fig. 4.4 illustrates a set of codewords, where any two codewords have the same distance D. The space satisfying $D_{\boldsymbol{x}_a\boldsymbol{y}} < D_{\boldsymbol{x}_i\boldsymbol{y}}$ can be divided into two subspace, where $D_{\boldsymbol{x}_a\boldsymbol{y}} < \frac{D}{2}$ and $D_{\boldsymbol{x}_a\boldsymbol{y}} \geq \frac{D}{2}$. As long as $D_{\boldsymbol{x}_a\boldsymbol{y}} < \frac{D}{2}$, the noisy sample $\boldsymbol{y}$ can always be correctly recognized. In some spaces of $D_{\boldsymbol{x}_a\boldsymbol{y}} \geq \frac{D}{2}$, the noisy sample $\boldsymbol{y}$ can still be correctly recognized, as shown by the noisy sample $\boldsymbol{y}$ in Fig. 4.4. Since there is no direct calculation of (4.9), we use the first subspace to calculate

a lower bound, which is expressed as

$$\begin{aligned} P_c &\geq P(D_{\boldsymbol{x}_a \boldsymbol{y}} < \frac{D}{2}) \\ &= P(||\boldsymbol{n}||_2 < \frac{D}{2}) \\ &= P(||\boldsymbol{n}||_2^2 < \frac{D^2}{4}) \end{aligned} \tag{4.10}$$

We denote the standard deviation of $\boldsymbol{n}$ for each dimension by c_n, which is a constant. The noise is assumed to be

$$\boldsymbol{n} = c_n \hat{\boldsymbol{n}}, \tag{4.11}$$

where $\hat{\boldsymbol{n}}$ is a L-dimensional white noise, and the noise in each dimension has a standard normal distribution. Substitute (4.11) into (4.10), we have

$$P_c \geq P(||\hat{\boldsymbol{n}}||_2^2 < \frac{D^2}{4c_n^2}). \tag{4.12}$$

Note that $||\hat{\boldsymbol{n}}||_2^2$ is a chi-square distribution with L degrees of freedom, and $\frac{D^2}{4c_n^2}$ depends on L and c_n. We can plot the lower bound of P_c with different dimensions and noise levels. When the noise levels is infinite large, the distances between a noisy sample and all the codewords are the same. Thus,

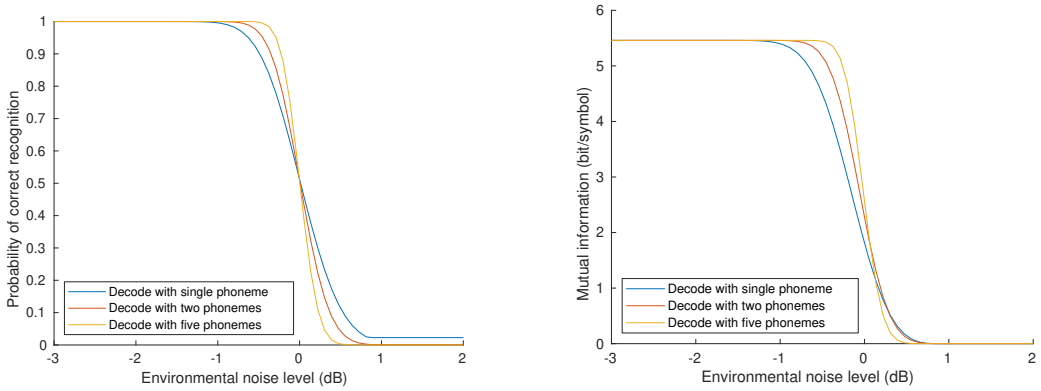
$$P_c = \frac{1}{N_{\text{cw}}}, \tag{4.13}$$

where N_{cw} is the number of codewords. Combining (4.12) and (4.13), we obtain the lower bound

$$P_c \geq \max\{P(||\hat{\boldsymbol{n}}||_2^2 < \frac{D^2}{4c_n^2}), \frac{1}{N_{\text{cw}}}\}. \tag{4.14}$$

We run the simulation based on the transmission of auditory spectrogram, which is assumed to carry all the linguistic information. The auditory spectrogram is a sequence of spectrum with the sampling frequency of 80 spectra/second (25 ms window with 50% overlap). We assume in speech the phoneme rate is 10 phonemes/second. We represent each spectrum with 30 ERB bands. Thus, each phoneme is represented by 8 spectra,

and the total dimensionality of a phoneme is $30 \times 8 = 240$. In the simulation, we set $D = 1$ for $L = 240$. Fig. 4.5a shows the lower bound of P_c with different dimensions and noise levels. The blue line shows the noisy samples are decoded every phoneme. The red line and the orange line consider the context information by using longer codeword length. We can see the curve of the probability of correct recognition becomes steeper with longer codeword length. Fig. 4.5b shows similar performance for the mutual information curve.



(a) Probability of correct recognition. (b) Mutual information.

Figure 4.5: Impact of the length of codeword.

As shown in Chapter 3, the psychometric curves are steep, which means they have a narrow range of SNR corresponding to the intelligibility from 0 to 1. Thus, if a SIP can produce a steep curve, then we consider it to be a good predictor. In speech interpretation, we use long codewords (a sequence of phonemes), instead of short codewords (a single phoneme). Using long codewords produces a steep curve, which is consistent to the psychometric curves.

4.4 Mutual Information Estimate for Discrete-valued Message

In this section, we first study the relationship between mutual information and speech intelligibility. Then, we estimate mutual information between the discrete message and the received speech.The transmission model consists of production channel and acoustic channel. The clean speech and the received speech are known. So the mutual information between them can be calculated. However, the discrete linguistic unit underlying the continuous clean speech cannot be immediately obtained. To solve this problem, we estimate the mutual information in each band between the linguistic unit and the clean speech via the theory of speech intelligibility index.

4.4.1 Relationship between speech intelligibility and mutual information metric

When developing a mutual information based speech intelligibility predictor, we presume that the mutual information and the speech intelligibility have a monotonic relationship. This assumption has not been verified yet. To verify this assumption, we study the relationship between the speech intelligibility and the mutual information based on a simple transmission model.

In this transmission model, speech intelligibility is measured as the percentage of word recognition. Mutual information is calculated between the transmitted word and the received word. We assume there are N_w words, and they are uniformly distributed. Thus, the entropy of the source is $\log_2 N_w$ bit/symbol. Let p denote the probability of correct transmission for each word, M and Z denote the transmitted and received word, respectively. The transmission matrix is

$$P(Z = w_j|M = w_i) = \begin{cases} p, & i = j \\ \frac{1-p}{N_w-1}, & i \neq j, \end{cases} \tag{4.15}$$

where w_i denotes a word. Note that in the worst case, where the listener cannot extract any information from the received signal, $p_{\min} = \frac{1}{N_w}$.

Since M is uniformly distributed and the transmission matrix is the same for all transmitted words, the probability of each received word is $\frac{1}{N_w}$. The entropy of the received word is

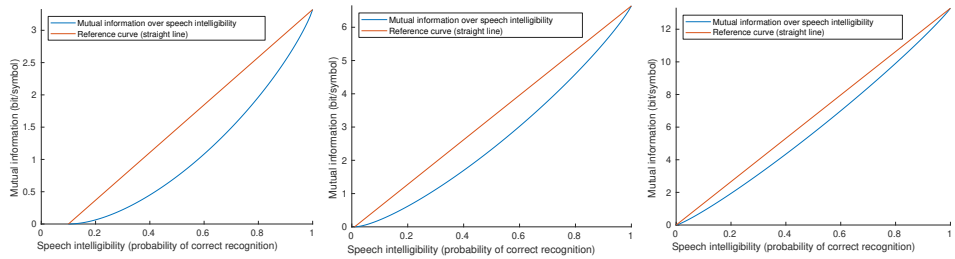
$$H(Z) = \log_2 N_w. \tag{4.16}$$

The conditional entropy $H(Z|M)$ is calculated as

$$\begin{aligned} H(Z|M) &= -\sum_{i,j} P(M = w_i, Z = w_j) \log_2 P(Z = w_j|M = w_i) \\ &= -\sum_i P(M = w_i) \sum_j P(Z = w_j|M = w_i) \log_2 P(Z = w_j|M = w_i) \\ &= -\sum_j P(Z = w_j|M = w_i) \log_2 P(Z = w_j|M = w_i) \\ &= -p \log_2 p - (1-p) \log_2 \frac{1-p}{N_w - 1} \end{aligned} \tag{4.17}$$

The mutual information between the transmitted word and the received word is calculated by

$$\begin{aligned} I(M;Z) &= H(Z) - H(Z|M) \\ &= \log_2 N_w + p \log_2 p + (1-p) \log_2 \frac{1-p}{N_w - 1}. \end{aligned} \tag{4.18}$$



(a) $N_w = 10$. (b) $N_w = 100$. (c) $N_w = 10000$.

Figure 4.6: Relationship between mutual information metric and speech intelligibility.

The blue curve in Fig. 4.6 illustrates the relationship between the mutual information metric and speech intelligibility for three sets of words. The red curve is used as reference, which is a straight line. For the lowest intelligibility of $\frac{1}{N_w}$, the mutual information is 0. For a 100% intelligibility, the mutual information is $\log_2 N_w$ bit/symbol. In figs. 4.6a to 4.6c, we can see the mutual information metric is a monotonically increasing function of speech intelligibility. As the set of words increases, the relationship becomes more linear.

4.4.2 Mutual information between the message and the received speech

To predict speech intelligibility, we desire to calculate the mutual information between the message and the received speech $I(M;Y)$. Recall that we have the Markov chain $M \rightarrow X \rightarrow Y$. Since the exact form of M is unknown for a given speech signal, we use $I(M;X)$ and $I(X;Y)$ to get an approximation of $I(M;Y)$. According to the data processing inequality, we have

$$I(M;Y) \leq \min(I(M;X), I(X;Y)). \tag{4.19}$$

As $I(M;X)$ represents the mutual information of linguistic information, we estimate $I(M;X)$ through the statistical data of linguistic unit, which

is a fixed term. For example, if we use $I(M;X)$ to measure the information per phoneme, it is about $\log_2 44 = 5.5$ bit/phoneme (44 phonemes for English).

The mutual information $I(X;Y)$ quantifies the quality of the acoustic channel. Recall that X and Y are multi-dimensional variables. If one channel has perfect condition, $I(X;Y)$ approaches infinite, which makes $I(M;Y)$ constant according to (4.19). Thus, we apply (4.19) for each channel, rather than the whole channels. As speech channels are frequency correlated, we have

$$\begin{aligned} I(M;Y) &\leq \sum_k I(M_k;Y_k) \\ &\leq \sum_k \min(I(M_k;X_k), I(X_k;Y_k)). \end{aligned} \tag{4.20}$$

We assume X_k and Y_k are continuous Gaussian distributed, then

$$I(X_k;Y_k) = -\frac{1}{2}\log_2(1-\rho^2_{M_kX_k}), \tag{4.21}$$

where $\rho_{M_kX_k}$ is the correlation coefficient between the clean speech and the received speech at channel k.

4.4.3 Mutual information for the production channel

Although M is unknown for a given speech signal, we can estimate $I(M;X)$ from the statistical data in linguistic study. However, we are interested in $I(M_k;X_k)$. Recall that speech intelligibility can be obtained from speech intelligibility index (SII) by applying a transfer function [55]. If we approximate the ratio of mutual information and speech intelligibility to be a constant, as shown in Fig. 4.6c, we can derive $I(M_k;X_k)$ from the speech intelligibility at each channel. The speech intelligibility at each channel is the intelligibility of a narrowband clean speech and we can derive this value by using the band importance function [55] and the transfer function [20].

Let us first estimate $I(M;X)$. In our study, M, X, and Y are multi-dimensional variables to represent the message, the speech, and the received speech for a phoneme. Although the smallest linguistic unit is a morpheme, we hypothesize that a better recognition of phoneme achieves a better recognition of linguistic unit. Thus, we can use the mutual information between the transmitted phoneme and the received phoneme to estimate speech intelligibility. Based on the probability of phonemes appearing in speech [3], the entropy of phoneme is estimated as

$$\begin{aligned} I(M;X) &= H(M) \\ &= -\sum_j p_i \log_2 p_i \\ &= 4.9 \text{ bit/phoneme}. \end{aligned} \tag{4.22}$$

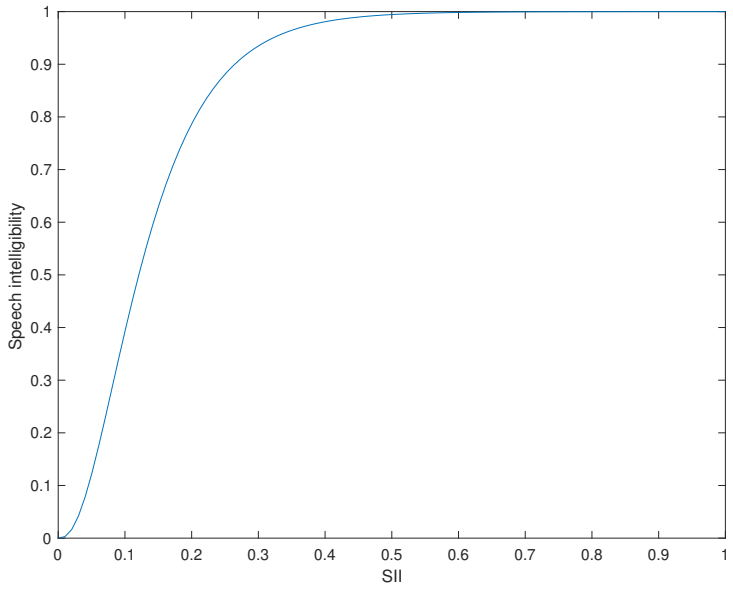


Figure 4.7: Transfer function for speech sentence.

Next, we estimate $I(M_k;X_k)$ based on speech intelligibility. Fig. 4.6c shows that the ratio of mutual information and speech intelligibility is almost a constant. Thus, we estimate $I(M_k;X_k)$ as

$$I(M_k;X_k) = p_k\, I(M;X), \tag{4.23}$$

where p_k is the speech intelligibility at channel k. The speech intelligibility can be derived from SII. The relationship between speech intelligibility and SII is represented by the transfer function. The exact form of the transfer function depends on speech material, e.g., words or sentence, and the proficiency of listeners [19, 102–104]. For the speech material of the sentence, the transfer function is expressed as [20]

$$p = (1 - 10^{-\frac{2.3}{0.428} SII})^{2.729}. \tag{4.24}$$

Fig. 4.7 shows the relationship between speech intelligibility and SII for speech sentence. To calculate p_k, we need to calculate SII at channel k. For clean speech, the SII at channel k is the same as the band importance function. The band importance function for standard SII can be found in [55]. To adjust the band importance function for the ERB-rate scale, we first calculate the band importance function density, i.e., band importance per Hz. Then, we integrate over the ERB band. Fig. 4.8 illustrates the SII for different ERB bands. The points on the curve represent the center frequency for each ERB band. Based on eqs. (4.23) and (4.24), and SII_k, we can calculate the mutual information $I(M_k; X_k)$ for each ERB band, which is shown in Fig. 4.9.

4.5 Evaluation

In this section, we evaluate the proposed transmission model based on a discrete-valued linguistic message. We estimated the mutual information between discrete linguistic unit and received speech, and compare the mutual information with the psychometric curve. If the discrete transmission model can represent the true transmission model, the mutual information curve based on the discrete model should give a better fit than the mutual information curve based on our continuous model of Section 3.2.

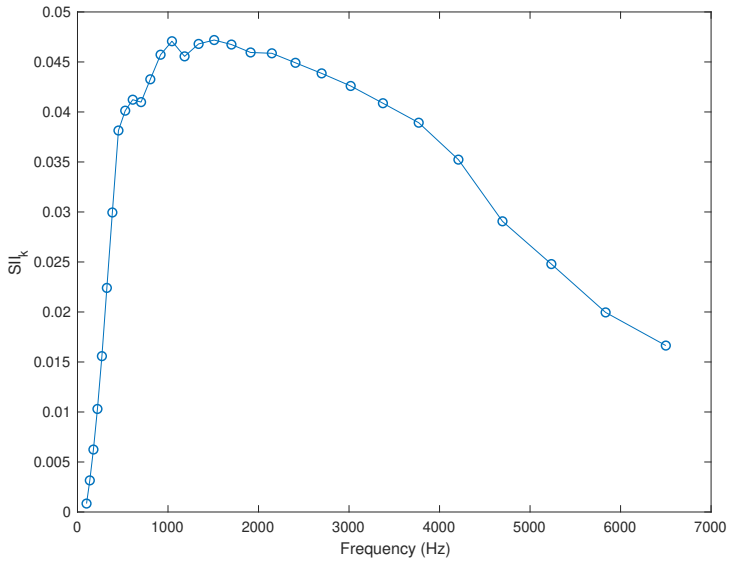


Figure 4.8: SII for different ERB band.

4.5.1 Data set

The data set is the same as the data set used for evaluating the continuous model, as introduced in Section 3.4.1.

4.5.2 Implementation

The mutual information between transmitted linguistic unit and received speech is calculated as (4.20). The log-auditory spectrogram was calculated as introduced in Section 3.3.1.

4.5.3 Results and discussion

The experimental results are shown in Fig. 4.10. Fig. 4.10a is the psychometric curve. Fig. 4.10b is the mutual information metric based on the discrete model. Fig. 4.10c is the mutual information metric based on the continuous model. Comparing Fig. 4.10b and Fig. 4.10c, we can see the SNR interval for the discrete model is more similar to the psychometric

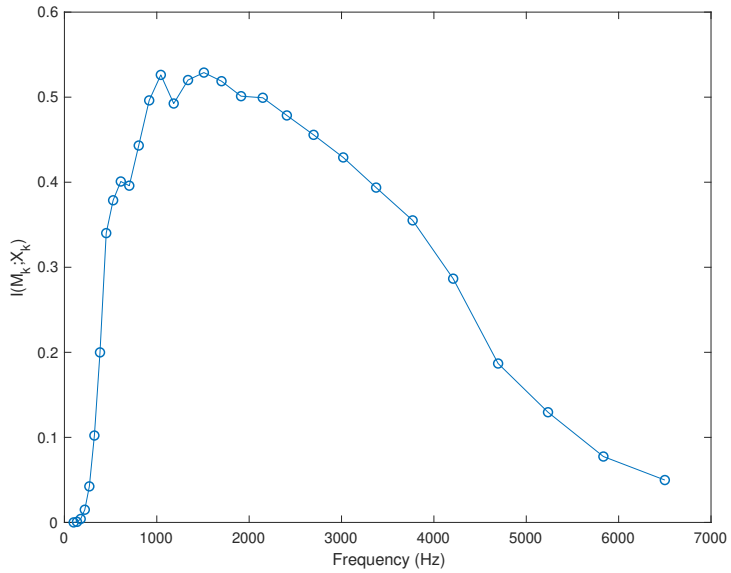
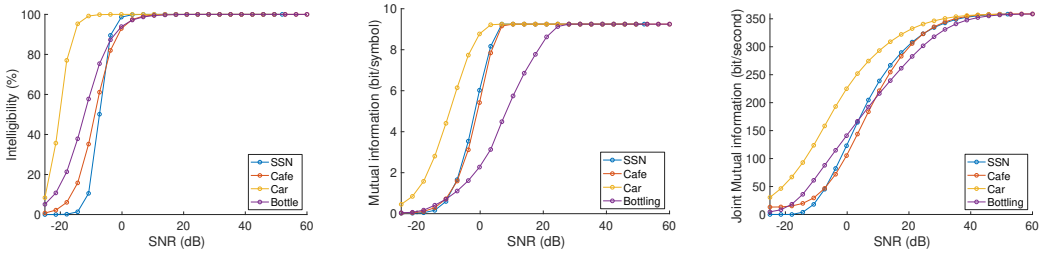


Figure 4.9: Mutual information for different ERB bands.

than the continuous model. This suggests that the linguistic unit has a discrete-valued distribution.

Although Fig. 4.10b suggests the modeling of discrete-valued linguistic unit, the mutual information curves do not match to the psychometric curves. For the same SNR, the mutual information curves of four conditions do not show consistent order as the psychometric curve. For the SSN and cafe noise, the mutual information curves almost overlap. Recall that in the discrete model, $I(X_k; Y_k)$ only depends on $\rho_{X_k Y_k}$, which does not distinguish stationary and non-stationary noise. In Fig. 4.10b, the location of bottling noise curve does not match to the bottling noise curve in the psychometric curve. However, in Fig. 4.10c the bottling noise curve matches better to the psychometric curve than the bottling noise curve in Fig. 4.10b. In Fig. 4.10b, we estimate the mutual information $I(M_k; X_k)$ for each band, which does not consider the frequency correlation, while in Fig. 4.10c frequency correlation is considered. This indicates that when calculating mutual information, taking into account of frequency correlation can make the mutual information curve match better to the psychometric

curve.



(a) Psychometric curve. (b) Discrete model. (c) Continuous model.

Figure 4.10: Subjective and objective results.

4.6 Summary

In this chapter, we proposed a transmission model based on discrete-valued linguistic unit. In the discrete model, the message goes through a channel codec. This explains the robustness of speech in a noisy environment, while the continuous model fails, because channel codec does not exist.

We verified the discrete model by comparing the psychometric curve and the mutual information curve. To obtain the mutual information curve, we need to calculate the mutual information. Unlike the continuous model, we cannot generate a pseudo-message based on the statistical characteristics of the message. Since the message transmission can be modeled as a Markov chain, we calculate the mutual information $I(M_k; X_k)$ and $I(X_k; Y_k)$ separately, and choose the minimal value as the mutual information $I(M_k; Y_k)$. The mutual information $I(M; Y)$ is the sum of $I(M_k; Y_k)$, which implies frequency correlation is not considered.

When comparing the psychometric curve and the mutual information curve, the SNR interval in the discrete model matches better to the psychometric curve than the continuous model. This suggests that the message is discrete-valued. However, in the discrete model, the order of the mutual information curves of four noise conditions does not match better to

the psychometric curve than the continuous model. The difference in the calculation of mutual information between these two models is that frequency correlation is considered in the continuous model, while not in the discrete model.

In conclusion, the proposed discrete model verifies that the discrete-unit model correctly predicts the steep transition from fully intelligible to not intelligible. To obtain a correct order of the mutual information curves, frequency correlation should be considered.

Chapter 5

Mutual Information based Speech Intelligibility Predictor

In this chapter, we propose a speech intelligibility predictor (SIP) for normal-hearing listeners. The proposed SIP is an extension of SIIB [13], which is the best intelligibility predictor in the mutual information (MI) criterion. SIIB assumes the underlying linguistic unit is a continuous-valued sound, and predicts speech intelligibility by calculating mutual information between the continuous-valued sound and the degraded speech. In Chapter 4, we showed that the linguistic unit is discrete-valued. Based on this finding, we propose a new mutual information based speech intelligibility predictor.

5.1 Background

Speech intelligibility is a key criterion when developing a speech communication system, such as mobile communications, public address systems, hearing instruments, etc. Speech intelligibility is defined as the proportion of linguistic units that can be correctly perceived by a listener in a listening test [105]. Speech intelligibility is measured by listening tests, which is time-consuming. To replace listening tests, the development of an objec-

tive speech intelligibility predictor (SIP) has received significant research interest.

Currently, SIIB and SIIB$^{\text{Gauss}}$ are the best SIPs based on a mutual information criterion. Since speech intelligibility is related to linguistic information, it is natural to predict speech intelligibility by calculating mutual information between linguistic units and degraded speech. In SIIB and SIIB$^{\text{Gauss}}$, the continuous-valued sound is viewed as linguistic information. However, in Chapter 3 and Chapter 4, we showed that the linguistic information consists of discrete-valued symbols, as the discrete-valued linguistic information gives a better fit for the mutual information curve with the psychometric curve.

In this chapter, we modify SIIB and SIIB$^{\text{Gauss}}$ in two aspects. First, the new SIP is modified based on the fact that linguistic information is discrete valued. As the linguistic information is unknown from a single clean speech signal, speech intelligibility is estimated by calculating the mutual information between the clean speech and degraded speech. Second, the new SIP considers frequency correlation in the degraded speech. When predicting speech intelligibility, the peripheral auditory system is usually used. A speech signal is decomposed into a set of narrow-band signals, as the human cochlea can be viewed as a filterbank. Linguistic information is encoded redundantly in these auditory bands [106–108]. For these two MI-based SIPs, frequency correlation among the auditory bands has not been fully considered. In [109, 110], mutual information is calculated for each one-third octave band independently and then summed up as the total mutual information. In SIIB and SIIB$^{\text{Gauss}}$, frequency correlation has been only considered for clean speech. The log-auditory spectra of clean and degraded speech are assumed Gaussian distributed. The Karhunen-Loève transform (KLT) is applied to the log-auditory spectra of clean speech to remove frequency correlation. Since the original signal is multivariate Gaussian distributed, the bands in the transformed log-auditory spectra of clean speech are independent. The same KLT matrix is applied to the

degraded speech. However, it cannot be guaranteed that the bands in the transformed log-auditory spectra of degraded speech are uncorrelated, because the covariance matrix of the degraded speech is not the same as or a scalar multiple of the covariance matrix of the clean speech. Instead of calculating mutual information for each original auditory band, SIIB and SIIB$^{\text{Gauss}}$ calculate mutual information for each transformed band. Although SIIB and SIIB$^{\text{Gauss}}$ only consider removing frequency correlation in the clean speech, they still achieve good prediction result when compared to the other intrusive SIPs [41].

5.2 A New Mutual Information based SIP

In this section, we first recap SIIB$^{\text{Gauss}}$. To solve the issues in SIIB$^{\text{Gauss}}$, we investigate the frequency correlation of the degraded speech and propose a new SIP based on mutual information.

5.2.1 Recap of SIIB$^{\text{Gauss}}$

In SIIB$^{\text{Gauss}}$, speech intelligibility is estimated by calculating the mutual information between the message and the degraded signal. Let $X_{j,t}$ and $Y_{j,t}$ denote the time-stacked log-auditory spectrograms of clean speech and degraded speech at band index j and time index t. A KLT matrix is obtained from the clean speech. This KLT matrix is applied on both $X_{j,t}$ and $Y_{j,t}$ to remove the frequency correlation. Thus, the transformed clean speech $X^H_{j,t}$ is frequency uncorrelated, while the transformed degraded speech $Y^H_{j,t}$ is approximately frequency uncorrelated.

SIIB$^{\text{Gauss}}$ has a two-stage process: the first is from message to clean speech, the second is from clean speech to degraded speech. To take the production channel into account, SIIB$^{\text{Gauss}}$ assumes a constant correlation coefficient of 0.75 between the transformed message and the transformed clean speech. The mutual information between the message and the de-

graded speech is calculated as

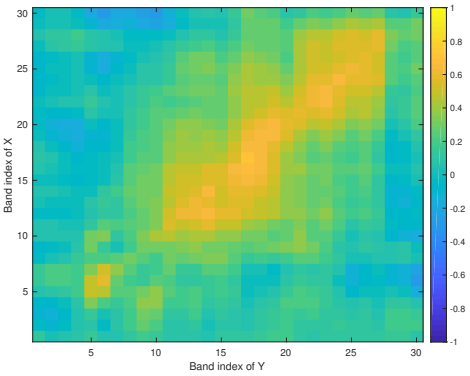
$$\begin{aligned} I(M;Y) &= \frac{1}{K}\sum_{j=1}^{KJ} \min(I(M_j^H;X_j^H), I(X_j^H;Y_j^H)) \\ &= \frac{1}{K}\sum_{j=1}^{KJ} \min(-\frac{1}{2}\log_2(1-0.75^2), -\frac{1}{2}\log_2(1-\rho^2_{X_j^H Y_j^H})), \end{aligned} \tag{5.1}$$

where J is the number of frequency channel of the original log-auditory spectrogram, K is the number of stacked frames, $\rho_{X_j^H Y_j^H}$ is the correlation coefficient between the transformed clean speech and the transformed degraded speech.

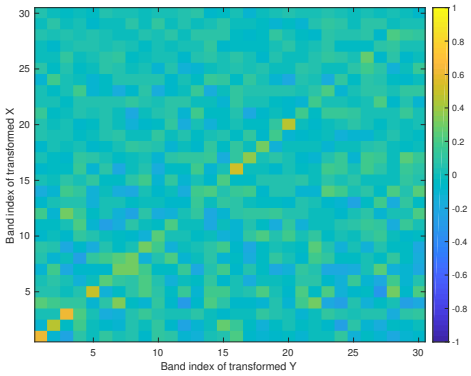
5.2.2 Frequency correlation in degraded speech

Frequency correlation is an important factor when developing a SIP. For many intrusive SIPs, the clean and degraded signals are processed through the peripheral auditory system, which can be viewed as a filterbank. The generated multi-dimensional signal is frequency correlated. In [41], it shows that the SIPs that consider frequency correlation perform better than the SIPs that do not consider frequency correlation. In SIIB$^{\text{Gauss}}$, the clean and degraded signals are log-auditory spectra. Since the KLT matrix is obtained based on the clean signal, the transformed clean signal is frequency uncorrelated. However, it cannot be guaranteed that the transformed degraded signal is frequency uncorrelated, as the covariance matrix of the degraded signal is not the same as the covariance matrix of the clean signal. Thus, the band j of the transformed clean speech is correlated to the band j and its adjacent bands of the transformed degraded speech.

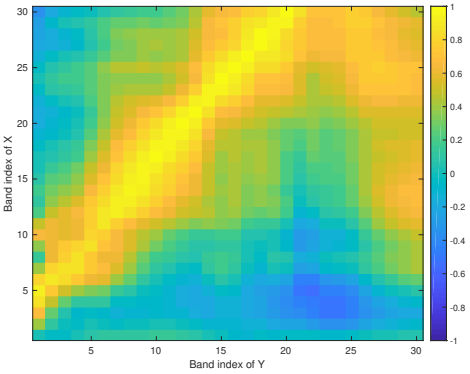
We illustrate this point with two examples. In the first example, the degraded signal is a noisy speech signal generated by a clean speech signal and a cafeteria noise with the SNR of -10 dB. Fig. 5.1 shows the correlation coefficients between the clean and the degraded signals. Since the two signals are multi-dimensional, the correlation coefficient in the matrices are



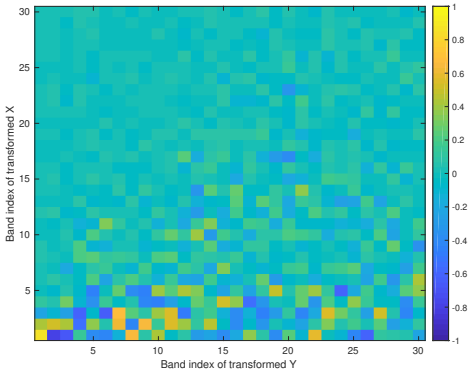
(a) Correlation coefficient between a clean signal and a noisy signal.



(b) Correlation coefficient between the corresponding transformed signals.



(c) Correlation coefficient between a clean signal and a frequency-lowered signal.



(d) Correlation coefficient between the corresponding transformed signals.

Figure 5.1: The left-side figures represent the correlation coefficient between the clean signal and the degraded signal before KLT. The right-side figures represent the correlation coefficient between the transformed clean signal and the transformed degraded signal.

not symmetric. Fig. 5.1a shows the band j of the clean signal is correlated with the band j and its adjacent bands of the degraded signal. After KLT, the correlation coefficient between the bands of the transformed clean signal and the transformed degraded signal is shown in Fig. 5.1b. We can see the diagonal line, but the values of the diagonal elements are similar to the off-diagonal elements. Thus, the off-diagonal elements should also be considered in the calculation of mutual information. In the second example, the degraded signal is a frequency lowered speech signal, which aims to improve high-frequency components of speech for hearing-impaired listeners. Fig. 5.1c shows the most correlated bands are off-diagonal, due to the frequency shift of the degraded signal. Fig. 5.1d does not show a diagonal line for the transformed signals. Since SIIB $^{\text{Gauss}}$ uses the diagonal elements to estimate speech intelligibility, the frequency-lowered method is not a suitable application for SIIB $^{\text{Gauss}}$.

5.2.3 Intelligibility prediction by mutual information

The new intelligibility predictor is modified to account for the facts that the message is a discrete linguistic unit and the transformed degraded signal is frequency correlated. Ideally, the speech intelligibility is estimated as $I(M;Y)$. Due to the Markov chain $M \to X \to Y$, we have

$$\begin{aligned} I(Y;X) &= I(Y;M,X) \\ &= I(Y;M) + I(Y;X|M), \end{aligned} \tag{5.2}$$

where M is the message, X is the clean speech, and Y is the degraded speech. Then,

$$I(M;Y) = I(X;Y) - I(Y;X|M). \tag{5.3}$$

As $I(Y;X|M) \geq 0$, we have

$$I(M;Y) \leq I(X;Y). \tag{5.4}$$

Calculating $I(Y;X|M)$ is impossible, because we do not know the exact distribution of M. Thus, we use the upper bound $I(X;Y)$ to replace

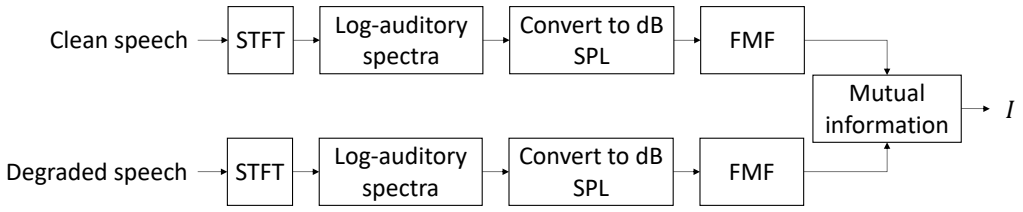


Figure 5.2: Diagram of the proposed SIP. FMF is the abbreviation of forward masking function.

$I(M;Y)$, and the error is $I(Y;X|M)$.

The diagram of the proposed SIP is shown in Fig. 5.2. The time-domain speech is first processed by the auditory peripheral processing, which is represented by the modules of STFT and log-auditory spectra. Then, the signal is converted from dB scale to dB SPL scale. The module of forward masking function (FMF) [75] takes into account the forward temporal masking. Finally, the mutual information is calculated between the clean speech and the degraded speech in this domain.

Now we study how the production noise can influence the error $I(Y;X|M)$. (5.3) can be rewritten as

$$I(Y;X|M) = I(X;Y) - I(M;Y). \tag{5.5}$$

Note that the production noise affects only $I(M;Y)$. When the production noise is much larger than the environmental noise,

$$\begin{aligned} I(M;Y) &\approx I(M;X) \\ &\ll I(X;Y). \end{aligned} \tag{5.6}$$

Then the error

$$\begin{aligned} I(Y;X|M) &= I(X;Y) - I(M;Y) \\ &\approx I(X;Y) - I(M;X) \\ &\approx I(X;Y). \end{aligned} \tag{5.7}$$

When the production noise is much smaller than the environmental noise,

$$I(M;Y) \approx I(X;Y). \tag{5.8}$$

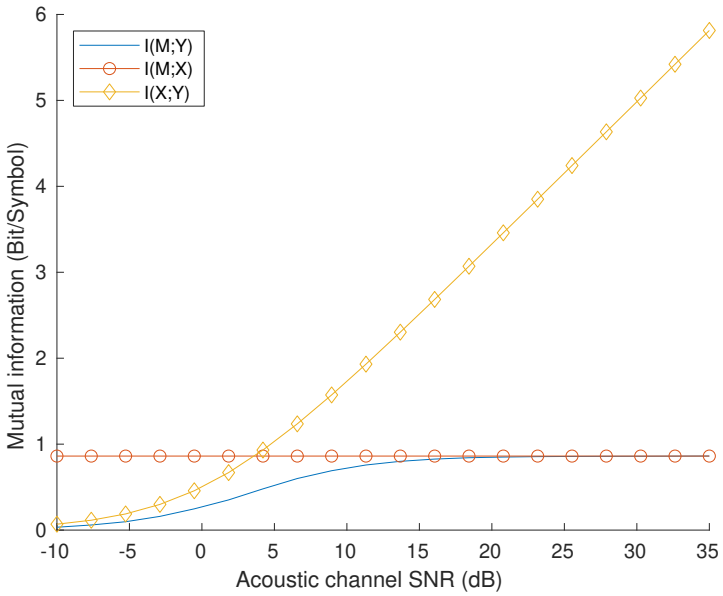


Figure 5.3: $I(M;Y)$, $I(M;X)$, and $I(X;Y)$ for different acoustic channel SNRs.

Then, the error

$$I(Y;X|M) \approx 0. \tag{5.9}$$

(5.7) and (5.9) show two extreme cases for $I(Y;X|M)$. Based on (5.5), $I(Y;X|M) \in (0, I(X;Y))$.

We use an example of single channel system to illustrate the relationship among $I(M;Y)$, $I(M;X)$, and $I(X;Y)$. The discrete messages were -1 and 1, which were uniformly distributed. The production noise was simulated as white noise and added to the discrete message to generate the clean speech. The correlation coefficient between M and X was 0.9. The acoustic noise was also simulated as white noise. We fixed the level of production noise, but varied the level of acoustic noise. The SNR of acoustic channel was varied from -10 dB to 35 dB. Fig. 5.3 shows the mutual information curves of $I(M;Y)$, $I(M;X)$, and $I(X;Y)$ for different acoustic channel SNRs. The entropy of the discrete message is $H(M) = 1$ bit/symbol. Due to the production noise, $I(M;X)$ is smaller than 1. At

high acoustic SNR, the production noise is much larger than the acoustic noise. We can see the blue line is overlapped with the red line, which is consistent with (5.6). At low acoustic SNR, the production noise is much smaller than the acoustic noise. We can see the blue line approaches the yellow line, which is consistent with (5.8).

5.3 Listening Test Data

In this section, we introduce four listening test data, which provide the results of subjective listening tests. These true intelligibility scores were used in the evaluation of the proposed SIP that calculates the mutual information between the clean speech and the degraded speech.

5.3.1 Kjems AN

Kjems AN is a data set that consists of noisy speech. 'AN' stands for additive noise, which means the noisy speech were constructed by adding noise directly on clean speech signals. In [94], Kjems *et al.* derived a psychometric curve, which relates speech intelligibility and acoustic channel SNR. Thus, based on the psychometric curve, we can obtain the intelligibility for any SNR. The clean speech was from the Dantale II corpus [95]. Four types of noise were used, which included speech shaped noise (SSN), cafeteria noise, car noise, and bottling noise. For each type of noise, 10 samples of the intelligibility data on the psychometric curve were selected. Nine samples were uniformly distributed over the intelligibility from 10% to 90%, and the tenth sample, which had the highest intelligibility value, was 99%. In total $10 \times 4 = 40$ conditions were used.

5.3.2 Kjems ITFS

Kjems ITFS is a data set that consists of ideal T-F segregation (ITFS) processed speech [94]. A noisy speech signal was generated by adding one

of four types of noise: SSN, cafeteria noise, car interior noise, or bottling noise. An ideal binary mask (IBM) and target binary mask (TBM) were used to generate ITFS processed speech. The mask was derived based on the SNR of the clean speech signal and a reference noise signal. The mask was set to 1, if the SNR was above a threshold. The mask was set to 0, if the SNR was below this threshold. By using the mask, useful speech components can be extracted from a noisy speech signal.

In IBM experiments, the noise reference was from each of the four types of noise. In TBM experiments, only speech shaped noise was used as the noise reference. Since the IBM experiment with speech shaped noise was equivalent to TBM, there were seven categories in total: IBM (TBM)/ssn, IBM/cafe, IBM/car, IBM/bottle, TBM/cafe, TBM/car, TBM/bottle. The clean speech were scaled and mixed with noises at three different SNRs, which correspond to 20% intelligibility, 50% intelligibility, and -60 dB. The data set uses -60 dB for the case, where speech cannot be understood at all. However, when applying appropriate binary mask on a noise spectrogram, a speech can still be reconstructed with good intelligibility. The appropriate binary mask can be obtained by choosing suitable LC (local criterion). When SNR>LC, the mask is set to 1, otherwise it is set to 0. RC stands for relative criterion, and RC = LC-SNR. It is the RC that determines the binary mask, not the absolute value of LC. The binary masks were calculated for eight different RC. Thus, in total there were $7 \times 3 \times 8 = 168$ conditions.

5.3.3 NELE Cooke data set

Near-end listening enhancement (NELE) Cooke data set evaluates near-end speech enhancement algorithms. The near-end refers to the side of listener. When people listen to a mobile phone in noisy environments, the background environment cannot be changed. However, speech intelligibility can still be increased by processing the speech signal before it is

played back. NELE Cooke data set evaluates nine algorithms, which are a subset of the original Cooke data set that evaluated 19 algorithms [111]. Since the entire sounds of the original Cooke data set were not available, NELE was created and used in the evaluation of SIIB [13].

We use the same data set as SIIB, which contains unprocessed noisy speech and enhanced speech by nine algorithms. The nine enhancement algorithms include AdaptDRC, F0-shift, IWFEMD, on/offset, OptimalSII, RESSYSMOD, SBM, SEO and SSS [111]. The noisy environment includes six conditions: babble noise from a female talker with SNR of -7 dB, -14 dB, and -21 dB; speech-shaped noise with SNR of 1 dB, -4 dB, and -9 dB. Thus, in total there are $6 \times (1 + 9) = 60$ conditions.

5.3.4 HuPost database

The HuPost database [112] evaluates the performance of eight enhancement algorithms. The sampling frequency is 8 kHz and the bandwidth is 300 - 3400 Hz, due to the use of the intermediate reference system (IRS) filter [113]. The clean speech signals were contaminated by four types of noise: babble noise, car noise, street noise, and train noise, with SNR of 0 dB and 5 dB. The unprocessed noisy speech and processed speech by eight enhancement algorithms were used in listening tests. Thus, in total there are $4 \times 2 \times (1 + 8) = 72$ conditions.

5.4 Evaluation

In this section, we evaluate the performance of the proposed SIP. The proposed SIP uses $I(X;Y)$ to estimate speech intelligibility. We first introduce the parameters that were used to calculate $I(X;Y)$. Then, we compare the performance of the proposed SIP with SIIB$^{\text{Gauss}}$, and its modified version, which uses a more accurate correlation coefficient for the production channel.

5.4.1 Parameter setup

All the speech signals were resampled to 16 kHz and were normalized to unit variance. A voice activity detector (VAD) was used to remove the non-speech segments in the clean speech and the degraded speech. The threshold of the VAD was set to the maximum power of the clean speech signal minus 30 dB. The spectrogram of speech was obtained by the STFT with 50% overlapped 25 ms Hann windows. For HuPost database, the auditory spectorgram was obtained through 19 gammatone filters [114] that were uniformly distributed on the ERB-rate scale from 300 Hz to 3400 Hz. For the other databases, the auditory spectrogram was obtained through 30 gammatone filters that were uniformly distributed on the ERB-rate scale from 100 Hz to 6.5 kHz.

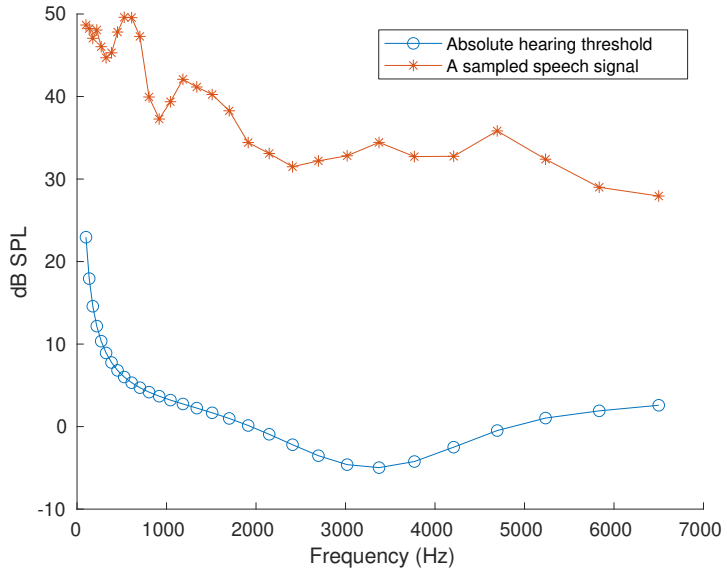


Figure 5.4: Loudness level (dB SPL) of a sampled speech signal and absolute hearing threshold.

Since the proposed SIP uses the temporal forward masking function, which is expressed in the unit of dB SPL, we need to covert the unit of the log-auditory spectrogram from dB to dB SPL. Let us denote the short-

time Fourier transform (STFT) of clean speech by $X_s(j,t)$, where s refers to STFT, j is the FFT bin index, and t is the time index. The log-auditory spectrum $X_k^{\log}(t)$ is calculated as

$$X_k^{\log}(t) = 10 \log_{10} \left(\sum_j W_k^2(j) \left| X_s(j,t) \right|^2 \right), \tag{5.10}$$

where $W_k(j)$ is the magnitude response of kth gammatone filter. To use the absolute hearing threshold in the temporal forward masking function, we convert the log-auditory spectrogram from dB to dB SPL by

$$X_k^{\text{SPL}}(t) = X_k^{\log}(t) + R, \tag{5.11}$$

where R is a constant across the ERB bands. We denote the logarithm of the average power of ERB bands between 800 Hz and 1200 Hz by $\bar{X}^{\log}$. Since the speech banana shows that the average speech energy between this frequency range is 40 dB SPL [115], the average sound pressure level of this narrow-band speech signal is set to 40 dB SPL. Thus, the constant R is calculated as

$$R = 40 - \bar{X}^{\log}. \tag{5.12}$$

Fig. 5.4 illustrates the loudness level of a sample speech signal in dB SPL and the absolute hearing threshold of normal-hearing people. The temporal forward masking function was used to consider the masking effect of the ear [75]. The signal after the processing of the temporal forward masking function is referred to as the log-auditory spectrogram. When calculating the mutual information, time correlation in the log-auditory spectrogram needs to be considered. As with SIIB, 15 continuous frames were stacked at an instant to form a stacked log-auditory spectrogram.

5.4.2 Performance criteria

The performance criteria measure the relationship between the estimated speech intelligibility produced by SIPs and the true speech intelligibility.

Good SIPs can generate positive relation between these two quantities. In our study, we choose Kendall rank correlation coefficient and Pearson correlation coefficient to measure the ordinal association and the linear relationship between them.

The range of Kendall rank correlation coefficient is between -1 and 1. A coefficient of 1 means for different degraded speech signals, the order of the estimated speech intelligibility is the same as the order of the true speech intelligibility. A coefficient of -1 means these two orders are reversed. The coefficient of 0 means the estimated intelligibility is independent of the true intelligibility. There are three types of Kendall rank correlation coefficient. In the evaluation, we use Kendall's Tau-b coefficient to consider the cases, where some degraded speech signals generate identical intelligibility or estimated intelligibility [116].

Pearson correlation coefficient measures the strength of linear relationship between two variables. In the evaluation, we do not measure the Pearson correlation coefficient directly between the true speech intelligibility and the estimated speech intelligibility, due to the following three reasons. First, Fig. 4.6 shows that the relationship between the speech intelligibility and the estimated mutual information is not linear when the number of linguistic units in a listening test is small. Second, speech intelligibility is also affected by the speech material used in a listening test. For example, in the same noisy environment, a listening test using speech sentences achieves higher speech intelligibility than a listening test using words. Third, we use $I(X;Y)$ rather than $I(M;X)$ to estimate speech intelligibility. To take into account of these factors, we use a logistic function expressed as

$$f(I) = \frac{100}{1 + \exp^{a(I-b)}}, \tag{5.13}$$

where $f(I)$ is the estimated speech intelligibility represented as the percent of correctly recognized words, I is the mutual information, a and b are fitted parameters, which minimize the mean square error between the estimated speech intelligibility and the true speech intelligibility.

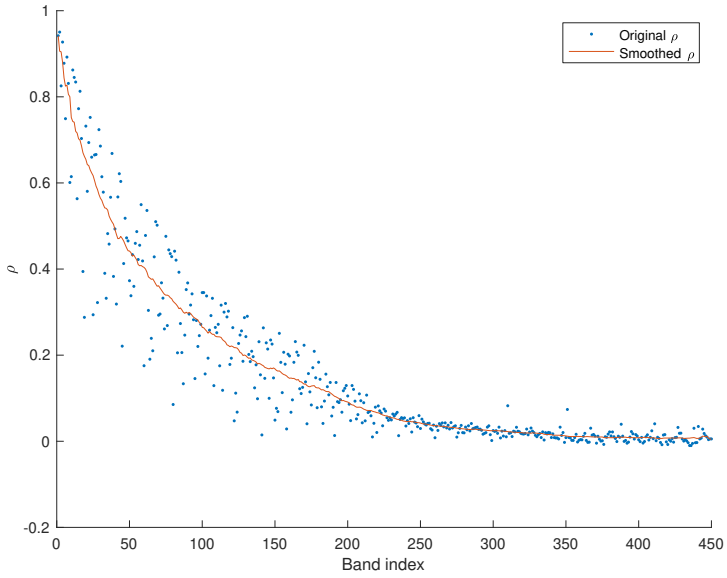


Figure 5.5: True correlation coefficients and its smoothed version for the production channel.

5.4.3 True correlation coefficients of SIIB$^{\text{Gauss}}$

In SIIB$^{\text{Gauss}}$, the log-auditory spectrogram is represented as a sequence of spectra. To consider time correlation, 15 continuous frames are stacked at an instant to form a stacked log-auditory spectrogram. To consider frequency correlation, the same KLT is applied on both the clean speech and the degraded speech. The correlation coefficient between the transformed message and the transformed clean speech is set as a constant of 0.75. However, as we discussed in chapter 3, the correlation coefficient between the transformed message and the transformed clean speech is not constant. In the continuous message-based SIP, we generate a pseudo message based on the correlation coefficient of the transformed production channel. To compare the performance of the discrete message-based SIP and the continuous message-based SIP, the correlation coefficient of the transformed production channel in SIIB$^{\text{Gauss}}$ is updated.

As with chapter 3, we used the CHAINS data set to extract message

Table 5.1: Kendall's rank coefficient for the proposed SIP.

	Kjems AN	Kjems ITFS	NELE Cooke	HuPost	Mean
Proposed SIP	0.89	0.77	0.79	0.74	0.80
$\text{SIIB}^{\text{Gauss}}$	0.80	0.73	0.77	0.72	0.76
$\text{SIIB}^{\text{Gauss}}$ with true correlation coefficients	0.81	0.74	0.68	0.73	0.74

Table 5.2: Pearson correlation coefficient for the proposed SIP.

	Kjems AN	Kjems ITFS	NELE Cooke	HuPost	Mean
Proposed SIP	0.97	0.92	0.95	0.93	0.94
$\text{SIIB}^{\text{Gauss}}$	0.92	0.89	0.95	0.92	0.92
$\text{SIIB}^{\text{Gauss}}$ with true correlation coefficients	0.93	0.91	0.86	0.92	0.91

by ensemble average. The frequency range of log-auditory spectrogram was from 100 Hz to 6500 Hz. Thirty gammatone filters were equally distributed on the corresponding ERB-rate scale. Then, we stacked the message and the clean speech and applied KLT on both. The blue dots in Fig. 5.5 represent the correlation coefficients between the transformed message and the transformed clean speech. To reduce the fluctuation of adjacent correlation coefficients, a moving average filter with the span of 50 data points was used. The red curve in Fig. 5.5 illustrates the smoothed correlation coefficients, which are used to generate the pseudo message in the evaluation of the continuous message-based SIP.

5.4.4 Experimental results

In the evaluation, we compare the performance of the proposed SIP, $\text{SIIB}^{\text{Gauss}}$ and $\text{SIIB}^{\text{Gauss}}$ with updated correlation coefficients. Kendall rank coefficient and Pearson correlation coefficient are shown in Table 5.1 and Table 5.2, respectively. We can see the proposed SIP performs best, followed by $\text{SIIB}^{\text{Gauss}}$ and $\text{SIIB}^{\text{Gauss}}$ with true correlation coefficients of the production channel.

In order to verify if the performance of the proposed SIP is better than the other two SIPs, we carried out a statistical hypothesis test. We only

compare the proposed SIP with SIIB$^{\text{Gauss}}$, as SIIB$^{\text{Gauss}}$ achieves a higher mean value than SIIB$^{\text{Gauss}}$ with true correlation coefficients. As each listening data set has a pair of coefficients that are from the proposed SIP and SIIB$^{\text{Gauss}}$, we can test the difference between two population means on the basis of such paired data [117]. If the difference is larger than 0, then the population mean from one SIP is larger than the other SIP, which means better intelligibility prediction is achieved. The difference of Kendall's rank coefficients is calculated as

$$\Delta\tau = \tau - \tau_{\text{SIIB}^{\text{Gauss}}}, \tag{5.14}$$

where τ and $\tau_{\text{SIIB}^{\text{Gauss}}}$ are Kendall's rank coefficients for the proposed SIP and SIIB$^{\text{Gauss}}$, respectively. Table. 5.3 shows the paired Kendall's rank coefficients and their differences for four data sets. We assume that the population of difference is approximately normal. As there are only four samples of the difference, we use the Student's t test. Let $\mu_{\Delta\tau}$ denote the mean of the population of differences. The null hypothesis and the alternative hypothesis can be formulated as

$$\begin{aligned} H_0 &: \mu_{\Delta\tau} \leq 0 \\ H_1 &: \mu_{\Delta\tau} > 0. \end{aligned} \tag{5.15}$$

The test statistic is

$$\begin{aligned} t_\tau &= \frac{\overline{\Delta\tau}}{s_{\Delta\tau}/\sqrt{N_{\text{Data}}}} \\ &= 2.57, \end{aligned} \tag{5.16}$$

where N_{Data} is the number of the listening test data, $\overline{\Delta\tau}$ and $s_{\Delta\tau}$ are the sample mean and corrected sample standard deviation of $\Delta\tau$, respectively. From the t table, the P-value for the test statistic t_τ is 0.04, which means the plausibility of the null hypothesis is low. Thus, we can confidently reject H_0.

Similarly, we carried out the hypothesis test for the Pearson correlation coefficient. The difference of Pearson correlation coefficients is calculated

Table 5.3: Hypothesis test for Kendall's rank coefficient.

	Kjems AN	Kjems ITFS	NELE Cooke	HuPost
Proposed SIP	0.89	0.77	0.79	0.74
$\text{SIIB}^{\text{Gauss}}$	0.80	0.73	0.77	0.72
Difference	0.09	0.04	0.02	0.02

Table 5.4: Hypothesis test for Pearson correlation coefficient.

	Kjems AN	Kjems ITFS	NELE Cooke	HuPost
Proposed SIP	0.97	0.92	0.95	0.93
$\text{SIIB}^{\text{Gauss}}$	0.92	0.89	0.95	0.92
Difference	0.05	0.04	0	0.01

as

$$\Delta\rho = \rho - \rho_{\text{SIIB}^{\text{Gauss}}}, \tag{5.17}$$

where ρ and $\rho_{\text{SIIB}^{\text{Gauss}}}$ are Pearson correlation coefficients for the proposed SIP and $\text{SIIB}^{\text{Gauss}}$, respectively. Table. 5.4 shows the paired Pearson correlation coefficients and their differences for four data sets. Let $\mu_{\Delta\rho}$ denote the mean of the population of differences. The null hypothesis and the alternative hypothesis can be formulated as

$$\begin{aligned} H_0 &: \mu_{\Delta\rho} \leq 0 \\ H_1 &: \mu_{\Delta\rho} > 0. \end{aligned} \tag{5.18}$$

The test statistic is

$$\begin{aligned} t_\rho &= \frac{\overline{\Delta\rho}}{s_{\Delta\rho}/\sqrt{N_{\text{Data}}}} \\ &= 2.03, \end{aligned} \tag{5.19}$$

where $\overline{\Delta\rho}$ and $s_{\Delta\rho}$ are the sample mean and corrected sample standard deviation of $\Delta\rho$, respectively. From the t table, the P-value is 0.07, which

means the plausibility of the null hypothesis is low. Thus, we can confidently reject H_0. Since the comparisons based on Kendall's rank coefficient and Pearson correlation coefficient are statistically significant, the proposed SIP can achieve a better intelligibility prediction than SIIB$^{\text{Gauss}}$.

5.5 Summary

In this chapter, we proposed a mutual information-based SIP for the discrete linguistic unit. The proposed SIP is different from SIIB and SIIB$^{\text{Gauss}}$ in two aspects. First, the proposed SIIB considers the fact that the speech message is discrete, while SIIB and SIIB$^{\text{arg}}$ assumes speech message is continuous sound. Second, the proposed SIP uses a more accurate mutual information equation for two Gaussian variables, while SIIB and SIIB$^{\text{Gauss}}$ use an approximation of the mutual information between two Gaussian variables.

In the proposed SIP, the log-auditory representation of the underlying discrete message cannot be obtained. Thus, it calculates mutual information between the log-auditory spectrograms of the clean speech and the degraded speech. This avoids introducing error in the generation of a pseudo message. In SIIB and SIIB$^{\text{Gauss}}$, the band k in the transformed clean speech is assumed only correlated to the band k in the transformed degraded speech, which is not true. For the enhancement algorithm of frequency lowering, the effect is more obvious. By assuming a joint Gaussian distribution for the clean and the degraded speech, the proposed SIP considers frequency correlation in both signals.

The proposed SIP shows higher Kendall's rank coefficient and Pearson correlation coefficient than SIIB and SIIB$^{\text{Gauss}}$. Although its performance is the best among mutual information-based SIPs, it is not as good as wSTMI [51], which is a data-driven SIP. This suggests a data-driven mutual information-based SIP can achieve a higher performance. A possible design is as follows. Note that the proposed SIP does not consider the pro-

duction channel, since the representation of the discrete message cannot be obtained. However, the mutual information for the production channel can still be estimated, as shown in Fig. 4.9. Thus, we can estimate the mutual information between the discrete message and the degraded speech for each critical band. A data-driven mutual information-based SIP can be created by assigning the trained weights to the mutual information of each critical band.

Chapter 6

Mutual Information based Frequency Lowering Fitting

In this chapter, we propose an automatic tool for the parameter fitting of a non-linear frequency lowering operator. The automatic tool is an objective mutual information-based SIP. It searches the parameters that maximize the mutual information between the message and the speech received by a hearing-impaired listener. To show the validity of the tool, the objective results, which are from the mutual information based metric, should be consistent to the subjective results, which are from listening tests. Since frequency lowering produces new patterns for speech sounds, the listeners may not be able to recognize speech sounds in the listening tests. To overcome this issue, we propose a new listening test mechanism that is based on sound distinction. The role of the sound distinction test is to obtain subjective speech intelligibility efficiently. The role of the objective measure is to find the optimal parameter in hearing instruments. We will show that the listening test results are consistent to the mutual information-based objective measure.

6.1 Background

Hearing instruments aim to enhance the intelligibility of speech received by hearing-impaired people. For each enhancement operator, its parameters need to be adjusted to maximize the intelligibility of processed speech. Although speech intelligibility can be reliably measured through listening tests, it is time-consuming. Thus, a good objective intelligibility measure can improve the efficiency of hearing instrument fitting.

For hearing-impaired listeners with a dead region at high frequencies, sound amplification in the dead region does not improve speech intelligibility. To recover the high-frequency information, a frequency lowering operator can be used. The main idea of frequency lowering is to present high frequency information at low frequency bands. A fitting procedure is required for the frequency lowering operator [118]. It has been shown that hearing instrument users with frequency lowering are likely to gain an improvement in speech intelligibility [119, 120].

In this chapter, we use a nonlinear frequency compression operator proposed in [58] as an example of a frequency lowering operator. This frequency lowering operator selects a cutoff frequency and compresses the signal above the cutoff frequency nonlinearly. The output frequency can be calculated as

$$f_{out} = \begin{cases} f_{in}, & f_{in} < f_c \\ f_c^{1-1/p} \cdot f_{in}^{1/p}, & f_{in} \geq f_c, \end{cases} \tag{6.1}$$

where f_{out} is the output frequency, f_{in} is the input frequency, f_c is the cutoff frequency, $p \geq 1$ is the compression ratio.

To use the objective measure for the fitting of frequency lowering operator, we need to show the objective measure and the subjective measure can provide consistent results. Since frequency lowering operators move high-frequency information to low frequencies, a high-frequency sound is represented by a new spectrum pattern, which is unfamiliar to the listeners. Thus, acclimatization is required for listeners to be able to identify

the new sounds of phonemes. Acclimatization is a process during which listeners start to learn the new patterns of a sound. Once the new patterns have been learned, the listeners are able to recognize frequency-lowered sounds by matching the input sound with the new patterns stored in the brain. Acclimatization often takes 4-6 weeks [58]. To improve the efficiency of the listening test, we propose a sound distinction test, which simulates the matching process in the brain but does not require acclimation.

6.2 Frequency Resolution of Human Auditory System

The human auditory system has a limited frequency resolution, which means the human ear cannot distinguish two tones whose frequencies are infinitely close. The frequency resolution of the human auditory system depends on the shape of the auditory filter [121]. If a signal has independent components that are very close in frequency, two different components cannot be distinguished, since they are within the same auditory filter. We do not consider the beating effect. When implementing frequency compression, different components get closer in the frequency. Although moving high-frequency components to the audible frequency range can increase the information in the received signal, extreme frequency compression may reduce the information. Thus, we need to find, to what degree the frequency lowering operator maximizes the received information.

The limited frequency resolution of the human auditory system can be interpreted as a limited time-width of a gammatone filter. The time-width describes the area in the time domain where the gammatone filter occupies. The time-width of a signal can be calculated via the inverse Fourier transform of the signal's frequency representation. The firing rate for the kth gammatone filter depends on the log-auditory spectra, which can be

calculated by (3.41). Since the calculation of the auditory spectra involves the summation of the energy of the speech signal, whose frequency components are within the shape of the gammatone filter, the auditory spectra can be viewed as a convolution of the speech spectra with the squared frequency response of the gammatone filter. In the time domain, this means after the peripheral auditory system, the time-width of the processed signal should always be within the time-width of the gammatone filter. As clean speech is fully intelligible, the time-width of the message must be within the time-width of the auditory filter. For speech with extreme frequency compression, its time-width exceeds the time-width of the gammatone filter. Thus, the exceeding part is lost, which means extreme frequency compression incurs information loss.

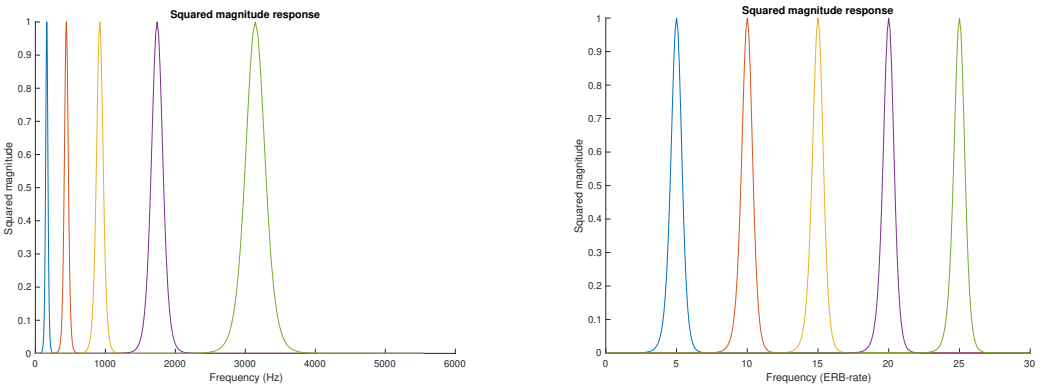
The time-width of the gammatone filter is determined by the shape of the filter. On the Hz scale of frequency, the shape of the gammatone filter varies across frequencies. With increasing of the center frequency, the bandwidth of the gammatone filter becomes gradually wider. However, on the ERB-rate scale of frequency, the shape of the gammatone filter remains identical for different center frequencies. In the following, we first introduce the procedure to derive the frequency response of the gammatone filter on the ERB-rate scale. Then, we plot the squared-magnitude response of the gammatone filters on both the Hz scale and the ERB-rate scale. Finally, we compare the time-width between speech and the human auditory system.

On the Hz scale, the frequency response of the gammatone filter at frequency f_0 can be represented by [114]

$$G(f) \approx [1 + j(f - f_0)/b]^{-n}, \tag{6.2}$$

where $j = \sqrt{-1}$, b is the bandwidth of the gammatone filter, n is the filter order. For a fourth-order gammatone filter, the bandwidth is calculated by [114]

$$b = 1.019\,\mathrm{ERB}(f), \tag{6.3}$$



(a) Hz scale.

(b) ERB-rate scale.

Figure 6.1: Squared-magnitude response of gammatone filters.

where $\mathrm{ERB}(f)$ is the equivalent rectangular bandwidth (ERB) of the auditory filter at the center frequency f in Hz. The ERB is calculated as [121]

$$\mathrm{ERB}(f) = 24.7\,(4.37\,f + 1). \tag{6.4}$$

The relationship between Hz scale and the ERB-rate scale can be expressed by [122]

$$\mathrm{ERBS}(f) = 21.4\,\log_{10}(1 + 0.00437 f). \tag{6.5}$$

Substituting eqs. (6.3) to (6.5) into (6.2), the frequency response of a fourth-order gammatone filter on the ERB-rate scale can be expressed by

$$\hat{G}(e) \approx [1 + j\,9.09\,(1 - 10^{\frac{e_0 - e}{21.4}})]^{-4}, \tag{6.6}$$

where e_0 corresponds to the center frequency f_0. Comparing (6.2) and (6.6), we find that (6.2) depends on f (b is a variable of f) and the distance $|f - f_0|$, while (6.6) only depends on the distance $|e - e_0|$. This indicates that the shape of the squared magnitude response of the gammatone filter changes on the Hz scale, but remains identical on the ERB-rate scale. Fig. 6.1 illustrates the squared magnitude responses of five gammatone filters.

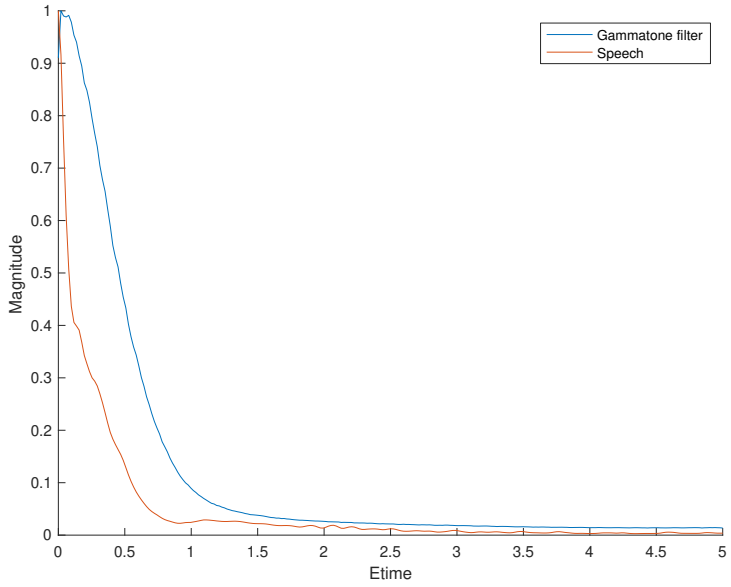


Figure 6.2: Time-width of speech and the squared-magnitude of a gammatone filter.

Since the shape of the gammatone filters remains unchanged on the ERB-rate scale, gammatone filters with different center frequencies have the same time-width. As long as the time-width of frequency-lowered speech is smaller than the time-width of the gammatone filter, frequency compression can increase mutual information of the received speech for person with a hearing impairment. As the frequency scale is not the usual Hz, we refer the time scale for the ERB rate to *Etime*. The time representation of the auditory spectra can be obtained via the inverse Fourier transform. Fig. 6.2 illustrates the time-width of speech and the gammatone filter. We can see that the time-width of clean speech is smaller than the time-width of the gammatone filter. Thus, moderate frequency compression can make inaudible high-frequency components audible, and does not lead to information loss for the already audible frequency components.

6.3 Mutual Information for Frequency-lowered Speech

As we use nonlinear frequency compression (NFC) to generate frequency-lowered speech, the frequency components below the cutoff frequency remain unchanged. This leads to an infinite mutual information between the clean speech and the frequency-lowered speech. In this section, we propose to use the pseudo message to calculate the mutual information. In addition, we estimate the maximal benefit of the NFC operator in terms of mutual information.

6.3.1 Mutual information calculation

In our study, we want to know if the automatic fitting tool can be used to find the optimal parameters of frequency lowering operators. In our proposed speech intelligibility predictor, the intelligibility is estimated as the mutual information between clean speech and degraded speech in the log-auditory domain. Production noise was omitted as it does not aid in performance. Thus, the proposed intelligibility predictor only measures the quality of the environmental channel that is between the clean speech and the degraded speech.

When searching for the optimal parameters of a frequency lowering operator, we need to consider the production noise as there is no environmental noise. Without production noise, the mutual information between the clean speech and the speech received by a hearing-impaired listener is infinite. Given a clean speech signal, we cannot obtain the discrete-valued message, since we do not have the codebook of speech. In Chapter 3, we found that when the message is assumed to be continuous-valued, we can generate a pseudo message, which has the same probability distribution as the true message. Although the pseudo message does not represent the true message, it is still able to generate the correct finite mutual informa-

tion between the pseudo message and the clean speech. Thus, we can generate the continuous-valued pseudo message and use (3.22) to calculate the mutual information between the message and the frequency-lowered speech.

6.3.2 Benefits of the NFC operator

The NFC operator can increase the mutual information between the message and the frequency-lowered speech. A natural question is that how much is the benefit of the NFC operator? To solve this problem, we need to know how much information is carried by the high-frequency bands. The aim of the NFC operator is to transmit high-frequency information through the low-frequency bands. Thus, the amount of information at high frequency is an upper bound on the benefits of the NFC operator.

Let us denote a K-dimensional message and clean speech by $M = [M_1, \cdots, M_K]^T$, and $X = [X_1, \cdots, X_K]^T$, respectively. Assume a hearing-impaired person can only hear signal components up to L dimension, where $L \leq K$. The information loss from high-frequency components is calculated by

$$I(M; X|X_1, \cdots, X_L) = I(M; X) - I(M; X_1, \cdots, X_L), \tag{6.7}$$

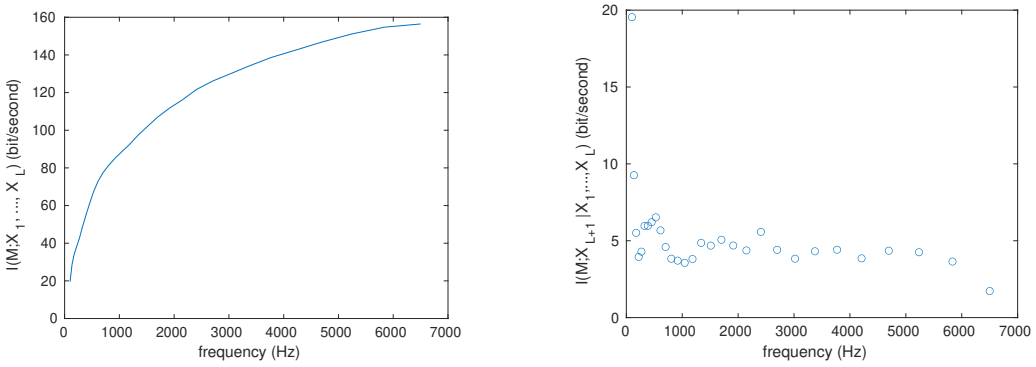
where $I(M; X)$ is the mutual information between the message and the clean speech, and $I(M; X_1, \cdots, X_L)$ is the mutual information between the message and the audible (low-frequency) speech. The information loss $I(M; X|X_1, \cdots, X_L)$ represents an upper bound on the benefit of the frequency lowering operator.

We used the CHAINS data set to calculate the upper bound of the benefit of the NFC operator. In the CHAINS data set, we randomly extracted 33 speech sentences. Each sentence was recorded by a different talker. Voice activity detection with the threshold of 30 dB was used to remove the silent segments in speech. The frequency range was from 100 Hz to 6500 Hz, which corresponded to 3.4 and 31.4 on the ERB-rate scale. Thirty

gammatone filters were used to generate the log-auditory spectrogram. In the received speech, we gradually increased the number of the gammatone filters. The pseudo message $\hat{M}$ was generated by using the smoothed correlation coefficients in Fig. 5.5. Fig. 6.3a illustrates the mutual information between the message and the clean speech for different audible bandwidths. The larger the audible bandwidth is, the more information a hearing-impaired person can receive. Fig. 6.3b illustrates the information gain for each band, which is calculated by

$$I(M; X_{L+1}|X_1, \cdots, X_L) = I(M; X_1, \cdots, X_{L+1}) - I(M; X_1, \cdots, X_L). \quad (6.8)$$

We can see that above 2 kHz, the information gain for each band is almost identical.



(a) Mutual information between the message and the audible speech components.

(b) Conditional mutual information for each band.

Figure 6.3: Mutual information over frequency.

6.3.3 Hearing instruments fitting based on mutual information

The goal of the hearing instruments fitting is to maximize the intelligibility of the received speech for hearing-impaired people. In our model, the

intelligibility is quantified by the mutual information. The Markov chain for the speech transmission can be expressed by $M \to S \to \tilde{S} \to \tilde{X}$, where $\tilde{S} = \mathcal{G}(S)$, with $\mathcal{G}$ denoting the NFC operator and $\tilde{X}$ being the received signal after the peripheral auditory processing of hearing-impaired people. The mutual information between the message and the received signal can be calculated as

$$I(M; \tilde{X}) = h(M) + h(\tilde{X}) - h(M, \tilde{X}), \tag{6.9}$$

where $h(\cdot)$ denotes differential entropy. Let θ denote the parameter of the NFC operator in hearing instruments. We choose θ such that

$$\theta^* = \arg\max_{\theta} I(M; \tilde{X}). \tag{6.10}$$

The log-auditory spectrogram is calculated as introduced in Section 3.3.1. The smoothed correlation coefficients in Fig. 5.5 are used to generate the pseudo message.

6.4 Sound Distinction Test

In this section, we design a listening test for frequency-lowered speech. As frequency-lowered speech is unfamiliar to the listener, time-consuming acclimatization is required for normal speech recognition listening test. To improve the efficiency of the listening test, we propose a sound distinction listening test which does not require the acclimatization process.

6.4.1 Concept of sound distinction test

In speech recognition, listeners are already familiar with the sounds of clean speech, which can be seen as templates stored in the brain of listeners. For an incoming sound, the listener compares it with the templates and selects the template that is closest to the incoming sound. As

frequency lowering operators change the pattern of clean speech, the listener has to acclimatize to the new templates (the frequency-lowered clean speech) in normal speech recognition test.

Unlike a speech recognition test, a sound distinction test does not require the listener to acclimatize to the new templates, since they are played back in the test. For an incoming frequency-lowered phoneme, the listener is presented with the templates that are frequency-lowered phonemes. One of these templates matches the incoming phoneme. We ask the listener to select the template that sounds most similar to the incoming phoneme.

In the listening test, we limit the number of the templates to two, because this makes it easier for the listener to respond. This is similar to the diagnostic rhyme test (DRT) [7], where a word recording is played back and the listener is asked to select the word from a given pair of two words. In the DRT, no recordings of the two words are played back, since the listener is assumed to be familiar with their pronunciations. In the sound distinction test, the recordings of the two templates are played back. To eliminate the cues from the talker in sound distinction, the incoming sound and the templates were recorded by different talkers.

6.4.2 Speech material

The speech material is in the form of /i/ + fricative. The fricative is one of the seven fricatives /ð, f, ʃ, s, θ, v, z/ [10], which have dominant components at high frequency. Their relative frequencies of occurrence in English are 2.83%, 1.75%, 0.54%, 4.59%, 0.6%, 1.95%, and 3.01%, respectively [3]. This amounts to 15.27% occurrence of all the phonemes. The speech material was recorded by three talkers. Each talker recorded each phoneme five times. Thus, in total there are $7 \times 3 \times 5 = 105$ sounds.

When comparing the incoming sound with the templates, the differences include spectral shape and phoneme length. Since we test the benefit of frequency lowering operators, the distinction of phonemes should

only be based on the spectral cue. Thus, we need to preprocess the speech material to ensure all speech sounds have the same length.

Table 6.1: Average duration of the phonemes [3].

Phonemes	ð	f	ʃ	s	θ	v	z
Length (ms)	119	122	118	129	119	78	85

This requires both sound /i/ and the phonemes have constant length for different phonemes. To realize that, we determine the starting point of the phoneme in each recording. Then, we calculated the lengths of /i/ and the phoneme, respectively. Based on the lengths of /i/ from all the recordings, we chose the minimal number, which was 0.19 seconds, as the new length of /i/. The remaining part of /i/ was removed in each recording. Similarly, we chose the maximal number from Table. 6.1, which was 129 milliseconds, as the new length of the phoneme. The remaining part of the phoneme was removed in each recording. If the original phoneme was shorter than the minimal number, this recording was removed from the speech material. After this preprocessing, three recordings of the phoneme /z/ and five recordings of the phoneme /ð/ were removed. The recordings of the other phonemes were all kept in the new speech material. The length of all the new speech material was 0.32 seconds. It should be noted that for each phoneme, each talker has more than one recording in the new speech material. Thus, we can always guarantee that the incoming sound and the template are from different talkers.

6.4.3 Classification of fricatives

In the sound distinction test, we ask the listener to distinguish the incoming sound and the templates. For some pairs of fricatives the spectra are very similar. For example, Fig. 6.4 illustrates the spectra of the pair /f/-/θ/. Even for the original recordings of the fricatives, the distinction between these pairs is not obvious. When evaluating the frequency lowering

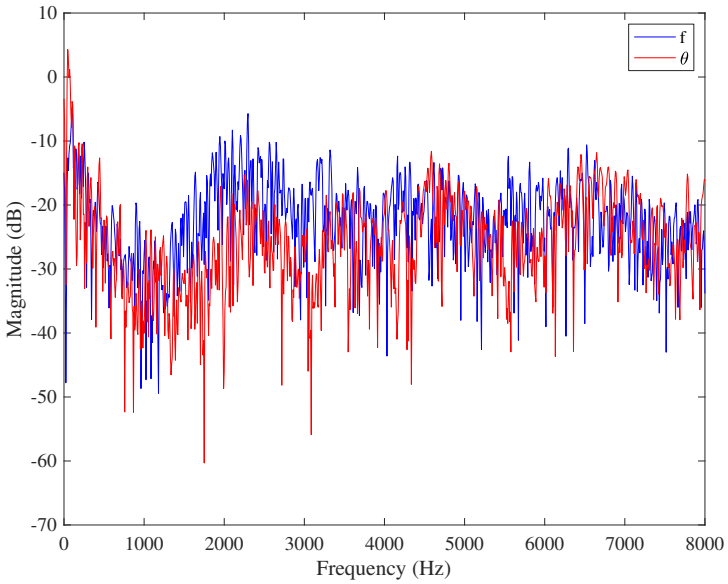
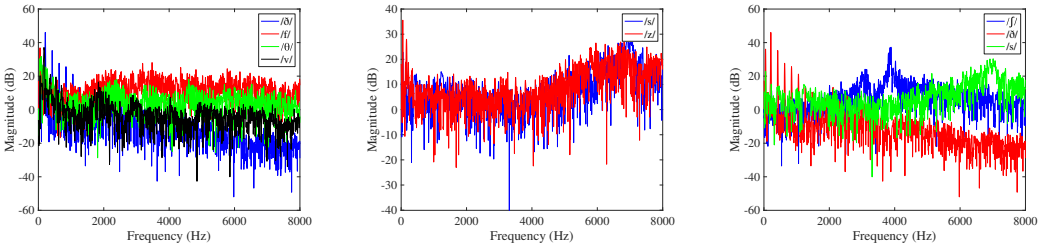


Figure 6.4: Spectra of /f/ and /θ/.

operators, the suitable speech material should be easily distinguished by normal-hearing listeners and hardly distinguished by hearing-impaired listeners. In other words, the pairs of fricatives should have clear difference at high frequency and similar spectrum at low frequency.

To generate suitable pairs of fricatives, we classify the seven fricatives into three types, which are shown in Fig. 6.5. Fig. 6.5a shows the spectra of the first type fricatives that include /ð/, /f/, /θ/, and /v/. Fig. 6.5b shows the spectra of the second type fricatives that include /s/ and /z/. Fig. 6.5c shows the spectrum of the third type fricative that includes only /ʃ/. The spectra of /ð/ from the first type and /s/ from the second type are also plotted for comparison. In Fig. 6.5, we can see the fricatives from the same type have similar spectra and the fricatives from different types have clearly different spectra.



(a) Type 1: /ð/, /f/, /θ/, and /v/.

(b) Type 2: /s/ and /z/.

(c) Type 3: /ʃ/. /ð/ and /s/ are also plotted for comparison.

Figure 6.5: Spectra of the seven fricatives.

Since the two templates should have clear difference at high frequency and share some similarity at low frequency, we can generate the pair of templates by selecting two fricatives from different types. A pair of templates can be from type 1 and type 2, or type 1 and type 3, or type 2 and type 3. When the pair is from type 1 and type 2, there are $C_4^1 \cdot C_2^1 = 8$ combinations. When the pair is from type 1 and type 3, there are $C_4^1 = 4$ combinations. When the pair is from type 2 and type 3, there are $C_2^1 = 2$ combinations. Thus, we can generate 14 eligible pairs of the templates, which is shown in Table 6.2.

Table 6.2: Template pairs for the sound distinction test.

No.	Template pair	No.	Template pair	No.	Template pair
1	ð - s	6	f - z	11	θ - ʃ
2	f - s	7	θ - z	12	v - ʃ
3	θ - s	8	v - z	13	s - ʃ
4	v - s	9	ð - ʃ	14	z - ʃ
5	ð - z	10	f - ʃ		

6.5 Evaluation

In this section, we first introduce the procedure of the sound distinction test. Then, we compare the subjective results from the listening tests with the objective results from the SIPs.

6.5.1 Listening test procedure

To carry out the listening test, we first need to determine the parameters of the NFC operator in (6.1). The cutoff frequency was chosen as 1.5 kHz, so we can keep the pitch cue and most vowel sounds unchanged. We assume hearing loss occurs above 2 kHz, so the hearing-impaired listener has difficulty in distinguishing fricatives, as shown in Fig. 6.5c. Different compression ratios were chosen in the listening test. As expressed in (6.1), the compression ratio is determined by the input frequency and the output frequency. We fixed the output frequency as 2 kHz and chose seven input frequencies which were 2 kHz, 3 kHz, 4 kHz, 5 kHz, 6 kHz, 7 kHz, and 8 kHz. Thus, seven different compression ratios were evaluated in the listening test. The input frequency of 2 kHz means no frequency compression, and the input frequency of 8 kHz means the extremest frequency compression. We assume a complete hearing loss above 2 kHz. Thus, the hearing loss was simulated by a low-pass filter with the cutoff frequency of 2 kHz. To reduce the low-frequency cues, which can be observed in Fig. 6.5c, in the sound distinction, we added a low-pass filtered speech-shaped noise (SSN) with the SNR of 0 dB on the clean speech material. The cutoff frequency was set to 2 kHz, so it did not mask the high frequency of the fricatives. Fig. 6.6 illustrates the generation procedure of a frequency-lowered speech received by the hearing-impaired listener.

Next, we need to determine the arrangement of the listening test. Specifically, we need to decide the content and the number of the speech material that are presented to the listeners. Recall that we have 14 pairs of templates. The incoming sound can be any fricative in one pairs. Thus,

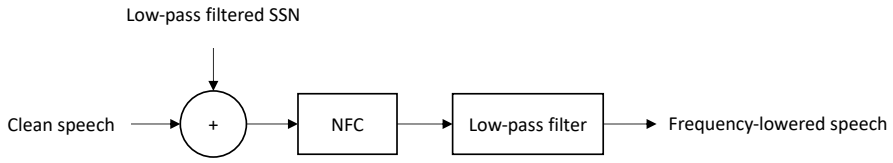


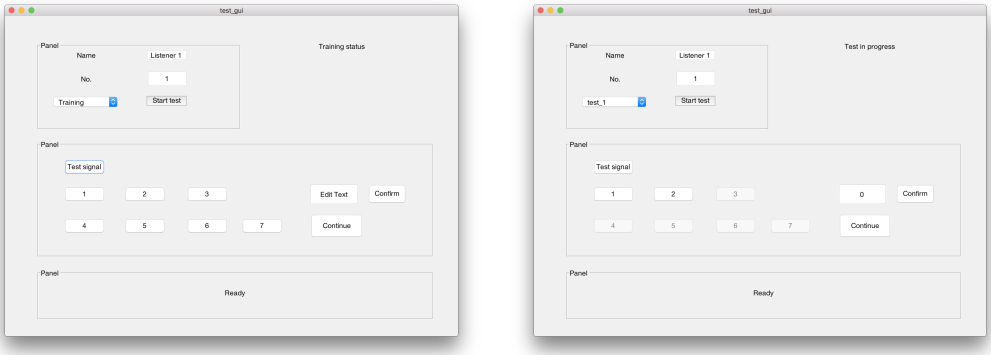
Figure 6.6: Diagram of the generation of a frequency-lowered speech signal.

we have 28 pairs of fricatives that need to be distinguished for one frequency lowering condition. Since seven compression ratios were chosen, there are $28 \times 7 = 196$ pairs of fricatives that need to be distinguished. This is the minimal amount of the pairs that are required to evaluate the performance of the NFC operator. Repeating the listening test can provide a more accurate result.

Carrying out 196 sound distinction tests is a laborious task for a single listener. To make the listening test doable, we divided the whole task into small sections and assigned them to eight listeners. So each listener only did a small part of the whole task. All of the listeners had normal-hearing ability and were paid for the participation. The listening test was approved by the Human Ethics Committee at Victoria university with the application number 0000025109. For each frequency lowering condition, each listener was asked to distinguish eight pairs of fricatives. The eight pairs were randomly selected from the whole set of pairs, which was 28 pairs. The total pairs from all the listeners should cover the whole set of pairs. Repetition is preferred, as more data can be collected. For example, four participants would do 32 pairs, which cover all the 28 pairs and have 4 repeated pairs. Seven participants would do 56 pairs, which repeat the whole set of pairs twice.

Before the actual listening test, the listeners were asked to go through a training session which aims to familiarize them with the speech material. In the training session, the clean speech signals were used. Similar to

the sound distinction test, the listeners were asked to match the incoming sound with a sound from the templates. However, the templates were all the seven fricatives, instead of two fricatives. Since each fricative can be the incoming sound, this sound distinction test was repeated seven times. We require the eligible listeners should have at least six successful matches in the training session. Seven participants achieved this goal, and one participant did not. So we only collect data from these seven eligible participants. In the test session, each listener went through the seven frequency lowering conditions. Thus, each listener did $8 \times 7 = 56$ sound distinction tests, and there were $56 \times 7 = 392$ sound distinction tests in total.



(a) Training session. (b) Test session.

Figure 6.7: The GUI for the listening test.

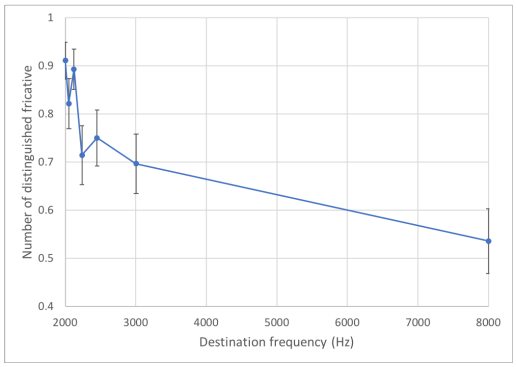
The listening test was carried out through GUI that was implemented via MATLAB. In the training session, as shown in Fig. 6.7a, the 'Test signal' is the incoming sound, which is one of the seven fricatives. The number 1-7 denote the templates. In the test session, the number of the templates is limited to two, as shown in Fig. 6.7b.

6.5.2 Subjective results

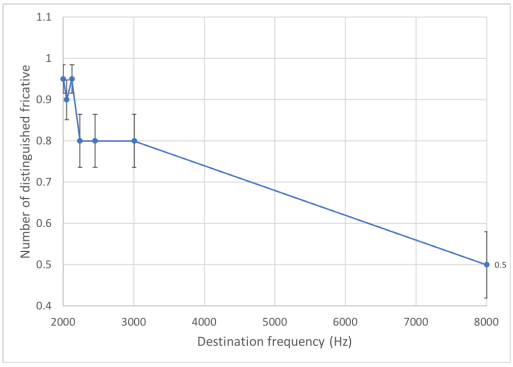
Although all seven participants went through the training session, we found that the performance of some participants was not very consistent, i.e., large fluctuation existed in their performance. Thus, we need to find the participants, whose performance was consistent. In the listening test, each listener did eight sound distinction tests for a specific frequency lowering condition. We can analyze the consistency of the performance by calculating the standard deviation of the eight data for each frequency lowering condition, and then averaging the standard deviations over all the seven frequency lowering conditions. Table 6.3 shows the standard deviation for each listener. We can see that participant 3 and participant 6 have large standard deviation in the listening test results. Thus, a more accurate listening test result can be obtained by removing the results from participant 3 and participant 6.

Table 6.3: Standard deviation of the results for each participant.

Participant No.	1	2	3	4	5	6	7	Mean
Standard deviation	0.23	0.21	0.50	0.38	0.31	0.48	0.33	0.35



(a) All the seven participants.



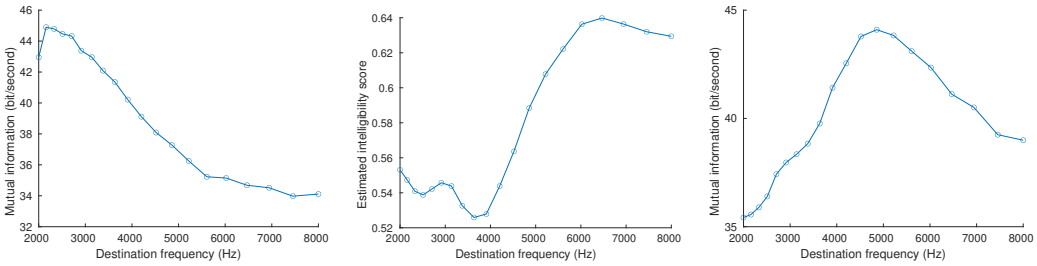
(b) Selected five participants.

Figure 6.8: Listening test results.

The listening test results from all the seven participants and the selected five participants are shown in Fig. 6.8. In Fig. 6.8a, each frequency lowering condition has $8 \times 7 = 56$ data. In Fig. 6.8b, each frequency lowering condition has $8 \times 5 = 40$ data. As introduced in Section 6.5.1, we chose seven frequency lowering conditions, where the signals from 2 kHz, 3 kHz, 4 kHz, 5 kHz, 6 kHz, 7 kHz, and 8 kHz were compressed to 2 kHz. This corresponds to compressing the signal at 8 kHz to 8 kHz, 3 kHz, 2.45 kHz, 2.24 kHz, 2.12 kHz, 2.05 kHz, and 2 kHz, which represent x-axis in Fig. 6.8. Both figures show that moderate frequency compression improves the sound distinction results. However, extreme frequency compression (destination frequencies below 2.12 kHz) does not provide further benefits. The standard error of the sample mean is plotted as error bar in both figures.

6.5.3 Objective results

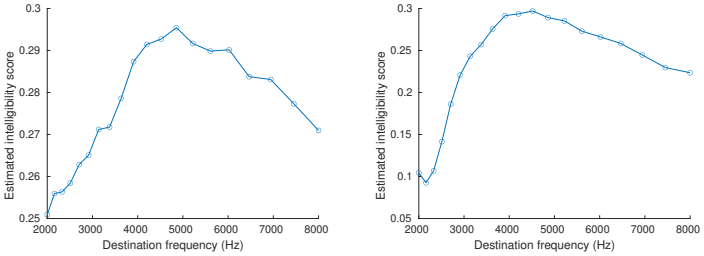
The objective result is the intelligibility estimated by SIPs. We evaluate the proposed fitting procedure by comparing the subjective result and the objective result. We again used the CHAINS data set to predict the speech intelligibility for different frequency lowering conditions. The generation of the speech material and the log-auditory spectrogram for the clean speech was the same as described in Section 6.3.2. The correlation coefficients between the transformed message and the transformed speech, as shown in Fig. 5.5, were used to generate the pseudo message. The frequency-lowered speech was generated through the diagram in Fig. 6.6, which was used in the listening test. In the simulation, we chose twenty frequency destinations which were uniformly distributed between 2 kHz and 8 kHz on the ERB-rate scale.



(a) Proposed intelligibility predictor for the NFC operator.

(b) wSTMI.

(c) $\text{SIIB}^{\text{Gauss}}$.



(d) ESTOI.

(e) HASPI.

Figure 6.9: Speech intelligibility estimated by SIPs.

The intelligibility of the frequency-lowered speech estimated by the proposed mutual information metric is shown in Fig. 6.9a. It shows the mutual information is maximized when the destination frequency is between 2 kHz and 2.3 kHz. This is consistent to the result of the sound distinction test in Fig. 6.8a, which shows the destination frequency for the highest distinction score is between 2 kHz and 2.24 kHz. By using (6.1), we can calculate the compression ratio for a given destination frequency. The compression ratios for the destination frequencies of 2 kHz, 2.24 kHz, and 2.3 kHz are 5.8, 4.2, and 3.9, respectively. Thus, the objective result shows that the optimal compression ratio is between 3.9 and 5.8, while the listening test result shows that the optimal compression ratio is between 4.2 and 5.8. This approximately matches the time-width of the squared-magnitude of the gammatone filter that is shown in Fig. 6.2. If we choose

the magnitude of 0.5 (-6 dB) as the threshold, the time-widths for speech and the gammatone filter are 0.09 and 0.45, respectively. This indicates that if the compression ratio is below $\frac{0.45}{0.09} = 5$, there is no intelligibility reduction, as the time-width of the compressed speech is still within the time-width of the ear.

Figures 6.9b to 6.9e show the intelligibility estimated by the other four state-of-the-art intelligibility predictors, which are wSTMI [51], SIIB$^{\text{Gauss}}$ [41], ESTOI [43], and HASPI [44]. The results from these four SIPs are not consistent to the result of the sound distinction test. The intelligibility estimated by wSTMI shows that the frequency destination should be between 6 kHz and 7 kHz. The intelligibility estimated by SIIB$^{\text{Gauss}}$, ESTOI, and HASPI shows that the frequency destination should be between 4.2 kHz and 5.2 kHz. SIIB$^{\text{Gauss}}$, ESTOI, and HASPI achieved similar results, because they use similar mechanism to predict speech intelligibility, i.e., the original spectro-temporal components are decomposed into mutually orthogonal subspaces. The intelligibility is then estimated based on the components in each subspace. Since SIIB$^{\text{Gauss}}$ calculates the mutual information for each transformed band individually, it applies KLT to remove correlation of signals at different frequency bands. Similarly, HASPI applies a set of cosine-based orthogonal functions to remove the correlation. ESTOI calculates the inner product of the two supervectors of the clean speech and the degraded speech. The supervectors include both spectral and temporal components. As inner product measures the orthogonality between two vectors, any additional orthogonal transform does not impact the result. However, each supervector can be viewed as a combination of basis vectors which are from mutually orthogonal subspaces. Thus, the intelligibility is based on the components projected onto the orthogonal subspaces.

6.6 Summary

We proposed a hearing instrument fitting method based on the mutual information-based intelligibility predictor. The goal of hearing instrument fitting is to find out the optimal parameter that maximizes the intelligibility of speech received by hearing-impaired people. Adjusting the parameter manually is time-consuming. We use the intelligibility predictor to improve the efficiency of the fitting process.

The NFC operator is one of frequency lowering operators that are used in some hearing instruments. We evaluated the proposed fitting method based on the NFC operator. Since the NFC operator does not modify the signal at low frequency, mutual information between the clean speech and the frequency-lowered speech is infinite. Thus, we used the speech transmission model based on the continuous-valued message to generate the pseudo message. The intelligibility of the frequency-lowered speech is estimated by the mutual information between the pseudo message and the frequency-lowered speech. Using the continuous-valued pseudo message makes it possible to estimate the intelligibility of the frequency-lowered speech, although the true message is discrete-valued.

To avoid the acclimatization in the speech recognition test for the frequency-lowered speech, we proposed a sound distinction test to replace speech recognition test. The logic of the sound distinction test is the same as the speech recognition test. In speech recognition test, the listener matches the recording of input speech with the templates stored in the brain. For frequency-lowered speech, the templates have not been learned by the listener. Thus, in the sound distinction test, we also present several recordings that are used as the templates. One of the templates matches the recording of input speech.

We evaluated the fitting method by comparing the result of the sound distinction test with the results of the objective SIPs. For the NFC operator, the cutoff frequency was 1.5 kHz. We needed to determine the

compression ratio which can be calculated through the bandwidth of the frequency-lowered speech. We assumed hearing loss occurs above 2 kHz. In the experiment, we chose different output bandwidths of the frequency-lowered speech (destination frequencies). The sound distinction test shows that the maximal intelligibility is achieved when the destination frequency is between 2 kHz and 2.24 kHz. This is consistent to the proposed mutual information-based metric that produces the maximal mutual information, when the destination frequency is between 2 kHz and 2.3 kHz. However, the subjective result is not consistent with the other four SIPs, which are wSTMI, $\text{SIIB}^{\text{Gauss}}$, ESTOI, and HASPI. For wSTMI, the maximal objective score is achieved between 6 kHz and 7 kHz. For $\text{SIIB}^{\text{Gauss}}$, ESTOI, and HASPI, the maximal objective score is achieved between 4.2 kHz and 5.2 kHz.

Finally, the optimal destination frequency range obtained from the sound distinction test matches the time-width of speech and the gammatone filter. The destination frequencies of 2 kHz and 2.24 kHz correspond to compression ratios of 4.2 and 5.8. Thus, the optimal compression ratio is between 4.2 and 5.8 for the NFC operator. Fig. 6.2 shows the time-width of the gammatone filter is approximately five times of the time-width of speech. This means theoretically the compressed speech does not lose any information, as long as the compression ratio is below five, which is within the optimal range of the compression ratio obtained from the sound distinction test.

Chapter 7

Conclusions and Future Work

The purpose of the objective SIPs is to replace the subjective listening test which is time-consuming and costly. The subjective listening test is required in the development of speech enhancement operators and the fitting of hearing instrument. Mutual information based SIPs estimate speech intelligibility by calculating the mutual information between the transmitted message and the received message. Different mutual information based SIPs use different representation forms of the message. There are three objectives for this thesis: (1) study the correct representation form of the message; (2) develop a mutual information based SIP according to the representation form of the message; (3) develop a fitting method for hearing instrument by using the mutual information based SIP. In this chapter, we will summarize the findings of these three objectives and discuss the possible directions in the future work.

7.1 Conclusions

7.1.1 Modeling of the Message

For the mutual information based SIPs, one important thing is to determine what signal is used as the transmitted message. Speech is contin-

uous sound, which includes the transmitted message and other information, such as talker information and environment information. SIIB and SIIB$^{\text{Gauss}}$ model the transmitted message as a continuous-valued variable, and the other information as additive noise. The received speech is considered as the received message. They achieved state-of-the-art prediction results among the mutual information based SIPs. However, the correlation coefficients between the transformed message and the transformed clean speech are considered constant across the transformed bands, which is not true.

In Chapter 3, we assume the transmitted message, the clean speech, and the received speech have Gaussian distribution, which is the same assumption as SIIB$^{\text{Gauss}}$. We improved SIIB$^{\text{Gauss}}$ by using more realistic correlation coefficients, which vary across the transformed bands. To evaluate the modified model, we compared the objective prediction results with the psychometric curves, which are subjective intelligibility results of four types of noisy speech signals over different SNRs for the acoustic channel. We found that the objective results did not match the subjective results well. The modified model has a much wider SNR interval between 0 intelligibility and 100% intelligibility than the psychometric curves.

We analyzed this phenomenon, and we think this mismatch is caused by the assumption that the transmitted message is continuous-valued. Note that for an additive white Gaussian noise (AWGN) channel and a fixed variance of the transmitted signal, the continuous-valued Gaussian distributed signal generates the maximal mutual information. Thus, the assumption of the continuous-valued message gives an over estimate of the mutual information for the production channel.

Chapter 4 studied the validity of the transmission model based on the discrete-valued message. The transmitted message is modeled as discrete-valued linguistic unit. There are three reasons for speech to have discrete-valued message. First, the discrete-valued message can make speech robust to acoustic noise. Speech is short-time stationary, which can be viewed

as a repetition code of the underlying message transmitted through an AWGN channel. The repetition coding makes it possible to reliably transmit the message from one end to the other end with any number of relays in between.

Second, due to the sphere packing, a high-dimensional discrete-valued message has its noisy realizations located on the sphere of a ball. If the noisy realization is closer to its corresponding message than to any other messages, the noisy realization can be correctly recognized. When the noise level just makes two spheres touch, increasing the noise a bit more will make the misrecognition suddenly happen. This is different to the continuous-valued message, which makes the probability of misrecognition gradually happen as the noise level increases. Thus, when the acoustic SNR increases from 0 intelligibility to 100% intelligibility (from misrecognition to recognition), the discrete-valued message produces a narrower SNR interval than the continuous-valued message.

The third reason why the message should be discrete-valued is that the discrete-valued message gives a lower estimation of the mutual information for the production channel than the continuous-valued message. This can also generate a narrower interval for the acoustic SNR, since the mutual information for the acoustic channel can quickly reach the same level as the mutual information for the acoustic channel. In the simulation, the mutual information for the production channel and the acoustic channel are calculated separately. The predicted intelligibility results based on the discrete-valued message matches the psychometric better than the predicted intelligibility results based on the continuous-valued message.

7.1.2 Speech Intelligibility Prediction based on Mutual Information

Based on the findings in Chapters 3 and 4, Chapter 5 proposed a new mutual information based SIP. As the transmitted message is discrete-valued,

we cannot generate a pseudo message from a clean speech signal. Thus, instead of calculating the mutual information between the message and the received speech, we calculate the mutual information between the clean speech and the received speech.

Since correlation exists in the frequency for both the clean speech and the degraded speech, the mutual information cannot be calculated as the sum of the mutual information for each band. The proposed SIP assumes they are jointly Gaussian distributed and use the relevant equation to calculate the mutual information. Unlike SIIB and SIIB$^{\text{Gauss}}$, the proposed SIP also takes into account the correlation in the frequency for the transformed received speech.

In addition, the proposed SIP improves the forward masking function. In SIIB and SIIB$^{\text{Gauss}}$, the minimal values in each band of the clean speech are used as the absolute hearing threshold. In the proposed SIP, the standard absolute hearing threshold is used. To convert the value of the log-auditory spectrogram from dB scale to dB SPL, we add the original log-auditory spectrogram by a constant, which is calculated as the level difference between the normal speech level (dB SPL) and the log-auditory spectrogram in 1 kHz.

The proposed SIP was evaluated based on four data sets, which are Kjems AN, Kjems ITFS, NELE Cooke and HuPost. Kendall rank correlation coefficient and Pearson correlation coefficient were used as the performance criteria. The proposed SIP shows higher correlation coefficients than SIIB$^{\text{Gauss}}$.

7.1.3 Mutual Information based Hearing Instrument Fitting

Chapter 6 proposed an automatic fitting method for hearing instruments. Hearing instrument fitting is necessary, since different hearing-impaired persons have different hearing losses. It is carried out by audiologists, who

adjust the parameters of hearing instruments to maximize the intelligibility of processed speech. Hearing instrument fitting is time-consuming, since listening tests are required. To solve this problem, the automatic fitting method uses an objective SIP to adjust the parameters.

Since hearing impairment usually starts at high frequency, our study uses the NFC operator, which is a frequency lowering operator in hearing instruments. The NFC operator can improve speech intelligibility, because the frequency resolution capacity of the human ear is above the frequency resolution of speech. Thus, the frequency-compressed speech can still be resolved by the human ear. The question is to what extent we should compress speech in frequency, such that the intelligibility is maximized. The automatic fitting method finds the parameters of the NFC operator by maximizing the mutual information between the transmitted message and the received speech. The NFC operator has two parameters, which are the cutoff frequency and the compression ratio. The signal below the cutoff frequency remains the same. We cannot use the proposed SIP in Chapter 5, since the unchanged low-frequency component makes the mutual information between the clean speech and the frequency-lowered speech infinite. To solve this problem, we use the modeling for the continuous-valued message and generate the pseudo message as the transmitted message. The mutual information is calculated between the pseudo message and the received speech.

The automatic fitting method was evaluated by comparing the predicted speech intelligibility, which is from the mutual information based SIP, with the listening test results. We proposed a sound distinction test for frequency-lowered speech to improve the test efficiency. We assume a hearing-impaired person cannot hear any signal above 2 kHz. The cutoff frequency of the NFC operator was 1.5 kHz. We compressed the frequency range [1.5 kHz 8 kHz] with different compression ratios. The objective metric shows moving the signal frequency from 8 kHz to somewhere between 2 kHz and 2.3 kHz maximizes speech intelligibility, which is consis-

tent to the listening test result that shows the frequency of 8 kHz should be moved to somewhere between 2 kHz and 2.24 kHz. The results from the other SIPs do not match the listening test result.

One remaining question is in what scenario we should use the pseudo message for the mutual information based SIP. In Chapter 5 the proposed SIP does not use the pseudo message, while in Chapter 6 the proposed SIP does. Using the pseudo message in Chapter 6 is to avoid infinite mutual information between the clean speech and the frequency-lowered speech. When using the pseudo message in the proposed SIP in Chapter 5, the intelligibility prediction result gets a bit worse than $\text{SIIB}^{\text{Gauss}}$. Note that the pseudo message is based on the modeling of the continuous-valued message, which is not the true modeling of the message. Thus, the pseudo message is necessary in the scenario where the mutual information between the clean speech and the received speech is infinite.

7.2 Future Work

All the SIPs use the knowledge from the human auditory system. A good understanding of the human auditory system helps in modeling the representation of the received speech. Human auditory system is complex. The question is how deep we should go to build a satisfied SIP, given the fact we are still exploring the mechanism of signal processing in the brain.

It seems data-driven SIPs can also produce good intelligibility prediction without going deep to the signal processing mechanism in the brain. Thus, in the future work, we can use the knowledge from the data-driven SIPs to improve the current mutual information based SIP. For example, we can calculate the mutual information between the message and the received speech for each critical band, and assign trained weights for each band to predict speech intelligibility.

The other future work relates to the development of a new frequency lowering operator. In the NFC operator, we need to decide the start fre-

quency. If we have a high start frequency, we have a good recognition of vowel sounds. However, less space is kept for high-frequency fricatives. Currently, we have frequency lowering operators that are based on linear/nonlinear frequency compression and frequency transposition on the Hz scale. A nonlinear frequency transposition on the Hz scale is missing. In speech, the sound at one instant can be either vowel-like sound or fricative-like sound. It cannot be both at the same time. Thus, a new frequency lowering operator based on nonlinear frequency transposition on the Hz scale (linear on the ERB-rate scale) would be worth considering.

Bibliography

[1] J. A. Galster, S. Valentine, J. A. Dundas, and K. Fitz, "Spectral iQ: Audibly improving access to high-frequency sounds," *Eden Prairie, MN: White paper, Starkey Laboratories Inc*, 2011.

[2] F. R. Moore, *Elements of computer music*. Prentice-Hall, Inc., 1990.

[3] H. T. Edwards and A. L. Gregg, *Applied phonetics*. Singular Publishing Group, 1992.

[4] K. H. Arehart, P. Souza, R. Baca, and J. M. Kates, "Working memory, age and hearing loss: Susceptibility to hearing aid distortion," *Ear and hearing*, vol. 34, no. 3, p. 251, 2013.

[5] W. H. Organization, "Deafness and hearing loss," 2015.

[6] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[7] W. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[8] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[9] J. M. Alexander, "20Q: Frequency lowering ten years later-new technology innovations," *AudiologyOnline (September, Article 18040)*, 2016.

[10] J. M. Alexander, "Nonlinear frequency compression: Influence of start frequency and input bandwidth on consonant and vowel recognition," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 938–957, 2016.

[11] J. Blauert, *Communication acoustics*, vol. 2. Springer, 2005.

[12] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An intelligibility metric based on a simple model of speech communication," in *International Workshop on Acoustic Signal Enhancement (IWAENC' 16)*, September 2016.

[13] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018.

[14] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *Signal Processing Magazine, IEEE*, vol. 32, no. 2, pp. 43–54, 2015.

[15] A. D. Weston and L. D. Shriberg, "Contextual and linguistic correlates of intelligibility in children with developmental phonological disorders," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 6, pp. 1316–1332, 1992.

[16] M.-Y. Jung, "The intelligibility and comprehensibility of world englishes to non-native speakers," *Journal of Pan-Pacific Association of Applied Linguistics*, vol. 14, no. 2, pp. 141–163, 2010.

[17] G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney, "Monosyllabic word recognition at higher-than-normal speech and noise levels," *The Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2431–2444, 1999.

[18] H. Fletcher, "The nature of speech and its interpretation," *The Bell System Technical Journal*, vol. 1, no. 1, pp. 129–144, 1922.

[19] H. Fletcher and R. H. Galt, "The perception of speech and its relation to telephony," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 89–151, 1950.

[20] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[21] V. Summers and M. T. Cord, "Intelligibility of speech in noise at high presentation levels: Effects of hearing loss and frequency regiona)," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1130–1137, 2007.

[22] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[23] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[24] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Speech reinforcement in noisy reverberant conditions under an approximation of the short-time SII," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4400–4404, IEEE, 2015.

[25] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[26] H. Puder, "Compensation of hearing impairment with hearing aids: Current solutions and trends," in *Voice Communication (SprachKommunikation), 2008 ITG Conference on*, pp. 1–4, VDE, 2008.

[27] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[28] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *Journal of speech and hearing research*, vol. 14, no. 4, pp. 677–709, 1971.

[29] S. A. Zollinger and H. Brumm, "The lombard effect," *Current Biology*, vol. 21, no. 16, pp. R614–R615, 2011.

[30] H. Brumm and S. A. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11-13, pp. 1173–1198, 2011.

[31] B. Sauert, *Near-End Listening Enhancement: Theory and Application*. Mainz, G, 2014.

[32] R. C. Hendriks, J. a. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 5, pp. 851–862, 2015.

[33] J. M. Kates, "Signal processing for hearing aids," in *Applications of Digital Signal Processing to Audio and Acoustics*, pp. 235–277, Springer, 2002.

[34] J. M. Kates, "Principles of digital dynamic-range compression," *Trends in Amplification*, vol. 9, no. 2, pp. 45–76, 2005.

[35] P. E. Souza, "Effects of compression on speech acoustics, intelligibility, and sound quality," *Trends in Amplification*, vol. 6, no. 4, pp. 131–165, 2002.

[36] D. Byrne, H. Dillon, T. Ching, R. Katsch, and G. Keidser, "Nal-nl1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures," *JOURNAL-AMERICAN ACADEMY OF AUDIOLOGY*, vol. 12, no. 1, pp. 37–51, 2001.

[37] L. E. Cornelisse, R. C. Seewald, and D. G. Jamieson, "The input/output formula: A theoretical approach to the fitting of personal amplification devices," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1854–1864, 1995.

[38] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery, "The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers," *International Journal of Audiology*, vol. 45, no. 10, pp. 563–579, 2006.

[39] A. H. Andersen, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

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

[41] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, vol. 26, no. 11, pp. 2153–2166, 2018.

[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[43] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[44] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[45] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The journal of the acoustical society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[46] S. Jørgensen and T. Dau, "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*, vol. 130, no. 3, pp. 1475–1487, 2011.

[47] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 2013.

[48] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, IEEE, 2020.

[49] M. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks," in *Interspeech 2020*, pp. 1151–1155, 2020.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[51] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speech intelligibility prediction using spectro-temporal modulation analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 210–225, 2020.

[52] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[53] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *The journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[54] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing*, pp. 313–319, Springer, 2005.

[55] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[56] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[57] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech.," in *ICASSP*, pp. 4214–4217, IEEE, 2010.

[58] A. Simpson, A. A. Hersbach, and H. J. McDermott, "Improvements in speech perception with an experimental nonlinear frequency compression hearing device," *International journal of audiology*, vol. 44, no. 5, pp. 281–292, 2005.

[59] W. E. Davis, "Proportional frequency compression in hearing instruments," *Hearing Review*, vol. 8, no. 2, pp. 34–43, 2001.

[60] J. M. Kates, "On using coherence to measure distortion in hearing aids," *The Journal of the Acoustical Society of America*, vol. 91, no. 4, pp. 2236–2244, 1992.

[61] G. C. Carter, C. H. Knapp, and A. H. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 4, pp. 337–344, 1973.

[62] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

[63] T. Houtgast and H. J. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica united with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[64] T. Houtgast and H. J. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[65] T. Houtgast, H. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function," *Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[66] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[67] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kay, "Extracting dynamic parameters from speech movement data," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1580–1588, 1993.

[68] S. Greenberg and B. E. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 3, pp. 1647–1650, IEEE, 1997.

[69] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS comput biol*, vol. 5, no. 3, p. e1000302, 2009.

[70] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[71] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech.," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[72] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information.," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 2, pp. 430–440, 2014.

[73] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing.," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 1, pp. 6–16, 2014.

[74] J. Taghia, R. Martin, J. Taghia, and A. Leijon, "An investigation on mutual information for the linear predictive system and the extrapolation of speech signals," in *Speech Communication; 10. ITG Symposium; Proceedings of*, pp. 1–4, VDE, 2012.

[75] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–3997, 2006.

[76] P. Lieberman, *The biology and evolution of language*. Harvard University Press, 1984.

[77] C. E. Shannon and W. Weaver, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, July, October 1948.

[78] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[79] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[80] K. K. Paliwal and L. D. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech communication*, vol. 45, no. 2, pp. 153–170, 2005.

[81] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term fourier spectrum for speech intelligibility," *The journal of the acoustical society of america*, vol. 127, no. 3, pp. 1432–1439, 2010.

[82] J. D. Robinson, T. Baer, and B. C. Moore, "Using transposition to improve consonant discrimination and detection for listeners

with severe high-frequency hearing loss: La utilización de la transposición para mejorar la discriminación consonántica y la detección en oyentes con hipoacusia severa para frecuencias agudas," *International Journal of Audiology*, vol. 46, no. 6, pp. 293–308, 2007.

[83] J. D. Robinson, T. H. Stainsby, T. Baer, and B. C. Moore, "Evaluation of a frequency transposition algorithm using wearable hearing aids," *International Journal of Audiology*, vol. 48, no. 6, pp. 384–393, 2009.

[84] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[85] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, 2015.

[86] G. Shi, M. M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1867–1874, 2006.

[87] K. Paliwal, B. Schwerin, and K. Wojcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Communication*, vol. 53, no. 3, pp. 327–339, 2011.

[88] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[89] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[90] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[91] D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel.," *Psychological review*, vol. 122, no. 2, p. 148, 2015.

[92] K. Beijering, C. Gooskens, and W. Heeringa, "Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm," *Linguistics in the Netherlands*, vol. 25, no. 1, pp. 13–24, 2008.

[93] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers," in *Proc of SPECOM*, vol. 6, pp. 431–435, 2006.

[94] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[95] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise," *International journal of audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[96] R. J. Muirhead, *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons, 2009.

[97] N. Misra, H. Singh, and E. Demchuk, "Estimation of the entropy of a multivariate normal distribution," *Journal of multivariate analysis*, vol. 92, no. 2, pp. 324–342, 2005.

[98] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, "A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula," *Human brain mapping*, vol. 38, no. 3, pp. 1541–1573, 2017.

[99] H. Goya, J. Cai, Q. Ding, and A. Fecher, "Development of vocabulary use in ESL composition," *INTESOL Journal*, vol. 8, no. 1, 2011.

[100] C. Coupé, Y. M. Oh, D. Dediu, and F. Pellegrino, "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche," *Science advances*, vol. 5, no. 9, p. eaaw2594, 2019.

[101] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "On the information rate of speech communication," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 5625–5629, IEEE, 2017.

[102] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials.," *Journal of experimental psychology*, vol. 41, no. 5, p. 329, 1951.

[103] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded cid w-22 word lists," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 2, pp. 427–438, 1991.

[104] A. M. Amlani, J. L. Punch, and T. Y. Ching, "Methods and applications of the audibility index in hearing aid selection and fitting," *Trends in amplification*, vol. 6, no. 3, pp. 81–129, 2002.

[105] D. Havelock, S. Kuwano, and M. Vorländer, *Handbook of signal processing in acoustics*. Springer Science & Business Media, 2008.

[106] H. Hermansky, "Coding and decoding of messages in human speech communication: Implications for machine recognition of speech," *Speech Communication*, vol. 106, pp. 112–117, 2019.

[107] H. H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, "Relevance of time–frequency features for phonetic and speaker-channel classification," *Speech communication*, vol. 31, no. 1, pp. 35–50, 2000.

[108] H. Hermansky, "Should recognizers have ears?," *Speech communication*, vol. 25, no. 1, pp. 3–27, 1998.

[109] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 6–16, 2014.

[110] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 430–440, 2014.

[111] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge.," in *Interspeech*, pp. 3552–3556, 2013.

[112] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[113] I. Rec, "P. 48, specification for an intermediate reference system," *International Telecommunication Union, CH-Geneva*, 1988.

[114] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[115] M. Ross, "The audiogram: Explanation and significance," *Hearing Loss Association of America*, vol. 25, no. 3, pp. 29–33, 2004.

[116] A. Agresti, *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons, 2010.

[117] W. C. Navidi, *Statistics for engineers and scientists*. McGraw-Hill Higher Education New York, NY, USA:, 2008.

[118] S. Scollie, D. Glista, and F. Richert, "Frequency lowering hearing aids: Procedures for assessing candidacy and fine tuning," A Sound Foundation Through Early Amplification conference. Chicago, Illinois, 2013.

[119] A. Simpson, "Frequency-lowering devices for managing high-frequency hearing loss: A review," *Trends in amplification*, vol. 13, no. 2, pp. 87–106, 2009.

[120] A. Simpson, A. Bond, M. Loeliger, and S. Clarke, "Speech intelligibility benefits of frequency-lowering algorithms in adult hearing aid users: a systematic review and meta-analysis," *International journal of audiology*, pp. 1–13, 2017.

[121] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[122] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.