Speech Communication from an Information Theoretical Perspective

by

Steven Van Kuyk

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
2019

Abstract

Throughout the last century, models of human speech communication have been proposed by linguists, psychologists, and engineers. Advancements have been made, but a theory of human speech communication that is both comprehensive and quantitative is yet to emerge. This thesis hypothesises that a branch of mathematics known as *information theory* holds the answer to a more complete theory. Information theory has made fundamental contributions to wireless communications, computer science, statistical inference, cryptography, thermodynamics, and biology. There is no reason that information theory cannot be applied to human speech communication, but thus far, a relatively small effort has been made to do so.

The goal of this research was to develop a quantitative model of speech communication that is consistent with our knowledge of linguistics and that is accurate enough to predict the intelligibility of speech signals. Specifically, this thesis focuses on the following research questions: 1) how does the acoustic information rate of speech compare to the lexical information rate of speech? 2) How can information theory be used to predict the intelligibility of speech-based communication systems? 3) How well do competing models of speech communication predict intelligibility?

To answer the first research question, novel approaches for estimating the information rate of speech communication are proposed. Unlike existing approaches, the methods proposed in this thesis rely on having a *chorus* of speech signals where each signal in the chorus contains the same linguistic message, but is spoken by a different talker. The advantage of this approach is that variability inherent in the production of speech can

be accounted for. The approach gives an estimate of about 180 b/s. This is three times larger than estimates based on lexical models, but it is an order of magnitude smaller than previous estimates that rely on acoustic signals.

To answer the second research question, a novel instrumental intelligibility metric called *speech intelligibility in bits* (SIIB) and a variant called SIIB$^{\text{Gauss}}$ are proposed. SIIB is an estimate of the amount of information shared between a talker and a listener in bits per second. Unlike existing intelligibility metrics that are based on information theory, SIIB accounts for *talker variability* and statistical dependencies between time-frequency units.

Finally, to answer the third research question, a comprehensive evaluation of intrusive intelligibility metrics is provided. The results show that SIIB and SIIB$^{\text{Gauss}}$ have state-of-the-art performance, that intelligibility metrics tend to perform poorly on data sets that were not used during their development, and show the advantage of reducing statistical dependencies between input features.

Acknowledgements

First and foremost, I thank Bastiaan Kleijn for his patience, guidance, supervision, and the many excellent opportunities that he provided to me throughout my studies. The majority of what I know about engineering I have learned from Bastiaan, and for that I will always be grateful towards him. His impact on the research in this thesis cannot be understated.

Second, I thank Richard Hendriks for his supervision. In particular, for hosting me at Delft University of Technology, for assisting with data collection, and for reading and critiquing the many revisions of my work.

I would also like to thank the following researchers for providing intelligibility data and MATLAB implementations of their intelligibility metrics: Asger Andersen, Fei Chen, Martin Cooke, Jesper Jensen, James Kates, Helia Relano-Iborra, João Santos, and Yan Tang. And also Ulrik Kjems, Kuldip Paliwal, Jalal Taghia, and Cees Taal, for making their materials publicly available. Without their generosity, Chapter 5 of this thesis would not have been possible.

I also thank Artemy Kolchinsky and Brendan Tracey from the Sante Fe Institute for collaborating with me and teaching me much about machine learning and information theory.

I thank the CoreAudio group at Apple Inc. for an engaging internship where I learned about the practical aspects of engineering.

Furthermore, I thank the Communication and Signal Processing (CaSP) group at Victoria University of Wellington and also the Victoria University of Wellington machine learning group, Festival of Doubt (FoD), for all the

useful presentations and discussions over the last three and a half years.

Finally, I thank my parents and Amber Kale for their continual love and support.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Articulation Index
ASR	Automatic Speech Recognition
BIF	Band Importance Function
bit	Binary Digit
CDP	Christiansen-Pedersen-Dau Metric
CVC	Consonant Vowel Consonant Syllable
CSII	Coherence Speech Intelligibility Index
DCT	Discrete Cosine Transform
ERB	Equivalent Rectangular Bandwidth
ESII	Extended Speech Intelligibility Index
ESTOI	Extended Short-Time Objective Intelligibility
GMM	Gaussian Mixture Model
HASPI	Hearing-Aid Speech Perception Index
HEGP	High Energy Glimpse Proportion
KLT	Karhunen-Loève Transform
KNN	K-Nearest Neighbour
GP	Glimpse Proportion
HSR	Human Speech Recognition
ITFS	Ideal Time-Frequency Segregation
MIKNN	KNN Mutual Information intelligibility measure
MMD	Maximum Mean Discrepency
MVDR	Minimum Variance Distortionless Response
NCM	Normalised Covariance Measure

PSD	Power Spectral Density
QSTI	Quasi-stationary Speech Transmission Index
SDR	Signal-to-Distortion Ratio
SEPSM	Speech-based Envelope Power Spectrum Model
SII	Speech Intelligibility Index
SIIB	Speech Intelligibility In Bits
SIMI	Speech Intelligibility predictor based on Mutual Information
SNR	Signal-to-Noise Ratio
SSN	Speech-Shaped Noise
STFT	Short-Time Fourier Transform
STI	Speech Transmission Index
STOI	Short-Time Objective Intelligibility Index
TFSS	Temporal Fine-Structure Spectrum based Index

Notation

Numbers and Sets

a or A	A real scalar
j	The imaginary unit (unless stated otherwise)
π	Pi
$\boldsymbol{a}$	A vector
$\mathbf{0}$	A vector of zeros
$\boldsymbol{A}$	A matrix
$\boldsymbol{I}$	Identity matrix with dimensionality implied
$\mathrm{diag}(\boldsymbol{a})$	A square diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\{\mathrm{a}_t\}$	A scalar stochastic process indexed by t
$\{\mathbf{a}_t\}$	A vector-valued stochastic process indexed by t
$\mathbb{A}$	A set
$\lvert\mathbb{A}\rvert$	The number of elements in the set $\mathbb{A}$
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^a$	The set of a-dimensional real vectors
$\mathbb{R}^{a\times b}$	The set of real-matrices with a rows and b columns
$\mathbb{C}$	The set of complex numbers
$\mathbb{C}^a$	The set of a-dimensional complex vectors
$\mathbb{Z}$	The set of integers
$\mathbb{E}_{++}$	The set of strictly positive even numbers
$\{a, b, c\}$	A set consisting of elements a, b, and c
$a \in \mathbb{A}$	a is a member of the set $\mathbb{A}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$
$A_{i,k}$	Element corresponding to the i'th row and k'th column of the matrix $\boldsymbol{A}$
a_i	Element i of the vector-valued random variable $\mathbf{a}$
$\mathbf{a}_t$	The t'th vector of the vector-valued stochastic process $\{\mathbf{a}_t\}$
$\mathrm{a}_{i,t}$	Element i of the vector-valued stochastic process $\{\mathbf{a}_t\}$ at time t
$\{\mathbf{a}_t^{[i]}\}$	A vector-valued stochastic process from the i'th source

Matrix and Vector Operations

$\boldsymbol{a}^*$	Transpose of the vector $\boldsymbol{a}$
$\boldsymbol{A}^*$	Transpose of the matrix $\boldsymbol{A}$
$\det(\boldsymbol{A})$	Determinant of the matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	Inverse of the matrix $\boldsymbol{A}$
$\|\boldsymbol{a}\|_2$	L2-norm of the vector $\boldsymbol{a}$

Sums and Integrals

$\sum_a$	sum over the domain of a
$\sum_{a \in \mathbb{A}}$	sum over the set $\mathbb{A}$
$\sum_{a=b}^{c}$	sum from $a = b$ to $a = c$ in steps of one

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a variable
$P(a)$	A probability distribution evaluated at $\mathrm{a} = a$
$\mathcal{N}$	The Gaussian distribution
$\mathcal{U}$	The uniform distribution
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}[\mathrm{a}]$	Expected value of the random variable a
$\mathrm{var}(\mathrm{a})$	The variance of the random variable a
$\mathbb{E}_{\mathrm{a}\sim P}[f(\mathrm{a})]$	Expected value of $f(a)$ with respect to $P(\mathrm{a})$
$H(a)$	Shannon information content of the outcome a
$H(\mathrm{a})$	Shannon entropy of the random variable a
$H(\{\mathrm{a}_t\})$	Entropy rate of the stochastic process $\{\mathrm{a}_t\}$
$h(\mathrm{a})$	Differential entropy of the random variable a
$I(\mathrm{a};\mathrm{b})$	Mutual information between the random variables a and b
$I(\{\mathrm{a}_t\};\{\mathrm{b}_t\})$	Mutual information rate between the stochastic processes $\{\mathrm{a}_t\}$ and $\{\mathrm{b}_t\}$
$\mathrm{a} \to \mathrm{b} \to \mathrm{c}$	A Markov chain

Functions

$\ln a$	Natural logarithm of a
$\log_b a$	Logarithm of a to base b
$\min(a, b)$	The minimum of a and b
$\max(a, b)$	The maximum of a and b
$\inf$	The infimum
$\sup$	The supremum
$\mathrm{sign}(a)$	The sign function ($\mathrm{sign}(a) = 1$ if $a > 0$ and $\mathrm{sign}(a) = -1$ otherwise.)
$\lfloor a \rfloor$	The floor function

Chapter 1

Introduction

Over the past 150 years several speech-based technologies have emerged that make possible what our ancestors believed impossible. Mobile phones are used to communicate from one side of the planet to the other, hearing-aids and cochlear implants restore hearing to the deaf, and automatic speech recognition (ASR) is increasingly used to instruct machines. Undoubtedly, speech-based technologies have contributed to the shaping of modern society and will continue to do so in the foreseeable future.

Because speech-based technologies are well established, it is tempting to think that there is little improvement to be made to their performance. However, in reality a range of challenges continue to persist. Mobile phone users have difficulty communicating in noisy environments such as subways and windy locations. Hearing-aid and cochlear implant users score worse at intelligibility listening tests than listeners with normal hearing. And when ASR is compared to human speech recognition (HSR), the performance of ASR is far worse in the presence of noise and reverberation.

This thesis takes the point of view that such limitations are the outcome of a poor understanding of human speech communication. What is missing is a model of speech communication that is derived from first-principles. Information theory may hold an answer.

1.1 Problem Statement

Key to the success of speech-based technology is an understanding of human speech communication. Broadly speaking, there are two approaches to understanding speech communication.

The first approach involves the analysis of acoustic speech signals and is typically referred to as *speech processing*. In speech processing, mathematical models that describe the characteristics, production, and transmission of acoustic speech signals are developed.

The second approach to understanding speech communication involves the analysis of human language and is referred to as *linguistics*. Linguistics focuses on how languages evolve and how humans assign perceptual meaning to symbols; both written and verbal. Typically, linguists break language into discrete lexical units such as sentences, words, letters, and phonemes, and analyse the syntactic and grammatical structure of sequences composed of such units.

Both perspectives of speech communication are valuable, but are often disjointed. The results and theories developed by speech processing engineers are highly specialised and can be misunderstood or unknown to linguists. Similarly, theories developed by linguists can be unknown to engineers and not obviously applicable to the design of the engineers' communication systems. Advancements have been made, but a unified theory of human speech communication that is both comprehensive and quantitative is yet to emerge.

Three problems can be identified when developing a model of speech communication. The first problem is that the basic units of speech and their relation to acoustic signals remain undiscovered. Speech is physically realised as a continuous acoustic signal but is perceived by a listener as a discrete sequence of lexical units. The traditional assumption in speech processing has been that if the right way of viewing acoustic speech signals was found, then a one-to-one relationship between acoustic cues and phonemes would become apparent (Denes, 1963). Evidence against this assumption was given by Liberman et al. (1957) who showed that the same segment of an acoustic speech signal can be perceived in different ways depending on preceding and succeeding sounds. In light of such experiments, there has been a wider acceptance of the view that a single acoustic cue carries information about successive linguistic units (e.g., Liberman et al., 1967; Nygaard and Pisoni, 1995; Denes, 1963). This concept is prevalent in ASR systems, which rely on sequential models such as *hidden Markov models* and *recurrent neural networks* to map acoustic signals to sentences (e.g., Rabiner, 1989; Graves et al., 2013).

The second problem to developing a model of speech communication is that there is variability inherent in the production of speech. The two main sources of talker variability are physiological differences (e.g., vocal tract length) and learned speech habits (e.g., foreign accents) (Huang et al., 2001). The consequence of talker variability is that two acoustic signals that contain the same lexical information (e.g., a sentence) can be different to one another. The problem of talker variability has received attention from the linguistic and ASR research communities, but it is often ignored by the speech enhancement and speech intelligibility research communities.

Finally, the third problem to consider is that speech communication is incredibly robust to signal distortions. Examples of distortions include additive environmental noise, interference from other talkers, and reverberation. An extreme example of a non-linear distortion is shown in Figure 1.1, which displays a short segment of an acoustic speech signal x_t as a function of time, where t is the time index, and a distorted signal $\mathrm{sign}(x_t)$, which is equal to 1 when $x_t > 0$ and equal to -1 otherwise. When $\mathrm{sign}(x_t)$ is played on a loudspeaker, the speech has poor quality, but surprisingly it is still intelligible.

Many believe that the robustness of speech communication is due to language context effects. While context certainly plays a significant role,

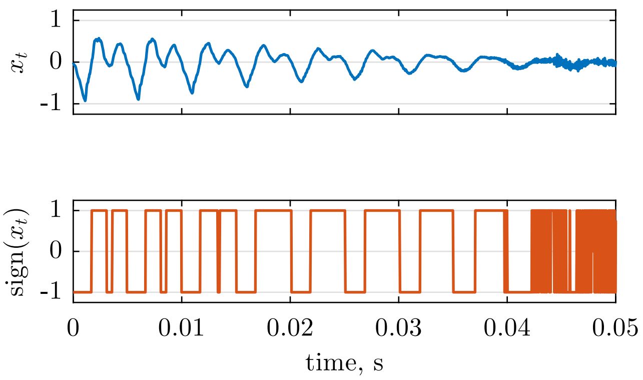

Figure 1.1: A short segment of an acoustic speech signal x_t and a distorted signal $\mathrm{sign}(x_t)$. Surprisingly, both signals are intelligible.

it cannot be the only source of robustness because language cannot be utilised without first recognising a sufficient proportion of the sounds that compose words and sentences (Allen, 2005a). Experiments performed by Miller and Nicely (1955) showed that with a signal-to-noise ratio (SNR) of -12 dB, humans can group individual phonemes into basic sound classes. For comparison, state-of-the-art ASR systems require a SNR of at least 0 dB to recognise sentences, even with the use of language models. The fact that humans can discriminate between individual phonemes in environments where ASR systems struggle highlights our present ignorance of speech communication.

How can the linguist make use of mathematical models of acoustic speech signals to better understand human speech perception? And how can the engineer make use of the knowledge of language to improve speech-based technologies? What is required is a unified understanding of speech communication. While this thesis does not offer a complete theory, the problem is approached in a new way. Specifically, this research attempts to develop a model of speech communication that is based on the mathematics of *information theory* (Shannon, 1948). The model attempts to link lexical information with the information of acoustic speech signals, incorporate talker variability, and be accurate enough to predict the intelligibility of speech in noisy environments.

1.2 Approaches to Modelling Speech Communication

Shannon's information theory (Shannon, 1948) provides a mathematical framework for analysing communication systems, regardless of the systems implementation. Information theory has made fundamental contributions to wireless communications, computer science, statistical inference, cryptography, thermodynamics, and biology (Cover and Thomas, 2012; MacKay, 2003). There is no reason that Shannon's theory cannot be applied to human speech communication. Surprisingly, a relatively small effort has been made to do so. Instead, the majority of existing speech communication models are based on *articulation index theory* and *deep learning*, both of which are discussed in the following.

Classical models of speech communication are largely based on *articulation index theory* (e.g., French and Steinberg, 1947; Fletcher and Galt, 1950; Allen, 1994). Articulation index theory was the outcome of a series of listening experiments conducted by Harvey Fletcher in the 1920's. Fletcher's experiments contributed to the understanding of speech communication by: (i) demonstrating that the intelligibility of nonsense syllables/words can be predicted from the intelligibility of their phonetic components, (ii) showing that acoustic queues are distributed over a range of frequency bands, and (iii) providing a basis for the development of hearing aids. Fletcher's results also gave rise to an algorithm called the articulation index (AI) (e.g., Kryter, 1962a; ANSI, 1969) that, given the SNR of various frequency bands and the channel bandwidth, can predict the intelligibility of a communication system.

Although articulation index theory has provided significant insight into the nature of speech communication, it is far from being an exhaustive theory. For example, it does not explain the effect of signal distortions other than band-pass filtering and additive noise, and it does not account for the variability of speech between talkers.

More recently, a branch of machine learning known as *deep learning* (see LeCun et al., 2015, for an overview) has delivered impressive results for speech-based technologies. Examples include automatic speech recognition (e.g., Graves et al., 2013; Amodei et al., 2016), text-to-speech synthesis (e.g., Van Den Oord et al., 2016), and speech coding (Kleijn et al., 2018).

Models based on deep learning rely on multi-layer *neural networks* that, given a data set of inputs and outputs, can be trained to learn a function that maps an input to an output. As an example, for ASR systems the input is an acoustic speech signal, the output is a sequence of words, and the mapping function is a conditional probability distribution of word sequences given acoustic signals. For text-to-speech synthesis the process is reversed: the input is a sequence of words, and the output is an acoustic signal.

Models based on deep learning have achieved impressive results in terms of their predictive power and fidelity. However, such models currently require thousands of hours of training data and rely on an exuberant amount of computational power. Moreover, such models typically consist of millions of parameters and thus can be difficult to interpret. As an example, *Deep Speech 2* (Amodei et al., 2016) is an ASR system that achieved human level performance for clean speech, but required about 12000 hours of training data and has lower performance than humans in noisy environments.

In this thesis, tools from both articulation index theory and deep learning are relied on. However, the research in this thesis differs from the mainstream by taking an information theoretical perspective. In this context, the effectiveness of communication is quantified using *mutual information*, which is a statistical measure of dependence between random variables. Low mutual information corresponds to poor communication, and high mutual information corresponds to good communication. Given a model and an accurate method for estimating mutual information, algorithms that explicitly maximise mutual information could be used to improve speech-based communication systems such as telephone, voice-over-IP, and hearing aids. Moreover, mutual information could be used to predict how well a communication system will perform in certain listening environments.

Some effort towards an information theoretic model of speech communication has been made. For example, Fano (1950) used a speech production model to estimate the information rate of speech communication. Allen (2005a,b) pointed out that the articulation index is a straight-line approximation of the normalised information capacity of a Gaussian channel. Jensen and Taal (2014) and Taghia and Martin (2014) hypothesised that the intelligibility of a degraded speech signal is related to the mutual information between the clean and degraded speech signal. Kleijn and Hendriks (2015) and Khademi et al. (2017) developed speech enhancement algorithms based on information theory. The research in this thesis can be viewed as an extension of those ideas.

1.3 Research Goals

The overall goal of this research is to approach the problem of speech communication from an information theoretical perspective. This thesis aims to develop a quantitative model that is consistent with our knowledge of linguistics and that is accurate enough to predict the intelligibility of speech signals. Naturally this could lead to algorithms that can be used to enhance the intelligibility of speech, particularly in the context of mobile telecommunications and hearing-aids. The research presented in this thesis helps to answer the following questions:

(i) *How does the acoustic information rate of speech compare to the lexical in-*

formation rate of speech?

The most basic question to ask in information theory is "how much information is transferred from the information source to the destination per unit of time"? For speech communication there are two ways to approach this problem: 1) from the linguistic perspective, and 2) from the speech processing perspective. The linguistic approach is to describe speech as a discrete sequence of lexical units. The information rate can then be computed by estimating the probability mass function of the lexical units. On the other hand, the speech processing approach is to describe speech communication using a statistical model of acoustic signals and to evaluate the information rate of the model. In the literature, estimates based on lexical sequences tend to be of the order of 50 b/s, whereas estimates based on acoustic signals tend to be of the order of 1×10^3 or 1×10^4 b/s. This thesis attempts to close the gap.

(ii) *How can information theory be used to predict the intelligibility of speech-based communication systems?*

When designing a speech-based communication system, it is important to understand how the system will affect intelligibility (i.e., the proportion of correctly identifiable words). Although formal listening tests can provide valid data, such tests are laborious and expensive to conduct. For this reason, algorithms that can predict intelligibility are of interest. In this thesis a new algorithm for predicting speech intelligibility that is based on information theory is proposed.

(iii) *How well do competing models of speech communication predict intelligibility?*

Over the past decade many algorithms for predicting intelligibility have been proposed, but have not been widely evaluated. For this reason, this thesis presents a comprehensive evaluation of intelligibility metrics. Additionally this thesis investigates why the top performing algorithms have high performance, and argues that some of the intelligibility metrics can be interpreted as approximations of mutual information.

1.4 Publications

As part of this research several papers have been peer-reviewed and published. They are:

- Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2018). An evaluation of intrusive instrumental intelligibility metrics. *IEEE Transactions on Audio, Speech, and Language Processing*.
- Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2018). An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*.
- Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2017). On the information rate of speech communication. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2016). An intelligibility metric based on a simple model of speech communication. *In Proceedings of the IEEE International Workshop on Acoustic Speech Enhancement*.

Material from the publications listed above form the basis of Chapters 3, 4 and 5 of this thesis. In addition, the following paper was peer-reviewed and published:

- Kolchinsky, A., Tracey, B. D., Van Kuyk, S. (2019). Caveats for information bottleneck in deterministic scenarios. *In Proceedings of the International Conference on Learning Representations*.

1.5 Thesis Outline and Contributions

This thesis consist of seven chapters. Chapter 2 presents an overview of the relevant literature, Chapter 3 develops a mathematical model of speech communication, Chapter 4 proposes two new algorithms for predicting intelligibility, Chapter 5 evaluates competing algorithms for predicting intelligibility, Chapter 6 proposes a new method for estimating the information rate of speech communication, and Chapter 7 concludes the work. Additional details and the contributions of each chapter are as follows.

Chapter 2: Literature Review

This chapter presents an overview of the literature surrounding information theory and speech communication. First, key ideas and quantities such as *entropy* and *mutual information* are introduced. Once these quantities have been defined, existing methods for estimating the information rate of speech communication are summarised. Two approaches are considered: the linguistic perspective, which represents speech as a sequence of lexical units, and the speech processing perspective, which represents speech as an acoustic signal. Next, the chapter introduces an important characteristic of speech communication: *talker variability*. Talker variability is largely caused by physiological differences between the vocal-tracts of different talkers, and may limit the maximum achievable information rate between talkers. The chapter then discusses the concept of *intelligibility*, why it is important, how it can be measured, and finally, summarises three existing information theoretic algorithms for predicting intelligibility.

Chapter 3: A Simple Model of Speech Communication

In this chapter, a simple mathematical model of speech communication is proposed. The model considers the transmission of a hypothetical message from a talker to a listener and quantifies the effectiveness of communication using the mutual information rate. In addition, this chapter introduces an auditory model for processing speech signals. The auditory model accounts for the frequency selectivity, dynamic range compression, upwards frequency masking, and forward temporal masking properties of the human auditory system. By combining the mathematical model of speech communication with the auditory model, a novel approach for estimating the information rate of speech communication is proposed. The method relies on having a *chorus* of speech signals, which consists of many talkers saying the same utterance. The statistics of talker variability are estimated using real-world data and then substituted into expressions for the mutual information rate.

Chapter 4: An Intelligibility Metric Based on Information Theory

This chapter proposes an intrusive intelligibility metric called *speech intelligibility in bits* (SIIB) and a variant called SIIB$^{\text{Gauss}}$. Intelligibility metrics are of importance as they can be used to predict the intelligibility of speech signals. SIIB and SIIB$^{\text{Gauss}}$ are based on the model of speech communication proposed in Chapter 3. Unlike competing intelligibility metrics, SIIB incorporates the effect of talker variability and partially accounts for statistical dependencies between time-frequency units of speech signals. The difference between SIIB and SIIB$^{\text{Gauss}}$ is that SIIB uses a non-parametric mutual information estimator based on k-nearest neighbours, whereas SIIB$^{\text{Gauss}}$ uses the capacity of a Gaussian communication channel.

Chapter 5: An Evaluation of Intelligibility Metrics

Chapter 5 evaluates the intelligibility metrics proposed in Chapter 4 and compares their performance to competing intelligibility metrics. In addition, this chapter investigates the ability of intelligibility metrics to generalise to new types of distortions and analyses why the top performing metrics have high performance. The intelligibility data were obtained from 11 listening tests described in the literature. Dutch, Danish, and English stimuli are included. The results show that (i) SIIB and SIIB$^{\text{Gauss}}$ have state-of-the-art performance, (ii) that intelligibility metrics tend to perform poorly on data sets that were not used during their development, and (iii) demonstrates the advantage of reducing statistical dependencies between input features.

Chapter 6: Estimating Mutual Information Using Siamese Networks

In this chapter, a novel approach for estimating the information rate of speech communication is proposed. Similarly to Chapter 3, the approach relies on having a chorus of speech signals. However, unlike Chapter 3, the approach does not make assumptions about the joint probability distribution of the hypothetical message and the speech signal. Instead, data-driven methods from machine learning are used: a *Siamese neural network*, and *Maximum Mean Discrepancy*. The approach is demonstrated on artificial examples, but is yet to be applied to real-world speech signals.

Chapter 7: Conclusions

In this final chapter the main conclusions of the thesis are given. Moreover, ideas for future research topics are discussed.

Chapter 2

Literature Review

This chapter reviews literature about speech communication and information theory. First, methods for estimating the information rate of speech communication are presented, second, the concept of talker variability is introduced, and third, existing information theoretic algorithms for predicting speech intelligibility are summarised. These three concepts lay the foundation for the research presented in this thesis.

2.1 Information Theory

Shannon's *information theory* (Shannon, 1948) provides a mathematical framework for analysing communication systems, regardless of the systems implementation. Information theory is fundamental to the design of wireless communications, cryptography, and data compression systems. There is no reason that information theory cannot be applied to models of human speech communication. Surprisingly, a relatively small effort has been made to do so.

There are two key concepts to information theory. The first concept is that the 'meaning' of a message is irrelevant to the engineering problem. Rather, the significant aspect is that a transmitted message is one selected from a set of possible messages. This view leads to a probabilistic

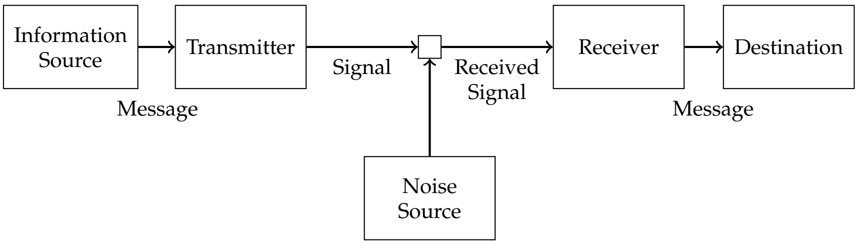

Figure 2.1: Shannon's general communication system.

approach.

The second concept of information theory is that a message of low probability contains more information than a message of high probability. This concept can be justified by noting that if there is no uncertainty at a receiver about what will be transmitted, then there is no information to be gained when the transmitted signal is received.

Figure 2.1 displays a diagram of Shannon's general communication system. The goal of the communication system is to reproduce at one point a message selected at another point. In the context of speech communication, the information source is the talker's brain, the transmitter includes the talker's vocal cords and vocal tract that encode the message into an acoustic signal, the channel is the physical medium that conducts the acoustic signal, the noise source characterises distortions introduced by the channel, the receiver is the listener's auditory system, and the destination is the listener's brain.

2.1.1 Definitions and properties

In this section, important definitions and properties from information theory are described. For more detail on information theory, see Cover and Thomas (2012) or MacKay (2003).

Definitions

Information theory quantifies information using a unit called a *binary digit* (bit). One bit is defined as the amount of information gained upon receiving one of two equally likely transmitted messages. Given a discrete random variable x with probability distribution $P(\mathrm{x})$, the *information content* of an outcome, $\mathrm{x} = x$, is defined as

$$H(x) = \log_2 \frac{1}{P(x)}. \tag{2.1}$$

In this way, a message of low probability contains more information than a message of high probability.

The information content only considers a single outcome. The *entropy* of a random variable is defined as the average information content over all possible outcomes $\mathbb{X}$:

$$\begin{aligned} H(\mathrm{x}) &= \mathbb{E}_{\mathrm{x}\sim P}[-\log_2 P(x)] && (2.2)\\ &= -\sum_{x\,\in\,\mathbb{X}} P(x)\log_2 P(x). && (2.3)\end{aligned}$$

The *mutual information* of two discrete random variables, x and y, quantifies the amount of information shared between x and y. Mutual information is defined as

$$I(\mathrm{x};\mathrm{y}) = \sum_{x,y\,\in\,\mathbb{X},\mathbb{Y}} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}, \tag{2.4}$$

where $\mathbb{Y}$ is the set of all possible outcomes of y. Mutual information can also be interpreted as the similarity between the joint distribution $P(\mathrm{x},\mathrm{y})$ and the product of the marginal distributions $P(\mathrm{x})P(\mathrm{y})$.

The *conditional entropy* of x given y quantifies the average uncertainty that remains about the outcome of x when the outcome of y is known. The conditional entropy of x given y is defined as

$$\begin{aligned} H(\mathrm{x}|\mathrm{y}) &= -\sum_{y\,\in\,\mathbb{Y}} P(y)\sum_{x\,\in\,\mathbb{X}} P(x|y)\log_2 P(x|y) && (2.5)\\ &= -\sum_{x,y\,\in\,\mathbb{X},\mathbb{Y}} P(x,y)\log_2 P(x|y). && (2.6)\end{aligned}$$

Properties

It can be shown that

$$H(\mathrm{x}) \leq \log_2 |\mathbb{X}| \qquad \text{with equality iff } \mathrm{x} \sim \mathcal{U}, \tag{2.7}$$

$$H(\mathrm{x}|\mathrm{y}) \leq H(\mathrm{x}) \qquad \text{with equality iff } P(\mathrm{x},\mathrm{y}) = P(\mathrm{x})P(\mathrm{y}), \tag{2.8}$$

$$0 \leq H(\mathrm{x}|\mathrm{y}), \tag{2.9}$$

$$0 \leq I(\mathrm{x};\mathrm{y}) \leq \min(H(\mathrm{x}), H(\mathrm{y})), \tag{2.10}$$

and that

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) \tag{2.11}$$

$$= H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) \tag{2.12}$$

$$= H(\mathrm{x}) + H(\mathrm{y}) - H(\mathrm{x},\mathrm{y}) \tag{2.13}$$

Furthermore, if and only if x and y are statistically independent, then

$$I(\mathrm{x};\mathrm{y}) = 0 \tag{2.14}$$

$$H(\mathrm{x},\mathrm{y}) = H(\mathrm{x}) + H(\mathrm{y}), \tag{2.15}$$

and if and only if a deterministic function, f, exists such that $\mathrm{x} = f(\mathrm{y})$, then

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) = H(\mathrm{y}). \tag{2.16}$$

For two invertible functions f and g,

$$I(f(\mathrm{x}); g(\mathrm{y})) = I(\mathrm{x};\mathrm{y}). \tag{2.17}$$

Continuous random variables

For continuous random variables, the entropy and conditional entropy are referred to as *differential entropy* and *conditional differential entropy*. Differential entropy, conditional differential entropy, and mutual information for continuous random variables, are defined analogously to the discrete case by replacing the summations in the above definitions with integrals.

However, some properties from the discrete case are lost. For example, differential entropy can be negative. To distinguish between entropy and differential entropy, H is used to denote entropy and h is used to denote differential entropy.

Differential entropy of the Gaussian distribution

The multivariate Gaussian distribution is defined by

$$P(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^* \Sigma^{-1}(\boldsymbol{x}-\mu)}, \tag{2.18}$$

where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{\mu}$ is the expected value of x, $\boldsymbol{\Sigma}$ is the covariance matrix of x, and $*$ denotes the transpose.

The differential entropy of the multivariate Gaussian distribution is

$$h(\mathbf{x}) = \frac{1}{2} \log_2 \det(2\pi e \boldsymbol{\Sigma}). \tag{2.19}$$

For the univariate case with with $\mathrm{var}(\mathrm{x}) = \sigma^2$, (2.18) simplifies to

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.20}$$

and (2.19) simplifies to

$$h(\mathrm{x}) = \frac{1}{2} \log_2 2\pi e \sigma^2. \tag{2.21}$$

2.2 The Lexical Information Rate of Speech

The most basic question to ask when applying information theory to speech communication is:

> *"How many bits of information are transferred from the talker to the listener each second"?*

In this section the above question is approached from the perspective of the linguist by providing a review of Shannon (1951), Fano (1950) and Flanagan et al. (2008), with some adaptations.

2.2.1 A word-based model of communication

Let $\mathbb{X}$ represent the set of all words in a given language, x be a random variable that represents a word spoken by a talker, and y represent the corresponding word received by a listener. Assuming a noiseless memoryless channel, $\mathrm{x} = \mathrm{y}$ and by (2.16) the mutual information between x and y is

$$\begin{aligned} I(\mathrm{x};\mathrm{y}) &= H(\mathrm{x}) && (2.22)\\ &= -\sum_{x\in\mathbb{X}} P(x)\log_2 P(x), && (2.23)\end{aligned}$$

where $P(x)$ is the probability that x is transmitted.

Shannon (1951) estimated (2.23) by computing word probabilities according to Zipf's law (Zipf, 1949). Zipf's law provides a reasonable approximation for word probabilities for a wide range of written and spoken languages (Piantadosi, 2014). Zipf's law states that if each word of $\mathbb{X}$ is ranked according to it's frequency of occurrence in speech, then the probability of the word with rank n, is given by

$$P(x_n) = 0.1/n, \tag{2.24}$$

where x_n denotes the n'th most frequent word in the language. Figure 2.2 plots (2.24) for $n < 50$ and compares the probabilities to the data tabulated in Davies and Gardner (2013) for English words.

Zipf's law cannot hold indefinitely since the sum of all probabilities must equal 1, while $\lim_{N\to\infty}\sum_{n=1}^{N} 0.1/n = \infty$. If we assume that Zipf's law is valid for $n \leq N$ such that $|1-\sum_{n=1}^{N} 0.1/n|$ is minimised, then $N = 12367$, which is comparable to the number of words in the spoken vocabulary of a child (Anglin et al., 1993). For $n > N$, we set $P(x_n) = 0$. The entropy of speech is then

$$\begin{aligned} H(\mathrm{x}) &= -\sum_{n=1}^{12367} \frac{0.1}{n}\log_2\frac{0.1}{n}\\ &= 9.72 \text{ bits per word.}\end{aligned} \tag{2.25}$$

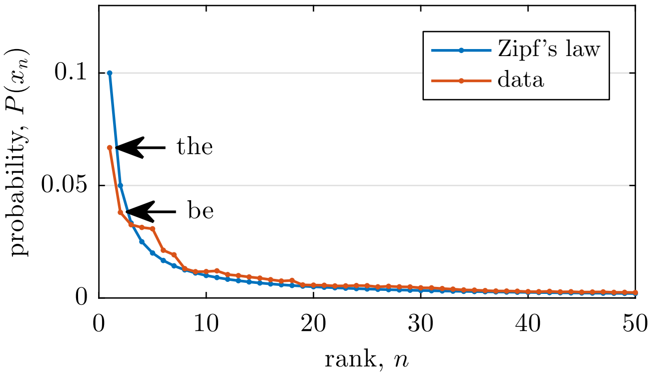

Figure 2.2: Word probability against frequency rank for the fifty most frequent English words. The red line shows data from Davies and Gardner (2013) and the blue line shows Zipf's law. The most frequent word is "the" with a probability of occurrence of 0.067. The second most frequent word is "be" with a probability of occurrence of 0.038.

Taking the average speaking rate of conversational speech to be 130 words per minute (Siegler and Stern, 1995; Levelt, 1999), the information rate of speech is 21 b/s. Note that this rate is for a noiseless memoryless channel. That is, the transmitted word is equal to the received word, and each transmitted word is selected independently from the previously transmitted words. If the dependencies between successive words were taken into account and if communication took place in a noisy environment, then the information rate would be reduced. Zipf's law is only an approximation of the word probabilities, thus a rate of 21 b/s is a crude estimate.

2.2.2 A phoneme-based model of communication

We now consider the case where x represents a phoneme rather than a word. For English, there are approximately 44 phonemes. The exact number depends on the dialect of the talker and the classification procedure used to distinguish between speech sounds. Assuming that each phoneme is selected for transmission independently and with equal probability, by

Table 2.1: Occurrence of English phonemes. Data from Denes (1963).

x	$P(x)$	x	$P(x)$
ə	0.090445	t	0.084033
ɪ	0.082537	n	0.070849
aɪ	0.028473	s	0.050893
e	0.028126	d	0.041767
iː	0.017878	l	0.036892
əʊ	0.017477	m	0.032890
ʌ	0.016701	ð	0.029927
ɒ	0.015330	k	0.028985
æ	0.015261	r	0.027697
eɪ	0.014956	w	0.025661
uː	0.014222	z	0.024927
ɔː	0.012007	b	0.020842
ɑː	0.007755	v	0.018515
aʊ	0.007741	p	0.017698
ʊ	0.007672	f	0.017283
ɜː	0.006661	h	0.016729
eə	0.004335	j	0.015303
ɪə	0.002867	ŋ	0.012436
ʊə	0.001426	ɡ	0.011619
ɔɪ	0.000872	ʃ	0.007021
		θ	0.005955
		dʒ	0.005138
		tʃ	0.003684
		ʒ	0.000512

(2.7) the entropy of English speech is $\log_2 44 = 5.46$ bits per phoneme. If the phonemes are selected independently and with probabilities equal to the observed frequencies tabulated in Denes (1963) and shown in Table 2.1, then the entropy reduces to $H(\mathrm{x}) = 4.91$ bits per phoneme. Taking the average speaking rate of conversational English to be 12 phonemes per second (Levelt, 1999; Tiffany, 1980), the information rate is about 60 b/s. Again, this estimate is for a noiseless memoryless channel and could be reduced further by accounting for statistical dependencies between phonemes.

For the word-based model of communication in Section 2.2.1, a speaking rate of 130 words per minute was used. When this rate is compared to the speaking rate of 12 phonemes per second that was used in the above calculation it implies that on average there are 5.5 phonemes per word, which is reasonable.

The information rate estimated using phonemes is larger than the information rate estimated using words by a factor of three, but this is reasonable considering the precision of Zipf's law. Additionally, because words are composed of multiple phonemes, the word probabilities implicitly account for some dependencies between successive phonemes. Thus, given that both models are memoryless, it is plausible that the information rate estimated using words is lower than the information rate estimated using phonemes. If more realistic models of communication were considered, i.e., if the phoneme-based model was not memoryless and instead considered the probability distribution of a sequence of phonemes, and likewise for the word-based model, then the information rate for words and phonemes would be equal.

In light of the above analysis, as a first approximation, we take 60 b/s as the *lexical* information rate of speech. This is similar to existing results in the literature such as Flanagan et al. (2008) and Fano (1950) that estimate the lexical information rate to be 50 b/s. The 10 b/s discrepancy with the former exists because Flanagan et al. (2008) used a different data set for calculating phoneme probabilities and a lower speech rate of 10 phonemes per second. The discrepancy with the latter exists because the analysis of Fano (1950) was based on the entropy of alphabetic letters rather than words or phonemes.

The lexical information rate does not include information about talker identification, emotional state, and prosody. However, these variables vary relatively slowly in time and contribute little to the overall information rate. As an example, Fano (1950) estimated that the total amount of talker-specific information (e.g., age, accent, sex) was of the order of 30 bits. This information only has to be transmitted once. Thus, if speech with a duration of one minute is considered, then accounting for talker-specific information increases the information rate by only 0.5 b/s.

2.2.3 The effect of noise and filtering

The phoneme-based model of communication from the previous section is now extended to include the effect of channel distortion. Specifically, the effect of additive stationary noise and linear filtering is considered. Additive stationary noise is a type of distortion commonly encountered in everyday life, and linear filtering is often used in telephony to reduce bandwidth. The analysis in this section follows that presented in Flanagan et al. (2008); however, we give our own interpretation of the results. In Chapter 5 the effect of other types of distortion such as reverberation, modulated noise, and enhancement algorithms is considered.

Noise could be introduced to the communication channel during the production of a phoneme by the talker, the transmission of the phoneme through the channel, or the decoding of the received phoneme by the listener. Because of noise, the phoneme selected by the talker may be different to the phoneme decoded by the listener. This means that the assumption from Section 2.2.1 that $\mathrm{x} = \mathrm{y}$ is not necessarily valid. For a noisy memoryless communication channel, the mutual information between x and y is

$$\begin{aligned} I(\mathrm{x}, \mathrm{y}) &= H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) && (2.26) \\ &= -\sum_{y \in \mathbb{X}} P(y) \log_2 P(y) + \sum_{x \in \mathbb{X}} P(x) \sum_{y \in \mathbb{X}} P(y|x) \log_2 P(y|x), && (2.27) \end{aligned}$$

where $P(y|x)$ is the conditional probability of perceiving phoneme y given that phoneme x was transmitted, and $\mathbb{X}$ is the set of all possible phonemes for the chosen dialect.

Miller and Nicely (1955) measured the conditional probabilities of 16 English consonants for a range of channel conditions. The experiment was performed by transmitting consonants one at a time over a telephone channel. Five normal-hearing female listeners were then asked to identify which consonant was transmitted.

Three types of channel distortions were tested: a channel with additive white noise[1], a low-pass filtered channel, and a high-pass filtered channel. For the additive white noise channel, the signal-to-noise ratio (SNR) was set to -18, -12, -6, 0, 6, and 12 dB and the bandwidth was set to 200-6500 Hz. For the low-pass filtered channel, the high-pass cut-off frequency was fixed at 200 Hz and the low-pass cut-off frequency was set to 300, 400, 600, 1200, 2500, and 5000 Hz. For the high-pass filtered channel, the low-pass cut-off frequency was fixed at 5000 Hz and the high-pass cut-off frequency was set to 200, 1000, 2000, 2500, 3000, and 4500 Hz. For both the low-pass filtered channel and the high-pass filtered channel, a fixed SNR that corresponded to 12 dB for unfiltered speech was used.

The results of the experiment were recorded for each channel condition in a 16×16 *confusion matrix* where a row corresponds to a transmitted phoneme and a column corresponds to a phoneme guessed by a listener. Table 2.2 shows a confusion matrix for a communication channel with a bandwidth of 200-6500 Hz and a SNR of 0 dB. The first row indicates the number of times each phoneme was guessed by a listener given that /p/ was spoken by the talker. For this communication channel we see that /p/ was most commonly confused with /t/ and /k/.

Using the confusion matrices tabulated in Miller and Nicely (1955), $P(\mathrm{x}), P(\mathrm{y}),$ and $P(\mathrm{y}|\mathrm{x})$ can be estimated for each channel condition. Consequently, the mutual information in (2.27) can be computed as a function of the SNR and channel bandwidth. Because the transmitted message was selected from sixteen possible phonemes, by (2.10) the mutual information cannot exceed $H(\mathrm{x}) = \log_2 16 = 4$ bits per phoneme. When this occurs we

[1] That is, a Gaussian signal with uniform power across the frequency band.

Table 2.2: A confusion matrix showing the perceptual confusions of consonants transmitted over a channel with a bandwidth of 200-6500 Hz and a SNR of 0 dB. Data from Miller and Nicely (1955).

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	150	38	88	7	13											
t	30	193	28	1												
k	86	45	138	4	1		1									1
f	4	3	5	199	46	4		1				1			1	
θ	11	6	4	85	114	10					2					
s		2	1	5	38	170	10			2						
ʃ		3	3			3	267									
b				7	4			235	4		34	27	1			
d									189	48		4	8	11		
g									74	161		4	8	25		
v				3	1			19		2	177	29	4	1		
ð								7		10	64	105	18			
z									17	23	4	22	132	26		
ʒ									2	3	1	1	9	191		1
m								1							201	6
n												3		1	8	240

say that the channel is *saturated*. Channel saturation is only achieved when all transmitted phonemes are correctly identified.

Figure 2.3 plots the mutual information in bits per phoneme against the SNR in dB. The mutual information is 0 bits per phoneme at -18 dB and increases approximately linearly with the SNR until 0 dB. From 0 dB to 12 dB the communication channel gradually approaches saturation.

Similarly, Figure 2.4 shows the effect of filtering on the mutual information. The blue curve indicates the mutual information for a channel with a bandwidth from 200 Hz to cut-off frequency f_c, and the red curve indicates the mutual information for a channel with a bandwidth from f_c to 5000 Hz. The curves intercept at 1250 Hz indicating that a channel from 200-1250 Hz transfers the same amount of information as a channel from 1250-5000 Hz. The amount of information at this cross-over point is 2.4 bits per phoneme. If the low-pass channel and high-pass channel were

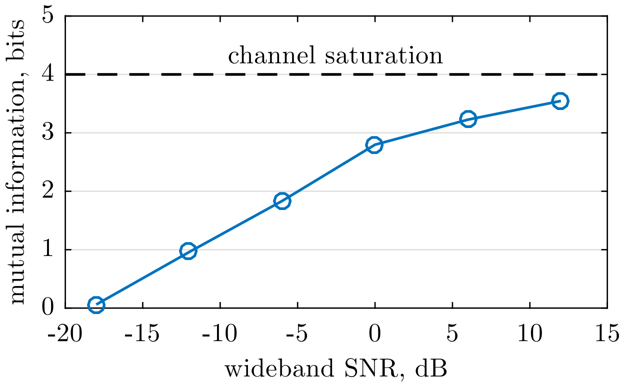

Figure 2.3: Mutual information against signal-to-noise ratio for phonemes transmitted over a noisy channel with a bandwidth of 200-6500 Hz. The black line shows the maximum achievable transfer of information corresponding to perfect phoneme recognition. Figure reproduced from Flanagan et al. (2008).

statistically independent, then the information for the wide-band channel from 200-5000 Hz would be $2.4 + 2.4 = 4.8$ bits per phoneme. However, this is not what we observe. The last data point of the blue curve indicates that the amount of information from the wide-band channel is 3.2 bits per phoneme. It follows that the low-pass and high-pass frequency channels are not independent; rather, some of the same information is transmitted in parallel over the separate frequency channels. This means that if the spectral content of a speech signal within a particular frequency band is distorted, then the information transferred through that frequency band is not completely lost. This 'frequency redundancy' is one reason for the robustness of human speech communication.

Limitations of Miller's experiment

Due to the design of the experiment in Miller and Nicely (1955), there are several limitations to the preceding analysis. One limitation is that a small subset of the 44 English phonemes was used. This subset makes up about

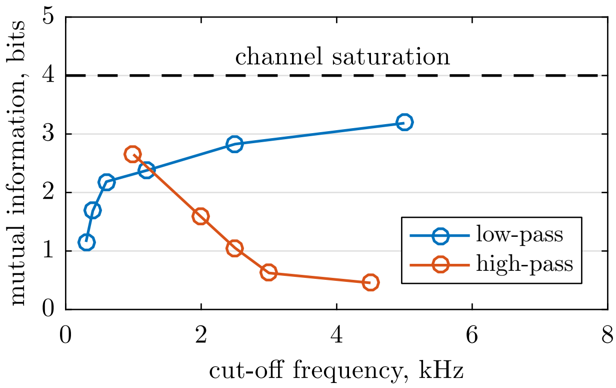

Figure 2.4: Mutual information against cut-off frequency for phonemes transmitted over a band-limited channel with a wide-band SNR of 12 dB. The black line shows the maximum achievable transfer of information corresponding to perfect phoneme recognition. Figure reproduced from Flanagan et al. (2008).

40% of English speech communication. Most importantly, the subset does not contain vowels, which tend to have more energy located below 3 kHz than consonants. This means that the effect of filtering consonants will differ to the effect of filtering vowels, so the results cannot be extrapolated to all human speech communication.

A second limitation is that the phonemes were transmitted with equal probability. If the true phoneme probabilities of conversational speech were used, then some spectral shapes would be more common than others and thus the effect of the filters would be different.

A third limitation is that the phonemes were presented in isolation. The acoustic realisation of a phoneme depends on preceding and succeeding sounds, thus the affect of noise and filtering on an isolated phoneme could differ to the affect of noise and filtering on a continuous stream of phonemes.

A fourth limitation is that a small selection of five listeners was used. In addition, all of the listeners were female, and the same five subjects that were used as listeners were also used as talkers. However, in Phatak et al. (2008) the SNR experiment was repeated using 10 males and 8 females and, in general, the consonant confusions were consistent with the original experiment.

2.3 The Acoustic Information Rate of Speech

Unlike the lexical information rate of speech that was discussed in Section 2.2, the *acoustic* information rate of speech is obtained without knowledge of language. Instead, observations of acoustic speech signals are relied upon. In this section, a review of the acoustic information rate of speech is presented. First a simple time-domain model of speech communication that is often included in textbooks (e.g., Flanagan et al. (2008)), is described, and then the analysis from Fano (1950) is described.

2.3.1 A simple time-domain model of speech communication

As a first approach to measuring the acoustic information rate of speech, consider the case where speech is represented by the samples of a time-domain acoustic speech signal. Let $\{\mathrm{x}_t\}$ be a stochastic process that represents the signal produced by the talker. Figure 2.5 shows the samples of an acoustic speech signal produced by an English speaking female.

The time-domain samples of an acoustic speech signal can be modelled as a stationary univariate Gaussian process (Jensen et al., 2005). Ideally a time-domain model would consider the joint distribution of a sequence of samples, rather than considering the marginal distribution only, but such an approach is not tractable. Figure 2.6 shows a histogram of x_t for the signal shown in Figure 2.5 and also includes a Gaussian probability density obtained using maximum likelihood estimation. We see that the Gaussian model provides a rough approximation. Replacing the Gaussian density

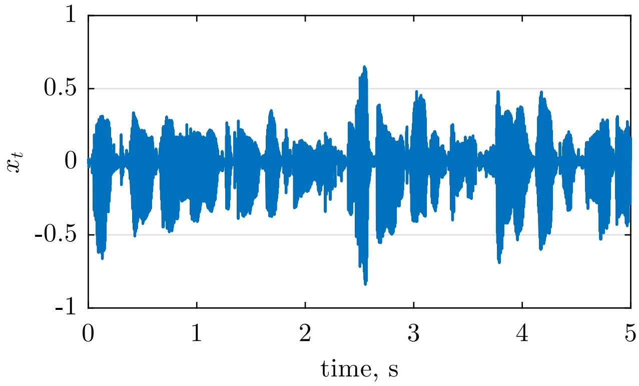

Figure 2.5: An acoustic speech signal sampled at 16 kHz.

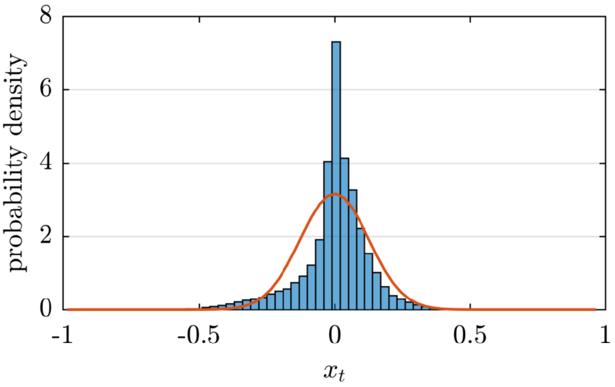

Figure 2.6: A histogram of an acoustic speech signal and the corresponding maximum-likelihood Gaussian probability density function.

with a Laplace density or a Generalised Gaussian density provides a better fit (Jensen et al., 2005), however, the Gaussian model is mathematically tractable and provides a useful starting point for further analysis.

Consider the transmission of $\{\mathrm{x}_t\}$ through a band-limited channel corrupted by additive Gaussian noise $\{\mathrm{n}_t\}$ that has been filtered to have the same power spectral density (PSD) as the PSD of $\{\mathrm{x}_t\}$. The mutual information rate between the transmitted Gaussian signal and the received

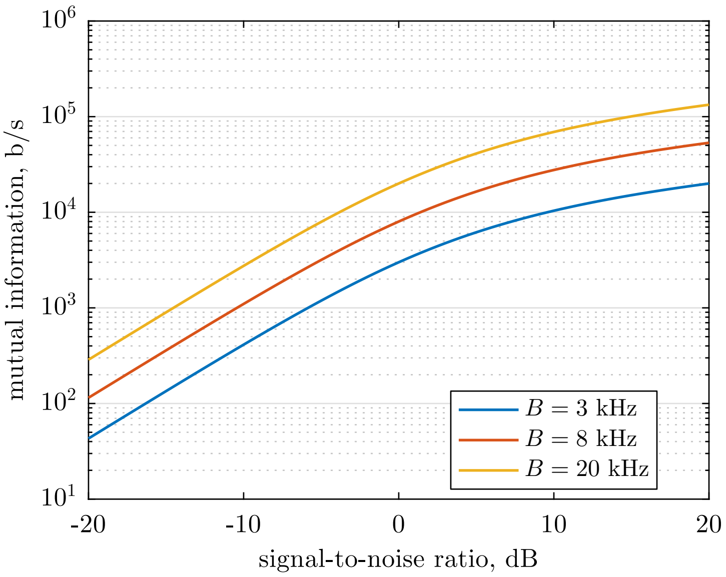

Figure 2.7: Mutual information for a Gaussian signal transmitted over a Gaussian channel.

signal $\{\mathrm{y}_t\}$ is given by (Shannon, 1948)

$$I(\{\mathrm{x}_t\};\{\mathrm{y}_t\}) = B\log_2(1+\mathrm{SNR}) \qquad \text{bits per second,} \tag{2.28}$$

where B is the bandwidth of the channel and SNR is the signal-to-noise ratio given by $\mathrm{var}(\mathrm{x}_t)/\mathrm{var}(\mathrm{n}_t)$.

The bandwidth required for speech communication can be taken as anywhere from 3 kHz to 20 kHz. The former corresponds to the bandwidth of telephone communication, which degrades speech signals but for the most part preserves intelligibility, and the later corresponds to the frequency range of the human ear.

The SNR required for speech communication varies depending on the complexity of the listening task. For example, the SNR required for distinguishing between a small set of words or syllables is lower than the SNR required for distinguishing between a large set of words or syllables. Similarly, due to the redundancy of language, the SNR required for conversational speech is lower than the SNR required for nonsense syllables

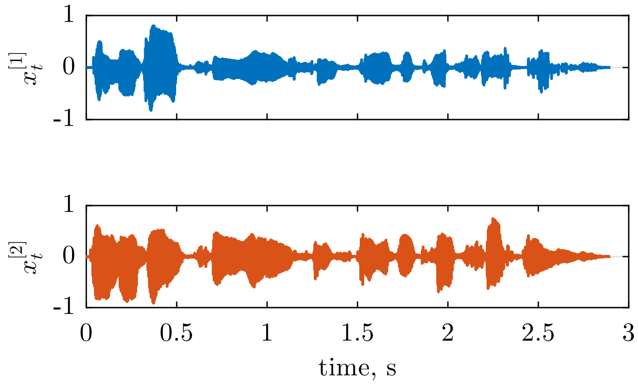

Figure 2.8: Time-domain speech signals for two talkers speaking in synchrony.

(Miller et al., 1951). Though the exact requirement depends on the listening task, intelligibility has not been known to improve beyond a SNR of 20 dB, and in most cases, intelligibility is at chance levels for a SNR below -20 dB.

Figure 2.7 plots (2.28) for the values discussed above. For ideal listening conditions we take $B = 8$ kHz and $\mathrm{SNR} = 20$ dB giving an information rate of 53266 bits per second.

Now consider the mutual information between two acoustic signals carrying the same lexical message but spoken by two different talkers. Let $\{\mathrm{x}_t^{[1]}\}$ denote the first signal and let $\{\mathrm{x}_t^{[2]}\}$ denote the second signal. If the signals are modelled as jointly Gaussian stationary processes, then the mutual information rate of the two signals is given by (Cover and Thomas, 2012)

$$I(\{\mathrm{x}_t^{[1]}\}, \{\mathrm{x}_t^{[2]}\}) = -B \log_2(1 - \rho^2) \qquad \text{bits per second,} \tag{2.29}$$

where ρ is the correlation coefficient between $\{\mathrm{x}_t^{[1]}\}$ and $\{\mathrm{x}_t^{[2]}\}$. Note that because of the stationary assumption, ρ is constant for all t. Figure 2.8 shows sample sequences of $\{\mathrm{x}_t^{[1]}\}$ and $\{\mathrm{x}_t^{[2]}\}$. The sample sequences were obtained by recording two female talkers speaking the same utterance in synchronous. Because the sampling rate is 16 kHz, the bandwidth is $B = 8$

kHz. The sample correlation coefficient for the data is $\hat{\rho} = 0.0215$, which results in $I(\{\mathrm{x}_t^{[1]}\}, \{\mathrm{x}_t^{[2]}\}) = 5.35$ bits per second.

From the above results we can identify three problems with the simple time-domain Gaussian model of speech communication. First, the acoustic information rate is three orders of magnitude larger than the lexical information rate from Section 2.2.2. This suggests that there is redundant information in the samples of acoustic speech signals that has not been accounted for. Second, the acoustic information rate given in (2.28) increases without bound with the SNR, whereas the lexical information rate in Figure 2.3 saturates at a particular value. This suggests that there is a physical limitation or internal noise source that has not been considered. Third, for two speech signals that contain the same lexical information we find that out of the 53266 bits per second encoded in each signal, only 5 bits of information per second are common to both signals. The fact that the time-domain model predicts that two signals carrying the same lexical information only share 5 bits of information per second, while the lexical information rate is 60 b/s, is an obvious flaw.

Clearly, the time-domain Gaussian model is a poor model of human speech communication. The reason the model fails is because the model is too simple: the samples of an acoustic speech signal do not accurately reflect the encoded lexical information. In order to get consistent results for the lexical information rate of speech and the acoustic information rate of speech, an appropriate representation of speech needs to be combined with a more powerful statistical model.

2.3.2 Fano's method for measuring the acoustic information rate

Fano (1950) proposed a method for measuring the acoustic information rate of speech that is based on *source-filter theory* (e.g., Stevens, 2000). Figure 2.9 shows a diagram of the source-filter model of speech production

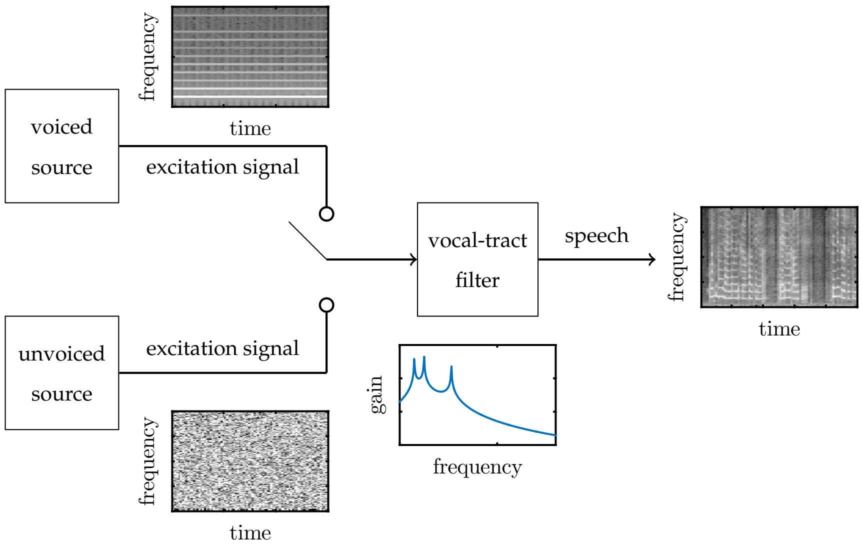

Figure 2.9: The source-filter model of speech production. The source produces an excitation signal and the vocal-tract filter spectrally shapes the excitation signal into a series of peaks and valleys. A *voiced* excitation signal is generated by forcing air through the vocal folds and an *unvoiced* excitation signal is generated by creating a turbulent flow of air in the airways. The time-domain waveform of a voiced excitation signal is roughly triangular and periodic, so in the time-frequency domain it is a harmonic line spectra with power that decreases as the frequency increases. The time-domain waveform of an unvoiced excitation signal is modelled by white Gaussian noise. The vocal-tract filter has a time-dependent transfer function that is modified throughout speech communication by moving the vocal articulators, e.g., the tongue, lips, and jaw. This model has gained wide use in speech enhancement (Loizou, 2013), speech coding (Kleijn and Paliwal, 1995), speech recognition (Rabiner and Juang, 1993), and evolutionary theories for the origin of speech (Fitch, 2000).

and gives a brief description of it's operation.

Fano's method models voiced speech signals in the time-frequency domain as a number of modulated carrier signals with harmonically related carrier frequencies. From the perspective of source-filter theory, the carrier signals are the frequency components of the excitation signal and the modulator is the time-varying vocal-tract filter. The information rate for voiced speech is governed by the bandwidth of the modulations, the number of carrier signals, and the SNR of the communication channel. Fano then advocates that the information rate of unvoiced speech should be comparable to the information rate of voiced speech.

Physical constraints limit the speed at which the tongue, mouth, and jaw can move, which has the effect that amplitude modulations are band-limited to about 10 Hz[2]. The exact number of carrier signals depends on the fundamental frequency of the excitation signal. A representative value for the fundamental frequency of a female talker is 250 Hz, resulting in 32 evenly spaced carriers over a 8000 Hz bandwidth. The SNR required for ideal listening conditions is 20 dB[3]. If the modulated carriers are modelled as Gaussian signals transmitted over independent frequency channels, then the information rate is given by $I(\{\mathrm{x_t}\}, \{\mathrm{y_t}\}) = 32 \times 10 \times \log_2(1+10^2) = 2130$ bits per second. Fano's analysis brings the acoustic information rate considerably closer to the lexical information rate, but the acoustic information rate is still 35 times larger.

There are two aspects of speech communication that Fano's method does not account for: 1) statistical dependencies in acoustic signals, and 2) *talker variability*. Fano's method estimates the information rate by summing the information over multiple frequency channels, however, as pointed out in Section 2.2.3, speech signals have 'frequency redundancy', meaning that some of the same information is transmitted in parallel over

[2] A modulation bandwidth of 10 Hz is an approximate value that is valid for vowels, but not for other speech sounds. For example, stop consonants have sharp transitions with a voice-onset time of 20 ms, which loosely corresponds to a bandwidth requirement of 50 Hz.

[3] The original analysis by Fano assumed that a bandwidth of 7000 Hz and a SNR of 24 dB was required for ideal listening conditions. For consistency with Section 2.3.1, we instead use a bandwidth of 8000 Hz and a SNR of 20 dB.

separate frequency channels. Because statistical dependencies between the modulator signals are not accounted for, Fano's method overestimates the information rate. The second reason that Fano's method over estimates the information rate is related to talker variability, which is discussed in the following section.

2.4 Talker Variability

Speech communication can be viewed as the transmission of lexical code words that are physically realised as acoustic signals. One aspect of speech communication that has been overlooked in this chapter thus far is that there is variability inherent in the production of speech signals. In other words, different acoustic signals can carry the same lexical message.

2.4.1 Example of talker variability

An example of talker variability is shown in Figure 2.10 which includes two spectrograms of the sentence "a fifth wheel caught speeding". The top utterance was produced by a female American talker and the bottom utterance was produced by a male Irish talker.

It can be seen that the male Irish talker tends to use more low frequency energy than the female American talker. For example, consider the speech segment from 0.25 s to 0.4 s. The male Irish talker concentrates energy between 0-3500 Hz, whereas the female American talker concentrates energy between 0-4500 Hz. Similarly, for the speech segment from 0.9 s to 1.0 s the male Irish talker concentrates energy upwards of 3000 Hz, whereas the female American talker concentrates energy upwards of 4000 Hz.

Another difference between the spectrograms is apparent at 0.45 s to 0.7 s where the phoneme /iː/ from "wheel" is pronounced. The spectrogram of the male Irish talker has a narrow band of concentrated energy that slowly fluctuates between 1000-2000 Hz, but the spectrogram of the

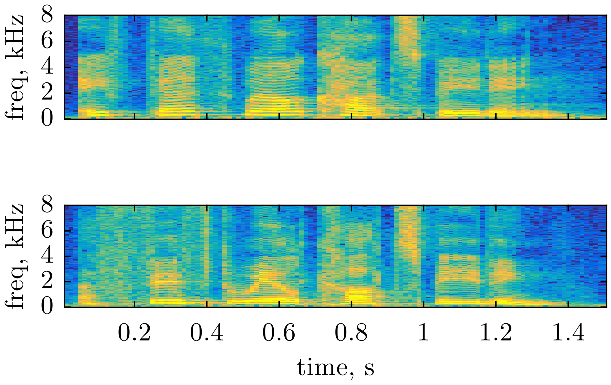

Figure 2.10: Spectrograms of an utterance produced by a female American talker (top) and a male Irish talker (bottom). The acoustic signals carry the same lexical information, but are different due to talker variability.

female American talker is missing this acoustic feature.

The two speech signals contain the same lexical message, but both talkers encode the lexical message in a unique way. This variability between talkers is a fundamental aspect of speech communication that should not be ignored.

2.4.2 Sources of talker variability

The two main sources of talker variability are physiological differences and learned speech habits (Huang et al., 2001). Physiological differences between talkers include differences in the length and shape of their vocal-tracts. The average vocal-tract length is 16.9 cm for adult males, 14.1 cm for adult females, and 7.9 cm for newborns (Goldstein, 1980). Talkers with longer vocal-tracts tend to have voices that sound 'deeper' than those with shorter vocal tracts. Other anatomical differences such as mouth shape, tongue length, and missing teeth also have an impact on the characteristics of the talker's vocal-tract filter and consequently affect the production of acoustic speech signals.

The second main source of talker variability is learned speech habits.

Learned speech habits are related to a talker's native-language, region of origin, and socio-economic status. Pronunciation, stress, and prosody are all affected. One consequence of learned speech habits is that non-native talkers sometimes replace unfamiliar phonemes of a target language with phonemes familiar to their native language (Flege et al., 2003). For example, native English talkers frequently replace the alveolar trill /r/ (i.e., a 'rolled r') that is common in Spanish and Arabic with the alveolar approximant /ɹ/ because it requires the talker less articulatory effort. Learned speech habits are also seen in young children that have not learned how to properly formulate the sounds of their native language (Stoel-Gammon, 1989).

Besides physiological differences and learned speech habits, other factors that contribute to talker variability include emotional state, speech impairments, speaking styles (e.g. whispering and shouting), and long-term habits (e.g. smoking and singing). For a more comprehensive discussion of factors that contribute to talker variability, see Benzeghiba et al. (2007).

2.4.3 An information theoretical perspective of talker variability

Fano (1950) identified that talker variability posed a problem to his communication model. He hypothesised that the differences between the acoustic signals of different talkers could be modelled as a type of noise that is inherent to the production of speech. This noise would bound the information rate of speech communication and lower his estimate of the acoustic information rate. Although Fano (1950) identified the problem, no attempt was made to address it.

The idea that talker variability could be modelled as noise was taken up again over 50 years later in Kleijn and Hendriks (2015), which coined the term *production noise*. It was shown, theoretically, that production noise would cause the effectiveness of communication to saturate at a particular

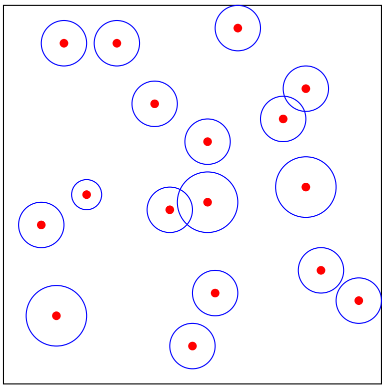

Figure 2.11: A channel coding interpretation of speech communication. The black box represents the set of all acoustic speech signals, the red dots represent linguistic code words, and the blue circles represent a subset of acoustic signals that are associated with a given code word.

bit rate. Thus, unlike (2.28), even if the SNR of the environmental channel is infinite, the information rate of speech communication is finite. Kleijn and Hendriks (2015) and Khademi et al. (2017) used this concept to design a state-of-the art speech enhancement algorithm. The enhancement algorithm redistributes power across frequency bands using a variant of the water-filling algorithm (Cover and Thomas, 2012). When production noise is accounted for, allocating power to frequency bands with high SNR has diminishing returns because the communication channel begins to saturate. Thus, compared to traditional water-filling, the speech enhancement algorithm allocates less power to high SNR frequency bands.

To make the notion of production noise more concrete, Figure 2.11 shows a channel-coding interpretation of production noise. The black box represents the set of all acoustic speech signals that can be articulated by humans; the red dots represent code words, which could be phonemes, words, sentences, or any other perceptual linguistic unit; and the blue circles represent a subset of acoustic signals that correspond to a particular code word. A change in language corresponds to a new set of red dots, a change in dialect or speaking habit corresponds to a slight perturbation of the red dots, and an increase in production noise corresponds to larger blue circles. In this model, production noise is thought of as the natural consequence of having many talkers with different vocal tracts and speech habits. Note that some of the acoustic subsets in Figure 2.11 overlap. This allows the same acoustic signal to be perceived differently depending on preceding and succeeding sounds, as was found in Liberman et al. (1957).

The production noise point-of-view is consistent with the observation that the ability to understand speech depends on the listener's familiarity with the talker's voice. When a listener encounters a talker with an unfamiliar speech habit or accent, it can take time for the listener to learn the talker's speech code.

The information rate of a language with a set of code words, $\mathbb{X}$, saturates at $\log_2 |\mathbb{X}|$ bits per code word. This means that increasing the number of code words increases the maximum achievable information rate. However, when production noise is considered, increasing the number of code words also increases the conditional probability of confusing one code word with another code word. Consequently there is a trade off between efficiency and robustness.

The only way to increase the robustness of speech communication without affecting it's efficiency is to decrease the production noise. That is, talkers must speak more similarly to one another. Note that this behaviour is observed when considering talkers that belong to a particular socio-economic class or geographical region. Talkers belonging to such groups typically develop similar speaking habits to one another, which makes communication more robust. Other factors that contribute to the formation of similar speaking habits include social bonding and self-identity, as advocated by *speech accommodation theory* (Gallois et al., 2005).

Talker variability plays a key role in this thesis. In Chapter 3, a method for estimating the statistics of production noise is developed, and in Chapter 4, an algorithm for predicting the intelligibility of speech signals that incorporates production noise is proposed.

2.5 Speech Intelligibility

A viable model of speech communication must be able to make quantifiable predictions that can be verified by experiment. The utility of the model developed in this thesis is evaluated by its ability to make predictions about the *intelligibility* of speech. For this reason, this section focuses on algorithms for predicting intelligibility. Such algorithms are called *instrumental intelligibility metrics*. The intelligibility metrics described in this section include: the articulation index (AI) (French and Steinberg, 1947), the speech intelligibility predictor based on mutual information (SIMI) (Jensen and Taal, 2014), and the k-nearest neighbour mutual information intelligibility measure (MIKNN) (Taghia and Martin, 2014). The AI, SIMI, and MIKNN are all based on information theory. Intelligibility metrics that are not based on information theory are described in Chapter 5.

2.5.1 What is speech intelligibility?

When designing a speech-based communication system (e.g., telephone, voice-over-IP, and public address systems), it is important to understand how the system will affect the intelligibility and quality of speech. Intelligibility is often defined as the proportion of words correctly identified by a listener (Allen, 2005a), whereas *speech quality* refers to the ′pleasantness′ or ′naturalness′ of the speech signal (Loizou, 2013).

Intelligibility and speech quality are not synonymous. As an example, telephone speech often has high intelligibility but is of poorer quality than clean speech due to bandwidth constraints. An increase in intelligibility does not always coincide with an increase in speech quality, and vice versa. This thesis focuses on intelligibility.

There are many factors that affect intelligibility. Examples include the complexity of the speech material, the talker's choice of words, the talker's speaking style (e.g., loud, conversational, whispered, clear, and fast), characteristics of the communication channel (e.g., quantisation, noise, and reverberation), the way that the communication system modifies user behaviour, the Lombard effect (Lombard, 1911), visual cues, the listener's familiarity with the talker, the listener's proficiency, and the listener's attentiveness. For a detailed discussion of these factors see French and Steinberg (1947), Miller et al. (1951), and Krause and Braida (2002).

2.5.2 Measuring speech intelligibility

In order to measure intelligibility, a variety of standardised listening tests have been developed. Examples include the Diagnostic Rhyme Test (ANSI, 1989), the Hagerman Test (Hagerman, 1982), and the Hearing in Noise Test (Nilsson et al., 1994). These tests have been carefully crafted to control for *context effects* and ensure that the test material is representative of everyday language.

A famous example (Bridle et al., 1983) that demonstrates the effect of context can be seen by comparing the sentence "how do humans recognise speech?" with the sentence "how do humans wreck a nice beach?". These sentences can be spoken such that the corresponding acoustic signals are effectively indistinguishable. In such a scenario only context can be relied on to correctly decode the lexical message. When left unaccounted for, this phenomenon complicates intelligibility testing, decreases the efficiency of the testing, and increases the variability of the results (Allen, 1994).

2.5.3 Instrumental intelligibility metrics

Although formal listening tests are capable of providing valid data, such tests are time-consuming, laborious, and expensive. This makes it unfeasible to continuously evaluate a communication system throughout the design process. For this reason, quantities that are fast to compute and correlated with intelligibility are of interest. Such quantities are referred to as *instrumental intelligibility metrics*.

Rather than using human subjects, instrumental intelligibility metrics may rely on knowledge of the clean speech, degraded speech, and the communication channel. There are two types of intelligibility metrics: *intrusive* and *non-intrusive*. Intrusive intelligibility metrics require knowledge of the clean speech and either the channel or the degraded speech, whereas non-intrusive intelligibility metrics require only the degraded speech. Although non-intrusive metrics are more widely applicable, they tend to be less correlated with intelligibility than intrusive metrics (Falk et al., 2015; Andersen et al., 2017). This thesis focuses on intrusive metrics.

Intelligibility metrics typically compute a number on a closed interval from 0 to 1, where 0 corresponds to low intelligibility and 1 corresponds to high intelligibility. Given the intelligibility results of a listening test and the corresponding values computed by an intelligibility metric, a *transfer function* that relates intelligibility to the intelligibility metric can be determined by fitting a curve to the data. In this way, the intelligibility of a signal that was not included in the listening test can be predicted. Figure 2.12 demonstrates this procedure.

For a given stimuli and degradation, an intelligibility metric always computes the same number, whereas the intelligibility of the degraded stimuli may vary depending on the experimental procedures of the listening test. For example, if the stimuli and listening environment are held constant and listeners are given a list of possible words to select from, then intelligibility will be higher than if no list is provided, however, the score computed by the intelligibility metric will be constant because the acoustic signals have not changed. Consequently, the transfer function also depends on the experimental procedures. This means that an intelligibility metric cannot be used to predict intelligibility without having at least some data from an intelligibility listening test.

In many cases, intelligibility metrics are useful even when a transfer

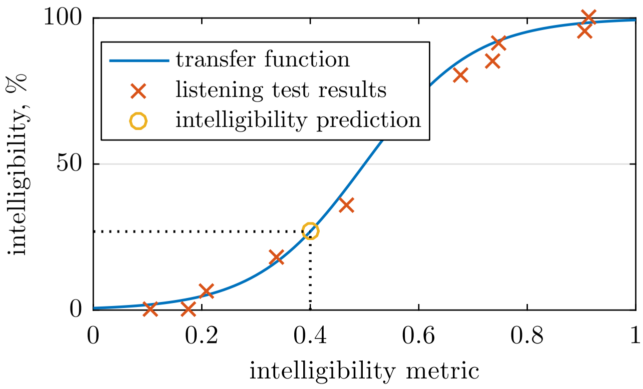

Figure 2.12: A transfer function relating scores from an instrumental intelligibility metric to intelligibility results from a listening test. The ten red crosses could be obtained by playing a stimuli on a loudspeaker in ten different listening environments and measuring the intelligibility using a standardised listening test. For each listening condition, an intelligibility metric could also compute a score. Given the ten measurements of intelligibility and the corresponding scores from an intelligibility metric, a transfer function can be fit to the data. The transfer function and intelligibility metric can then be used to predict the intelligibility of stimuli that were not included in the listening test.

function is not provided. If a communication system is modified and scores from an intelligibility metric increase, then an increase in intelligibility can also be expected. However, without any listening test data, it is not possible to predict exactly how much the intelligibility will increase by.

2.5.4 Articulation index

The development of one of the first intelligibility metrics begun in the 1920's and is called the articulation index (AI). The AI is the outcome of a series of experiments conducted by Harvey Fletcher. The aim of the experiments was to investigate the effect of additive noise and filtering on speech intelligibility. To do so, the SNR and bandwidth of a communication channel were varied. Nonsense *consonant vowel consonant syllables* (CVC) were transmitted through the channel and the intelligibility in terms of the percentage of correctly identified CVC's was recorded.

One of the experiments (Fletcher and Galt, 1950) involved filtering speech with complimentary low-pass and high-pass filters, which split the acoustic speech signals into a low-band signal and a high-band signal. Fletcher set out to find a relationship between the wide-band and narrow-band intelligibility. Let s, s_L, and s_H denote the intelligibility of the wide-band signal, the low-band signal, and high-band signal, respectively. It was found that the data could be modelled by

$$\log(1-s) = \log(1-s_L) + \log(1-s_H), \tag{2.30}$$

or in terms of error probabilities

$$e = e_L e_H, \tag{2.31}$$

where $e = 1 - s$, $e_L = 1 - s_L$ and $e_H = 1 - s_H$. Eventually the analysis was extended from two frequency bands to F frequency bands resulting in the equation

$$e = e_1 e_2 \dots e_F, \tag{2.32}$$

where e_f denotes the intelligibility error for speech in the f'th frequency band. e_f is referred to as the *band articulation error*.

Using the above model, an intelligibility metric known as the articulation index (AI) was developed (French and Steinberg, 1947). The AI can be described by the function (Allen, 1994)

$$\mathrm{AI}(s) = \frac{\log(1-s)}{\log(e_{\mathrm{min}})}, \tag{2.33}$$

where e_{min} is the minimum intelligibility error under ideal listening conditions, (i.e., when there is no filtering or noise). The function is normalised by $\log(e_{\mathrm{min}})$ so that $\mathrm{AI}(s)$ is confined to the closed interval $0 \leq \mathrm{AI}(s) \leq 1$.

Intelligibility is at its minimum when $\mathrm{AI}(s) = 0$ and at its maximum when $\mathrm{AI}(s) = 1$. Substituting (2.30) into (2.33), gives

$$\mathrm{AI}(s) = \sum_{f=1}^{F} \mathrm{AI}(s_f), \tag{2.34}$$

where s_f denotes the intelligibility of speech in the f'th frequency band. Thus, the AI is a transformation of intelligibility such that each frequency band provides an independent contribution to the overall score.

French and Steinberg (1947) used AI theory to relate intelligibility to the SNR of each frequency band. Let $\boldsymbol{p}$ be a vector where each element p_f corresponds to the variance of the clean speech in a particular frequency band and similarly let $\boldsymbol{n}$ be a vector where each element n_f corresponds to the noise variance in a particular frequency band. The relationship between the band articulation error and SNR for frequency band f can be written as (Allen, 1994)

$$e_f = e_{\mathrm{min}}^{\overline{\mathrm{SNR}}_f/F} \tag{2.35}$$

where,

$$\overline{\mathrm{SNR}}_f = \begin{cases} 0, & \text{if } 10\log_{10}(p_f/n_f) < 0 \\ 10\log_{10}(p_f/n_f)/30, & \text{if } 0 \leq 10\log_{10}(p_f/n_f) \leq 30 \\ 1, & \text{if } 10\log_{10}(p_f/n_f) > 30, \end{cases} \tag{2.36}$$

is the normalised SNR for band f. From here on $\overline{\mathrm{SNR}}_f$ is referred to as an *audibility function*.

The audibility function clips and normalises the SNR so that $0 \leq \overline{\mathrm{SNR}}_f \leq 1$. The motivation for the audibility function is as follows. Experiments showed that the average listener possessed a 30 dB speech detection threshold. If the SNR of a speech signal was below 0 dB, then the signal would not be detected and would not contribute to the total intelligibility. On the other hand, when the SNR exceeded 30 dB, there was no noticeable increase in intelligibility.

Combining (2.32) with (2.35), we have that

$$
\begin{aligned}
e &= e_1 e_2 \dots e_F && (2.37)\\
&= e_{\text{min}}^{\overline{\text{SNR}}_1/F} e_{\text{min}}^{\overline{\text{SNR}}_2/F} \dots e_{\text{min}}^{\overline{\text{SNR}}_F/F} && (2.38)\\
&= e_{\text{min}}^{\frac{1}{F}\sum_{f=1}^{F}\overline{\text{SNR}}_f}. && (2.39)
\end{aligned}
$$

Taking the logarithm of (2.39) and rearranging, we have that:

$$
\begin{aligned}
\log(e) &= \log\big(e_{\text{min}}^{\frac{1}{F}\sum_{f=1}^{F}\overline{\text{SNR}}_f}\big), && (2.40)\\
\log(1-s) &= \big(\frac{1}{F}\sum_{f=1}^{F}\overline{\text{SNR}}_f\big)\log(e_{\text{min}}), && (2.41)\\
\frac{\log(1-s)}{\log(e_{\text{min}})} &= \frac{1}{F}\sum_{f=1}^{F}\overline{\text{SNR}}_f. && (2.42)
\end{aligned}
$$

Finally, applying (2.33) to (2.42), we have that

$$\text{AI}(s) = \frac{1}{F}\sum_{f=1}^{F}\overline{\text{SNR}}_f. \tag{2.43}$$

Thus, the AI can be computed as a normalised SNR averaged over F non-overlapping frequency bands.

Originally, 20 frequency bands were selected such that under ideal listening conditions each band contributed equally to the total intelligibility. Since then, for practical reasons, sets of frequency bands that do not contribute equally to intelligibility have been proposed. Consequently, *band-importance functions* (BIF) that describe the relative importance of each frequency band have been developed (Pavlovic, 1987). In this case, the AI becomes a weighted-average of the audibility functions. That is,

$$\text{AI}(s) = \sum_{f=1}^{F} w_f \cdot \overline{\text{SNR}}_f, \tag{2.44}$$

where $0 \leq w_f \leq 1$ is the relative importance of frequency band f and satisfies $\sum_{f=1}^{F} w_f = 1$. The most commonly used formulation of the AI is (2.44).

In practical applications, $\boldsymbol{p}$ and $\boldsymbol{n}$ are computed by estimating the power spectral density (PSD) of the speech signal and the PSD of the noise signal. A popular algorithm for estimating the PSD is the Welch method (Welch, 1967). The variance of the speech signal in the fth frequency band, p_f, is found by integrating the speech PSD over the frequency values corresponding to the fth band. Likewise, n_f, is calculated from the noise PSD. (2.36) is then used to calculate the band audibility function and the AI is computed according to (2.44). The percentage of identifiable CVC's can then be predicted according to (2.39). Specifically,

$$s = 1 - e_{\mathrm{min}}^{\mathrm{AI}(s)}. \tag{2.45}$$

This procedure for calculating the AI and predicting intelligibility was standardised in ANSI (1969).

The AI has been shown to be successful at predicting intelligibility for noisy band-limited channels (Kryter, 1962b); however, the AI has a number of limitations. First, AI theory was developed using experiments based on nonsense CVC speech material. As discussed in Section 2.5.1 and Section 2.5.2, context effects and the complexity of the speech material can affect intelligibility. For a given SNR and bandwidth, the intelligibility of meaningful sentences is higher than the intelligibility of nonsense CVC's. The AI, as formulated in (2.44), has no functionality for adapting to different speech materials. Second, the AI is based on long-term statistics. Consequently, the AI cannot accurately account for distortions caused by noise sources that fluctuate over time such as competing talkers and wind (Rhebergen and Versfeld, 2005). Lastly, the AI cannot account for nonlinear distortions commonly introduced by enhancement algorithms and hearing-aids. This is because the SNR is not clearly defined in such scenarios (Loizou and Ma, 2011).

The articulation index and mutual information

As Allen (2005b) pointed out, it is interesting that the average in (2.43) is applied to a logarithmic measure of SNR rather than a linear measure. This suggests that the AI is closely related to the information capacity of a Gaussian communication channel.

Recall that the information capacity of a univariate Gaussian channel with signal variance p_f and noise variance n_f is (Shannon, 1948)

$$C(p_f, n_f) = \frac{1}{2}\log_2\left(1 + \frac{p_f}{n_f}\right). \tag{2.46}$$

For a single frequency band with $0 < 10\log_{10}(\frac{p_f}{n_f}) < 30$, the AI is given by

$$\begin{aligned}\mathrm{AI}(s_f) &= \overline{\mathrm{SNR}}_f &(2.47)\\ &= 10\log_{10}(p_f/n_f)/30 &(2.48)\\ &= \left(\frac{2\log_{10}2}{3}\right)\left(\frac{1}{2}\log_2\frac{p_f}{n_f}\right). &(2.49)\end{aligned}$$

Notice the similarity between (2.46) and (2.49). This leads to the approximation,

$$\mathrm{AI}(s_f) \approx \frac{2\log_{10}2}{3}C(p_f, n_f), \tag{2.50}$$

thus, the AI can be interpreted as a scaled channel capacity. Figure 2.13 plots $\mathrm{AI}(s_f)$ and the scaled $C(p_f, n_f)$ against the SNR. We see that the approximation is reasonable for $10\log_{10}(p_f/n_f) < 30$. However, the AI saturates at 1, whereas the information capacity increases without bound almost linearly with the SNR in dB.

Note that AI theory is analogous to the simple-time domain model of speech communication discussed in Section 2.3.1. Specifically, if the noise PSD is a scalar multiple of the speech PSD, then the AI in (2.43) is a scaled approximation of (2.28).

Finally, note that development of the AI began at Bell Labs in the 1920's, whereas information theory was not developed until about 1950, also at Bell Labs. Shannon's information theory can thus be viewed as

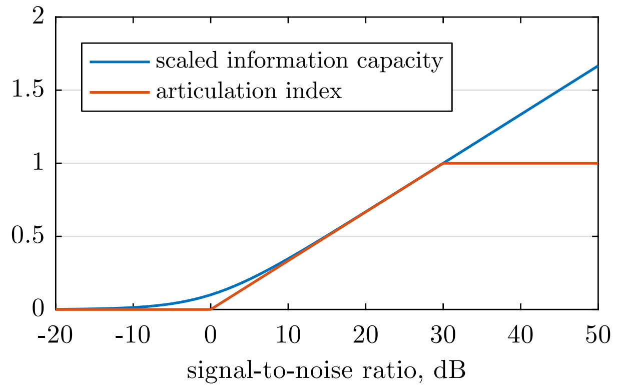

Figure 2.13: Comparison of the articulation index, $\mathrm{AI}(s_f)$, and the scaled capacity of a Gaussian channel, $C(p_f, n_f)$, for a single frequency band.

a more general and theoretical model of communication that is consistent with Fletcher's early empirical research on speech communication and the articulation index.

Speech intelligibility index

In the years following the initial standardisation of the AI (ANSI, 1969), new experiments were conducted that resulted in a revised version of the ANSI standard (ANSI, 1997a). Additionally a new name was given to the revised intelligibility metric: the speech intelligibility index (SII).

The SII extended the AI by modifying the audibility function (2.36) to account for self-masking, reverberation, and vocal effort. Additionally, new band-importance functions and transfer functions[4] were included so that intelligibility predictions could be made for speech materials other than nonsense CVC's. Overall the SII provides a more general framework than the AI.

[4] (2.45) is an example of a transfer function. See also Figure 2.12.

2.5.5 Speech intelligibility predictor based on mutual information

Jensen and Taal (2014) hypothesised that intelligibility is related to the mutual information between the clean and distorted temporal envelopes of narrow-band time-domain signals. This idea is subtly different to the AI, which effectively computes the mutual information between narrow-band time-domain signals, as opposed to their temporal envelopes. The motivation for using the temporal envelopes is that representing signals in this way approximates the signal processing of human auditory system. The resulting intelligibility metric was named SIMI and is described below.

SIMI approximates the signal processing of the human auditory system by splitting the acoustic signals into non-overlapping 1/3 octave frequency bands and extracting the temporal envelope for each frequency band. Let $\{\tilde{\mathrm{x}}_i\}$ be a real-valued stochastic process that represents the samples of a clean acoustic speech signal where i is the sample index. Similarly let $\{\tilde{\mathrm{y}}_i\}$ represent the samples of a distorted signal received by a listener. The short-time single-sided discrete Fourier Transform of $\{\tilde{\mathrm{x}}_i\}$ is a complex vector-valued random process denoted $\{\hat{\mathbf{x}}_t\}$ where $\hat{\mathbf{x}}_t \in \mathbb{C}^{\frac{N}{2}-1}$ has elements given by

$$\hat{\mathrm{x}}_{\omega,t} = \sum_{n=0}^{N-1} w_n \tilde{\mathrm{x}}_{\Delta t+n} e^{-j2\pi n\omega/N}. \tag{2.51}$$

Each vector element indexed by $\omega \in \{0, 1, \ldots, \frac{N}{2} - 1\}$ corresponds to a frequency bin, t is the frame index, $N \in \mathbb{E}_{++}$ is the frame length and also the discrete Fourier transform size, Δ is the step size, and w_n is an analysis window.

The temporal envelope for the f'th frequency band is computed according to

$$\mathrm{x}_{f,t} = \sqrt{\sum_{\omega\in\mathbb{F}_f} |\hat{\mathrm{x}}_{\omega,t}|^2} \tag{2.52}$$

where $\mathbb{F}_f$ is a set of frequency bin indices representing the f'th one-third octave band and $f \in \{1, 2, \ldots, F\}$ is the frequency band index. The clean temporal envelopes form a real vector-valued random processes denoted $\{\mathbf{x}_t\}$. The distorted temporal envelopes, $\{\mathbf{y}_t\}$, are defined similarly.

SIMI is based on the hypothesis that intelligibility is related to the mutual information rate of $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$. Assuming that the elements in $\{\mathbf{x}_t\}$ are statistically independent, and likewise for $\{\mathbf{y}_t\}$, the mutual information rate decomposes into a summation of mutual information terms:

$$I(\{\mathbf{x}_t\}; \{\mathbf{y}_t\}) = \frac{1}{FT} \sum_f \sum_t I(\mathrm{x}_{f,t}; \mathrm{y}_{f,t}), \tag{2.53}$$

where T is the sequence length.

SIMI combines a parametric model with a lower bound to estimate $I(\mathrm{x}_{f,t}; \mathrm{y}_{f,t})$. Specifically, if $\mathrm{x}_{f,t}$ is modelled as a Chi-distributed random variable with k' degrees of freedom, then it can be shown that

$$\begin{aligned} I(\mathrm{x}_{f,t}; \mathrm{y}_{f,t}) \geq \ln\Gamma(k'/2) &+ \frac{1}{2}\Big(k' - \ln 2 - (k'-1)\psi(k'/2)\Big) \\ &- \frac{1}{2} \ln 2\pi e(k' - 2\frac{\Gamma^2((k'+1)/2)}{\Gamma^2(k'/2)}) \\ &- \frac{1}{2} \ln(1 - \rho_{f,t}^2), \end{aligned} \tag{2.54}$$

where Γ and ψ denote the gamma and digamma function, respectively, and $\rho_{f,t}$ is the correlation coefficient between $\mathrm{x}_{f,t}$ and $\mathrm{y}_{f,t}$ given by

$$\rho_{f,t} = \frac{\mathbb{E}[\mathrm{x}_{f,t}\mathrm{y}_{f,t}] - \mathbb{E}[\mathrm{x}_{f,t}]\mathbb{E}[\mathrm{y}_{f,t}]}{\sqrt{(\mathbb{E}[\mathrm{x}_{f,t}^2] - \mathbb{E}[\mathrm{x}_{f,t}]^2)(\mathbb{E}[\mathrm{y}_{f,t}^2] - \mathbb{E}[\mathrm{y}_{f,t}]^2)}}. \tag{2.55}$$

In practice, SIMI uses a sampling rate of 10 kHz, a 256-point Hann analysis window, $\Delta = 128$, and $F = 15$ 1/3 octave-bands with centre frequencies between 150 Hz and 4.3 kHz. Additionally, silent frames (i.e., frames with energy less than 30 dB below the frame with maximum energy) are not included in the summation of (2.53). The time-varying moments in

(2.55) are computed using a single pole IIR filter. For example, $\mathbb{E}[\mathrm{x}_{f,t}\mathrm{y}_{f,t}]$ is estimated as

$$\hat{\mu}_{\mathrm{x}_{f,t}\mathrm{y}_{f,t}} = \alpha\hat{\mu}_{\mathrm{x}_{f,t-1}\mathrm{y}_{f,t-1}} + (1-\alpha)x_{f,t}y_{f,t}, \tag{2.56}$$

where $\alpha = 0.95$ corresponds to a time-constant of 250 ms. In Chapter 5 the performance of SIMI is evaluated for a wide range of real-world listening conditions.

2.5.6 K-nearest neighbour mutual information intelligibility measure

Taghia and Martin (2014) proposed the k-nearest neighbour mutual information intelligibility measure (MIKNN). Similarly to SIMI, MIKNN is based on the hypothesis that intelligibility is related to the mutual information between the clean and distorted temporal envelopes of narrowband time-domain signals. MIKNN uses the same representation of speech as SIMI. Concretely, (2.51) and (2.52) are used to extract the clean and distorted temporal envelopes, $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$.

The key difference between MIKNN and SIMI is that instead of using the parametric lower bound in (2.54), MIKNN uses a non-parametric mutual information estimator based on k-nearest neighbours (KNN) (Kraskov et al., 2004). One advantage of the KNN mutual information estimator is that it can account for non-linear dependencies between the clean and distorted temporal envelopes. This contrasts with (2.54), which is a function of Pearson's correlation coefficient only. Pearson's correlation coefficient cannot quantify non-linear dependencies.

A second advantage of the KNN mutual information estimator is that it directly estimates mutual information, as opposed to estimating a lower bound that may not be tight for every listening environment. However, the advantages of the KNN mutual information estimator come at the cost of additional computational complexity.

Similarly to SIMI, MIKNN assumes that the 1/3-octave bands are statistically independent and sums the mutual information of each frequency band. However, unlike SIMI, mutual information is not estimated on a short-time scale of 250 ms like in (2.56). Instead, mutual information is estimated using the entire utterance. In Chapter 5 the performance of MIKNN is evaluated for a wide range of real-world listening conditions.

2.6 Summary of Literature Review

This chapter reviewed the relevant literature regarding speech communication and information theory. First, the key ideas of Shannon's information theory were discussed: 1) all communication is probabilistic, and 2) a message of low probability contains more information than a message of high probability. Based on these concepts, the effectiveness of communication can be quantified using mutual information, which is a function of the joint probability distribution of the transmitted message and the received message.

The chapter then reviewed methods that use simple language models to estimate the information rate of speech communication. When Zipf's law is used to model the probability distribution of words, the lexical information rate of speech is 21 b/s. When modelling speech as a sequence of phonemes, the lexical information rate is 60 b/s. Both of these estimates assume that each lexical unit is selected for transmission independently from the previously transmitted lexical units. If the dependencies between successive words or phonemes were accounted for, then the rate would decrease.

Using the listening test results from Miller and Nicely (1955), this chapter estimated the information rate of phonemes as a function of signal-to-noise ratio and bandwidth. The mutual information increases with SNR almost linearly over a 12 dB range, but then begins to saturate. For band-limited speech, the sum of the information in the low-pass band and the high-pass band is greater than the information in the wide-band. This means that some information is shared between the frequency bands, i.e., there is ’frequency redundancy’.

Next, the chapter considered measuring the information rate of speech communication without knowledge of language. Instead of using language models, observations of acoustics signals were relied on. It was shown that modelling the time samples of an acoustic speech signal as a Gaussian process gives results that are not consistent with our understanding of linguistics. The Gaussian time-domain model predicts that the acoustic information rate of speech in ideal listening conditions is approximately 53000 b/s, and that two acoustic signals carrying the same lexical message only have 5 bits of information in common each second. Moreover, the acoustic information rate increases without bound with the environmental SNR. The reason these errors occur is because the time-domain samples of an acoustic signal do not accurately reflect the encoded lexical information. What is required is an appropriate representation of speech and a more powerful statistical model.

The chapter then presented Fano’s method for measuring the information rate of speech communication. Fano’s method is based on the source-filter theory of speech production. In this case, the information rate is determined by the number of carrier signals, the modulation bandwidth, and the SNR of the communication channel. Fano’s analysis results in an estimate of 2130 b/s, which is still 35 times larger than the lexical information rate. Fano hypothesised that the difference might be caused by talker variability.

Talker variability is the natural consequence of having many talkers with physiologically different vocal tracts and differing speech habits. This variability can be thought of as a type of production noise that is inherent in all speech communication. Production noise places a constraint on the maximum number of lexical code words that can be used without causing some acoustic sounds to be confused with others, and thus could limit the information rate.

Next, the chapter focused on intelligibility. Intelligibility is defined as the proportion of correctly identified words, and is an important characteristic of speech-based communication systems. Intelligibility can be measured using formal listening tests, but such tests are time-consuming to conduct. For this reason, algorithms that can predict intelligibility are of interest. Such algorithms are referred to as instrumental intelligibility metrics. One of the first intelligibility metrics was developed by Harvey Fletcher and is called the articulation index. It turns out that Shannon's information theory can be viewed as a generalisation of Fletcher's early empirical work.

Lastly, the chapter summarised two modern intelligibility metrics that are based on information theory: SIMI and MIKNN. Both of these intelligibility metrics use an auditory model to extract the temporal envelopes of acoustic speech signals and then estimate the mutual information between clean and degraded temporal envelopes. SIMI estimates mutual information using a parametric model, whereas MIKNN uses a non-parametric mutual information estimator. Neither of these intelligibility metrics account for talker variability or frequency redundancy, and thus are likely to overestimate the amount of information shared between a talker and a listener.

Chapter 3

A Simple Model of Speech Communication

Information theory provides mathematical tools for quantifying the effectiveness of communication. In this chapter, a simple model of speech communication that is based on information theory is developed. The model considers the transmission of a message from a talker to a listener. Speech signals are processed by an auditory model that accounts for the frequency and temporal masking of the cochlea and applies non-linear dynamic range compression. It is hypothesised that talker variability limits the maximum transfer of information between the talker and listener. Therefore, using real-world data, the variability between talkers is measured and used in combination with the proposed speech communication model to estimate the information rate of speech communication.

3.1 Model of Speech Communication

A message, $\{\mathbf{m}_t\}$, speech signal, $\{\mathbf{x}_t\}$, and degraded speech signal, $\{\mathbf{y}_t\}$, are represented by ergodic, stationary, discrete-time vector-valued stochastic processes where $t \in \mathbb{Z}$ is the time index. It is assumed that the outcomes of $\mathbf{m}_t$, $\mathbf{x}_t$, and $\mathbf{y}_t$ can be represented within $\mathbb{R}^D$. While this is not part of our formalism, which is not based on linguistics, the message may be thought of as a sequence of latent variables that represent, for example, a sequence of sentences, phonemes, or neural states[1]. The clean speech signal and the degraded speech signal could be represented by spectrograms, discrete-time acoustic waveforms, or the output of an auditory model.

The talker encodes the message into a speech signal according to a conditional probability distribution $P(\mathbf{x}_t|\mathbf{m}_t)$. In this way the variability of different talkers encoding the same message into different speech signals is incorporated into the model.

The speech signal is transmitted to a listener through a communication channel that may degrade the signal. Examples of degradation may include noise, reverberation, speech coding algorithms, and speech enhancement algorithms. Overall, the communication process at each t is described by a Markov chain:

$$\mathbf{m}_t \rightarrow \mathbf{x}_t \rightarrow \mathbf{y}_t. \tag{3.1}$$

$\mathbf{m}_t \rightarrow \mathbf{x}_t$ is called the *speech production channel* and $\mathbf{x}_t \rightarrow \mathbf{y}_t$ is called the *environmental channel*.

The Markov condition means that $\mathbf{m}_t$ and $\mathbf{y}_t$ are conditionally independent given $\mathbf{x}_t$. That is,

$$P(\mathbf{m}_t, \mathbf{y}_t|\mathbf{x}_t) = P(\mathbf{m}_t|\mathbf{x}_t)P(\mathbf{y}_t|\mathbf{x}_t). \tag{3.2}$$

Equivalently,

$$\begin{aligned} P(\mathbf{m}_t|\mathbf{y}_t, \mathbf{x}_t) &= \frac{P(\mathbf{m}_t, \mathbf{y}_t|\mathbf{x}_t)}{P(\mathbf{y}_t|\mathbf{x}_t)} \\ &= \frac{P(\mathbf{m}_t|\mathbf{x}_t)P(\mathbf{y}_t|\mathbf{x}_t)}{P(\mathbf{y}_t|\mathbf{x}_t)} \\ &= P(\mathbf{m}_t|\mathbf{x}_t). \end{aligned} \tag{3.3}$$

This means that all the information that $\{\mathbf{y}_t\}$ contains about $\{\mathbf{m}_t\}$ is obtained from $\{\mathbf{x}_t\}$.

[1] More details regarding the message are given in Section 3.1.3.

3.1.1 Information rate of the communication channel

The effectiveness of the communication channel is described by the mutual information rate between $\{\mathbf{m}_t\}$ and $\{\mathbf{y}_t\}$. Let

$$\mathbf{m}^K = [\mathbf{m}_1^*, \mathbf{m}_2^*, \cdots, \mathbf{m}_K^*]^* \tag{3.4}$$

be a vector obtained by stacking K consecutive message vectors and similarly for $\mathbf{x}^K$ and $\mathbf{y}^K$. The mutual information rate between $\{\mathbf{m}_t\}$ and $\{\mathbf{y}_t\}$ is defined by

$$I(\{\mathbf{m}_t\}; \{\mathbf{y}_t\}) = \lim_{K\to\infty} \frac{1}{K} I(\mathbf{m}^K; \mathbf{y}^K), \tag{3.5}$$

where $I(\mathbf{m}^K; \mathbf{y}^K)$ is the mutual information between $\mathbf{m}^K$ and $\mathbf{y}^K$. The mutual information is defined by

$$I(\mathbf{m}^K; \mathbf{y}^K) = \int P(\boldsymbol{m}^K, \boldsymbol{y}^K) \log_2 \frac{P(\boldsymbol{m}^K, \boldsymbol{y}^K)}{P(\boldsymbol{m}^K)P(\boldsymbol{y}^K)} d\boldsymbol{m}^K d\boldsymbol{y}^K, \tag{3.6}$$

where $P(\mathbf{m}^K, \mathbf{y}^K)$ is the joint probability distribution of $\mathbf{m}^K$ and $\mathbf{y}^K$, $P(\mathbf{m}^K)$ is the marginal probability distribution of $\mathbf{m}^K$, and $P(\mathbf{y}^K)$ is the marginal probability distribution of $\mathbf{y}^K$.

Note that the mutual information rate in (3.5) is defined using infinite length vectors. In practice, such vectors do not exist. Thus, a finite approximation must be made when estimating the mutual information rate given two sample sequences. Section 4.2.1 takes such an approach and Figure 4.5 demonstrates the effect of K.

An upper bound on the information rate

In some situations the integral in (3.6) could be intractable. This is mainly due to the difficulty of modelling $P(\mathbf{m}^K, \mathbf{y}^K)$, which varies depending on the environment of the talker and the listener. In this case, it can be easier to evaluate a bound. An upper bound for the mutual information rate can be obtained by applying the data processing inequality (Cover and Thomas, 2012) twice:

$$I(\{\mathbf{m}_t\}; \{\mathbf{y}_t\}) \leq I(\{\mathbf{m}_t\}; \{\mathbf{x}_t\}) \tag{3.7}$$

and

$$I(\{\mathbf{m}_t\};\{\mathbf{y}_t\}) \leq I(\{\mathbf{x}_t\};\{\mathbf{y}_t\}), \tag{3.8}$$

which leads to the upper bound

$$I(\{\mathbf{m}_t\};\{\mathbf{y}_t\}) \leq \min(I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}), I(\{\mathbf{x}_t\};\{\mathbf{y}_t\})). \tag{3.9}$$

The data processing inequality can be applied due to the Markov condition of (3.1). The advantage of (3.9) is that the speech production channel and the environmental channel are 'decoupled'. Thus, the effectiveness of each channel can be analysed separately and then combined together.

3.1.2 The univariate Gaussian channel

For illustrative purposes, consider the specific case where all processes are jointly Gaussian, univariate, memoryless, and stationary. In this case, the speech vector, $\mathbf{x}_t$, and the message vector, $\mathbf{m}_t$, are scalar random variables and thus are written as x_t and m_t, respectively.

Let p_t be *production noise* and n_t be *environmental noise*. For each time-step, t, let the following equations hold:

$$\mathrm{x}_t = \mathrm{m}_t + \mathrm{p}_t \tag{3.10}$$

$$\mathrm{y}_t = \mathrm{x}_t + \mathrm{n}_t. \tag{3.11}$$

Given (3.10) we have that $\mathrm{var}(\mathrm{x}) = \mathrm{var}(\mathrm{m}) + \mathrm{var}(\mathrm{p})$. Note that due to the stationary assumption, it is reasonable to remove the time subscripts. Furthermore, note that (3.10) and (3.11) satisfy the Markov condition in (3.1). This additive Gaussian model was first proposed as a model of speech communication in Kleijn and Hendriks (2015), where it was used to develop an algorithm to increase the intelligibility of speech degraded by additive environmental noise.

For the above communication model it can be shown that (Appendix A.1)

$$I(\{\mathrm{m}_t\};\{\mathrm{x}_t\}) = -\frac{1}{2}\log_2(1-\rho_{\mathrm{mx}}^2), \tag{3.12}$$

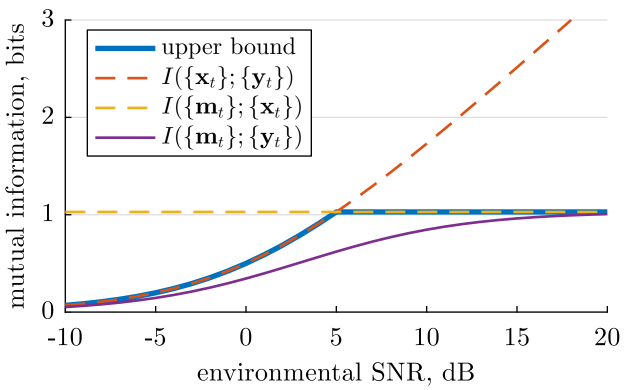

Figure 3.1: Mutual information for a univariate Gaussian channel with a production SNR of 5 dB.

$$I(\{\mathrm{x}_t\};\{\mathrm{y}_t\}) = -\frac{1}{2}\log_2(1-\rho_{\mathrm{xy}}^2), \tag{3.13}$$

and (Appendix A.2)

$$I(\{\mathrm{m}_t\};\{\mathrm{y}_t\}) = -\frac{1}{2}\log_2(1-\rho_{\mathrm{mx}}^2\rho_{\mathrm{xy}}^2), \tag{3.14}$$

where ρ_{mx} and ρ_{xy} are the correlation coefficients between m and x, and x and y, respectively, and are given by (Appendix A.3)

$$\rho_{\mathrm{mx}} = \sqrt{\frac{\mathrm{var}(\mathrm{m})/\mathrm{var}(\mathrm{p})}{1+\mathrm{var}(\mathrm{m})/\mathrm{var}(\mathrm{p})}} \tag{3.15}$$

and

$$\rho_{\mathrm{xy}} = \sqrt{\frac{\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})}{1+\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})}}. \tag{3.16}$$

The *production SNR*, $\mathrm{var}(\mathrm{m})/\mathrm{var}(\mathrm{p})$, can be thought of as a fixed value inherent to the nature of speech production. On the other hand, the *environmental SNR*, $\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})$, can vary with the environment of the talker and the listener.

It is informative to plot $I(\{\mathrm{m}_t\};\{\mathrm{y}_t\})$ against the environmental SNR for a fixed value of the production SNR. Figure 3.1 plots $I(\{\mathrm{m}_t\};\{\mathrm{y}_t\})$, $I(\{\mathrm{x}_t\};\{\mathrm{y}_t\})$, $I(\{\mathrm{m}_t\};\{\mathrm{x}_t\})$, and (3.9) against the environmental SNR for a production SNR of $10\log_{10}(\mathrm{var(m)}/\mathrm{var(p)}) = 5$ dB. We see that the usefulness of the communication channel saturates at the information rate of the speech production channel and that increasing the environmental SNR offers diminishing returns. Recall that this behavior is consistent with the lexical model of speech communication discussed in Section 2.2.3. The upper bound is tight for very low and very high values of environmental SNR. In Chapter 4, (3.9) is used to develop an intelligibility metric.

Interpretation noise

When Kleijn and Hendriks (2015) proposed (3.10) and (3.11) as a model of speech communication, a third equation was included:

$$\mathrm{v}_t = \mathrm{y}_t + \mathrm{i}_t, \tag{3.17}$$

where v_t is the received message perceived by the listener's brain and i_t is *interpretation noise*. This third equation is motivated by the observation that the same acoustic signal can be perceived differently by different listeners. Interpretation noise is particularly important when a listener has little experience with the talkers accent or when hearing-impaired listeners are considered. This thesis focuses only on improving the modelling of the production channel and the environmental channel. The inclusion of an interpretation channel is left as a direction for future work.

3.1.3 Defining the message using the information bottleneck principle

In the previous sections the message $\{\mathbf{m}_t\}$ was presented as an abstract stochastic process. In the present section, the concept of a message is further developed.

As advocated in Van Kuyk et al. (2017), a natural way to define the message of a speech signal without prior knowledge of human language or neural science is to apply the *information bottleneck principle* (Tishby et al.,

1999) to a *chorus* of speech signals. The basic concept of the approach is to extract the information that is consistent between talkers who speak the same utterance.

Let $\{\mathbf{x}_t^{[1]}\}$, $\{\mathbf{x}_t^{[2]}\}$, ..., $\{\mathbf{x}_t^{[B]}\}$ denote B speech signals where each speech signal contains the same lexical information but is spoken by a different talker. The superscript $[b]$ denotes the talker. A chorus, $\{\boldsymbol{\chi}_t\}$, is a vector-valued random process created by concatenating the vectors of all B talkers at each time-step. That is,

$$\{\boldsymbol{\chi}_t\} = \{(\left[(\mathbf{x}^{[1]})^*, (\mathbf{x}^{[2]})^*, \ldots, (\mathbf{x}^{[B]})^*\right]^*)_t\}, \tag{3.18}$$

where we recall that $*$ denotes the transpose.

The message can be defined as a stochastic function of the chorus,

$$\{\mathbf{m}_t\} = \psi(\{\boldsymbol{\chi}_t\}) + \{\mathbf{u}_t\}, \tag{3.19}$$

where $\mathbf{u}_t$ is independently identically distributed multivariate Gaussian noise and ψ is a deterministic function that minimises the information bottleneck:

$$\underset{\psi}{\operatorname{argmin}} \;\; I(\{\mathbf{m}_t\}; \{\boldsymbol{\chi}_t\}) - \beta I(\{\mathbf{x}_t\}; \{\mathbf{m}_t\}). \tag{3.20}$$

The speech $\{\mathbf{x}_t\}$ in (3.20) is independently drawn from the same distribution that was used to generate speech signals for the chorus. That is, $\{\mathbf{x}_t\}$ is spoken by another talker that is not in the chorus. The noise, $\{\mathbf{u}_t\}$, ensures that the mutual information between the message and the chorus is bounded from above.

On the one hand, optimising the information bottleneck leads to a function ψ that creates a compressed description of the chorus as it minimises $I(\{\mathbf{m}_t\}; \{\boldsymbol{\chi}_t\})$. Minimising this term gives ψ a disincentive to simply accumulate in the message all speech signals in the chorus, or even to select a single speech signal as the message. On the other hand, optimising the information bottleneck maximises the information shared between the message and speech carrying the message, i.e., $I(\{\mathbf{x}_t\}; \{\mathbf{m}_t\})$, which ensures that the message encoded by a talker can be predicted from a speech signal. β is a positive Lagrange multiplier that controls the trade-off between compression and prediction.

The mutual information rate $I(\{\mathbf{x}_t\};\{\mathbf{m}_t\})$ cannot exceed the true information rate of speech communication as the speech $\{\mathbf{x}_t\}$ in (3.20) and the speech signals $\{\mathbf{x}_t^{[1]}\}$, $\{\mathbf{x}_t^{[2]}\}$, . . ., $\{\mathbf{x}_t^{[B]}\}$ in the chorus $\{\boldsymbol{\chi}_t\}$ are independently drawn from the same conditional probability distribution. Thus, over-weighting $I(\{\mathbf{x}_t\};\{\mathbf{m}_t\})$ (i.e., $\beta >> 1$) does not result in an increase in its value. Such over-weighting of the second term also prevents ψ from being the trivial function that always maps to zero.

Due to the invariance of mutual information to reparameterisation, i.e., (2.17), the optima of (3.20) for a fixed β is not unique. Any one-to-one transformation of a message sequence would result in another valid message sequence. This is not unreasonable as a given utterance can be represented in many forms. For example, a particular utterance could be equally well described as a sequence of phonemes, a sequence of words, or a sequence of letters. Moreover, a given utterance can be represented in many different languages.

In summary, the message is a stochastic process that contains features that are consistent between talkers who speak the same utterance and does not contain features that are not shared between talkers who speak the same utterance (e.g., phase, timbre, loudness, and pitch). Unfortunately, (3.20) is difficult to solve due to the high number of dimensions of $\{\boldsymbol{\chi}_t\}$ along with the complicated statistical dependencies between the speech signals. For this reason, in this thesis, rather than explicitly solving (3.20), knowledge of the speech production process, the auditory periphery, and several simplifying assumptions are relied on to estimate the information rate of speech communication without relying on observations of $\mathbf{m}_t$. In particular, the remainder of Chapter 3 and Chapter 4 develop a methodology that assumes that $\mathbf{m}_t$ and $\mathbf{x}_t$ are jointly Gaussian, and Chapter 6 develops a methodology that uses a *Siamese neural network* (Bromley et al., 1994; Chopra et al., 2005).

3.2 An Auditory Representation of Speech

In the previous sections the clean speech $\{\mathrm{x}_t\}$ and the degraded speech $\{\mathbf{y}_t\}$ were presented as abstract stochastic processes. In the present section, $\{\mathrm{x}_t\}$ and $\{\mathbf{y}_t\}$ are made more concrete. To do so, acoustic speech signals are processed by a simple auditory model that approximates the frequency resolution, temporal resolution, and dynamic range compression of the human auditory system. The output of the auditory model is a sequence of *auditory log-spectra*. $\{\mathrm{x}_t\}$ represents a sequence of auditory log-spectra for a clean speech signal and $\{\mathbf{y}_t\}$ represents a sequence of auditory log-spectra for a degraded speech signal.

It is well known that the human auditory system is a lossy system; that is, it discards information from acoustic signals (Dau et al., 1996). This fact can be easily verified by an experiment where listeners are asked to distinguish between two tones, one at frequency f and another at frequency $f + \delta$. If δ is small enough, then the two tones are not perceptually different from one another. This is one reason why it is important to consider the effect of the auditory system when developing a model of speech communication. In the following, equations for computing $\{\mathrm{x}_t\}$ and $\{\mathbf{y}_t\}$ are described. More sophisticated auditory models exist (e.g., Dau et al., 1996; Kates and Arehart, 2014; Lyon, 2017), but the additional accuracy comes at the expense of additional computation.

Computing the short-time Fourier transform

Let $\{\tilde{\mathrm{x}}_i\}$ be a real-valued random process that represents the samples of an acoustic speech signal where i is the sample index and let $\{\hat{\mathrm{x}}_t\}$ be the short-time single-sided discrete Fourier transform (STFT) of $\{\tilde{\mathrm{x}}_i\}$ where

$\hat{\mathbf{x}}_t \in \mathbb{C}^{\frac{N}{2}-1}$ has elements given by

$$\hat{\mathrm{x}}_{\omega,t} = \sum_{n=0}^{N-1} w_n \tilde{\mathrm{x}}_{\Delta t+n} e^{-j2\pi n\omega/N}. \tag{3.21}$$

Each vector element indexed by $\omega \in \{0, 1, \dots, \frac{N}{2} - 1\}$ corresponds to a frequency bin, t is the frame index, $N \in \mathbb{E}_{++}$ is the frame length and also the discrete Fourier transform size, Δ is the step size, w_n is an analysis window, and j is the imaginary unit. In this thesis, $N = 400$, $\Delta = 200$, and a sampling rate of $f_s = 16$ kHz is used. These values correspond to a frame length of 25 ms and a frame rate of $R = 80$ frames/s. The periodic Hann window defined by $w_n = \frac{1}{2}(1 - \cos\frac{2\pi n}{N})$ is used. The rationale for the above STFT parameter values is given in Appendix B.

The *spectrogram*, denoted $\{|\hat{\mathbf{x}}|_t^2\}$, is obtained from the STFT by computing the squared magnitude of each element in $\{\hat{\mathbf{x}}_t\}$. That is,

$$|\hat{\mathbf{x}}|_t^2 = \left[|\hat{\mathrm{x}}_{0,t}|^2, \dots, |\hat{\mathrm{x}}_{\omega,t}|^2, \dots, |\hat{\mathrm{x}}_{N/2-1,t}|^2\right]^*. \tag{3.22}$$

Applying an auditory filterbank

Let $\{\mathbf{x}'_t\}$ represent an *auditory spectrogram* of $\{\tilde{\mathrm{x}}_i\}$, where $\mathbf{x}'_t \in \mathbb{R}^F$ is given by

$$\mathbf{x}'_t = \ln \boldsymbol{G}|\hat{\mathbf{x}}|_t^2. \tag{3.23}$$

Here, the logarithm is applied element-wise and $\boldsymbol{G} \in \mathbb{R}^{F\times\frac{N}{2}-1}$ represents an auditory filterbank where each row corresponds to the squared magnitude frequency response of an auditory filter. The logarithm roughly approximates the non-linear dynamic range compression of the human cochlea (Lyon, 2017). In this thesis, gammatone filters evenly spaced on the equivalent rectangular bandwidth (ERB) rate scale are used (Slaney, 1993). The ERB gammatone filterbank roughly models the frequency response of the human cochlea. The squared magnitude frequency response of the f'th filter is well approximated by (Holdsworth et al., 1988)

$$G_{f,\omega} = \frac{1}{\left(1 + \frac{(F_\omega - c_f)^2}{a^2 b_f^2}\right)^\eta} \tag{3.24}$$

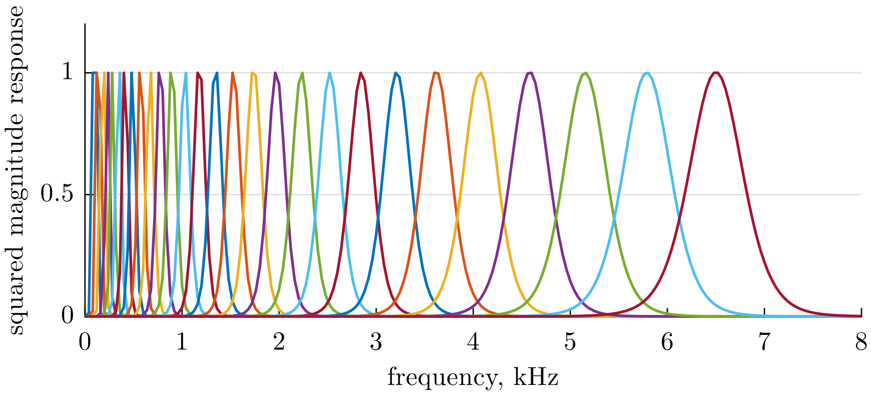

Figure 3.2: Gammatone filters evenly spaced on the ERB-rate scale.

where $\eta = 4$ is the filter order, F_ω is the frequency in Hz that corresponds to the frequency index ω, c_f is the centre frequency of the f'th filter in Hz, $b_f = 24.7(0.00437c_f + 1)$ is the bandwidth of the f'th filter in Hz, and $a = \frac{(\eta-1)!^2}{\pi(2\eta-2)!2^{-(2\eta-2)}}$ is a normalisation factor. Given minimum and maximum centre frequencies, c_1 and c_F, the centre frequencies for the remaining filters are computed according to

$$\log_{10} c_f = \frac{c_f^{\mathrm{ERBS}}/21.4 - 1}{0.00437} \tag{3.25}$$

where the values for c_f^{ERBS} are evenly spaced between $21.4\log_{10}(0.00437c_1 + 1)$ and $21.4\log_{10}(0.00437c_F + 1)$. In this thesis $F = 28$, $c_1 = 100$ Hz, and $c_F = 6500$ Hz. F was selected using the ERB-rate scale (Slaney, 1993) and c_1 and c_F were selected such that the filters' squared magnitude responses span frequencies from 0-8 kHz, without causing the response of the first filter or the F'th filter to be severely truncated.

The resulting filterbank is shown in Figure 3.2. The asymmetric overlapping nature of the filterbank means that a signal located at a low frequency may mask a signal at a higher frequency. This phenomena of human hearing is referred to as *upwards frequency masking* (Wegel and Lane, 1924).

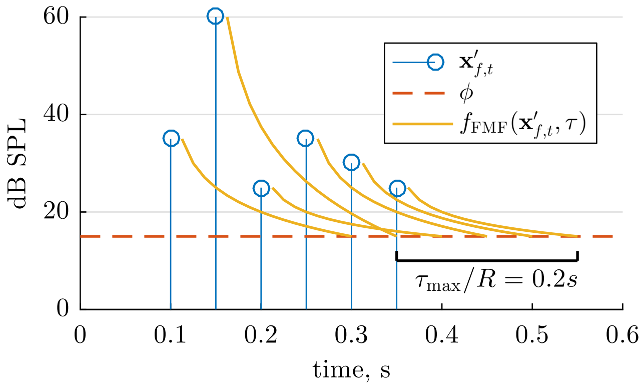

Figure 3.3: Forward temporal masking functions generated by a sequence of impulses. The third impulse, which is relatively quiet, is masked by the second impulse.

Applying a forward temporal masking function

Finally, to compute the auditory representation of speech used in this thesis, i.e., $\{\mathbf{x}_t\}$, a *forward temporal masking function* is applied to $\{\mathbf{x}'_t\}$. A forward temporal masking function (Oxenham, 2001) describes how a sound at time t can mask another sound at time $t + \tau$ where $\tau \geq 1$. This effect is important for degraded speech signals. For example, immediately after a gunshot, a listener will be unable to hear quiet speech sounds. Rhebergen et al. (2006) proposed the following forward masking function:

$$f_{\text{FMF}}(\mathbf{x}'_t, \tau) = \mathbf{x}'_t - \frac{\ln \tau}{\ln \tau_{\max}}(\mathbf{x}'_t - \boldsymbol{\phi}), \tag{3.26}$$

where $\boldsymbol{\phi} \in \mathbb{R}^F$ is the threshold of hearing for each frequency band and $\tau_{\max} = \lfloor 0.2R \rfloor$ is the maximum duration that one sound can mask subsequent sounds. Figure 3.3 displays the forward masking functions that result from several successive observations of $\mathbf{x}'_t$ for a single frequency band f. The forward masking functions decay logarithmically with time, but the rate of decay depends on the masking functions initial value. After $\tau_{\max}$ time steps, the function falls below the threshold of hearing and has no further effect.

To take forward temporal masking into account for all t, the maximum between x'_t and the forward masking functions produced by all previous inputs is computed:

$$\mathrm{x}_t = \max_{\tau=1,2,\ldots,\tau_{\max}} \big(f_{\mathrm{FMF}}(\mathrm{x}'_{t-\tau}, \tau), \mathrm{x}'_t\big). \tag{3.27}$$

In this way, unlike traditional exponential smoothing (i.e., a single pole infinite impulse-response filter), sharp onsets are followed instantaneously, but decay gradually. Note that for $\tau > \tau_{\max}$, all previous forward masking functions are below the threshold of hearing and thus have no effect. An 'overlap-max' algorithm can be used to efficiently compute (3.27) for all t by considering a sliding window of size $\tau_{\max}$.

For a degraded speech signal, $\{\tilde{\mathrm{y}}_i\}$, $\{|\hat{\mathbf{y}}|^2_t\}$, $\{\mathbf{y}'_t\}$, and $\{\mathbf{y}_t\}$, are defined analogously to their clean speech counterparts. Figure 3.4 displays examples of outcomes $\{\tilde{y}_i\}$, $\{|\hat{\boldsymbol{y}}|^2_t\}$, $\{\boldsymbol{y}'_t\}$, and $\{\boldsymbol{y}_t\}$. For Figure 3.4, a clean speech signal was degraded by additive white noise for t corresponding to 0 s to 2/3 s, a sinusoid for t corresponding to 2/3 s to 4/3 s, and a delta train for t corresponding to 4/3 s to 2 s. The upwards frequency masking property of the auditory model can be seen by examining $\{\boldsymbol{y}'_t\}$ from 2/3 s to 4/3 s, and the effect of the forward temporal masking function can be seen by examining $\{\boldsymbol{y}_t\}$ from 4/3 s to 2 s. $\{\mathrm{x}_t\}$ and $\{\mathrm{y}_t\}$ are referred to as sequences of auditory log-spectra.

3.3 Estimating the Information Rate of the Speech Production Channel

In this section, the information rate of the speech production channel $I(\{\mathbf{m}_t\}, \{\mathrm{x}_t\})$ is estimated by combining the speech communication model described in Section 3.1 with the auditory model described in Section 3.2. Recall that x_t is represented as temporally smoothed auditory log-spectra defined in (3.27) and $\{\mathbf{m}_t\}$ represents the underlying message of the speech signal.

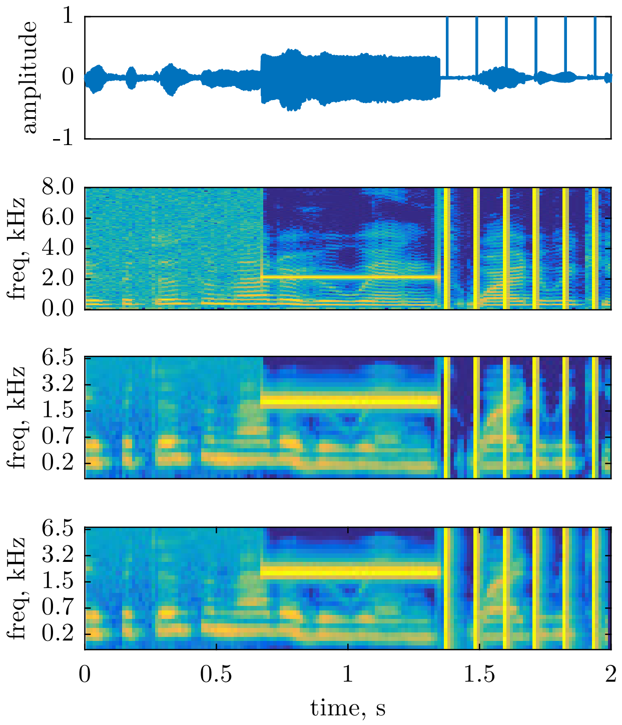

Figure 3.4: Comparison of speech representations for a speech signal degraded by additive white noise, a sinusoid, and a delta train, consecutively. From top to bottom: samples of a degraded time-domain acoustic signal $\{\tilde{y}_i\}$, the spectrogram $\{|\hat{\boldsymbol{y}}|_t^2\}$, the signal after applying an ERB gammatone filterbank $\{\boldsymbol{y}'_t\}$, and lastly the output of the auditory model $\{\boldsymbol{y}_t\}$, which includes a forward temporal masking function.

To estimate the information rate of the speech production channel, it is necessary to make further assumptions about the joint distribution $P(\mathbf{m}^K, \mathbf{x}^K)$. In this section the assumptions are described. In Section 3.3.3 and Section 3.3.4 the validity and limitations of the assumptions are discussed.

First, it is assumed that all processes are memoryless. That is, $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$ are statistically independent for all $\tau \neq 0$, and likewise for $\mathbf{m}_t$ and

$\mathbf{m}_{t+\tau}$. Equivalently,

$$P(\mathbf{x}_t, \mathbf{x}_{t+\tau}) = P(\mathbf{x}_t)P(\mathbf{x}_{t+\tau}) \tag{3.28}$$

and

$$P(\mathbf{m}_t, \mathbf{m}_{t+\tau}) = P(\mathbf{m}_t)P(\mathbf{m}_{t+\tau}), \tag{3.29}$$

for all t. In this case, the mutual information rate simplifies to the mutual information:

$$\begin{aligned} I(\{\mathbf{m}_t\}; \{\mathbf{x}_t\}) &= \lim_{K\to\infty} \frac{1}{K} I(\mathbf{m}^K; \mathbf{x}^K) && (3.30) \\ &= \lim_{K\to\infty} \frac{1}{K} \sum_{t=1}^{K} I(\mathbf{m}_t; \mathbf{x}_t) && (3.31) \\ &= \lim_{K\to\infty} \frac{1}{K} K I(\mathbf{m}_t; \mathbf{x}_t) && (3.32) \\ &= I(\mathbf{m}_t; \mathbf{x}_t). && (3.33) \end{aligned}$$

(3.31) follows from the memoryless assumption and (3.32) follows from the stationary assumption made in Section 3.1.

Second, it is assumed that the vector elements of $\mathbf{x}_t$ (i.e., the ERB frequency bands) are statistically independent, and likewise for $\mathbf{m}_t$. In this case, the mutual information decomposes into a summation over the vector elements:

$$I(\mathbf{m}_t; \mathbf{x}_t) = \sum_{f=1}^{F} I(\mathrm{m}_{f,t}, \mathrm{x}_{f,t}). \tag{3.34}$$

Third, let us consider the nature of speech production. As discussed in Section 2.3.2, the production of an acoustic speech signal can be modelled by the convolution of a vocal-tract filter impulse response and an excitation signal. Furthermore, recall from Section 2.4.2 that one of the main causes of talker variability is physiological differences between vocal tracts. Under such a speech production model it is natural to model production noise as convolutional noise in the time-domain, which implies that production noise is multiplicative in the frequency-domain. Furthermore, because (3.23) applies a logarithm, for the representation of speech considered in this thesis, it is natural to model production noise, $\mathbf{p}_t$, as additive zero-mean noise. That is,

$$\mathbf{x}_t = \mathbf{m}_t + \mathbf{p}_t, \tag{3.35}$$

where $\mathbf{m}_t$ and $\mathbf{p}_t$ are statistically independent. This is a major divergence from the preliminary model proposed by Kleijn and Hendriks (2015), which modelled production noise as additive in the time-domain and additive in the frequency-domain.

The final assumption is that $\mathbf{x}_t$, $\mathbf{m}_t$, and $\mathbf{p}_t$ are multivariate Gaussian random variables. Specifically, $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_\mathrm{x}, \boldsymbol{R}_\mathrm{x})$, $\mathbf{m}_t \sim \mathcal{N}(\boldsymbol{\mu}_\mathrm{m}, \boldsymbol{R}_\mathrm{m})$, and $\mathbf{p}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}_\mathrm{p})$. Because of the assumption that the vector elements are statistically independent, the covariance matrices must be diagonal matrices. That is, $\boldsymbol{R}_\mathrm{x} = \mathrm{diag}(\mathrm{var}(\mathbf{x}_t))$ and likewise for $\boldsymbol{R}_\mathrm{m}$ and $\boldsymbol{R}_\mathrm{p}$.

Given the above set of assumptions, the mutual information rate of the speech production channel is

$$\begin{aligned} I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}) &= I(\mathbf{m}_t;\mathbf{x}_t) \\ &= \sum_{f=1}^{F} I(\mathrm{m}_{f,t}, \mathrm{x}_{f,t}) \\ &= -\sum_{f=1}^{F} \frac{1}{2}\log_2(1 - r_f^2). \end{aligned} \tag{3.36}$$

We call r_f the *speech production correlation coefficient* for frequency band f. The speech production correlation coefficient describes the efficiency of encoding a message into a speech signal as determined by the natural variation between the vocal tracts of different talkers. Using the model of speech communication developed thus far, it can be shown that the speech production correlation coefficient for frequency band f is given by (Appendix A.3):

$$r_f^2 = \frac{\mathrm{var}(\mathrm{x}_{f,t}) - \mathrm{var}(\mathrm{p}_{f,t})}{\mathrm{var}(\mathrm{x}_{f,t})}. \tag{3.37}$$

Note that due to the stationary assumption, r_f^2, $\mathrm{var}(\mathrm{x}_{f,t})$, and $\mathrm{var}(\mathrm{p}_{f,t})$ are constant for all t. Consequently, in the following section $\mathrm{var}(\mathrm{x}_{f,t})$ and $\mathrm{var}(\mathrm{p}_{f,t})$ are replaced by $\mathrm{var}(\mathrm{x}_f)$ and $\mathrm{var}(\mathrm{p}_f)$, respectively.

3.3.1 Estimating the statistics of speech and production noise

In order to estimate the information rate of the speech production channel, $\mathrm{var}(\mathrm{x}_f)$ and $\mathrm{var}(\mathrm{p}_f)$ need to be estimated for each frequency band. To do so we rely on a chorus, $\{\boldsymbol{\chi}_t\}$, which was defined in Section 3.1.3. A chorus is a collection of B speech signals, $\{\mathbf{x}_t^{[1]}\}$, $\{\mathbf{x}_t^{[2]}\}$, ..., $\{\mathbf{x}_t^{[B]}\}$, where each signal contains the same lexical information, but is spoken by a different talker. The superscript $[b]$ denotes the talker. At each time step $\mathbf{x}_t^{[b]}$ is independently sampled from $P(\mathbf{x}_t|\mathbf{m}_t = \boldsymbol{m}_t)$ for some message $\{\boldsymbol{m}_t\}$.

Note that because production noise is modelled as zero-mean, at each time-step the expected value of the chorus over the talkers is $\boldsymbol{m}_t$. Thus, an estimate of the message can be easily obtained and then subtracted from each speech signal in the chorus to obtain an estimate of the production noise. In practice, the sample mean is used instead of the expected value operator. Concretely, the production noise of talker b at time t can be estimated as

$$\hat{\boldsymbol{p}}_t^{[b]} = \boldsymbol{x}_t^{[b]} - \frac{1}{B-1} \sum_{\beta=1,\ \beta \neq b}^{B} \boldsymbol{x}_t^{[\beta]}. \tag{3.38}$$

Removing the b'th observation from the sample mean results in an unbiased estimator.

Given a chorus, $\mathrm{var}(\mathrm{p}_f)$ can be estimated as

$$\hat{\sigma}_{\mathrm{p}_f}^2 = \frac{B-1}{B} \frac{1}{TB} \sum_{b=1}^{B} \sum_{t=1}^{T} \left(\hat{p}_{f,t}^{[b]} \right)^2, \tag{3.39}$$

where T is the sequence length of the speech signals in the chorus and $\frac{B-1}{B}$ is a bias reduction factor. The bias exists because $\hat{p}_{f,t}^{[b]}$ is used instead of the true production noise $p_{f,t}^{[b]}$.

Similarly $\mathrm{var}(\mathrm{x}_f)$ can be estimated as

$$\hat{\sigma}^2_{\mathrm{x}_f} = \frac{1}{TB-1} \sum_{b=1}^{B} \sum_{t=1}^{T} \left(x^{[b]}_{f,t} - \frac{1}{TB} \sum_{b=1}^{B} \sum_{t=1}^{T} x^{[b]}_{f,t} \right)^2 . \tag{3.40}$$

For both (3.39) and (3.40) it is reasonable to estimate the variance by summing over t due to the ergodic assumption from Section 3.1.

3.3.2 Implementation

In this section, an experiment for estimating the information rate of the speech communication channel is described. To do so, the estimators from the previous section are applied to real-world data.

Speech corpus

To estimate production noise, data from the CHAINS speech corpus was used. The CHAINS corpus includes easy reading material spoken by $B = 36$ talkers consisting of 16 females, and 20 males. 28 of the talkers were from Eastern Ireland, 3 were from the UK, and 5 were from the USA. The corpus contains recordings for six different speaking styles. We used the SOLO speaking style, where each talker read a prepared text at a comfortable rate. The speech material consisted of nine phonetically rich phrases and 24 TIMIT sentences. See Cummins et al. (2006) for more details.

Dynamic time-warping

For the production noise estimate of (3.38) to be valid, the speech signals in the chorus need to be time-aligned. Aligning the signals ensures that $\boldsymbol{m}_t$ is the same for each talker at each time-step. To align the utterances, a dynamic time-warping algorithm (Müller, 2007) was applied.

Given two sequences, $\{\boldsymbol{x}^{[1]}_t\}$ and $\{\boldsymbol{x}^{[2]}_t\}$, the goal of dynamic time-warping is to find a monotonic function, f_{dtw}, that compresses and

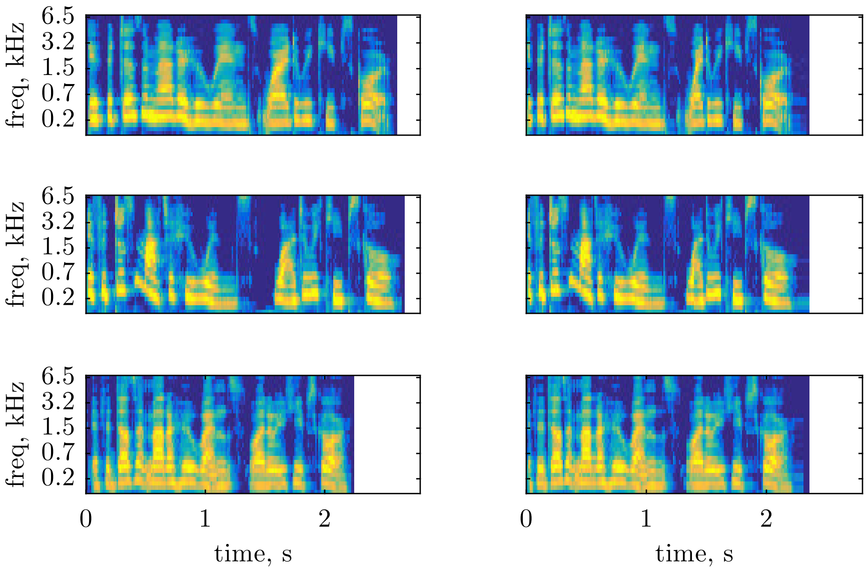

Figure 3.5: Left: sequences of auditory log-spectra for a sentence spoken by three different talkers. Each utterance contains the same sequence of phonemes, but the duration of each phoneme differs across the talkers. Right: the same sequences after using dynamic time-warping to align the phonemes.

stretches time such that the average Euclidean distance is minimised. That is,

$$\underset{f_{\text{dtw}}}{\arg\min} \; \frac{1}{T} \sum_{t} \|\boldsymbol{x}_t^{[1]} - \boldsymbol{x}_\tau^{[2]}\|_2, \tag{3.41}$$

where $\tau = f_{\text{dtw}}(t)$ is the time-warped time index and T is the sequence length. (3.41) can be efficiently solved using a linear programming algorithm (Müller, 2007).

For each sentence in the CHAINS corpus, the utterances spoken by different talkers were time-warped onto the utterance with the median duration. Doing so reduces the overall amount of time-warping required to align the utterances. After time-warping, for each talker, all utterances were concatenated together. This resulted in a sequence length of $T = 6690$ for each talker, which corresponds to 84 seconds of material for each talker.

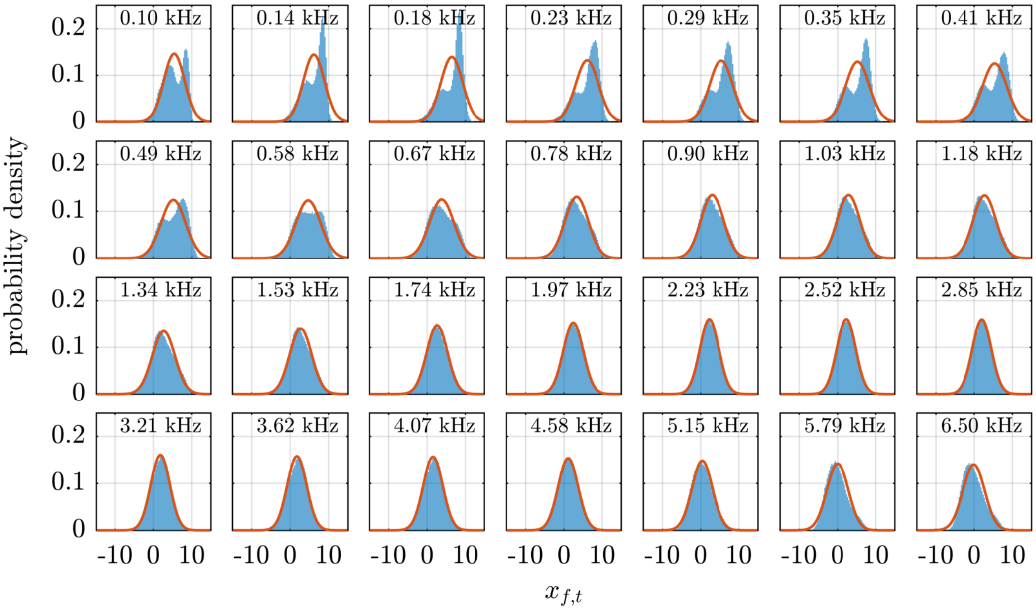

Figure 3.6: Histograms and maximum likelihood Gaussian density functions for the log-energy of clean speech x_f for each ERB frequency band. The centre frequency of each frequency band is indicated.

Figure 3.5 shows several utterances from the CHAINS corpus before and after applying dynamic time-warping.

3.3.3 Results

In order to check the validity of the Gaussian assumption from Section 3.3, Figure 3.6 plots histograms of $\boldsymbol{x}_t$ for each frequency band using data from the CHAINS corpus. Gaussian probability density functions that were obtained using maximum likelihood estimation are also shown. We see that $\mathrm{x}_{f,t}$ is approximately Gaussian, however, there are some differences, particularly for low frequency bands.

In the literature, (e.g., Gerkmann and Martin, 2009; Jensen and Taal, 2014), the temporal envelopes of speech signals are often modelled using right-skewed distributions such as chi and gamma distributions. How-

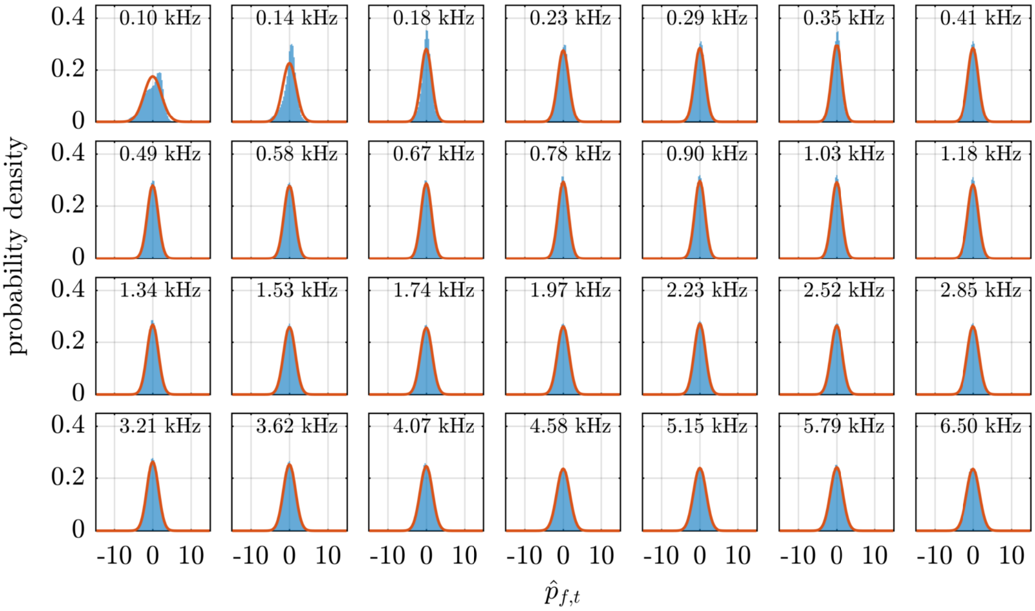

Figure 3.7: Histograms and maximum likelihood Gaussian density functions for production noise $\hat{\mathrm{p}}_f$ for each frequency band. The centre frequency of each frequency band is indicated.

ever, the logarithm in (3.23) transforms the right-skewed distribution into a distribution closer to a Gaussian distribution, which makes a Gaussian model for $\mathrm{x}_{f,t}$ a reasonable approximation.

Similarly, Figure 3.7 plots histograms of $\hat{\boldsymbol{p}}_t$ for each frequency band. Note that in (3.38), $\hat{\boldsymbol{p}}_t$ is calculated as a summation of $\boldsymbol{x}_t$ terms. The central limit theorem states that a sum of independent identically distributed random variables tends towards a Gaussian random variable (Navidi, 2008). This partially explains why the model of production noise developed in this chapter is well approximated by a Gaussian random variable.

Although it is reasonable to approximate $\mathrm{x}_{f,t}$ and $\mathrm{p}_{f,t}$ as Gaussian, in practice neither variable is truly Gaussian. All of the histograms shown in Figure 3.6 and Figure 3.7 fail the Kolmogorov-Smirnov test for normality (Massey, 1951) at the 5% significance level. For low frequency bands, the clean speech histograms are bimodal, and for the high frequency bands,

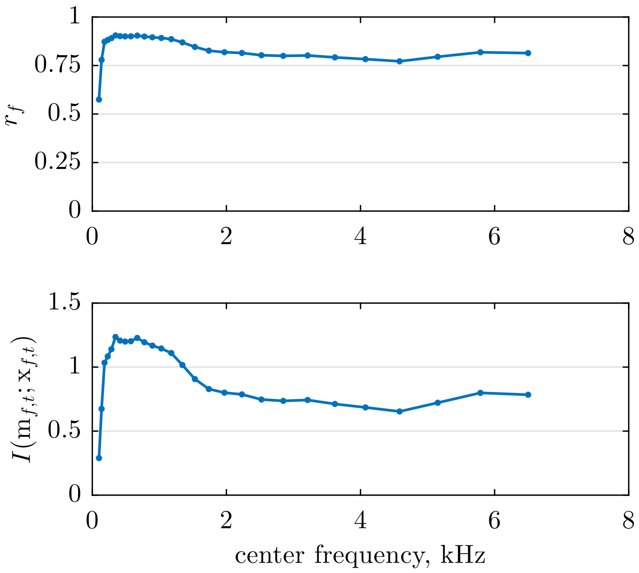

Figure 3.8: Top: the production noise correlation coefficient for each frequency band. Bottom: the mutual information rate for each frequency band.

the clean speech histograms have a small, but statistically significant, positive skew. Furthermore, the kurtosis of the production noise histograms is too large to be Gaussian. The observation that $\mathrm{x}_{f,t}$ and $\mathrm{p}_{f,t}$ are not truly Gaussian is worth keeping in mind, however, as shown in Chapter 5, the Gaussian model is sufficient for predicting the intelligibility of speech.

Figure 3.8 plots the speech production correlation coefficient and the mutual information rate of the speech production channel for the CHAINS corpus. Note that the speech production correlation coefficient is slightly higher than what we reported in Van Kuyk et al. (2016). This is because the experiment in Van Kuyk et al. (2016) used a simpler auditory model.

Summing over the frequency bands and multiplying by the frame rate (recall that $R = 80$), the data from Figure 3.8 gives an information rate of $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}) = 2070$ b/s. This is much lower than the acoustic information rate computed in Section 2.3.1, but an order of magnitude larger than the lexical information rate of 60 b/s from Section 2.2.2.

Note that our estimate is similar to that obtained by Fano's method in Section 2.3.2, even though our estimate considers production noise, whereas Fano's method does not. The reason for this is due to a discrepancy between the assumed bandwidth of the vocal-tract modulations. For the calculation in Section 2.3.2, a modulation bandwidth of 10 Hz was assumed. According to the Nyquist theorem, the minimum frame rate required for capturing such modulations is 20 Hz. However, the auditory model in this chapter uses a frame rate of 80 Hz. If the frame rate for the auditory model was reduced to 20 Hz, then our estimate of the information rate would also reduce by about a factor of four to approximately 500 b/s. Thus, accounting for production noise further closes the gap between the acoustic information rate of speech and the lexical information rate.

One may ask whether a modulation bandwidth of 10 Hz is reasonable. In Elliott and Theunissen (2009) log-spectrograms of speech signals were temporally filtered to artificially restrict the modulation bandwidth of the signals. Acoustic signals were then resynthesised and a listening test was conducted. It was found that intelligibility was significantly impaired only when modulations were restricted below 12 Hz, thus a 10 Hz modulation bandwidth is not unreasonable. Further evidence can be found in Kates and Arehart (2015), which used mutual information to determine the relative importance of different modulation frequencies for intelligibility, and found that modulation frequencies below 12.5 Hz were the most important.

The experiments of Elliott and Theunissen (2009) and Kates and Arehart (2015) imply that the auditory model in the present chapter may over-sample the signal. In Chapter 4 the statistical dependencies between successive speech frames are partially accounted for, in which case estimates of the information rate are insensitive to over-sampling.

3.3.4 Limitations of the communication model

The set of assumptions made in this chapter leads to a simple model of speech communication that is mathematically tractable and theoretically motivated. Unlike the models in Section 2.3, the information rate of the speech production channel saturates due to the inclusion of production noise. Additionally, the model further closes the gap between the lexical and acoustic information rate of speech.

In practice, however, some of the assumptions made in this chapter do not hold. Consider the fact that, on occasion, it is possible to predict the next word uttered by a talker before they say it. This implies that the message is not memoryless. Furthermore, due to the overlapping nature of the filterbank used in (3.23) and shown in Figure 3.2, the elements of $\mathbf{x}_t$ cannot be statistically independent. The main reason that the model developed thus far overestimates the information rate of speech is because statistical dependencies between the time-frequency units are not accounted for. In later sections of this thesis these issues are addressed and it is shown that despite the simplicity of the communication model, with a few improvements the model can accurately predict the intelligibility of speech signals for a wide range of distortions.

3.4 Summary of Chapter

This chapter developed a simple model of speech communication. The model considers the transmission of a message from a talker to a listener. There is a speech production channel, which describes the encoding of a message into a speech signal, and an environmental channel, which describes the transmission of a speech signal to a listener. All signals are modelled by ergodic stationary multivariate stochastic processes that represent sequences of auditory log-spectra.

To compute auditory log-spectra, acoustic speech signals are processed by an auditory model that accounts for the upwards frequency masking and forward temporal masking of the human auditory system. To account for upwards frequency masking, an ERB gammatone filterbank is used. To account for forward temporal masking, a function that responds to sudden changes instantaneously but decays gradually is used. In addition, logarithmic dynamic range compression is applied.

The communication model assumes that, for each frequency band, auditory log-spectra are distributed according to a Gaussian distribution. By inspecting the histograms of auditory log-spectra for each ERB frequency band, it was found that the Gaussian model provides a reasonable approximation, but in reality is not true. All of the histograms failed the Kolmogorov-Smirnov test for normality at the 5% significance level.

The speech production channel incorporates the effect of talker variability by modelling production noise as additive Gaussian noise. This is reasonable when considering a source-filter model of speech production.

Using real-word data, the information rate of the speech communication model was estimated. To do so, the variance of production noise was estimated using a chorus of speech signals that consisted of many talkers saying the same utterance. The experiment resulted in an estimate of the information rate of speech of the order of 500 b/s. This is not as low as the lexical information rate of speech, which is about 60 b/s, but is significantly closer to the lexical information rate than other methods that are based on acoustic speech signals. The remaining difference between the lexical information rate and the acoustic information rate is likely due to an assumption that the ERB frequency bands are statistically independent. In the following chapter this issue is addressed.

Chapter 4

An Intelligibility Metric Based on Information Theory

The previous chapter proposed a model of speech communication that consists of a speech production channel and an environmental channel. In the present chapter, methods for estimating the information rate of the environmental channel are discussed and the communication model is extended to partially account for time-frequency dependencies. This leads to a new intelligibility metric called *speech intelligibility in bits* (SIIB), and a variation called SIIB$^{\text{Gauss}}$. Both algorithms estimate the amount of information shared between a talker and a listener in bits per second.

Recall from Chapter 3 that $\{\mathbf{m}_t\}$ is the message, $\{\mathbf{x}_t\}$ is a clean speech signal produced by a talker and $\{\mathbf{y}_t\}$ is a degraded speech signal received by a listener. The present chapter uses the same representation of speech as that described in Section 3.2 (i.e., auditory log-spectra). In addition, the present chapter maintains the assumption that $\mathbf{m}_t$ and $\mathbf{x}_t$ are multivariate Gaussian random variables. However, unlike Chapter 3, the present chapter does not assume that $\{\mathbf{m}_t\}$, $\{\mathbf{x}_t\}$, or $\{\mathbf{y}_t\}$ are memoryless. Furthermore, the present chapter does not assume that the elements of $\mathbf{m}_t$ are statistically independent, and likewise for $\mathbf{x}_t$ and $\mathbf{y}_t$. Finally, recall from Section 3.1.1 that $\mathbf{m}^K$, $\mathbf{x}^K$, and $\mathbf{y}^K$ denote vectors obtained by stacking K

consecutive vectors of $\mathbf{m}_t$, $\mathbf{x}_t$, and $\mathbf{y}_t$, respectively.

4.1 Information Rate of the Communication Channel

The intelligibility metrics proposed in the present chapter are based on the hypothesis that intelligibility is a function of the mutual information rate between the message, $\{\mathbf{m}_t\}$, and the degraded speech, $\{\mathbf{y}_t\}$. Recall from Section 3.1.1 that the mutual information rate between the message and the degraded speech is given by

$$I(\{\mathbf{m}_t\};\{\mathbf{y}_t\}) = \lim_{K\to\infty} \frac{1}{K} I(\mathbf{m}^K;\mathbf{y}^K), \tag{4.1}$$

where $I(\mathbf{m}^K;\mathbf{y}^K)$ is the mutual information between $\mathbf{m}^K$ and $\mathbf{y}^K$ defined in (3.6).

In the previous chapter the information rate of the production channel $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\})$ was estimated. Note that for the production channel, the joint distribution $P(\mathbf{m}^K,\mathbf{x}^K)$ is fixed. That is, given a language, the human physiology, and the physics of our reality, $P(\mathbf{m}^K,\mathbf{x}^K)$ does not change. The same cannot be said about $P(\mathbf{m}^K,\mathbf{y}^K)$. For example, the marginal distribution $P(\mathbf{y}^K)$ for speech degraded by reverberation is different to the marginal distribution $P(\mathbf{y}^K)$ for noisy speech processed by an enhancement algorithm. Likewise, $P(\mathbf{m}^K,\mathbf{y}^K)$ also varies depending on the communication system and the environment of the talker and the listener. This makes the task of estimating $I(\{\mathbf{m}_t\};\{\mathbf{y}_t\})$ difficult. Instead of relying on a parametric model, bounds and non-parametric methods may be able to better handle the wide variety of possible degradation.

To estimate (4.1), realisations of $\{\mathbf{m}_t\}$ and $\{\mathbf{y}_t\}$ are required. In Section 3.3.1 it was shown that estimating a realisation of $\{\mathbf{m}_t\}$ requires a chorus of speech signals spoken by different talkers. In typical applications of intelligibility prediction, such a chorus is not available, so instead the upper bound from (3.9) is relied on:

$$I(\{\mathbf{m}_t\};\{\mathbf{y}_t\}) \leq \min(I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}), I(\{\mathbf{x}_t\};\{\mathbf{y}_t\})). \tag{4.2}$$

This upper bound effectively applies hard-clipping to $I(\{\mathbf{x}_t\};\{\mathbf{y}_t\})$ as determined by the natural variability between talkers.

Given a clean speech signal and a distorted speech signal, $I(\{\mathbf{x}_t\};\{\mathbf{y}_t\})$ can be estimated using a non-parametric mutual information estimator and $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\})$ can be computed using a parametric model of $P(\mathbf{m}^K, \mathbf{x}^K)$. The following sections describe this approach more concretely.

4.1.1 Information rate of the environmental channel

The mutual information rate of the environmental channel is given by

$$I(\{\mathbf{x}_t\};\{\mathbf{y}_t\}) = \lim_{K\to\infty} \frac{1}{K} I(\mathbf{x}^K;\mathbf{y}^K). \tag{4.3}$$

Estimating the mutual information between vectors of high dimensionality is a challenging task (Doquire and Verleysen, 2012) particularly when the vector elements have strong statistical dependencies (Gao et al., 2015). For this reason, an invertible transform q that aims to remove the dependencies between the vector elements is introduced.

Let $\tilde{\mathbf{x}}^K = q(\mathbf{x}^K)$ and $\tilde{\mathbf{y}}^K = q(\mathbf{y}^K)$. In the following it is assumed that the elements of $\tilde{\mathbf{x}}^K$ can be approximated as statistically independent, and likewise for $\tilde{\mathbf{y}}^K$. Then (4.3) can be decomposed into a summation:

$$\begin{aligned} I(\{\mathbf{x}_t\};\{\mathbf{y}_t\}) &= \lim_{K\to\infty} \frac{1}{K} I(\mathbf{x}^K;\mathbf{y}^K) && (4.4)\\ &= \lim_{K\to\infty} \frac{1}{K} I(\tilde{\mathbf{x}}^K;\tilde{\mathbf{y}}^K) && (4.5)\\ &= \lim_{K\to\infty} \frac{1}{K} \sum_{\lambda} I(\tilde{\mathrm{x}}^K_\lambda;\tilde{\mathrm{y}}^K_\lambda), && (4.6) \end{aligned}$$

where λ denotes the element index of the transformed vectors. (4.5) follows because q is invertible and (4.6) follows because q removes the statistical dependencies between the stacked-vector elements.

Finding an invertible q that simultaneously removes the dependencies in both $\mathbf{x}^K$ and $\mathbf{y}^K$ is difficult. Early speech recognition systems used the discrete cosine transform (DCT), which results in Mel-frequency cepstral coefficients (Davis and Mermelstein, 1980). It can be shown that the DCT approximates the Karhunen-Loève Transform (KLT) for stationary signals (Rao and Yip, 1990). The KLT is the transformation used for q in the present chapter and is given by:

$$\tilde{\mathbf{x}}^K = \boldsymbol{U}(\mathbf{x}^K - \mathbb{E}[\mathbf{x}^K]) \tag{4.7}$$

and

$$\tilde{\mathbf{y}}^K = \boldsymbol{U}(\mathbf{y}^K - \mathbb{E}[\mathbf{y}^K]), \tag{4.8}$$

where $\boldsymbol{U} \in \mathbb{R}^{KF \times KF}$ is a matrix with rows equal to the unit-magnitude eigenvectors of the covariance matrix of $\mathbf{x}^K$. In this context, the vector elements indexed by λ are referred to as *eigenchannels*. The KLT ensures that the elements of $\tilde{\mathbf{x}}^K$ are statistically uncorrelated, and if $\mathbf{x}_t$ is Gaussian, which is a reasonable approximation, then the elements are also statistically independent.

The KLT does not guarantee the same properties for $\tilde{\mathbf{y}}^K$ unless $\mathbf{y}^K$ is also Gaussian and has a covariance matrix with the same eigenvectors as the covariance matrix of $\mathbf{x}^K$. In practice, the environmental channel can result in non-Gaussian $\mathbf{y}^K$ or can introduce statistical dependencies in $\mathbf{y}^K$ that are not present in $\mathbf{x}^K$. An example of the latter is a reverberant channel, which increases the statistical time dependencies between successive auditory log-spectra. In this case, the statistical dependencies in the transmitted signal are accounted for by the KLT, but the statistical dependencies in the received signal are not accounted for. The consequence is that (4.6) underestimates the mutual information rate. To see this note that because the conditional differential entropy obeys

$h(\tilde{\mathbf{x}}^K|\tilde{\mathbf{y}}^K) = \sum_\lambda h(\tilde{\mathrm{x}}^K_\lambda|\tilde{\mathbf{y}}^K) \leq \sum_\lambda h(\tilde{\mathrm{x}}^K_\lambda|\tilde{\mathrm{y}}^K_\lambda)$, we have that

$$\begin{aligned} I(\tilde{\mathbf{x}}^K;\tilde{\mathbf{y}}^K) &= h(\tilde{\mathbf{x}}^K) - h(\tilde{\mathbf{x}}^K|\tilde{\mathbf{y}}^K) && (4.9)\\ &= \left(\sum_\lambda h(\tilde{\mathrm{x}}^K_\lambda)\right) - h(\tilde{\mathbf{x}}^K|\tilde{\mathbf{y}}^K) && (4.10)\\ &\geq \sum_\lambda h(\tilde{\mathrm{x}}^K_\lambda) - h(\tilde{\mathrm{x}}^K_\lambda|\tilde{\mathrm{y}}^K_\lambda) && (4.11)\\ &= \sum_\lambda I(\tilde{\mathrm{x}}^K_\lambda;\tilde{\mathrm{y}}^K_\lambda). && (4.12) \end{aligned}$$

Although the KLT does not meet all of the requirements for q, we have found that it is effective at *reducing* the dependencies and significantly improves the performance of SIIB and SIIB$^{\text{Gauss}}$. In fact, in Chapter 5 we show that the KLT can also improve the accuracy of other intelligibility metrics in the literature.

Recall that for some applications of the KLT it is common not to use all of the eigenchannels. Instead, only the eigenchannels with the largest eigenvalues are used. The rationale for doing so is that, for many real-world signals, a large proportion of the variance can be captured using a small number of basis vectors. This approach was taken when estimating the information rate of speech communication in Van Kuyk et al. (2017). That is, some of the eigenchannels were excluded from the summation in (4.6). However, during the development of SIIB it was found that retaining all of the eigenchannels led to more accurate predictions of intelligibility. Thus, SIIB does not attempt to reduce the number of dimensions by discarding eigenchannels.

Estimating mutual information

Equation (4.6) shows that the mutual information rate of the environmental channel can be decomposed into a summation of mutual information terms; one for each eigenchannel. The implementation of SIIB proposed in this thesis estimates each mutual information term in (4.6) using a popular mutual information estimator that was proposed by Kraskov et al.

(2004) and is based on k-nearest neighbours (KNN). However, combining the KLT with the KNN mutual information estimator is not the only way to estimate the mutual information rate of the environmental channel. In the following, several other mutual information estimators are discussed.

When developing the MIKNN intelligibility metric described in Section 2.5.6, Taghia and Martin (2014) considered fitting a Gaussian Mixture Model (GMM) to their representation of speech and then substituted the resulting joint probability distribution into the definition of mutual information. However, they report little benefit to this approach when compared with the KNN mutual information estimator.

Another approach for estimating mutual information is to use kernel-based estimators such as that proposed by Kolchinsky and Tracey (2017) and used in Kolchinsky et al. (2017). This type of mutual information estimator is differentiable, which would be advantageous for applications such as speech enhancement where mutual information could be optimised using gradient-based methods.

Another mutual information estimator was recently proposed by Belghazi et al. (2018) and is called *mutual information neural estimation* (MINE). MINE estimates mutual information by training a neural network via stochastic gradient ascent to maximise a lower bound of the mutual information. The lower bound is based on dual representations of Kullback-Leibler divergence. They report good results for high-dimensional data such as images, which may make MINE suitable for estimating the mutual information between clean and distorted auditory log-spectra. Methods for estimating mutual information other than the KNN mutual information estimator could result in a better algorithm for predicting intelligibility and is left as a direction for future work.

One final consideration when estimating mutual information is the sequence length. If the number of observations is too few, then the mutual information estimator may have significant bias or variance that could lead to a poor intelligibility prediction. For this reason it is worth pointing out that all of the stimuli used in the evaluation of SIIB in Chapter 5 have a duration of at least 20 s. The performance results in Chapter 5 suggest that SIIB is reliable for stimuli of such duration. For stimuli with much shorter duration, it could be that SIIB is unreliable, however, in most scenarios this outcome can be avoided simply by concatenating multiple short utterances into a single utterance of a longer duration.

4.1.2 Information rate of the speech production channel

Approximating $\{\mathbf{m}_t\}$ and $\{\mathbf{x}_t\}$ as jointly Gaussian multivariate processes, the information rate of the speech production channel is

$$\begin{aligned} I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}) &= \lim_{K\to\infty} \frac{1}{K} I(\mathbf{m}^K;\mathbf{x}^K) && (4.13) \\ &= \lim_{K\to\infty} \frac{1}{K} I(\tilde{\mathbf{m}}^K;\tilde{\mathbf{x}}^K) && (4.14) \\ &= \lim_{K\to\infty} \frac{1}{K} \sum_{\lambda=1}^{KF} I(\tilde{\mathrm{m}}_\lambda^K;\tilde{\mathrm{x}}_\lambda^K) && (4.15) \\ &= \lim_{K\to\infty} -\frac{1}{K} \sum_{\lambda=1}^{KF} \frac{1}{2}\log_2(1-r_\lambda^2), && (4.16) \end{aligned}$$

where $\tilde{\mathbf{m}}^K$ is defined similarly to $\tilde{\mathbf{x}}^K$ and r_λ is the production noise correlation coefficient from Chapter 3. The production noise correlation coefficient describes the efficiency of encoding a message into a speech signal according to $P(\mathbf{x}^K|\mathbf{m}^K)$.

In Chapter 3, r_f, described the correlation coefficient for each ERB frequency band, whereas in the present chapter r_λ is the correlation coefficient for each eigenchannel. Similarly to Chapter 3, r_λ could be estimated using a chorus of speech signals, however, for simplicity, in the present chapter we set $r_\lambda = 0.75$ for all λ. This value for r_λ was selected according to a line search from $r_\lambda = 0$ to $r_\lambda = 1$ in steps of 0.05. For each value of r_λ, the performance of SIIB was measured using the criteria and listening test data sets described in Chapter 5.

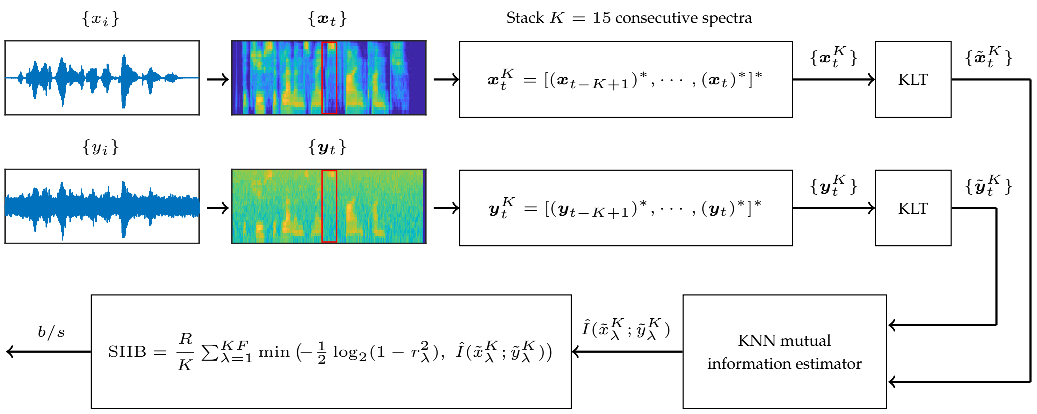
$\{x_i\}$
$\{\boldsymbol{x}_t\}$
Stack $K = 15$ consecutive spectra
$\boldsymbol{x}_t^K = [(\boldsymbol{x}_{t-K+1})^*, \cdots, (\boldsymbol{x}_t)^*]^*$
$\{\boldsymbol{x}_t^K\}$
KLT
$\{\tilde{\boldsymbol{x}}_t^K\}$
$\{y_i\}$
$\{\boldsymbol{y}_t\}$
$\boldsymbol{y}_t^K = [(\boldsymbol{y}_{t-K+1})^*, \cdots, (\boldsymbol{y}_t)^*]^*$
$\{\boldsymbol{y}_t^K\}$
KLT
$\{\tilde{\boldsymbol{y}}_t^K\}$
b/s
$\mathrm{SIIB} = \frac{R}{K}\sum_{\lambda=1}^{KF} \min\left(-\frac{1}{2}\log_2(1-r_\lambda^2),\ \hat{I}(\tilde{x}_\lambda^K;\tilde{y}_\lambda^K)\right)$
$\hat{I}(\tilde{x}_\lambda^K;\tilde{y}_\lambda^K)$
KNN mutual information estimator

Figure 4.1: Speech intelligibility in bits (SIIB)

4.2 Proposed Algorithms

4.2.1 SIIB

SIIB combines (4.3), (4.16), and (4.2) to give an estimate of the amount of information shared between $\{\mathbf{m}_t\}$ and $\{\mathbf{y}_t\}$ in bits per second. It is given by

$$\mathrm{SIIB} = \frac{R}{K}\sum_{\lambda=1}^{KF} \min\left(-\frac{1}{2}\log_2(1-r_\lambda^2), I(\tilde{\mathrm{x}}_\lambda^K;\tilde{\mathrm{y}}_\lambda^K)\right). \tag{4.17}$$

Recall that R is the frame rate in Hz, F is the number of ERB bands used by the auditory model, K is the number of stacked auditory log-spectra, and r_λ is the production noise correlation coefficient for the λth eigenchannel.

The implementation of SIIB used in this thesis is now described. An estimate of $I(\tilde{\mathrm{x}}_\lambda^K;\tilde{\mathrm{y}}_\lambda^K)$ for each eigenchannel is computed by applying a k-nearest neighbour mutual information estimator (Kraskov et al., 2004) to observed sample sequences $\{\tilde{x}_{\lambda,t}^K\}$ and $\{\tilde{y}_{\lambda,t}^K\}$. To obtain $\{\tilde{x}_{\lambda,t}^K\}$ and $\{\tilde{y}_{\lambda,t}^K\}$, a clean acoustic speech signal and a distorted acoustic signal are resampled to a sampling rate of $f_s = 16$ kHz. An energy-based voice activity detector with a 40 dB threshold is applied to remove silent segments. Subsequently, the acoustic signals are converted into sequences of auditory log-spectra,

$\{\boldsymbol{x}_t\}$ and $\{\boldsymbol{y}_t\}$, using the auditory model from Chapter 3. Specifically, (i) the acoustic signals are transformed to the STFT domain using a 400-point Hann window with 50% overlap, (ii) a gammatone filterbank that includes $F = 28$ filters linearly spaced on the ERB-rate scale between $c_1 = 100$ Hz and $c_F = 6.5$ kHz is used, and (iii) the forward masking temporal function from (3.26) is applied. This gives a frame rate of $R = 80$ Hz.

A sequence of stacked auditory log-spectra for the clean speech is formed by stacking $K = 15$ consecutive vectors:

$$\boldsymbol{x}_t^K = [(\boldsymbol{x}_{t-K+1})^*, (\boldsymbol{x}_{t-K+2})^*, \cdots, (\boldsymbol{x}_t)^*]^*. \tag{4.18}$$

$\boldsymbol{y}_t^K$ is defined similarly. Setting $K = 15$ means that time dependencies spanning 187.5 ms are considered. For comparison, the mean duration of a phoneme is 80 ms (Crystal and House, 1988). The sample covariance matrix of $\{\boldsymbol{x}_t^K\}$ is computed and the KLT in (4.7) and (4.8) is applied to obtain the sequences $\{\tilde{\boldsymbol{x}}_t^K\}$ and $\{\tilde{\boldsymbol{y}}_t^K\}$, which consist of the elements $\tilde{x}_{\lambda,t}^K$ and $\tilde{y}_{\lambda,t}^K$, respectively. A diagram of the above implementation of SIIB is shown in Figure 4.1.

4.2.2 SIIB$^{\text{Gauss}}$

SIIB$^{\text{Gauss}}$ is a variant of SIIB based on a fully parametric model. Rather than using a non-parametric mutual information estimator to estimate the mutual information of the environmental channel, it is assumed that $\{\mathrm{x}_t\}$ and $\{\mathrm{y}_t\}$ are jointly Gaussian multivariate processes. As pointed out in Section 4.1.1, this may not be a realistic assumption because the environment of the talker and listener can vary dramatically. For this reason SIIB$^{\text{Gauss}}$ can be viewed as a simplification of SIIB.

Given the above assumption, the upper bound in (4.2) is not required. Instead, (4.17) can be replaced by the capacity of a Gaussian channel:

$$\text{SIIB}^{\text{Gauss}} = -\frac{R}{2K}\sum_{\lambda=1}^{KF}\log_2(1 - r_\lambda^2\rho_\lambda^2), \tag{4.19}$$

where r_λ is the production noise correlation coefficient for each eigenchannel and ρ_λ is the correlation coefficient between $\tilde{\mathrm{x}}^K_{\lambda,t}$ and $\tilde{\mathrm{y}}^K_{\lambda,t}$ for each eigenchannel. SIIB$^{\text{Gauss}}$ computes $\{\tilde{\mathbf{x}}^K_t\}$ and $\{\tilde{\mathbf{y}}^K_t\}$ using the same auditory model and stacking procedure as SIIB, uses the same values for the parameters as SIIB, and also uses the KLT to reduce statistical dependencies.

Note that if the environmental channel were truly Gaussian and had independent eigenchannels, then $I(\{\mathbf{x}_t\};\{\mathbf{y}_t\}) = -\frac{R}{2K}\sum_\lambda \log_2(1-\rho^2_\lambda)$. In this case, SIIB in (4.17) can be written as

$$\frac{R}{2K}\sum_{\lambda=1}\min\big(-\log_2(1-r^2_\lambda),-\log_2(1-\rho^2_\lambda)\big).$$

Comparing this expression to (4.19), we see that for a Gaussian environmental channel SIIB applies hard-clipping, whereas SIIB$^{\text{Gauss}}$ essentially applies soft-clipping.

One advantage of SIIB$^{\text{Gauss}}$ is that it is based on the correlation coefficient and thus is fast to compute. Moreover, SIIB$^{\text{Gauss}}$ does not rely on an upper bound like SIIB. However, these advantages come at the cost of an additional assumption: that $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$ are jointly Gaussian, which may not be valid in general. In Chapter 5 the performance of SIIB and SIIB$^{\text{Gauss}}$ are compared.

4.3 Comparison with Existing Algorithms

Recall that Section 2.5 outlined two pre-existing intelligibility metrics called SIMI and MIKNN, which are both based on information theory. SIIB most closely resembles MIKNN because both of these metrics rely on a KNN mutual information estimator, whereas SIIB$^{\text{Gauss}}$ most closely resembles SIMI because SIIB$^{\text{Gauss}}$ is dependent on the correlation coefficient and SIMI is dependent on the 'short-time' correlation coefficient. There are several important differences between these pre-existing intelligibility metrics and those proposed in the present chapter. In this section, the key differences between the intelligibility metrics are highlighted.

First, SIIB and SIIB$^{\text{Gauss}}$ represent speech signals using a more realistic auditory model than SIMI and MIKNN. To account for the frequency masking of the auditory system, SIIB and SIIB$^{\text{Gauss}}$ use a gammatone filterbank with centre frequencies linearly spaced on the ERB-rate scale up to a maximum centre frequency of 6.5 kHz. To account for the temporal masking of the auditory system, SIIB and SIIB$^{\text{Gauss}}$ use the forward temporal masking function in (3.26). This contrasts with the auditory model used by SIMI and MIKNN, which uses a rectangular non-overlapping filterbank consisting of $F = 15$ $1/3$ octave-bands between 0.15-4.3 kHz, and does not use a forward temporal masking function. In addition, the auditory model used by SIIB and SIIB$^{\text{Gauss}}$ applies logarithmic dynamic range compression, whereas the auditory model used by SIMI and MIKNN does not. Figure 4.2 compares the output of the auditory model used by SIIB and SIIB$^{\text{Gauss}}$ with the output of the auditory model used by SIMI and MIKNN. In Chapter 5 the impact that the auditory model has on intelligibility prediction is investigated.

In theory, mutual information is invariant under the reparameterisation of the marginal variables (Kraskov et al., 2004) (e.g., see (2.17)). Because $\ln$ is a deterministic invertible function, for two random variables a and b, we have that $I(\mathrm{a}; \mathrm{b}) = I(\ln \mathrm{a}; \ln \mathrm{b})$. With this result in mind, it may seem that applying logarithmic dynamic range compression to sequences of auditory spectra would have little effect on estimating the mutual information rate between the sequences. However, recall from Section 3.3.3 that the logarithm makes $\mathbf{x}_t$ more like a Gaussian random variable. Furthermore, recall from Section 4.1.1 that the KLT is only effective at removing statistical dependencies when the random variables are Gaussian. For this reason the logarithmic dynamic range compression used by SIIB and SIIB$^{\text{Gauss}}$ plays an important role.

In addition to the above reason for including logarithmic dynamic range compression, Kraskov et al. (2004) provide further motivation. Regarding the KNN mutual information estimator, which is used by SIIB,

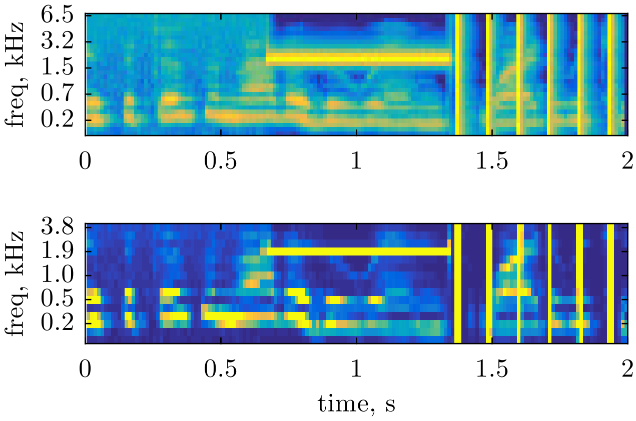

Figure 4.2: Comparison of auditory models. Top: a sequence of auditory log-spectra computed using the auditory model used by SIIB and SIIB$^{\text{Gauss}}$. Bottom: a sequence of auditory spectra computed using the auditory model used by SIMI and MIKNN. For both cases, the stimuli was created by degrading a clean speech signal with additive white noise, a sinusoid, and a delta train, consecutively. For SIIB, the energy of the sinusoid and delta-train leaks into nearby time-frequency units due to SIIB's auditory model. For SIMI and MIKNN, the energy from the sinusoid and delta-train is more localised, which effectively decreases the degradation and could lead to an overestimate of intelligibility.

Kraskov et al. (2004) state that "if the marginal distributions are skewed, it might be a good idea to transform them such as to be more uniform (or at least single-humped and more or less symmetric)" as this often improves the accuracy of the KNN mutual information estimator.

A second difference between the pre-existing intelligibility metrics and those proposed in the present chapter is that SIMI and MIKNN assume that the 1/3 octave bands are statistically independent, whereas SIIB and SIIB$^{\text{Gauss}}$ use the KLT to reduce the statistical dependencies between the time-frequency units. To show the importance of reducing statistical dependencies, Figure 4.3 displays a sample correlation matrix (i.e., the nor-

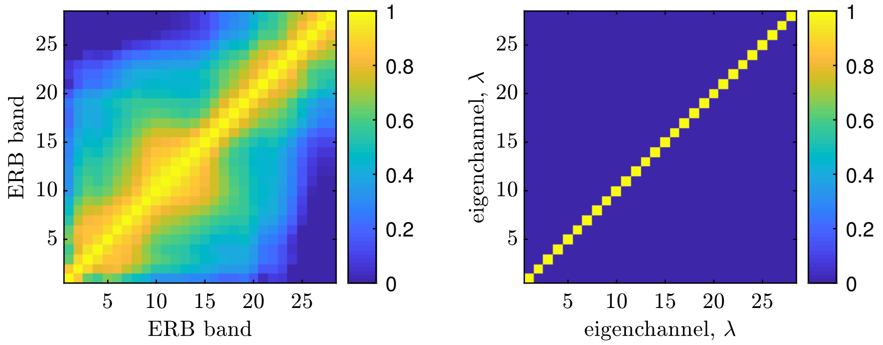

Figure 4.3: Correlation matrices for auditory log-spectra computed by SIIB and SIIB$^{\text{Gauss}}$. Left: the correlation between each ERB frequency band with every other ERB frequency band. Right: the correlation matrix after applying the KLT to the auditory log-spectra.

malised covariance matrix) for $\{\mathbf{x}_t^K\}$ and $\{\tilde{\mathbf{x}}_t^K\}$ for $K = 1$ using speech signals from the CHAINS corpus described in Chapter 3.

Given a sequence of stacked auditory log-spectra $\{\boldsymbol{x}_t^K\}$ that consists of elements $x_{i,t}^K$, the sample correlation between the i'th and j'th element is given by:

$$C_{i,j} = \frac{\sum_t (x_{i,t}^K - \frac{1}{T}\sum_t x_{i,t}^K)(x_{j,t}^K - \frac{1}{T}\sum_t x_{j,t}^K)}{\sqrt{\sum_t (x_{i,t}^K - \sum_t x_{i,t}^K)^2}\sqrt{\sum_t (x_{j,t}^K - \sum_t x_{j,t}^K)^2}}. \tag{4.20}$$

The sample correlation matrix $\boldsymbol{C} \in \mathbb{R}^{FK \times FK}$ is formed by computing the correlation for all pairs of i and j. The sample correlation matrix for $\{\tilde{\mathbf{x}}_t^K\}$ is computed in the same way, except that the KLT is first applied. For Figure 4.3, $K = 1$ (i.e., the vectors are not stacked), thus Figure 4.3 shows the correlation between the ERB frequency bands only (i.e., correlations between successive auditory log-spectra are not shown). It is clear that the ERB frequency bands display strong correlation, and thus cannot be statistically independent. In addition we see that the KLT is effective at reducing the correlation, and if $\{\mathbf{x}_t\}$ is Gaussian, it follows that the statistical dependencies are removed.

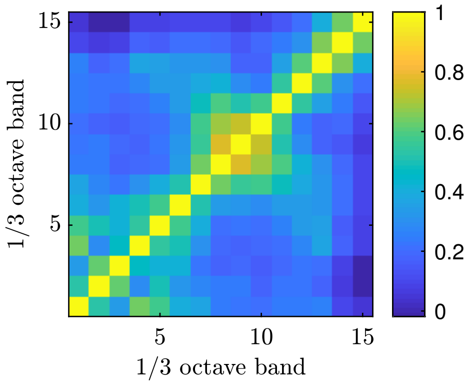

Figure 4.4: A correlation matrix for auditory spectra computed using SIMI and MIKNN. The correlation between each 1/3 octave band with every other 1/3 octave band is displayed.

One reason that there are statistical dependencies between the ERB bands is because the gammatone filters used by SIIB and SIIB$^{\text{Gauss}}$ overlap (e.g., see Figure 3.2). However, this is not the only source of the statistical dependencies. It is not unreasonable to believe that due to imperfect muscular control of the vocal-tract and the physics that govern acoustics, the energy of a speech signal in one frequency band may be constrained to be similar to the energy in a neighbouring frequency band. To demonstrate this, we also compute the sample correlation matrix for $\{\mathbf{x}_t\}$ using the auditory model used by SIMI and MIKNN, which do not use an overlapping filter bank or apply dynamic range compression. Figure 4.4 shows that for the auditory model used by SIMI and MIKNN, the resulting representation of speech also displays correlations between the frequency bands. Note that because the 1/3 octave bands sample the frequency spectrum more coarsely and do not overlap, the correlation between the 1/3 frequency bands is less than correlation between the ERB frequency bands. However, the correlation is still not negligible.

The use of the KLT is an important point of difference between the intelligibility metrics proposed in the present chapter and pre-existing metrics found in the literature. In Chapter 5 the impact that the KLT has on intelligibility prediction is investigated.

Lastly, SIIB and SIIB$^{\text{Gauss}}$ incorporate the effect of talker variability, whereas SIMI and MIKNN do not. Importantly, this means that SIIB and SIIB$^{\text{Gauss}}$ saturate at $-\frac{R}{2K}\sum_{\lambda}\log_2(1-r_{\lambda}^2)$ bits per second.

4.4 Revisiting the Information Rate of Speech

Recall that Section 3.3 proposed a method for estimating the information rate of the speech production channel. The method relies on having a chorus of speech signals. When the method is applied to the CHAINS corpus, an information rate of $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\}) = 2070$ b/s is obtained. After accounting for over-sampling, the rate reduces to about 500 b/s.

One limitation of the method in Section 3.3 is that it assumes that the elements of $\mathbf{m}_t$ are statistically independent, and likewise for $\mathbf{x}_t$. Moreover, Section 3.3 assumed that $\mathbf{m}_t$ and $\mathbf{m}_{t+\tau}$ are statistically independent for all $\tau \neq 0$, and likewise for $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$. The consequence of these assumptions is that the method overestimates the information rate. However, the present chapter argued that if $\mathbf{x}_t$ is Gaussian, then by stacking K successive auditory log-spectra and applying the KLT, the statistical dependencies can be removed.

In this section, the experiment from Section 3.3 is repeated, however, for the estimates of the production noise, production noise variance, and speech variance in (3.38), (3.39), and (3.40), respectively, estimators are applied to stacked auditory log-spectra that have been processed by the KLT, that is, $\tilde{\mathbf{x}}^K$, as defined in (4.7). This contrasts with Section 3.3, where estimators were applied to auditory log-spectra $\mathbf{x}_t$, not $\tilde{\mathbf{x}}^K$.

Figure 4.5 plots the estimate of $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\})$ against K for data from the CHAINS corpus after using the KLT to remove statistical dependencies. As K increases, time dependencies spanning greater durations are accounted for and so the information rate decreases. The lowest observed

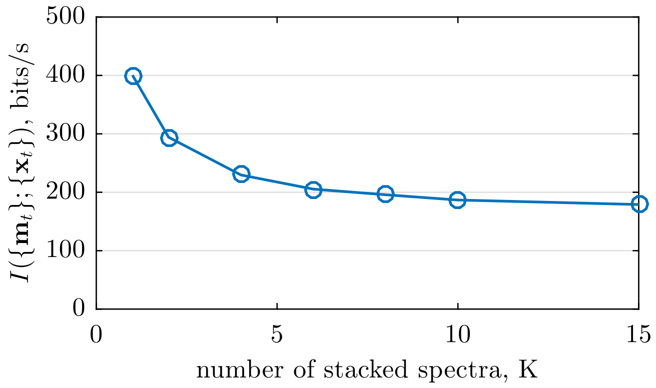

Figure 4.5: Estimate of the mutual information rate of the speech production channel against the number of stacked auditory log-spectra. $K = 15$ corresponds to 187.5 ms of speech for each stacked spectra.

information rate is about 180 b/s. This is significantly closer to the lexical information rate of speech of 60 b/s from Section 2.2. The remaining difference is likely due to the assumption that the auditory log-spectra are Gaussian. Because the auditory log-spectra are only approximately Gaussian, the KLT does not remove all the statistical dependencies.

4.5 Summary of Chapter

This chapter developed two new intelligibility metrics called SIIB and SIIB$^{\text{Gauss}}$. The intelligibility metrics are based on the hypothesis that intelligibility is a function of the mutual information rate of a message selected by a talker and a signal received by a listener. Both intelligibility metrics rely on an auditory model and use the KLT to reduce statistical dependencies between time-frequency units. The difference between SIIB and SIIB$^{\text{Gauss}}$ is that SIIB estimates mutual information using a non-parametric estimator based on k-nearest neighbours, whereas SIIB$^{\text{Gauss}}$ uses the capacity of a Gaussian communication channel. It was hypothesised that because the non-parametric mutual information estimator used by SIIB is valid for a wide range of degradation, SIIB will have higher performance than SIIB$^{\text{Gauss}}$.

SIIB and SIIB$^{\text{Gauss}}$ show similarities to existing intelligibility metrics in the literature such as SIMI and MIKNN. The three main differences are 1) SIIB and SIIB$^{\text{Gauss}}$ use a more accurate auditory model that accounts for the upwards frequency masking, forward temporal masking, and dynamic range compression of the human auditory system, 2) SIIB and SIIB$^{\text{Gauss}}$ use the KLT to reduce statistical dependencies between time-frequency units, and 3) SIIB and SIIB$^{\text{Gauss}}$ account for talker variability. Importantly, talker variability causes the information rate of the communication channel to saturate.

Lastly, the information rate of speech communication was revisited. Chapter 3 proposed a method for estimating the information rate of speech communication that is based on a chorus of speech signals. In Chapter 3 the experiment assumed that the time-frequency units were statistically independent. In the present chapter, the experiment from Chapter 3 was repeated, however, the KLT was used to reduce the statistical dependencies. Doing so reduced our estimate of the information rate of speech to about 180 b/s. This is still larger than the lexical information rate of speech, which is 60 b/s. The remaining discrepancy could be the result of assuming that the signals are Gaussian. If the signals are not Gaussian, then the KLT does not remove all statistical dependencies. If a more powerful transform for removing statistical dependencies were used, the information rate may decrease further.

Chapter 5

An Evaluation of Intelligibility Metrics

As discussed in Section 2.5, a key component to the design of speech-based communication systems is an understanding of how they affect intelligibility. Although formal listening tests can provide valid data, such tests are time-consuming, laborious, and expensive. For this reason, quantities that are fast to compute and correlated with intelligibility are of interest. Such quantities are referred to as *instrumental intelligibility metrics*. In Chapter 4, two new intelligibility metrics called SIIB and SIIB$^{\text{Gauss}}$ were proposed. In the present chapter, the accuracy of SIIB and SIIB$^{\text{Gauss}}$ is evaluated and compared to existing intelligibility metrics from the literature.

The present chapter is largely reproduced from Van Kuyk et al. (2018a). For consistency with the original publication, SIIB is considered to be a ’pre-existing’ intelligibility metric, whereas SIIB$^{\text{Gauss}}$ is considered to be a ’modified’ intelligibility metric, even though both metrics were developed as a part of this thesis.

5.1 Motivation

Over the past decade many intrusive intelligibility metrics have been proposed. Examples include the coherence SII (CSII) (Kates and Arehart, 2005), the extended SII (ESII) (Rhebergen and Versfeld, 2005), the quasi-stationary STI (QSTI) (Schwerin and Paliwal, 2014), the normalised covariance measure (NCM) (Koch, 1992; Goldsworthy and Greenberg, 2004), the temporal fine-structure spectrum based index (TFSS) (Chen et al., 2013), the hearing-aid speech perception index (HASPI) (Kates and Arehart, 2014), the Christiansen-Pedersen-Dau metric (CPD) (Christiansen et al., 2010), those based on the short-time objective intelligibility measure (STOI) (e.g., Taal et al., 2011a; Jensen and Taal, 2016), those based on the speech-based envelope power spectrum model (sEPSM) (e.g., Jørgensen and Dau, 2011; Jørgensen et al., 2013; Relaño-Iborra et al., 2016)), and those based on the glimpse proportion metric (GP) (e.g., Cooke, 2006; Barker and Cooke, 2007; Tang and Cooke, 2016). Many of these metrics have not been extensively tested on data sets other than those used during their development. Additionally, the above metrics are often heuristically motivated, which suggests that they may not generalise well to new environments and speech enhancement strategies.

Motivated by the fact that many intrusive intelligibility metrics have been recently proposed but have not been widely evaluated, this chapter presents a study on the accuracy of 12 existing monaural intrusive intelligibility metrics. To assess the accuracy of each metric, the strength of the relationship between intelligibility and the metric is measured. The intelligibility data were obtained from 11 experiments described in the literature. The data include Dutch, Danish, and English speech that was degraded by additive noise, reverberation, and competing talkers, and subjected to speech enhancement.

The majority of the intelligibility metrics in this chapter were developed with Germanic languages in mind, however, the studies of Jin (2014);

Wong et al. (2007); Xia et al. (2012); Chen and Loizou (2011) have suggested that many intelligibility metrics can obtain good performance for Mandarin, Cantonese, and Korean.

In addition to evaluating the accuracy of pre-existing intelligibility metrics, this chapter analyses why the top performing metrics have high performance. Specifically, the effect of decorrelating input features, the effect of the auditory model, and the effect of using different distortion measures is investigated.

Previous evaluations of intrusive intelligibility metrics exist. For example, Ma et al. (2009) and Taal et al. (2010) evaluated the accuracy of intelligibility metrics for noise-reduced speech, and Taal et al. (2011b) evaluated the accuracy of intelligibility metrics for speech processed by ideal time-frequency segregation (ITFS). Those evaluations each considered a single type of degradation, whereas the evaluation in this chapter considers data from many real-word scenarios.

Evaluations can also be found in publications that propose new intelligibility metrics, but in terms of the number of intelligibility metrics and the number of data sets, the scope of such evaluations is smaller than the present study. Two advantages of considering a broader scope are 1) it is easier to determine why some intelligibility metrics perform better than others, and 2) it is possible to investigate the ability of intelligibility metrics to generalise to new types of distortion. To our knowledge, in terms of the number of listening tests and intelligibility metrics, the evaluation in this chapter is the most comprehensive evaluation of intelligibility metrics for speech in noise to date.

The remainder of this chapter is organised as followed. Section 5.2 describes the listening test data and Section 5.3 describes intelligibility metrics from the literature. Modified intelligibility metrics are proposed in Section 5.4. Performance criteria are described in Section 5.5 and results are presented in Section 5.6. Finally, Section 5.7 concludes the chapter.

5.2 Listening Test Data

The evaluation in the present chapter considers the results of 11 intelligibility studies. From these studies, 13 data sets were created. In this section, each data set is described. Table 5.1 summarises the data sets, while the accompanying references provide additional details. The naming convention for the data sets includes the first author of the publication that describes the data set in full, and an abbreviation that indicates the type of degradation or processing. The order that the data sets are presented in is such that similar data sets are grouped together.

Many of the data sets in this section include stimuli processed by speech enhancement algorithms. There are two main approaches to speech enhancement: 1) the speech signal can be modified prior to degradation (e.g., optimal energy redistribution (Taal et al., 2014) and dynamic range compression (Zorila et al., 2012)), or 2) the speech signal can be modified after degradation has been introduced (e.g., spectral subtraction (Boll, 1979) and Wiener filters (Wiener, 1949)). In this chapter, the former type of algorithm is referred to as a *pre-processing* algorithm and the latter as a *post-processing* algorithm.

JensenMOD

The first data set consists of speech degraded by noise with strong temporal modulations. In Jensen and Taal (2016) phrases from the Dantale II corpus (Wagener et al., 2003) were degraded by ten types of noise. Four of the noise types included Track 1, 4, 6, and 7 from the ICRA noise corpus (Dreschler et al., 2001). The ICRA signals are synthetic signals with spectral and temporal properties similar to speech. Four of the noise types were constructed by multiplying speech-shaped noise (SSN) (i.e., Gaussian noise with a long-term power-spectrum that is similar to the power spectrum of clean speech) with $1 + \sin(2\pi f t + \phi)$ where ϕ is uniformly distributed between $\pm\pi$, t is the sample index, and $f = 2, 4, 8,$ or 16 Hz. The

Table 5.1: Summary of listening test data sets. B, is the bandwidth, m is the number of listeners and n is the number of listening conditions.

Name	Degradation	Enhancement strategy	B, kHz	m	n
JensenMOD (Jensen and Taal, 2016)	Modulated noise	None	10.0	12	60
SantosREV (Santos et al., 2014)	Noise & reverb	None	8.0	10	17
KjemsAN (Kjems et al., 2009)	Noise	None	7.7	15	40
KjemsITFS (Kjems et al., 2009)	Noise	Ideal time-frequency segregation.	7.7	15	168
TaalPOST (Taal et al., 2011a)	Noise	Minimum mean-squared error estimate of the short-time spectral amplitude.	8.7	15	15
JensenPOST (Jensen and Hendriks, 2012)	Noise	Minimum mean-squared error estimate of the short-time spectral amplitude.	4.0	13	20
HuPOST (Hu and Loizou, 2007)	Noise	Spectral subtractive, sub-space, statistical model based, and Wiener-type algorithms.	3.5	40	72
HendriksPRE (Hendriks et al., 2015)	Noise & reverb	Optimal energy redistribution.	8.0	8	20
KleijnPRE (Kleijn and Hendriks, 2015)	Noise	Optimal energy redistribution.	8.0	9	32
CookePRE (Cooke et al., 2013)	Noise & competing talker	Nine pre-processing enhancement algorithms.	8.0	175	60
KhademiJOINT (Khademi et al., 2017)	Noise	MVDR beamformer, Wiener filter, & optimal energy redistribution.	8.0	7	24
DutchMRG	-	JensenPOST, HendriksPRE, KleijnPRE, and KhademiJOINT merged into a single data set.	-	-	-
DantaleMRG	-	KjemsAN, KjemsITFS, and TaalPOST merged into a single data set.	-	-	-

final two noise sources were machine-gun noise and destroyers-operation-room noise from the NOISEX corpus (Varga and Steeneken, 1993). Six SNRs were chosen for each noise source so that some stimuli were unintelligible and others were perfectly intelligible. In total there are 10 noise sources $\times$ 6 SNRs = 60 conditions. Stimuli were presented to 12 normal-hearing listeners. For each word in a given sentence, the listeners were instructed to identify the correct word from a list of ten possibilities. See Jensen and Taal (2016) for more details.

SantosREV

The second data set consists of speech corrupted by noise and reverberation. In Santos et al. (2014), IEEE sentences (Rothauser et al., 1969) were degraded by three types of distortion: 1) additive noise, 2) reverberation, and 3) additive noise and reverberation. For the additive noise distortion, SSN and babble noise at SNRs of -5, 0, 5, and 10 dB were used. For the reverberant distortion, IEEE sentences were convolved with a room impulse response with T60 = 0.3, 0.6, 0.8, 1, and 1.4 s. For the additive noise and reverberant distortion the sentences were convolved with room impulse responses with T60 = 0.3 and 0.6 s and mixed with SSN at SNRs of 5 dB and 10 dB. In total there are 8 noise + 5 reverberant + 4 noise and reverberant = 17 conditions. Stimuli were presented to ten normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. See Santos et al. (2014) for more details.

Originally, the distorted stimuli in SantosREV were offset in time from the clean stimuli. However, time-alignment is a requirement for many intrusive intelligibility metrics. For this chapter, the signals in SantosREV were aligned by finding the time-offset that maximised the cross-correlation of the clean and distorted stimuli. This resulted in significantly higher performance scores than those reported in (Santos et al., 2014).

KjemsAN

The third data set consists of speech degraded by additive noise. In Kjems et al. (2009) phrases from the Dantale II corpus (Wagener et al., 2003) were degraded by four types of noise: SSN, cafeteria noise, noise from a bottling factory hall, and car interior noise. The stimuli were presented to 15 normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. Based on the listening test results, Kjems *et al.* derived psychometric curves that relate intelligibility to SNR for each noise type.

For this chapter, KjemsAN was created by adding the noise signals to the clean Dantale II sentences at ten SNRs. The SNRs were selected by sampling the psychometric curves at intervals of 10% intelligibility from 10% to 100%. In total there are 4 noise types $\times$ 10 SNRs = 40 conditions.

KjemsITFS

The fourth data set consists of speech subjected to ideal time-frequency segregation processing (ITFS) (Brungart et al., 2006). ITFS processing aims to eliminate the energy of a speech signal at particular time-frequency locations by multiplying the short-time Fourier transform of the speech signal with a binary gain function. Similarly to KjemsAN, the listening experiment was conducted by Kjems et al. (2009), used phrases from the Dantale II corpus (Wagener et al., 2003), involved 15 normal-hearing listeners, and used the same four types of noise. For each noise type, the noisy phrases were processed by two types of ITFS called an ideal binary mask and a target binary mask. Three SNRs were used (-60 dB, and SNRs corresponding to 20% and 50% intelligibility) and eight variants of each ITFS algorithm were considered. In total there are 168 conditions. See Kjems et al. (2009) for more details.

TaalPOST

The fifth data set consists of speech subjected to post-processing enhancement. In Taal et al. (2011a) phrases from the Dantale II corpus were degraded by SSN at SNRs of 8.9, 7.7, 6.5, 5.2, and 3.1 dB. The MMSE-STSA enhancement algorithm (Ephraim and Malah, 1984) and an improved version (Erkelens et al., 2007) were applied to the noisy phrases. In total there are 5 SNRs × (2 algorithms + 1 unprocessed) = 15 conditions. Stimuli were presented to 15 normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information, and the proportion of correctly identified words was recorded.

JensenPOST

The sixth data set consists of speech subjected to post-processing enhancement. In Jensen and Hendriks (2012) phrases from the Dutch version of the Hagerman test (Houben et al., 2014) were degraded by SSN at SNRs of -8, -6, -4, -2, and 0 dB and processed by three enhancement algorithms. The three algorithms compute a minimum mean-squared error estimate of the clean speech by multiplying the short-time spectral amplitude of the noisy speech with a gain function. In total there are 5 SNRs × (3 algorithms + 1 unprocessed) = 20 conditions. Stimuli were presented to 13 normal-hearing listeners. For each word in a given sentence, the listeners were shown ten candidate words from which they were instructed to select from.

HuPOST

The seventh data set consists of speech subjected to post-processing enhancement. In Hu and Loizou (2007) IEEE sentences (Rothauser et al., 1969) were filtered by a simulated telephone channel, degraded by four noise types: babble, car, street, and train, at SNRs of 0 and 5 dB, and processed by eight enhancement algorithms encompassing spectral subtractive, sub-space, statistical model based and Wiener-type algorithms. In total there are 4 noise types × 2 SNRs × (8 algorithms + 1 unprocessed)= 72 conditions. Stimuli were presented to 40 normal-hearing listeners where ten listeners were used for each of the four noise types. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. See Hu and Loizou (2007) for more details.

HendriksPRE

The eighth data set consists of speech subjected to pre-processing enhancement and degraded by reverberation and noise. In Hendriks et al. (2015) phrases from the Dutch version of the Hagerman test (Houben et al., 2014) were processed by four enhancement algorithms, convolved with a room impulse response with a T60 time of 1 s, and then degraded by SSN at SNRs of -2, 0, 2, and 4 dB. Three of the enhancement algorithms optimally redistribute the energy of the clean speech according to a distortion criterion. The fourth algorithm uses steady-state suppression to reduce degradation caused by reverberation. In total there are 4 SNRs × (4 algorithms + 1 unprocessed) = 20 conditions. Stimuli were presented to eight normal-hearing listeners. For each word in a given sentence, the listeners were instructed to identify the correct word from a list of ten possibilities. See Hendriks et al. (2015) for more details.

KleijnPRE

The ninth data set consists of speech subjected to pre-processing enhancement and degraded by noise. In Kleijn and Hendriks (2015) phrases from the Dutch version of the Hagerman test (Houben et al., 2014) were subjected to three pre-processing enhancement algorithms and then degraded either by SSN at SNRs of $-15, -12, -9,$ and -6 dB, or car noise at SNRs of $-23, -20, -17,$ and -14 dB. The three enhancement algorithms optimally redistribute the energy of the clean speech according to a distortion criterion. In total there are 2 noise types × 4 SNRs × (3 algorithms + 1 unprocessed) = 32 conditions. Stimuli were presented to nine normal-hearing listeners. For each word in a given sentence, the listeners were instructed to identify the correct word from a list of ten possibilities. See Kleijn and Hendriks (2015) for more details.

CookePRE

The tenth data set consists of speech subjected to pre-processing enhancement and degraded by noise. In Cooke et al. (2013) IEEE sentences (Rothauser et al., 1969) were processed by 19 pre-processing enhancement algorithms and degraded either by SSN at SNRs of 1, -4, and -9 dB, or by speech from a competing talker at SNRs of -7, -14, and -21 dB. Stimuli were presented to 175 normal-hearing listeners. The listeners were instructed to transcribe sentences without any additional information and the proportion of correctly identified words was recorded. Short words (e.g., a, the, in, to) were not scored.

For this chapter, a subset of the data in Cooke et al. (2013) was considered because the entire data set was not available. Ten of the IEEE sentences for each condition and nine of the enhancement algorithms were used. The algorithms are referred to in Cooke et al. (2013) as AdaptDRC, F0-shift, IWFEMD, on/offset, OptimalSII, RESSYSMOD, SBM, SEO, and SSS. In total there are 2 noise sources × 3 SNRs × (9 algorithms + 1 unprocessed) = 60 conditions.

KhademiJOINT

The eleventh data set consists of speech that has been jointly processed by far-end and near-end enhancement algorithms. In Khademi et al. (2017), four enhancement strategies were considered, all of which used a minimum variance distortionless response (MVDR) beamformer at the far-end.

The first strategy used no near-end enhancement, the second used blind optimal energy redistribution at the near-end, the third used blind optimal energy redistribution at the near-end and an additional Wiener filter at the far-end, and the fourth used jointly optimal energy redistribution at the near-end. Three near-end SNRs (-7.5, 0, and 5 dB) and two far-end SNRs (-10 and 2.5 dB) were used. In total there are 4 enhancement strategies × 3 near-end SNRs × 2 far-end SNRs = 24 conditions. For each condition phrases from the Dutch version of the Hagerman test (Houben et al., 2014) were presented to seven normal-hearing listeners. For each word in a given sentence, the listeners were instructed to identify the correct word from a list of ten possibilities. See Khademi et al. (2017) for more details.

DutchMRG

The twelfth data set was created by merging JensenPOST, HendriksPRE, KleijnPRE, and KhademiJOINT. It is reasonable to merge these data sets because the associated listening tests all used phrases from the Dutch version of the Hagerman test (Houben et al., 2014) and were conducted using the same procedures by the Circuits and Systems Group at Delft University of Technology. Note, that the number of subjects differed for the four experiments. DutchMRG was included in the evaluation to test if the intelligibility metrics give consistent measurements for different enhancement strategies.

DantaleMRG

The thirteenth data set was created by merging KjemsAN, KjemsITFS, and TaalPOST. It is reasonable to merge these data sets because the associated listening tests all used phrases from the Dantale II corpus. To prevent KjemsITFS from dominating the other data sets, 60 out of the 168 conditions from KjemsITFS were randomly selected, and all of the conditions for KjemsAN and TaalPOST were selected. Note that the listening tests were conducted by different laboratory groups. Similarly to DutchMRG, this data set was included to test if the intelligibility metrics give consistent measurements for different enhancement strategies. JensenMOD also used the Dantale II corpus, but was not included in DantaleMRG because the listening test for JensenMOD presented listeners with ten candidate words to select from, whereas the listening tests for KjemsAN, KjemsITFS, and TaalPOST did not.

5.3 Pre-Existing Intelligibility Metrics

Over the past decade a large number of intrusive intelligibility metrics have been proposed. In this section, 12 metrics from the literature, which are considered in this evaluation, are summarised. An overview of the metrics can be found in Table 5.2. See the accompanying references for more detailed descriptions. Additionally, detailed descriptions of SIIB and SIIB$^{\text{Gauss}}$ can be found in Chapter 4, and detailed descriptions of SIMI and MIKNN can be found in Chapter 2. Unless stated otherwise, all parameters were selected according to those recommended in the original publications.

Speech Intelligibility Index

The speech intelligibility index (SII) (ANSI, 1997b) is based on the idea that intelligibility is related to audibility. To compute the SII, a bandpass filterbank is applied to the clean speech and the noise signal, and a weighted average of the long-term SNR of each frequency band is calculated. The weights define a band-importance function (BIF) that characterises the relative importance of each frequency band. Prior to averaging, the SNR is clipped to be between ± 15 dB and normalised to be between 0 and 1. This reflects the idea that below -15 dB the speech signal is inaudible and

Table 5.2: Pre-existing intelligibility metrics considered in this study.

Abbreviation	Description
SII	The speech intelligibility index (ANSI, 1997b).
HEGP	The high-energy glimpse proportion metric (Tang and Cooke, 2016).
CSII-MID	The mid-level coherence SII (Kates and Arehart, 2005).
HASPI	The hearing-aid speech perception index (Kates and Arehart, 2014).
NCM-BIF	The normalised covariance measure with signal-dependent band-importance functions (Ma et al., 2009).
QSTI	The quasi-stationary speech transmission index (Schwerin and Paliwal, 2014).
STOI	The short-time objective intelligibility measure (Taal et al., 2011a).
ESTOI	The extended STOI measure (Jensen and Taal, 2016).
MIKNN	The k-nearest neighbour mutual information intelligibility measure (Taghia and Martin, 2014).
SIMI	Speech intelligibility prediction based on a mutual information lower bound (Jensen and Taal, 2014).
SIIB	Speech intelligibility in bits (Van Kuyk et al., 2018b).
sEPSM$^{\text{corr}}$	The speech-based envelope power spectrum model with short-time correlation (Relaño-Iborra et al., 2016).

above 15 dB the intelligibility is at its maximum. The SII is known to perform well for speech degraded by stationary additive noise, but poorly for speech degraded by modulated noise sources (Rhebergen and Versfeld, 2005).

In this chapter, the SII was only evaluated using JensenMOD, KjemsAN, and CookePRE. For the remaining data sets, either the noise signal was not available, or noise was not the main cause of distortion. The implementation of the SII was obtained from the Acoustical Society of America (`http://sii.to`) and used the 1/3 octave band procedure with the BIF tabulated in Table 3 of (ANSI, 1997b).

High-Energy Glimpse Proportion Metric

The glimpse proportion metric (GP) is the initial stage of the glimpsing model of speech perception (Cooke, 2006) and has been used as an intelligibility metric in various studies (e.g., (Barker and Cooke, 2007; Tang and Cooke, 2016)). The GP is defined as the proportion of spectro-temporal regions where the clean speech has energy greater than the noise signal by a pre-defined threshold. The GP shares similarities with the SII in that both metrics assume that audibility is the determining factor of intelligibility. The difference is that the SII averages the long-term SNR of each frequency band, whereas the GP is the proportion of short-time frequency-local SNRs above a threshold.

In Tang and Cooke (2016) a variation of the GP called the high-energy GP (HEGP) was shown to be more highly correlated with intelligibility than the original GP. The main difference between the metrics is that HEGP only uses spectro-temporal regions where the noisy speech has above average energy. Similarly to the SII, HEGP can only quantify distortion caused by additive noise signals. For this reason, HEGP was evaluated using KjemsAN, JensenMOD, and CookePRE only.

The implementation of HEGP used in this chapter was obtained from its developers. Note that CookePRE is a subset of a data set that was used during the development of HEGP.

Coherence Speech Intelligibility Index

The coherence speech intelligibility index (CSII) (Kates and Arehart, 2005) is based on the SII, but replaces the SNR of each frequency band with a signal-to-distortion ratio (SDR). The SDR is estimated from the coherence function (Carter et al., 1973) of the clean and distorted speech signal. For the case of speech degraded by additive noise, the SDR and SNR are equivalent, making the CSII a generalisation of the SII that can be applied to a wider range of distortions. In Kates and Arehart (2005) it was found that the performance of the CSII could be improved by calculating the CSII separately for low, mid, and high-energy speech segments.

The implementation of the CSII used in this chapter was obtained from (Loizou, 2013) and is described in (Ma et al., 2009), where it is referred to as $\mathrm{CSII}_{\mathrm{mid}}$. Note that the implementation in (Loizou, 2013) differs to that originally proposed in (Kates and Arehart, 2005) in that (Loizou, 2013) averages the CSII over short-time segments. For this chapter, the implementation in (Loizou, 2013) was modified to make it more similar to that originally proposed (i.e., it does not use short-time segments) because we found that the original method had higher overall performance. In this chapter the algorithm is referred to as CSII-MID.

Hearing-Aid Speech Perception Index

The hearing-aid speech perception index (HASPI) (Kates and Arehart, 2014) is based on an elaborate auditory model where the shape and bandwidth of the cochlear filters depend on the speech signal intensity and the outer hair-cell damage of the listener. Dynamic range compression is applied to the output of each cochlear filter in accordance with physiological measurements of compression in the cochlea and psychophysical estimates of compression in the human ear. Additionally, a time-alignment stage is included. The auditory model has two outputs: a sequence of short-time log-spectra, and a basilar membrane vibration signal for each frequency band.

From the outputs of the auditory model the cepstral correlation and auditory coherence are computed. To compute cepstral correlation, the log-spectra are converted to an approximation of Mel-frequency cepstral coefficients (Davis and Mermelstein, 1980) by taking the inner product between the log-spectra and a set of cosine functions. Pearson's correlation coefficient between the cepstra of the clean and distorted speech is then computed for each cepstral dimension and the resulting coefficients are averaged.

The auditory coherence is computed by splitting the basilar membrane vibration signals into three sets that contain low, mid, and high-energy segments. For each set and each frequency band, short-time correlation coefficients between the clean vibration signals and the distorted vibration signals are computed and then averaged over the time dimension and the frequency dimension. This results in three auditory coherence terms corresponding to low, mid, and high energy segments.

HASPI is computed as a linear combination of the cepstral correlation and the three auditory coherence terms. The relative importance of each term depends on the type of distortion and thus is fitted to the intelligibility data. In this chapter the weights of the cepstral correlation and auditory coherence terms were computed for each data set such that the mean squared error between the predicted and measured intelligibility scores was minimised. However, it was found that similar performance could be obtained simply by summing the cepstral correlation and high-energy auditory coherence. The implementation of HASPI used in this chapter was obtained from its developers.

Normalised Covariance Measure

The normalised covariance measure (NCM) (Koch, 1992; Goldsworthy and Greenberg, 2004) is a variant of the STI that uses clean speech as the probe signal. To compute the NCM, a band-pass filterbank is applied to the clean and distorted speech signals, and the temporal envelope of the output of each filter is extracted. Subsequently, the normalised covariance (i.e., Pearson's correlation coefficient) between the clean and distorted envelopes is calculated and converted to an apparent SNR for each frequency band. Similarly to the SII, the apparent SNR is clipped before a weighted average over the frequency bands is computed.

In Ma et al. (2009) it was found that the NCM is strongly correlated with intelligibility for speech subjected to post-processing enhancement. The correlation was particularly strong when new signal dependent BIFs were used. The implementation of the NCM used in this chapter was obtained from (Loizou, 2013) and is described in (Ma et al., 2009) where it is referred to as NCM $W_i^{(1)}, p = 1.5$. In this chapter the algorithm is referred to as NCM-BIF. Note that HuPOST was used during the development of NCM-BIF.

Quasi-Stationary Speech Transmission Index

The quasi-stationary speech transmission index (QSTI) was proposed in (Schwerin and Paliwal, 2014). The QSTI is a variation of the STI that uses clean speech as the probe signal and averages the score over short-time segments. In Schwerin and Paliwal (2014) the QSTI was reported to be more strongly correlated with intelligibility than the traditional STI.

The implementation of the QSTI used in this chapter was obtained from its developers webpage. Note that HuPOST, TaalPOST, and KjemsITFS were used during the development of QSTI.

Short-Time Objective Intelligibility Measure

The short-time objective intelligibility measure (STOI) was proposed in (Taal et al., 2011a) as an algorithm for predicting the intelligibility of time-frequency weighted noisy speech. To compute STOI, a simple model of the human auditory system is used to extract temporal envelopes of the clean speech and the distorted speech for various frequency bands. The temporal envelopes are segmented into short-time frames with a duration of 386 ms and a clipping procedure is used to ensure that the SDR of each frame is greater than -15 dB. STOI is calculated by computing Pearson's correlation coefficient between the clean and distorted envelopes for each short-time frame and each frequency band and then taking the mean.

The implementation of STOI used in this chapter was obtained from its developer's webpage. Note that TaalPOST and KjemsITFS were used during the development of STOI.

Extended Short-Time Objective Intelligibility Measure

The extended short-time objective intelligibility measure (ESTOI) was proposed in (Jensen and Taal, 2016) to address the finding that STOI performs poorly for modulated noise sources (e.g., Gaussian noise that is amplitude modulated by a sinusoid). Rather than computing the correlation of the clean and distorted envelopes for short-time segments, ESTOI computes the correlation between clean and distorted spectra so that 'glimpses of clean speech' can be detected. Additionally, the clipping procedure in STOI was removed to make the new model more mathematically tractable.

The implementation of ESTOI used in this chapter was obtained from its developer's webpage. Note that JensenPOST, JensenMOD, KjemsITFS, and a data set similar to KjemsAN were used during the development of ESTOI.

K-Nearest Neighbour Mutual Information Intelligibility Measure

The k-nearest neighbour (KNN) mutual information intelligibility measure (MIKNN) was proposed in (Taghia and Martin, 2014) while investigating the use of information theoretical techniques for intelligibility prediction. MIKNN uses the same representation of speech as STOI, however, rather than using the short-time correlation coefficient to quantify distortion, MIKNN estimates the mutual information between the clean and distorted temporal envelopes using a non-parametric estimator based on k-nearest neighbours (Kraskov et al., 2004). One advantage of mutual information is that unlike Pearson's correlation coefficient, mutual information can account for non-linear dependencies.

The implementation of MIKNN used in this chapter was obtained from its developer's webpage. Note that TaalPOST and KjemsITFS were used during the development of MIKNN.

Speech Intelligibility Prediction Based on Mutual Information

Similarly to MIKNN, the speech intelligibility prediction based on mutual information measure (SIMI) (Jensen and Taal, 2014) is based on the hypothesis that intelligibility is related to the mutual information between the clean and distorted temporal envelopes. In contrast to MIKNN, SIMI estimates a lower bound on the mutual information by assuming a parametric statistical model. Another important difference between SIMI and MIKNN is that SIMI operates on short-time segments of 250 ms, whereas MIKNN uses whole utterances. In Jensen and Taal (2014) SIMI was used to justify some of the heuristic design decisions of STOI.

The implementation of SIMI used in this chapter was obtained from its developer's webpage. Note that JensenPOST, KjemsITFS, and a data set similar to KjemsAN were used during the development of SIMI.

Speech Intelligibility in Bits

Speech intelligibility in bits (SIIB) is an information theoretic intelligibility metric that was recently proposed in (Van Kuyk et al., 2018b). Similar to MIKNN, a non-parametric mutual information estimator (Kraskov et al., 2004) is used to estimate the information shared between a clean and distorted speech signal.

There are three main differences between SIIB and MIKNN. First, SIIB uses the Karhunen-Loève transform (KLT) (Karhunen, 1947) to reduce statistical dependencies between spectro-temporal regions, and thus reduces overestimation of the information rate.

Second, SIIB accounts for 'production noise', which incorporates differences in pronunciation between talkers. Importantly, production noise causes the information rate of the communication channel to saturate (Kleijn and Hendriks, 2015).

Third, SIIB uses an auditory model that more accurately accounts for the frequency masking (Wegel and Lane, 1924) and temporal masking (Oxenham, 2001) of the human auditory system. To account for frequency masking, the temporal envelopes are extracted using an equivalent rectangular bandwidth (ERB) gammatone filterbank (Slaney, 1993). To account for temporal masking, the forward masking function suggested in (Rhebergen et al., 2006) is used. Additionally, logarithmic compression is applied to the envelopes.

The end result of SIIB is an estimate of the information shared between a talker and a listener in bits per second. Note that all of the data sets considered in this chapter were used during the development of SIIB.

Speech-Based Envelope Power Spectrum Model with Short-Time Correlation

The speech-based envelope power spectrum model forms the basis of three intelligibility metrics: sEPSM (Jørgensen and Dau, 2011), mr-sEPSM (Jørgensen et al., 2013), and sEPSM$^{\text{corr}}$ (Relaño-Iborra et al., 2016). All of the sEPSM metrics use the Hilbert transform and a gammatone filterbank to extract temporal envelopes for different frequency bands. A second bandpass filterbank called a modulation filterbank is then applied to each envelope signal. This results in a multi-dimensional representation that includes a time, frequency, and modulation dimension. Within this multi-dimensional domain, sEPSM and mr-sEPSM quantify distortion using a SNR metric, whereas sEPSM$^{\text{corr}}$ quantifies distortion using short-time correlation coefficients similarly to STOI. In this chapter only the most recent metric is considered: sEPSM$^{\text{corr}}$.

Note that the output of sEPSM$^{\text{corr}}$ increases as the duration of the stimulus increases. This is a consequence of the ‘multiple looks’ strategy that sEPSM$^{\text{corr}}$ uses to integrate information over the time dimension. For this reason, when comparing results from multiple data sets (i.e., for the merged data sets), it is important that the duration of the stimuli is held constant. In this chapter, when evaluating sEPSM$^{\text{corr}}$, all stimuli were truncated to have a duration of 20 seconds.

The implementation of sEPSM$^{\text{corr}}$ used in this chapter was obtained from its developers. Note that KjemsITFS was used during the development of sEPSM$^{\text{corr}}$.

5.4 Modified Intelligibility Metrics

One of the goals of this chapter is to investigate why some intelligibility metrics have higher performance than others. In this section we modify existing intelligibility metrics so that effective strategies can be identified.

5.4.1 Investigating the effect of decorrelating input features

The majority of the intelligibility metrics in the previous section quantify distortion by comparing time and/or frequency local features. SIIB and HASPI are exceptions to this. SIIB decorrelates log-spectra over the time and frequency dimension using the KLT, and HASPI decorrelates log-spectra over the frequency dimension using a cosine expansion similar to the type-1 discrete cosine transform (DCT) (Rao and Yip, 1990). Recall that for stationary signals the DCT asymptotically approximates the KLT.

To investigate the effect of decorrelating input features, SIIB and STOI were modified to produce two intelligibility metrics denoted SIIB$^{\text{noKLT}}$ and STOI$^{\text{KLT}}$. To compute SIIB$^{\text{noKLT}}$, the implementation of SIIB described in (Van Kuyk et al., 2018b) was used, but the KLT was not applied. To compute STOI$^{\text{KLT}}$ three changes are made to the original STOI implementation (Taal et al., 2011a):

1. Instead of using temporal envelopes to represent speech signals, log-temporal envelopes are used. To prevent singularities, a small amount of uniformly distributed noise is added to the envelopes before applying the logarithm.

2. The KLT is used to decorrelate the log-temporal envelopes over the frequency dimension. To do so, the eigenvectors of the covariance matrix of the clean log-temporal envelopes are computed.

3. Short-time correlation coefficients for the eigenchannels are computed and then averaged to produce a final value. The short-time segmentation approach in (Taal et al., 2011a) is used, but the clipping procedure is not.

By comparing the performance of STOI with STOI$^{\mathrm{KLT}}$, and SIIB with SIIB$^{\mathrm{noKLT}}$ the effect of decorrelating input features can be investigated.

5.4.2 Investigating the effect of the auditory model

The auditory model that is used to extract features could have a significant impact on performance. To investigate this effect, the auditory model used for STOI$^{\mathrm{KLT}}$ (i.e., STOIs auditory model) was replaced with the auditory model used by SIIB. The differences between the auditory models are: 1) SIIB uses an ERB gammatone filterbank, whereas STOI uses a 1/3 octave band rectangular filterbank, 2) SIIB considers frequencies up to 8 kHz, whereas STOI considers frequencies up to 5 kHz, and 3) SIIB includes a forward temporal masking function, whereas STOI does not. The resulting intelligibility metric is denoted STOI$^{\mathrm{KLT}}_{\mathrm{gamma}}$.

5.4.3 Investigating the effect of mutual information estimation

The majority of the intelligibility metrics in the previous section rely on the correlation coefficient to quantify distortion. On the other hand, SIIB and MIKNN use a non-parametric mutual information estimator. Recall that if the clean and degraded signals are jointly Gaussian, then the mutual information is a function of the correlation coefficient only. In Jensen and Taal (2014) this observation was used to justify the use of the correlation coefficient. However, a direct comparison between the performance obtained using a non-parametric mutual information estimator and the performance obtained using the capacity of a Gaussian channel has not been made.

To investigate the effect of mutual information estimation, SIIB was modified to produce a simpler metric called SIIB$^{\text{Gauss}}$. The original SIIB algorithm (Van Kuyk et al., 2018b) quantifies distortion using a KNN mutual information estimator, whereas SIIB$^{\text{Gauss}}$ uses the information capacity of a Gaussian channel. Concretely,

$$\text{SIIB}^{\text{Gauss}} = -\frac{R}{2K}\sum_{\lambda}\log_2(1 - r^2\rho_\lambda^2), \tag{5.1}$$

where R is the frame rate, $K = 15$ is the number of stacked log-spectra, $r = 0.75$ is the production noise correlation coefficient, λ is the eigenchannel index, and ρ_λ is the correlation coefficient between the λth clean eigenchannel and the λth distorted eigenchannel. The values for R, K and r are the same as those in (Van Kuyk et al., 2018b).

5.5 Performance Criteria

The key requirement of an intelligibility metric is that it has a strong monotonic increasing relationship with intelligibility. This chapter uses two performance criteria to quantify the strength of the relationship: Kendall's tau coefficient, τ, and Pearson's correlation coefficient, ρ. Both performance criteria are discussed below.

In the following, p_c is the intelligibility in terms of percentage of words correctly identified for condition c in a particular data set and $d(x_c, y_c)$ is the corresponding score computed by an intelligibility metric. The clean signal x_c is formed by concatenating all available clean sentences for condition c and likewise for the distorted signal y_c.

5.5.1 Kendall's tau coefficient

Kendall's tau coefficient (Kendall, 1938), τ, measures the ordinal association between two quantities. Let i and j be two conditions in a data set where $i \neq j$. The pair formed by $(w_i, d(x_i, y_i))$ and $(w_j, d(x_j, y_j))$ is concordant if both $w_i > w_j$ and $d(x_i, y_i) > d(x_j, y_j)$, or if both $w_i < w_j$ and $d(x_i, y_i) < d(x_j, y_j)$. The pair is disconcordant if $w_i > w_j$ and $d(x_i, y_i) < d(x_j, y_j)$ or if $w_i < w_j$ and $d(x_i, y_i) > d(x_j, y_j)$. Kendall's tau coefficient is given by

$$\tau = \frac{n_C - n_D}{n(n-1)/2}, \tag{5.2}$$

where n_C is the number of concordant pairs, n_D is the number of disconcordant pairs, and n is the number of conditions in the data set. Kendall's tau coefficient ranges between -1 and 1. If $\tau = -1$ then p_c and $d(x_c, y_c)$ have a monotonic decreasing relationship, if $\tau = 1$ they have a monotonic increasing relationship, and if they are statistically independent then $\tau = 0$.

5.5.2 Pearson's correlation coefficient

Pearson's correlation coefficient, ρ, is defined as the normalised covariance between two quantities. To use ρ effectively, the relationship between the quantities must be linear. For this reason, a monotonic function f is applied to $d(x_c, y_c)$ to linearise the relationship before computing ρ. The function f can be thought of as a mapping from the metric to predicted intelligibility scores, but more generally it is simply a tool for quantifying the strength of the relationship between $d(x_c, y_c)$ and p_c.

In the literature f is commonly assumed to be a logistic function, e.g., (Gordon-Salant and Fitzgibbons, 1995; Kates and Arehart, 2005; Taal et al., 2011a):

$$f(d(x_c, y_c)) = \frac{100}{1 + e^{a(d(x_c, y_c) - b)}}, \tag{5.3}$$

where b is the midpoint and a is the slope at the midpoint. These parameters are fitted to the data to minimise the mean squared error between p_c and $f(d(x_c, y_c))$.

In the literature ρ is sometimes also computed without applying a mapping function. However, we believe that such a measure is misleading because without f, a metric with a strong non-linear relationship between p_c and $d(x_c, y_c)$ will have a small value for ρ, but could also have a monotonic increasing relationship with intelligibility.

Note that p_c depends on the experimental procedures used to measure intelligibility, but that $d(x_c, y_c)$ does not. For example, the intelligibility of a given stimulus can be increased by changing an open listening test to a closed listening test[1]. It follows that the relationships between intelligibility and intelligibility metrics also depend on experimental procedures. For this reason, f is fit individually to each data set. Pearson's correlation coefficient is calculated according to

$$\rho = \frac{\sum_c (p_c - \bar{w}_c)(f(d(x_c, y_c)) - \bar{f}(d(x_c, y_c)))}{\sqrt{\sum_c (p_c - \bar{w}_c)^2 \sum_c (f(d(x_c, y_c)) - \bar{f}(d(x_c, y_c)))^2}}, \tag{5.4}$$

where the overbar is used to denote the mean over all conditions in the data set. Finally, because the relationship between an intelligibility metric and intelligibility should be monotonically increasing, negative values of ρ and τ are set to zero.

5.6 Results

Scatter plots for all data sets described in Section II and all pre-existing intelligibility metrics described in Section III are displayed in Figure 5.1.

[1] In a closed listening test, subjects are given a list of possible speech sounds, e.g., phones or words, and are asked to identify the sounds that they heard. In an open listening test, no list is provided, which makes the test more difficult.

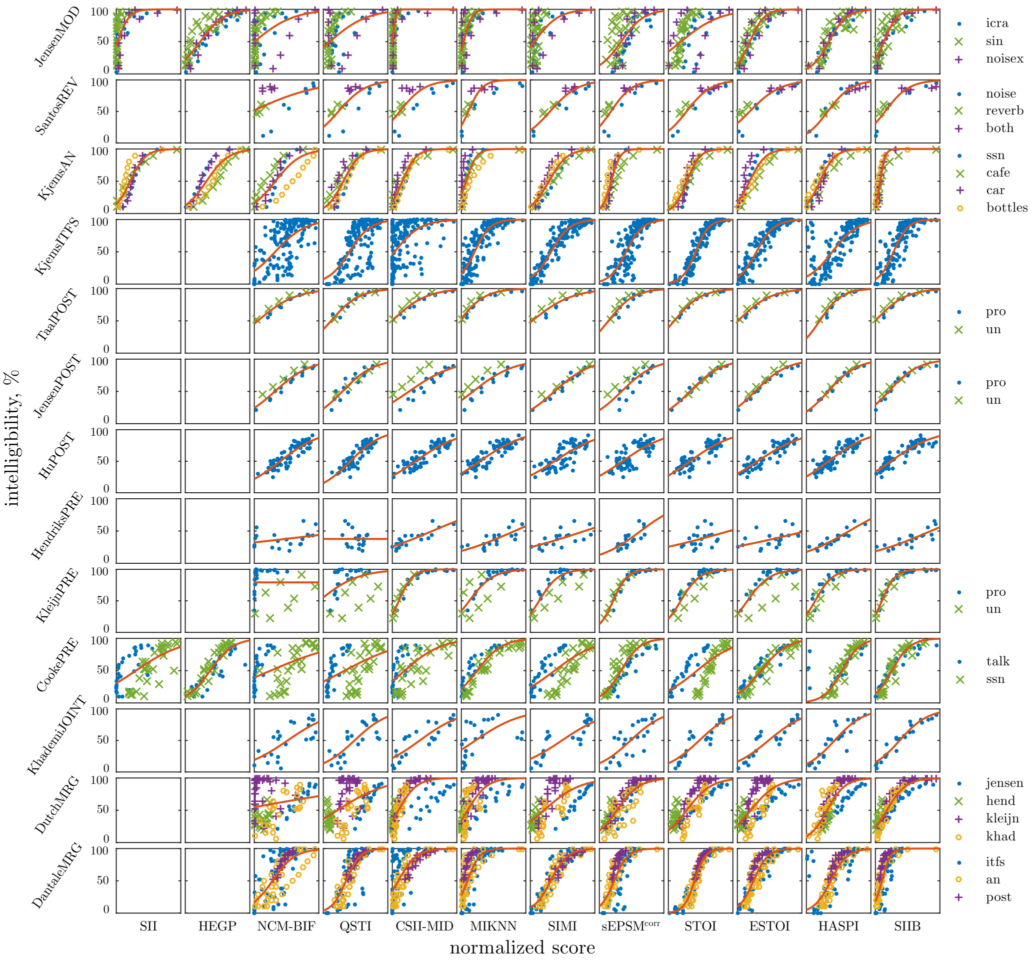

Figure 5.1: Scatter plots for all data sets and pre-existing intelligibility metrics. The vertical axis is the ′ground-truth′ intelligibility in terms of the percentage of words correctly identified during listening tests, and the horizontal axis is the score computed by an intelligibility metric. The horizontal axis of each plot has been normalised to be between 0 and 1. Each data point corresponds to a processing condition. The mapping function in (5.3) is also shown.

Each row of plots corresponds to a data set and each column of plots corresponds to an intelligibility metric. The vertical axis of each scatter plot is the 'ground-truth' intelligibility in terms of the percentage of words correctly identified during listening tests, and the horizontal axis is the score computed by an intelligibility metric. To facilitate an easy visual comparison, the horizontal axis of each scatter plot is normalised to be between 0 and 1. Each point on a scatter plot corresponds to a condition in the respective data set. The function in (5.3) that was used to linearise the relationship between the intelligibility scores and the metric for each data set is also shown. For an ideal intelligibility metric, all points would fall exactly on top of the fitted curve.

The labels 'icra', 'sin', 'noisex', 'noise', 'reverb', 'both', 'ssn', 'cafe', 'car', 'bottles', 'talk', and 'ssn' in Figure 5.1 indicate the type of environmental degradation in the data set. The labels 'pro' and 'un' indicate whether a stimulus was processed by an enhancement algorithm or was unprocessed. The labels 'jensen', 'hend', 'kleijn', 'khad', 'itfs', 'an', and 'post' refer to individual data sets within the merged data sets.

Table 5.3 displays Kendall's tau coefficient for all data sets and intelligibility metrics and, similarly, Table 5.4 displays Pearson's correlation coefficient. In both tables, an asterisk is used to indicate when a data set was used during the development of an intelligibility metric. For the remainder of the chapter, 'unseen' refers to a data set that was not used during development, and 'seen' refers to a data set that was used during development. The mean performance of each intelligibility metric and a confidence interval, [$\mathrm{CI}_{\mathrm{low}}$, $\mathrm{CI}_{\mathrm{high}}$], with 95% coverage of the mean performance is also included. The confidence intervals were calculated using the non-parametric BC_{a} bootstrap approach (Efron, 1987). To do so, 5000 bootstrap sample sequences of p_c and $d(x_c, y_c)$ were generated for each data set and intelligibility metric. The sample distribution of the mean performance of each intelligibility metric was then estimated from the bootstrap sample sequences.

From here on, subscripts are used to indicate performance criteria for particular intelligibility metrics. For example, ρ_{SIIB}, refers to the correla-

Table 5.3: Performance in terms of Kendall's tau coefficient, τ, for all data sets and intelligibility metrics. The intelligibility metrics are listed in order of mean performance and are grouped by pre-existing metrics (left) and modified metrics (right).

	SII	HEGP	NCM-BIF	QSTI	CSII-MID	MIKNN	SIMI	sEPSM$^{\text{corr}}$	STOI	ESTOI	HASPI	SIIB	SIIB$^{\text{noKLT}}$	STOI$^{\text{KLT}}$	STOI$^{\text{KLT}}_{\text{gamma}}$	SIIB$^{\text{Gauss}}$
JensenMOD	0.52	0.71	0.41	0.34	0.57	0.55	0.34	0.51	0.38	0.75*	0.75	0.74*	0.59	0.72	0.71	0.74
SantosREV	–	–	0.38	0.61	0.57	0.70	0.72	0.72	0.82	0.79	0.85	0.82*	0.82	0.79	0.79	0.80
KjemsAN	0.76	0.75	0.65	0.78	0.80	0.65	0.81*	0.74	0.81	0.74*	0.79	0.82*	0.74	0.74	0.76	0.84
KjemsITFS	–	–	0.48	0.51*	0.41	0.71*	0.80*	0.70*	0.82*	0.81*	0.66	0.73*	0.69	0.73	0.74	0.73
TaalPOST	–	–	0.85	0.87*	0.81	0.83*	0.81	0.79	0.92*	0.96	0.83	0.87*	0.87	0.79	0.79	0.87
JensenPOST	–	–	0.81	0.80	0.60	0.68	0.92*	0.66	0.89	0.83*	0.95	0.92*	0.65	0.82	0.82	0.94
HuPOST	–	–	0.67*	0.68*	0.63	0.64	0.55	0.44	0.59	0.69	0.61	0.74*	0.39	0.72	0.70	0.73
HendriksPRE	–	–	0.30	0.00	0.69	0.56	0.52	0.59	0.26	0.43	0.78	0.66*	0.72	0.53	0.62	0.60
KleijnPRE	–	–	0.13	0.20	0.86	0.71	0.57	0.88	0.70	0.58	0.79	0.86*	0.77	0.78	0.88	0.86
CookePRE	0.44	0.72*	0.38	0.38	0.46	0.72	0.52	0.71	0.56	0.77	0.75	0.76*	0.71	0.87	0.84	0.77
KhademiJOINT	–	–	0.50	0.51	0.71	0.53	0.74	0.60	0.79	0.80	0.77	0.89*	0.90	0.82	0.87	0.90
DutchMRG	–	–	0.13	0.29	0.57	0.54	0.44	0.68	0.59	0.46	0.64	0.75*	0.58	0.54	0.67	0.74
DantaleMRG	–	–	0.54	0.64	0.53	0.61	0.80	0.66	0.83	0.75	0.67	0.68*	0.58	0.70	0.73	0.71
Mean	0.57	0.73	0.48	0.51	0.63	0.65	0.66	0.67	0.69	0.72	0.76	0.79	0.69	0.73	0.76	0.79
CI$_{\text{low}}$	0.50	0.68	0.43	0.46	0.59	0.61	0.61	0.63	0.65	0.68	0.72	0.75	0.66	0.70	0.73	0.76
CI$_{\text{high}}$	0.64	0.77	0.52	0.55	0.67	0.69	0.69	0.70	0.73	0.75	0.78	0.81	0.72	0.76	0.79	0.81

Table 5.4: Performance in terms of Pearson's correlation coefficient, ρ, for all data sets and intelligibility metrics. The intelligibility metrics are listed in order of mean performance and are grouped by pre-existing metrics (left) and modified metrics (right).

	SII	HEGP	NCM-BIF	$\mathrm{\hat{Q}STI}$	CSII-MID	MIKNN	SIMI	$\mathrm{sEPSM}^{\mathrm{corr}}$	STOI	ESTOI	HASPI	SIIB	$\mathrm{SIIB}^{\mathrm{noKLT}}$	$\mathrm{STOI}^{\mathrm{KLT}}$	$\mathrm{STOI}^{\mathrm{KLT}}_{\mathrm{gamma}}$	$\mathrm{SIIB}^{\mathrm{Gauss}}$
JensenMOD	0.65	0.88	0.45	0.43	0.65	0.72	0.51	0.68	0.47	0.92*	0.92	0.89*	0.78	0.90	0.88	0.89
SantosREV	–	–	0.46	0.76	0.72	0.90	0.94	0.87	0.94	0.91	0.97	0.93*	0.98	0.93	0.95	0.93
KjemsAN	0.89	0.89	0.80	0.90	0.92	0.78	0.93*	0.87	0.93	0.87*	0.93	0.94*	0.88	0.88	0.89	0.94
KjemsITFS	–	–	0.67	0.72*	0.49	0.88*	0.95*	0.84*	0.96*	0.95*	0.78	0.89*	0.83	0.89	0.91	0.89
TaalPOST	–	–	0.95	0.95*	0.93	0.95*	0.92	0.90	0.98*	0.97	0.95	0.96*	0.96	0.92	0.92	0.96
JensenPOST	–	–	0.95	0.93	0.78	0.86	0.97*	0.80	0.99	0.97*	0.99	0.98*	0.77	0.95	0.96	0.98
HuPOST	–	–	0.89*	0.89*	0.89	0.88	0.77	0.73	0.87	0.90	0.88	0.92*	0.65	0.91	0.92	0.92
HendriksPRE	–	–	0.29	0.00	0.86	0.76	0.66	0.78	0.35	0.47	0.92	0.82*	0.91	0.65	0.77	0.73
KleijnPRE	–	–	0.00	0.34	0.98	0.82	0.87	0.98	0.92	0.81	0.94	0.97*	0.97	0.91	0.99	0.98
CookePRE	0.62	0.90*	0.47	0.49	0.65	0.90	0.69	0.89	0.70	0.94	0.86	0.94*	0.90	0.96	0.97	0.95
KhademiJOINT	–	–	0.74	0.80	0.87	0.53	0.84	0.75	0.90	0.90	0.87	0.96*	0.96	0.91	0.97	0.95
DutchMRG	–	–	0.19	0.49	0.74	0.72	0.65	0.85	0.82	0.69	0.81	0.92*	0.77	0.75	0.87	0.91
DantaleMRG	–	–	0.72	0.81	0.68	0.76	0.94	0.78	0.96	0.90	0.77	0.82*	0.72	0.86	0.89	0.85
Mean	0.72	0.89	0.58	0.66	0.78	0.80	0.82	0.82	0.83	0.86	0.89	0.92	0.85	0.88	0.91	0.92
$\mathrm{CI}_{\mathrm{low}}$	0.65	0.85	0.53	0.61	0.74	0.76	0.79	0.79	0.80	0.82	0.86	0.90	0.83	0.86	0.89	0.89
$\mathrm{CI}_{\mathrm{high}}$	0.78	0.92	0.63	0.70	0.81	0.83	0.85	0.85	0.86	0.89	0.91	0.93	0.87	0.90	0.93	0.93

tion coefficient that SIIB achieved on some data set.

5.6.1 Remarks for the pre-existing metrics

It is clear that out of the pre-existing metrics SIIB and HASPI have the highest performance overall, on average achieving τ_{SIIB} = 0.79 and ρ_{SIIB} = 0.92, and τ_{HASPI} = 0.76 and ρ_{HASPI} = 0.89. This performance is followed closely by ESTOI, which has an average score of τ_{ESTOI} = 0.72 and ρ_{ESTOI} = 0.86. HEGP has high performance for data sets distorted by additive noise achieving an average score of τ_{HEGP} = 0.73 and ρ_{HEGP} = 0.89, but its usefulness is limited to situations where noise is the main source of degradation and where the noise signal is available.

The top performance rating of SIIB may be criticized on the grounds that SIIB has been 'over-designed' for the data sets in this evaluation. Although the parameters of SIIB were not intentionally optimised for the data sets in this chapter, the developers of SIIB were the only researchers with access to all the data sets and thus had greater opportunity to redesign their algorithm when weaknesses were exposed during SIIBs development.

Many of the intelligibility metrics performed poorly on HendriksPRE. This is likely due to the large T60 time of the room impulse response that causes severe reverberant distortion. As shown in Figure 5.2, the large T60 time somewhat 'blurs' the time-alignment of clean and degraded temporal envelopes. Many intrusive intelligibility metrics require that the clean and degraded signals are strictly time-aligned, and thus are over-sensitive to temporal blurring. Out of all the intelligibility metrics in this evaluation, HASPI achieved the highest performance for HendriksPRE (τ_{HASPI} = 0.78, ρ_{HASPI} = 0.92) and is also the only intelligibility metric that included time-alignment processing.

Recall that HASPI is computed as a linear combination of four terms: the cepstral correlation, and three auditory coherence terms. The weights

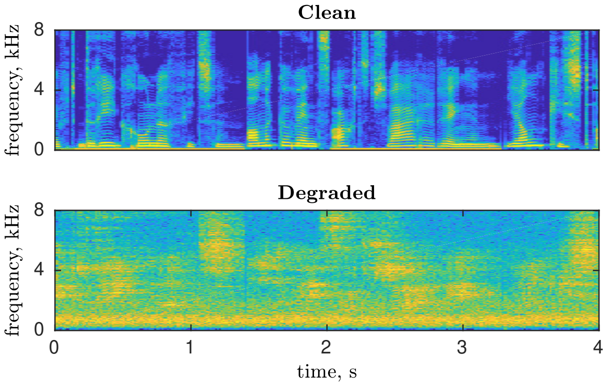

Figure 5.2: An example of a clean and degraded stimulus from HendriksPRE. The severe reverberant distortion 'blurs' the time-alignment between the stimuli.

in the linear combination were optimised for each data set to maximise performance. None of the other intelligibility metrics modify their parameters based on the data, suggesting that the high performance of HASPI may be attributed to overfitting. To test this hypothesis, HASPI was computed simply by summing the cepstral correlation term and the high-energy auditory coherence term with equal weight. Doing so reduced the mean performance of HASPI to $\tau_{\text{HASPI}} = 0.73$ and $\rho_{\text{HASPI}} = 0.88$, which is still very high. Thus, the high performance of HASPI is unlikely the result of overfitting.

Another criteria that can be used to evaluate performance is whether a metric gives consistent predictions across classes of distortions. For example, CookePRE has two distinct classes: stimuli degraded by a competing talker, and stimuli degraded by SSN. Metrics may give consistent intelligibility predictions within a class, but could give inconsistent predictions between classes. An example of this can be seen in the scatter plot corresponding to STOI and DutchMRG. STOI gives consistent predictions for JensenPOST, KleijnPRE, and KhademiJOINT, but when the data sets are merged together we see distinct clusters corresponding to each data set. This means that for a given clean stimulus, a STOI score of 0.5 for noise-

Table 5.5: Mean performance of pre-existing intelligibility metrics for 'seen' and 'unseen' data sets.

	SII	HEGP	NCM-BIF	QSTI	CSII-MID	MIKNN	SIMI	$\text{sEPSM}^{\text{corr}}$	STOI	ESTOI	HASPI	SIIB
mean τ^{seen}	–	0.72	0.67	0.69	–	0.77	0.84	0.70	0.87	0.78	–	0.79
mean τ^{unseen}	0.57	0.73	0.46	0.45	0.63	0.63	0.60	0.66	0.66	0.69	0.76	–
mean ρ^{seen}	–	0.90	0.89	0.86	–	0.92	0.95	0.84	0.97	0.93	–	0.92
mean ρ^{unseen}	0.72	0.88	0.56	0.60	0.78	0.78	0.78	0.82	0.80	0.84	0.89	–

reduced speech and a STOI score of 0.5 for pre-processed speech could correspond to different intelligibility scores.

5.6.2 Investigating the performance in terms of generalisation

Considering only entries in Table 5.3 and Table 5.4 that have an asterisk, the mean performance of all such entries for all pre-existing metrics and data sets is $\tau = 0.78$ and $\rho = 0.92$. Considering only entries that do not have an asterisk, the mean performance for all pre-existing metrics and data sets is $\tau = 0.62$ and $\rho = 0.76$. This result demonstrates that, in general, intelligibility metrics have high performance for seen data sets, and poor performance for unseen data sets.

To further investigate the performance of intelligibility metrics in terms of their ability to generalise, Table 5.5 displays the mean performance for unseen data sets and seen data sets for each pre-existing intelligibility metric. HASPI has the highest performance for unseen data sets achieving $\tau^{\text{unseen}}_{\text{HASPI}} = 0.76$ and $\rho^{\text{unseen}}_{\text{HASPI}} = 0.89$. HEGP also has high performance for unseen data sets, however, recall that HEGP was evaluated exclusively on data sets with additive noise degradation.

STOI and SIMI both have outstanding performance for seen data sets

($\tau_{\mathrm{STOI}}^{\mathrm{seen}}$ = 0.87, $\rho_{\mathrm{STOI}}^{\mathrm{seen}}$ = 0.97, and $\tau_{\mathrm{SIMI}}^{\mathrm{seen}}$ = 0.84, $\rho_{\mathrm{SIMI}}^{\mathrm{seen}}$ = 0.95), but poor performance for unseen data sets ($\tau_{\mathrm{STOI}}^{\mathrm{unseen}}$ = 0.66, $\rho_{\mathrm{STOI}}^{\mathrm{unseen}}$ = 0.80, and $\tau_{\mathrm{SIMI}}^{\mathrm{unseen}}$ = 0.60, $\rho_{\mathrm{SIMI}}^{\mathrm{unseen}}$ = 0.78). This is because STOI and SIMI were specifically designed for speech processed by ITFS and noise-reduction algorithms, whereas the data sets in this evaluation include degradation caused by reverberation and modulated noise sources. Similarly, NCM-BIF was designed specifically for speech processed by noise-reduction algorithms. Observe that in Figure 5.1 NCM-BIF has good performance for the data sets with noise-reduction: HuPOST, JensenPOST, and TaalPOST, but poor performance for the remaining data sets. These results show the danger of using intelligibility metrics outside of their intended domain.

In light of the above paragraphs, to ensure that future intelligibility metrics generalise to new data sets and give consistent predictions between classes, it may be more beneficial to gather data points with different types of degradation than to collect many data points for a single type of degradation. This notion is consistent with the high performance of HASPI, which considered six types of degradation during development: additive noise, envelope-clipping, ITFS processing, frequency-compression, noise reduction, and vocoded-speech.

5.6.3 Remarks for the modified intelligibility metrics

In general, removing the KLT from SIIB significantly reduced performance (on average $\tau_{\mathrm{SIIB}^{\mathrm{no\_KLT}}} = 0.69$ and $\rho_{\mathrm{SIIB}^{\mathrm{no\_KLT}}} = 0.85$). Furthermore, introducing the KLT to STOI improved performance (on average $\tau_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.73$ and $\rho_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.88$). The increase in overall performance for STOI$^{\mathrm{KLT}}$ is mainly due to large increases in performance for JensenMOD, HendriksPRE, and CookePRE. Note that STOI$^{\mathrm{KLT}}$ performs worse than STOI for KjemsITFS and TaalPOST, however, these are the same data sets that were used to tune the parameters of STOI during STOIs development.

The five intelligibility metrics with the highest performance: SIIB, SIIB$^{\mathrm{Gauss}}$, STOI$^{\mathrm{KLT}}_{\mathrm{gamma}}$, HASPI, and STOI$^{\mathrm{KLT}}$ are also the only metrics that decorrelate log-spectra. This outcome clearly demonstrates the advantage that can be obtained by reducing the statistical dependencies between input features.

Recall that ESTOI was proposed as an extension to STOI that can 'listen to glimpses of clean speech'. Interestingly, for the data sets that contain modulated noise, STOI$^{\mathrm{KLT}}$ has similar performance to ESTOI (for JensenMOD, $\tau_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.72$, $\rho_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.90$, and for CookePRE, $\tau_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.87$, $\rho_{\mathrm{STOI}^{\mathrm{KLT}}} = 0.96$). SIIB and SIIB$^{\mathrm{Gauss}}$, which are based on long-term statistics, also have good performance for JensenMOD and CookePRE. Such results contest the idea that short-time segmentation is necessary for predicting the intelligibility of modulated noise sources.

On average STOI$^{\mathrm{KLT}}_{\mathrm{gamma}}$ achieved $\tau_{\mathrm{STOI}^{\mathrm{KLT}}_{\mathrm{gamma}}} = 0.76$ and $\rho_{\mathrm{STOI}^{\mathrm{KLT}}_{\mathrm{gamma}}} = 0.91$. Thus, by introducing the KLT to STOI and using a more realistic auditory model, performance competitive with SIIB could be obtained. This means that for some representations of speech signals, the correlation coefficient and the KNN mutual information estimator can quantify distortion equally well. A partial explanation for this result can be found by considering the high performance of SIIB$^{\mathrm{Gauss}}$ ($\rho_{\mathrm{SIIB}^{\mathrm{Gauss}}} = 0.92$ and $\tau_{\mathrm{SIIB}^{\mathrm{Gauss}}} = 0.79$), which suggests that the Gaussian communication channel is a reasonable approximation of the true communication channel for many real-word distortions.

Finally, recall that SIIB$^{\mathrm{Gauss}}$ $= -\frac{F}{2K}\sum_j \log_2(1 - r^2\rho_j^2)$. Since r and ρ_j are between -1 and 1, the product of their squares is likely to be small, particularly for challenging listening environments. Using the approximation $\log_2(1+a) \approx a/\ln(2)$ for small a, we have that SIIB$^{\mathrm{Gauss}}$ $\approx \frac{F}{2K\ln(2)} r^2 \sum_j \rho_j^2$. This approximation strongly resembles the distortion measure used by STOI$^{\mathrm{KLT}}$ and STOI$^{\mathrm{KLT}}_{\mathrm{gamma}}$, which can be written as $\sum_j \sum_t \rho_{j,t}$, where t is the short-time segment index.

5.7 Summary of Chapter

In this chapter, the accuracy of 12 intelligibility metrics from the literature was evaluated using the results of 11 listening tests. The stimuli included pre-processing enhancement, post-processing enhancement, and environmental distortions such as noise and reverberation. In order to analyse why the top performing metrics have high performance, four new intelligibility metrics were proposed. The main conclusions are as follows.

1. Out of the pre-existing metrics, SIIB and HASPI had the highest overall performance.

2. Many intrusive metrics struggle with severe reverberant distortion. This may be because they are over-sensitive to the time-alignment of clean and distorted temporal envelopes.

3. In general, intelligibility metrics perform more poorly on unseen data sets than on seen data sets. For this reason, caution should be taken when using intelligibility metrics outside of their intended domain.

4. For unseen data sets, HASPI had the highest performance. This suggests that HASPI is appropriate for situations where many types of potentially new speech material and distortions are likely. Additionally, unlike the other metrics, HASPI has built-in time-alignment processing and can account for hearing impairments.

5. The five intelligibility metrics with the highest overall performance are also the only metrics that decorrelate log-spectra. On average, introducing the KLT to STOI improved performance and removing the KLT from SIIB reduced performance. These results demonstrate the advantage of removing statistical dependencies between input features.

6. The high performance of SIIB$^{\text{Gauss}}$ suggests that the Gaussian communication channel is a reasonable approximation of the true communication channel for many real-world distortions. Additionally, SIIB$^{\text{Gauss}}$ has performance similar to SIIB, but takes less time to compute by two orders of magnitude.[2]

7. It was shown that STOI$^{\text{KLT}}$ and STOI$^{\text{KLT}}_{\text{gamma}}$ can be interpreted as approximations of SIIB$^{\text{Gauss}}$.

[2] MATLAB implementations of SIIB$^{\text{Gauss}}$ and SIIB are available at www.stevenvankuyk.com/MATLAB_code

Chapter 6

Estimating Mutual Information Using Siamese Networks

In Chapter 3 and Chapter 4 the information rate of the speech production channel $I(\{\mathbf{m}_t\};\{\mathbf{x}_t\})$ was estimated, where $\{\mathbf{m}_t\}$ is a hypothetical message and $\{\mathbf{x}_t\}$ is speech produced by a talker that has been processed by an auditory model. For the proposed communication model, the mutual information rate depends only on the mutual information between $\mathbf{m}^K$ and $\mathbf{x}^K$, which are obtained by stacking K consecutive vectors of $\{\mathbf{m}_t\}$ and $\{\mathbf{x}_t\}$, respectively. To estimate the mutual information $I(\mathbf{m}^K;\mathbf{x}^K)$, several assumptions about the probability distribution $P(\mathbf{m}^K,\mathbf{x}^K)$ were made. Specifically, $\mathbf{m}^K$ and $\mathbf{x}^K$ were assumed to be jointly Gaussian, in which case the production noise is modelled as additive Gaussian noise. To account for statistical dependencies between time-frequency units, an invertible transform q was introduced. For Gaussian random variables, the KLT is a reasonable function for q.

The proposed communication model led to a state-of-the-art intelligibility metric that can accurately predict intelligibility for a wide range of real-world distortions. However, the resulting estimate of the information rate of speech communication remains larger than the lexical information rate described in Section 2.2. This suggests that some of the assumptions made during the development of the model may over-simplify speech communication.

The approach proposed in the present chapter extends the work of the previous chapters by removing the assumption that $\mathbf{m}^K$ and $\mathbf{x}^K$ are jointly Gaussian. Instead, this chapter considers a family of functions for the transform q and finds a q such that $q(\mathbf{x}^K) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Doing so facilities estimation of the information rate of speech communication. To find an appropriate function for q, techniques from deep learning are relied on. Specifically, a *Siamese neural network* (Bromley et al., 1994; Chopra et al., 2005) and *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2007) are used.

Unlike the previous chapters, the work in this chapter has not been published or peer-reviewed. In addition, this chapter does not use real-world data. Instead, theory is developed and the approach is demonstrated on artificial examples. The work in this chapter should thus be viewed as a preliminary study that could be refined and then applied to real-world data as a future research topic.

Finally, this chapter uses simplified notation. First, all time indices are removed, which is reasonable because the stochastic processes are stationary. Second, this chapter does not consider the mutual information rate. Instead, this chapter focuses on mutual information. As shown in the previous chapters, it is easy to extend the analysis from mutual information to mutual information rate by stacking consecutive vectors. Thus, in this chapter, the message vector is denoted by $\mathbf{m}$, the speech vector is denoted by $\mathbf{x}$, and the mutual information of the speech production channel is denoted $I(\mathbf{m}; \mathbf{x})$.

6.1 Mutual Information Estimation as an Optimisation Problem

It is well known that given some target probability distribution P_{target} and a continuous random variable $\mathbf{x}$, a deterministic function q exists such that $q(\mathbf{x}) \sim P_{\text{target}}$ (Kallenberg, 2006). In other words a random variable can be transformed such that the transformed random variable is distributed according to some target distribution.

The fundamental principle of the method proposed in this chapter is to convert the speech vector, $\mathbf{x}$, into a zero-mean multivariate Gaussian random variable with statistically independent vector elements. We call the converted speech the *latent variable* and it is given by $\mathbf{z} = q(\mathbf{x})$ for some deterministic function q. Given that $\mathbf{z}$ is Gaussian, the estimation of mutual information can be considerably simplified.

Because q is deterministic, $\mathbf{m} \to \mathbf{x} \to \mathbf{z}$ forms a Markov chain. Furthermore, due to the data processing inequality (Cover and Thomas, 2012), we have that $I(\mathbf{m}; \mathbf{x}) \geq I(\mathbf{m}; \mathbf{z})$, with equality if q is invertible. Thus, estimating the mutual information of the message and the latent variable provides a way to estimate the mutual information of the message and the speech without placing assumptions on the joint distribution $P(\mathbf{x}, \mathbf{m})$.

Concretely, our goal is to solve the following optimisation problem:

$$\begin{aligned} &\underset{q}{\text{maximise}} && I(\mathbf{m}; \mathbf{z}) \\ &\text{subject to} && \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \end{aligned} \tag{6.1}$$

where $\mathbf{z} = q(\mathbf{x})$. In practice, q is selected from a family of functions parametised by a vector $\boldsymbol{\theta}$. To this end, q is implemented by a *feedforward neural network* with weights and biases denoted by $\boldsymbol{\theta}$. From here on, q_θ denotes the neural network.

In the following, the above optimisation problem is reformulated and a differentiable Lagrangian function is derived. The Lagrangian is used to find local optima by using *stochastic gradient descent* and *back-propagation* to learn the values for $\boldsymbol{\theta}$.

6.1.1 Maximum Mean Discrepancy

One way to enforce the constraint in (6.1) that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is to use *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2007). MMD is distance measure for probability distributions that was originally proposed as a non-parametric test statistic for the *two-sample problem*. Given two sets of observations from two probability distributions, the two-sample problem is a statistical test of the null hypothesis that the two distributions are equal against the alternative hypothesis that the distributions are different.

More recently, MMD has been used in deep learning to constrain a latent variable to have a desired target distribution (e.g., Dziugaite et al., 2015; Zhao et al., 2018; Braithwaite and Kleijn, 2018). Let $P(\mathbf{z})$ be the probability distribution of the latent variable $\mathbf{z}$ and let P_{target} be the target distribution. For the optimisation problem in (6.1) the target distribution is $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. It can be shown that MMD is zero if and only if $\mathbf{z} \sim P_{\text{target}}$. Thus, the constraint in (6.1) can be replaced with the constraint that $\text{MMD} = 0$. In practice, it is not possible to make MMD exactly zero using neural networks because there is always some statistical noise, but this is unlikely to cause major problems.

MMD is based on functions in the unit ball of a reproducing kernel Hilbert space. Let $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(N)}\}$ denote a set of N independent samples from $P(\mathbf{z})$ and let $\{\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(N)}\}$ denote a set of N independent samples from P_{target}. An unbiased empirical estimate of MMD^2 is given by

(Dziugaite et al., 2015)

$$\begin{aligned}\mathrm{MMD}^2 = & \frac{1}{N(N-1)}\sum_{n\neq n'} k(\boldsymbol{z}^{(n)}, \boldsymbol{z}^{(n')}) \\ & + \frac{1}{N(N-1)}\sum_{m\neq m'} k(\boldsymbol{v}^{(m)}, \boldsymbol{v}^{(m')}) \\ & - \frac{2}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} k(\boldsymbol{z}^{(n)}, \boldsymbol{v}^{(m)}),\end{aligned} \tag{6.2}$$

where k is a kernel function. As is conventional, in this chapter a Gaussian kernel is used:

$$k(\boldsymbol{z}_n, \boldsymbol{z}_{n'}) = \exp\left(\frac{-||\boldsymbol{z}_n - \boldsymbol{z}_{n'}||_2^2}{\sigma^2}\right), \tag{6.3}$$

where σ^2 is the *kernel bandwidth*. For more information on the theoretical properties and motivation of MMD see Gretton et al. (2007), and for more details on training neural networks using MMD see Dziugaite et al. (2015).

6.1.2 Mutual Information

We are interested in maximising $I(\mathbf{m};\mathbf{z})$ because, due to the data processing inequality, $I(\mathbf{m};\mathbf{z})$ is a lower bound for $I(\mathbf{m};\mathbf{x})$. In this section, a lower bound for $I(\mathbf{m};\mathbf{z})$ is developed. By maximising the lower bound of $I(\mathbf{m};\mathbf{z})$, $I(\mathbf{m};\mathbf{x})$ can be estimated.

The mutual information between $\mathbf{m}$ and $\mathbf{z}$ is given by

$$I(\mathbf{m};\mathbf{z}) = h(\mathbf{z}) - h(\mathbf{z}|\mathbf{m}), \tag{6.4}$$

where $h(\mathbf{z})$ is the differential entropy of the latent variable and $h(\mathbf{z}|\mathbf{m})$ is conditional differential entropy of the latent variable given the message.

Because $\mathbf{z}$ is constrained to be Gaussian with statistically independent vector elements, the differential entropy $h(\mathbf{z})$ is given by

$$h(\mathbf{z}) = \sum_{\lambda}\frac{1}{2}\log_2 2\pi e \mathrm{var}(\mathrm{z}_\lambda), \tag{6.5}$$

where z_λ is the λ'th vector element of $\mathbf{z}$. Note that when the constraint in (6.1) is satisfied, $\mathrm{var}(\mathrm{z}_\lambda) = 1$ for all λ.

Because the elements of $\mathbf{z}$ are constrained to be statistically independent, the conditional differential entropy $h(\mathbf{z}|\mathbf{m})$ is given by

$$h(\mathbf{z}|\mathbf{m}) = \sum_\lambda h(\mathrm{z}_\lambda|\mathbf{m}). \tag{6.6}$$

From the definition of conditional differential entropy, we have that:

$$\begin{aligned} h(\mathrm{z}_\lambda|\mathbf{m}) &= \int_{\boldsymbol{m}} P(\boldsymbol{m}) h(\mathrm{z}_\lambda|\boldsymbol{m}) d\boldsymbol{m} & (6.7)\\ &= -\int_{\boldsymbol{m}} P(\boldsymbol{m}) \int_{z_\lambda} P(z_\lambda|\boldsymbol{m}) \log_2 P(z_\lambda|\boldsymbol{m}) dt_\lambda d\boldsymbol{m}. & (6.8)\end{aligned}$$

Recall that, for a fixed variance, $-\int_{z_\lambda} P(z_\lambda|\boldsymbol{m}) \log_2 P(z_\lambda|\boldsymbol{m}) dt_\lambda$ is maximised when $P(\mathrm{z}_\lambda|\boldsymbol{m})$ is Gaussian; in which case it is equal to $\frac{1}{2}\log_2\big(2\pi e\ \mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})\big)$ (Cover and Thomas, 2012). Moreover, because $\log_2$ is a concave function, Jensen's inequality (Cover and Thomas, 2012) can be used to move the expectation over $\mathbf{m}$ inside the logarithm. Concretely, an upper bound for $h(\mathbf{z}|\mathbf{m})$ is given by

$$\begin{aligned} h(\mathbf{z}|\mathbf{m}) &= \sum_\lambda h(\mathrm{z}_\lambda|\mathbf{m}) & (6.9)\\ &\le \sum_\lambda \int_{\boldsymbol{m}} P(\boldsymbol{m}) \frac{1}{2} \log_2\big(2\pi e\ \mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})\big) d\boldsymbol{m} & (6.10)\\ &\le \sum_\lambda \frac{1}{2}\log_2\left(2\pi e \int_{\boldsymbol{m}} P(\boldsymbol{m}) \mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m}) d\boldsymbol{m}\right). & (6.11)\end{aligned}$$

Combining (6.4), (6.5), and (6.11) gives the following lower bound for the mutual information:

$$\begin{aligned} I(\mathbf{m};\mathbf{x}) &\ge I(\mathbf{m};\mathbf{z}) & (6.12)\\ &\ge \sum_\lambda \frac{1}{2}\log_2 \frac{\mathrm{var}(\mathrm{z}_\lambda)}{\int_{\boldsymbol{m}} P(\boldsymbol{m}) \mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m}) d\boldsymbol{m}}, & (6.13)\\ &\triangleq I_{q_{\boldsymbol{\theta}}} & (6.14)\end{aligned}$$

where we recall that the first inequality follows from the data processing inequality. In Section 6.2 it is shown that, given an appropriate data set, a *Siamese neural network* architecture can be used to easily estimate the expected conditional variance $\int_{\boldsymbol{m}} P(\boldsymbol{m})\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})d\boldsymbol{m}$, which makes the lower bound $I_{q_{\boldsymbol{\theta}}}$ convenient for optimisation.

6.1.3 The Lagrangian

Using the derivations from the preceding sections, the optimisation problem in (6.1) can be reformulated as:

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimise}} \quad & -\sum_{\lambda} \frac{1}{2} \log_2 \frac{\mathrm{var}(\mathrm{z}_\lambda)}{\int_{\boldsymbol{m}} P(\boldsymbol{m})\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})d\boldsymbol{m}} \\ \text{subject to} \quad & \mathrm{MMD}^2 = 0, \end{aligned} \tag{6.15}$$

where we recall that $\mathbf{z} = q_\theta(\mathbf{x})$, and z_λ is λ'th element of $\mathbf{z}$.

The Lagrangian function for (6.15) is given by

$$\mathcal{L}(\boldsymbol{\theta}, \beta) = \beta\mathrm{MMD}^2 - \sum_{\lambda} \frac{1}{2} \log_2 \frac{\mathrm{var}(\mathrm{z}_\lambda)}{\int_{\boldsymbol{m}} P(\boldsymbol{m})\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})d\boldsymbol{m}}, \tag{6.16}$$

where $\beta > 0$ is a *Lagrange multiplier*. If (6.15) were a convex optimisation problem, then the Lagrangian dual problem $\sup_\beta \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \beta)$ would give a global optima for (6.15) (Boyd and Vandenberghe, 2004). However, (6.15) is not a convex problem because the equality constraint is not affine. Even so, $\inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \beta)$ can be used to find local optima for some value of β. In this case, β can be interpreted as a hyper-parameter that controls the penalty of using a solution where the probability distribution of $\mathbf{z}$ is not $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (e.g., Fletcher, 1975; Smith and Coit, 1997).

6.2 Implementation

This section describes our implementation of the proposed approach for estimating mutual information. In order to compute an empirical estimate of the mutual information lower bound in (6.15), an empirical estimate of $\int_{\boldsymbol{m}} P(\boldsymbol{m})\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})d\boldsymbol{m}$ is computed.

To estimate the conditional variance, $\mathrm{var}(\mathrm{t}_\lambda|\boldsymbol{m})$, a *Siamese neural network* architecture (Bromley et al., 1994; Chopra et al., 2005) is relied on. A Siamese neural network consists of two-identical sub-networks, i.e., two exact copies of q_θ with the same weights and biases. During training, two samples are independently drawn from $P(\mathbf{x}|\boldsymbol{m})$ for some message $\mathbf{m} = \boldsymbol{m}$. Let $\boldsymbol{x}^{[1]}$ and $\boldsymbol{x}^{[2]}$ denote the two samples. One of the sub-networks is applied to $\boldsymbol{x}^{[1]}$ and the other sub-network is applied to $\boldsymbol{x}^{[2]}$. The output of the two sub-networks are denoted $\boldsymbol{z}^{[1]}$ and $\boldsymbol{z}^{[2]}$. Specifically,

$$\boldsymbol{z}^{[1]} = q_{\boldsymbol{\theta}}(\boldsymbol{x}^{[1]}) \tag{6.17}$$

and

$$\boldsymbol{z}^{[2]} = q_{\boldsymbol{\theta}}(\boldsymbol{x}^{[2]}). \tag{6.18}$$

By definition, the conditional variance $\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})$ is

$$\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m}) = \mathbb{E}_{\mathrm{z}_\lambda}[(\mathrm{t}_\lambda - \mathbb{E}_{\mathrm{z}_\lambda}[\mathrm{z}_\lambda|\,\boldsymbol{m}])^2 \mid \boldsymbol{m}]. \tag{6.19}$$

Using $\boldsymbol{z}^{[1]}$ and $\boldsymbol{z}^{[2]}$, an empirical estimate of $\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})$ is

$$\begin{aligned} \hat{\sigma}^2_{\mathrm{z}_\lambda|\boldsymbol{m}} &= \frac{1}{2}\left(t_\lambda^{[1]} - \frac{z_\lambda^{[1]} + z_\lambda^{[2]}}{2}\right)^2 + \frac{1}{2}\left(z_\lambda^{[2]} - \frac{z_\lambda^{[1]} + z_\lambda^{[2]}}{2}\right)^2 &(6.20)\\ &= \frac{1}{2}\left(z_\lambda^{[1]} - z_\lambda^{[2]}\right)^2, &(6.21) \end{aligned}$$

where $z_\lambda^{[1]}$ and $z_\lambda^{[2]}$ are the λ'th elements of $\boldsymbol{z}^{[1]}$ and $\boldsymbol{z}^{[2]}$, respectively. This shows that the conditional variance of the latent variable given a particular message can be computed as half the squared error of the two outputs of the Siamese network, where the two inputs to the Siamese network are generated using the same message. If more than two copies of q_θ were used, then a more accurate estimate of $\mathrm{var}(\mathrm{z}_\lambda|\boldsymbol{m})$ could be obtained, however, this would place a larger burden on collecting training data.

Let $\{\boldsymbol{x}^{[1](1)}, \boldsymbol{x}^{[2](1)}, \dots, \boldsymbol{x}^{[1](n)}, \boldsymbol{x}^{[2](n)}, \dots, \boldsymbol{x}^{[1](N)}, \boldsymbol{x}^{[2](N)}\}$ denote a mini-batch of N training examples where the superscript (n) denotes the n'th pair of training examples in the mini-batch. Similarly, let $\{\boldsymbol{z}^{[1](1)}, \boldsymbol{z}^{[2](1)}, \dots, \boldsymbol{z}^{[1](n)}, \boldsymbol{z}^{[2](n)}, \dots, \boldsymbol{z}^{[1](N)}, \boldsymbol{z}^{[2](N)}\}$ denote the output of the neural network for each training example in the mini-batch. Using (6.2), and (6.21), a differentiable empirical estimate of the Lagrangian in (6.16) is given by

$$\begin{aligned}\hat{\mathcal{L}}(\boldsymbol{\theta}, \beta) = &\frac{\beta}{N(N-1)} \sum_{n \neq n'} k(\boldsymbol{z}^{[1](n)}, \boldsymbol{z}^{[1](n')}) \\ &+ \frac{\beta}{N(N-1)} \sum_{n \neq n'} k(\boldsymbol{v}^{(n)}, \boldsymbol{v}^{(n')}) \\ &- \frac{2\beta}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} k(\boldsymbol{z}^{[1](n)}, \boldsymbol{v}^{(n')}) \\ &- \sum_{\lambda} \frac{1}{2} \log_2 \left(\frac{1}{N} \sum_{n} \big(z_{\lambda}^{[1](n)} - \frac{1}{N} \sum_{n'} z_{\lambda}^{[1](n')}\big)^2 \right) \\ &+ \sum_{\lambda} \frac{1}{2} \log_2 \left(\frac{1}{N} \sum_{n} \frac{1}{2} \big(z_{\lambda}^{[1](n)} - z_{\lambda}^{[2](n)}\big)^2 \right),\end{aligned} \tag{6.22}$$

where we recall that $\boldsymbol{\theta}$ denotes the parameters of the neural network, β is a Lagrange multiplier, k is a Gaussian kernel, each $\boldsymbol{v}^{(n)}$ is independently sampled from P_{target}, λ is the latent vector index, $\frac{1}{N}\sum_n (z_{\lambda}^{[1](n)} - \frac{1}{N}\sum_{n'} z_{\lambda}^{[1](n')})^2$ is an empirical estimate of $\text{var}(\text{z}_{\lambda})$, and $\frac{1}{N}\sum_n \frac{1}{2}(z_{\lambda}^{[1](n)} - z_{\lambda}^{[2](n)})^2$ is an empirical estimate of $\int_{\boldsymbol{m}} P(\boldsymbol{m}) \text{var}(\text{z}_{\lambda}|\boldsymbol{m}) d\boldsymbol{m}$. In the following section, several experiments that demonstrate our approach are provided.

6.3 Experiments

This section describes three artificial experiments that demonstrate the proposed approach for estimating mutual information. For all experiments, training data sets were created by sampling 40000 message vectors $\boldsymbol{m}$ from $P(\text{m})$. For each message vector, 20 speech vectors were sampled from $P(\mathbf{x}|\boldsymbol{m})$. This simulates having 20 talkers uttering each of the 40000 messages. The conditional distribution $P(\mathbf{x}|\mathbf{m})$ that was used to generate the data set was different for each experiment and is described in the following sections.

For all the experiments, the *Adam* variant (Kingma and Ba, 2014) of stochastic gradient descent was applied to the empirical estimate of the Lagrangian in (6.22). At the start of each training epoch, the training data was randomly split into $\lfloor 40000/N \rfloor$ mini-batches, where for each of the N messages in a mini-batch two of the 20 speech vectors were randomly selected as inputs to the Siamese network. The mini-batch size was $N = 1024$, the learning rate was 0.0005, and the number of epochs was 2000, except for Experiment 3, where the number of epochs was increased to 10000. The neural network q_θ consisted of two fully connected feed-forward layers each with 200 ReLU, and a third fully connected linear layer with D output units. The Lagrange multiplier was set to $\beta = 5000$. Using a large value for β is necessary to enforce the constraint that the latent vector is Gaussian. The kernel bandwidth for MMD was $\sigma^2 = 2D$.

6.3.1 Experiment 1

For the first experiment, the message vectors $\boldsymbol{m} \in \mathbb{R}^4$ were sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The speech vectors $\boldsymbol{x} \in \mathbb{R}^{10}$ were created by linearly embedding the message vectors into $\mathbb{R}^{10}$ and then adding Gaussian production noise $\boldsymbol{p} \in \mathbb{R}^{10}$. Specifically, the following model was used:

$$\mathbf{x} = \boldsymbol{E}\mathbf{m} + \mathbf{p} \tag{6.23}$$

where the embedding matrix $\boldsymbol{E} \in \mathbb{R}^{10\times 4}$ was randomly selected at the start of the experiment and $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. At the start of the experiment, each element of $\boldsymbol{E}$, denoted $E_{i,j}$, was selected by sampling $\mathcal{N}(0, \eta_i^2)$ where the variance for each row i is given by the i'th element of the vector

$$\boldsymbol{\eta}^2 = \frac{1}{4}[0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5]^*.$$

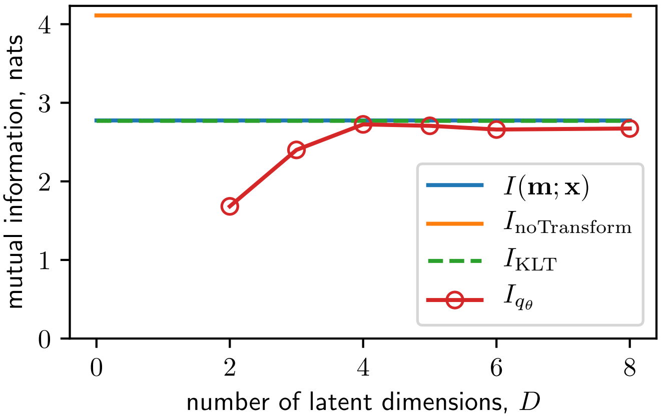

Figure 6.1: Comparison of approaches for estimating mutual information for Experiment 1. The proposed approach, I_{q_θ}, gives a reasonable estimate provided that the number of latent dimensions is large enough.

It can be shown that the mutual information for this statistical model is given by (Appendix A.4)

$$I(\mathbf{m};\mathbf{x}) = \frac{1}{2}\log_2 \det(\boldsymbol{E}\boldsymbol{E}^* + \boldsymbol{I}). \tag{6.24}$$

In Experiment 1, six neural networks were trained where the size of the latent vector $\mathbf{z}$ was $D = 2, 3, 4, 5, 6$ and 8. Figure 6.1 plots the mutual information against the size of the latent vector D for each network after training. $I(\mathbf{m};\mathbf{x})$ is the true mutual information, I_{q_θ} is an estimate of the lower bound in (6.13) that q_θ is trained to maximise, $I_{\text{noTransform}}$ is an estimate of the mutual information obtained by applying the method from Chapter 3 to the training data, and I_{KLT} is an estimate of the mutual information obtained using the KLT for q, like in Chapter 4.

We see that if the size of the latent vector is smaller than the number of message features, i.e., $D < 4$, then the approach proposed in the present chapter underestimates $I(\mathbf{m};\mathbf{x})$. This behaviour is to be expected. If the number of independent features in the latent space is less than the number

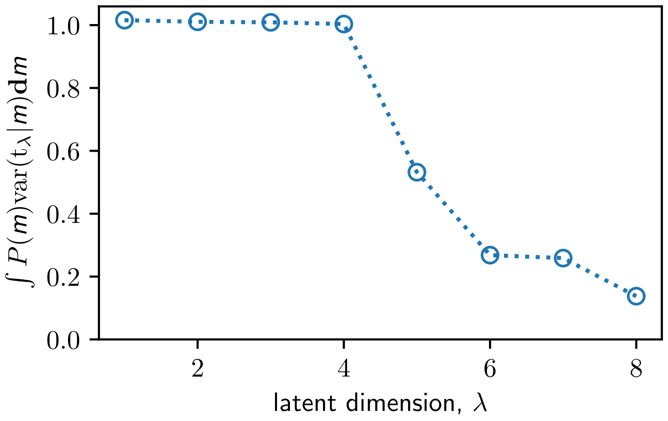

Figure 6.2: Expected value of the conditional variance for each latent dimension for a trained neural network. Latent dimensions where the expected value of the conditional variance is 1 provide no information about the message.

of independent message features, then q_θ does not have enough capacity to retain all the information about $\mathbf{m}$. For $D \geq 4$, q_θ is effective at removing the statistical dependencies in $\mathbf{x}$, which leads to an accurate estimate of $I(\mathbf{m};\mathbf{x})$.

Note that $I_{\text{noTransform}}$ overestimates $I(\mathbf{m};\mathbf{x})$. That is because no transform is used to account for the statistical dependencies in $\mathbf{x}$.

Also note that I_{KLT} provides an accurate estimate of $I(\mathbf{m};\mathbf{x})$. That is because for this experiment, $\mathbf{x}$ and $\mathbf{m}$ are jointly Gaussian, in which case the KLT provides the optimal solution.

Further insight can be gained by plotting the expected value of the conditional variance for each dimension of the latent variable. Figure 6.2 shows such a plot for the neural network with $D = 8$. For four of the latent dimensions, the expected value of the conditional variance is close to it's maximum value of 1. That means that these dimensions of the latent variable provide almost no information about the message. Specifically, they do not contribute to the sum in (6.13). The remaining four latent dimensions have low expected conditional variance, which means they do provide information about the four message features.

6.3.2 Experiment 2

For the second experiment, the same statistical model that was used to generate the training data for Experiment 1 was used. However, for Experiment 2, the size of the latent variable was $D = 6$ and $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \phi^2 \boldsymbol{I})$, where the production noise variance was $\phi^2 = 0.5, 1.0, 2.0, 4.0, 8.0$, and 16.0. For each value of ϕ^2, a new training data set was generated and a new neural network was trained. The same embedding matrix $\boldsymbol{E} \in \mathbb{R}^{10\times 4}$ was used for all of the data sets. It can be shown that the mutual information for this statistical model is (Appendix A.4)

$$I(\mathbf{m}; \mathbf{x}) = \frac{1}{2} \log_2 \frac{\det(\boldsymbol{E}\boldsymbol{E}^* + \phi^2 \boldsymbol{I})}{(\phi^2)^{10}}. \tag{6.25}$$

Figure 6.3 plots $I(\mathbf{m}; \mathbf{x})$, $I_{q_{\boldsymbol{\theta}}}$, $I_{\text{noTransform}}$, and I_{KLT} against the production noise variance ϕ^2. We see that the approach proposed in the present chapter, $I_{q_{\boldsymbol{\theta}}}$, provides an accurate estimate of $I(\mathbf{m}; \mathbf{x})$ for all the values of ϕ^2.

6.3.3 Experiment 3

For the third experiment, the message vectors $\boldsymbol{m} \in \mathbb{R}^4$ were sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. However, unlike Experiment 1 and Experiment 2, the message vectors were embedded into $\boldsymbol{x} \in \mathbb{R}^{10}$ using a non-linear embedding. Specifically, the following model was used:

$$\mathbf{a} = \boldsymbol{E}\mathbf{m} + \mathbf{p} \tag{6.26}$$

$$\mathbf{b} = [\mathrm{a}_1^2, \mathrm{a}_2^2, \mathrm{a}_3^2, \mathrm{a}_4^2, \mathrm{a}_5^2]^* \tag{6.27}$$

$$\mathbf{x} = [\mathbf{a}^*, \mathbf{b}^*]^*, \tag{6.28}$$

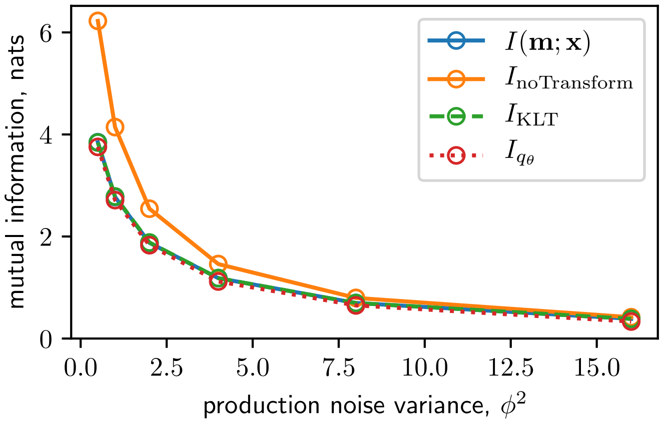

Figure 6.3: Comparison of approaches for estimating mutual information for Experiment 2. The proposed approach, $I_{q_{\boldsymbol{\theta}}}$, gives a reasonable estimate of mutual information for all values of production noise that were considered.

where $\boldsymbol{E} \in \mathbb{R}^{5\times 4}$ is an embedding matrix, $\mathbf{b}$ is obtained by squaring each element in $\mathbf{a}$, $\mathbf{x}$ is obtained by concatenating $\mathbf{a}$ and $\mathbf{b}$, and $*$ denotes the transpose. For this model, some of the elements of $\mathbf{x}$ are non-linearly dependent with the other elements and are non-Gaussian.

Similarly to the previous experiments, at the start of the Experiment 3, each element of $\boldsymbol{E}$, denoted $E_{i,j}$, was selected by sampling $\mathcal{N}(0, \eta_i^2)$ where the variance for each row i is given by the i'th element of the vector

$$\boldsymbol{\eta}^2 = \frac{1}{4}[1, 2, 3, 4, 5]^*.$$

The size of the latent variable was $D = 6$ and $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \phi^2 \boldsymbol{I})$, where the production noise variance was $\phi^2 = 0.5, 1.0, 2.0,$ and 4.0. It can be shown that the mutual information for this statistical model is (Appendix A.4)

$$I(\mathbf{m};\mathbf{x}) = \frac{1}{2}\log_2 \frac{\det(\boldsymbol{E}\boldsymbol{E}^* + \phi^2 \boldsymbol{I})}{(\phi^2)^5}. \qquad (6.29)$$

Figure 6.4 plots $I(\mathbf{m};\mathbf{x})$, $I_{q_{\boldsymbol{\theta}}}$, $I_{\text{noTransform}}$, and I_{KLT} against the production

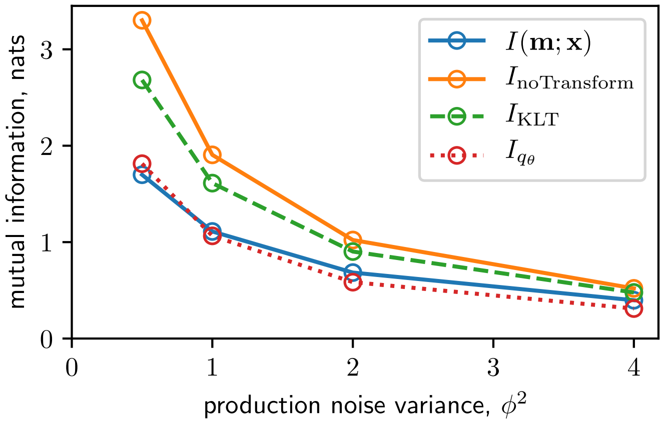

Figure 6.4: Comparison of approaches for estimating mutual information for Experiment 3. $I_{\text{noTransform}}$ and I_{KLT} overestimate the mutual information because they cannot account for non-linear dependencies.

noise variance ϕ^2. We see that the approach proposed in the present chapter, $I_{q_{\boldsymbol{\theta}}}$, provides a reasonable estimate of $I(\mathbf{m};\mathbf{x})$, whereas $I_{\text{noTransform}}$, and I_{KLT} overestimate the mutual information because they do not account for the non-linear dependencies in the speech vector and are based on the false assumption that $P(\mathbf{m};\mathbf{x})$ is Gaussian.

6.4 Summary of Chapter

This chapter proposed a new method for estimating the mutual information of the speech production channel $I(\mathbf{m};\mathbf{x})$ that does not make assumptions about the joint distribution $P(\mathbf{m};\mathbf{x})$ and does not require realisations of $\mathbf{m}$. Instead, the proposed approach considers mutual information estimation as an optimisation problem. Maximum mean discrepancy is used to apply a Gaussian constraint to the latent variable, and a lower bound of mutual information is estimated using a Siamese neural network. The lower bound is maximised using stochastic gradient ascent.

In this chapter good results were obtained for artificial examples. In situations where the KLT provides the optimal estimate of mutual information, the proposed approach also performs well. When non-linear message embeddings are used, the KLT fails, whereas the proposed approach maintains reasonable performance.

One limitation of the evaluation in this chapter is that relatively simple statistical models for $\mathbf{m}$ and $\mathbf{x}$ were considered. For more sophisticated statistical models, such as messages generated from Gaussian mixture models, a more complicated neural network architecture may be required. Additionally, the evaluation in this chapter only considered inputs with 10 dimensions. It is not immediately clear that the proposed approach will scale well to higher dimensions. For the auditory representation of speech in this thesis, stacking $K = 15$ consecutive spectra that each consist of 28 ERB frequency bands would result in input vectors with 420 dimensions. A *recurrent neural network* (Goodfellow et al., 2016) is a natural network architecture for speech signals and may provide better results for real-world data.

Chapter 7

Summary, Future Work, and Conclusion

7.1 Summary

Shannon's information theory provides mathematical tools for quantifying the effectiveness of communication systems, regardless of the systems implementation. The goal of this thesis was to develop a mathematical model of speech communication that is based on information theory. Specifically, the research in this thesis focused on the following three research questions:

1. How does the acoustic information rate of speech compare to the lexical information rate of speech?
2. How can information theory be used to predict the intelligibility of speech-based communication systems?
3. How well do competing models of speech communication predict intelligibility?

In this section, the work in this thesis is summarised with respect to each research question.

How does the acoustic information rate of speech compare to the lexical information rate of speech?

Chapter 2 presented an overview of existing methodologies for estimating the information rate of speech communication. There are two approaches to estimating the information rate of speech: 1) the linguistic perspective, where models of language are used, and 2) the speech processing perspective, where models of acoustic speech signals are used.

When speech is modelled as a sequence of phonemes, the lexical information rate of speech is 60 b/s. When the time-domain samples of an acoustic speech signal are modelled as a Gaussian process, the acoustic information rate of speech is approximately 53000 b/s. When models of speech production are considered (i.e., Fano's method), the acoustic information rate of speech is 2130 b/s. This thesis hypothesised that the discrepancies between estimates based on language models and the much larger estimates based on acoustic models are caused by two factors: 1) talker variability, and 2) statistical dependencies in acoustic signals.

In Chapter 3, a simple model of speech communication that is based on information theory was developed. The communication model includes a speech production channel that accounts for talker variability, and an environmental channel that accounts for environmental disturbances. The effectiveness of communication saturates either at the mutual information rate of the speech production channel, or the mutual information rate of the environmental channel, whichever is lowest. By modelling auditory log-spectra as Gaussian, modelling talker variability as additive production noise, and using a chorus of talkers, a novel method for estimating the acoustic information rate of speech was developed. When the proposed method is applied to real-world data, an estimate of 2070 b/s is obtained. After accounting for oversampling, the rate reduces to about 500 b/s. Thus, it is concluded that accounting for talker variability reduces the gap between estimates of the information rate that are based on language models, and estimates of the information rate that are based on acoustic models.

One limitation of the experiment in Chapter 3 is that the proposed method assumes that auditory log-spectra are memoryless, and that the signals in different ERB frequency bands are statistically independent. The consequence of these assumptions is that the method overestimates the information rate. For this reason, in Chapter 4 the communication model was extended to account for statistical dependencies. To do so, the KLT is used. The KLT is effective at removing statistical dependencies between time-frequency units of auditory log-spectra, provided the auditory log-spectra are Gaussian. When the KLT is applied, the acoustic information rate of speech reduces to about 180 b/s.

The remaining discrepancy between the lexical information rate of speech communication and the acoustic information rate of speech communication is likely due to an assumption that the joint probability distribution of the message and the auditory log-spectra is Gaussian. In practice, auditory log-spectra are only approximately Gaussian, in which case the KLT cannot remove all statistical dependencies. For this reason, Chapter 6 proposed a method for estimating the information rate of speech communication that does not make assumptions about the joint distribution of the message and the speech. Instead, of making assumptions about the joint distribution, the proposed method relies on techniques from deep learning. In particular, a Siamese neural network and maximum mean discrepancy are used. Although the proposed approach has not been applied to real-word data, good results are achieved for synthetic data sets, making the proposed approach a promising direction for future research.

How can information theory be used to predict the intelligibility of speech-based communication systems?

Chapter 2 explained that intelligibility is an important characteristic of speech-based communication systems and that intelligibility can be measured using formal listening tests. However, such tests are time-consuming, laborious, and expensive. For this reason, instrumental intelligibility metrics that can predict the intelligibility of a communication system are of importance.

Chapter 2 summarised three existing intelligibility metrics that are based on information theory: AI, SIMI, and MIKNN. None of these intelligibility metrics account for talker variability or statistical dependencies between time-frequency units of acoustic speech signals. For this reason, Chapter 4 proposed a novel intelligibility metric called speech intelligibility in bits (SIIB) and a variant called SIIB$^{\text{Gauss}}$, which are both based on the speech communication model developed in Chapter 3.

SIIB and SIIB$^{\text{Gauss}}$ both rely on a parametric model for the speech production channel, however, SIIB and SIIB$^{\text{Gauss}}$ differ in how they estimate the mutual information rate of the environmental channel. SIIB uses a non-parametric mutual information estimator that is based on k-nearest neighbours, whereas SIIB$^{\text{Gauss}}$ uses the information capacity of a Gaussian communication channel. The main differences between pre-existing intelligibility metrics based on information theory, and SIIB and SIIB$^{\text{Gauss}}$ is that SIIB and SIIB$^{\text{Gauss}}$ 1) use a more realistic auditory model, 2) use the KLT, and 3) account for talker variability. In Chapter 5, it was found that both SIIB and SIIB$^{\text{Gauss}}$ have state-of-the-art performance.

How well do competing models of speech communication predict intelligibility?

To answer the third research question, Chapter 5 presented a comprehensive evaluation of 12 monaural instrumental intelligibility metrics from the literature. To assess the accuracy of each metric, intelligibility data from 11 listening tests were obtained. The data include Dutch, Danish and English speech that was degraded by additive noise, reverberation, and competing talkers, and subjected to speech enhancement. To our knowledge, in terms of the number of intelligibility metrics and number of listening tests, the evaluation in Chapter 5 is the most comprehensive evaluation of monaural intrusive intelligibility metrics for speech in noise to date.

In addition to evaluating the accuracy of intelligibility metrics, Chapter 5 investigated why the top performing metrics have high performance. Specifically, the effect of decorrelating input features, the effect of the auditory model, and the effect of using different distortion measures was investigated. Furthermore, the ability of intelligibility metrics to generalise to new types of distortion was considered.

It was concluded that SIIB and SIIB$^{\text{Gauss}}$ have state-of-the-art performance, that many intelligibility metrics struggle with severe reverberant distortion, that many intelligibility metrics do not generalise well to new types of distortion, that the intelligibility metrics with the highest performance are also the only metrics that attempt to decorrelate input features, and that information theory provides an explanation for the success of the correlation coefficient as a distortion measure for intelligibility metrics.

7.2 Directions for Future Work

Based on the results in this thesis, there are several research topics that would be interesting to explore. The most obvious direction to take would be to refine the mutual information estimator proposed in Chapter 6 and then apply it to real-world data. Finding a suitable data set for training the neural network may prove to be an obstacle. At this stage, it is not clear how much training data is required for an accurate estimate of the information rate, but the TIMIT speech corpus (Garofolo et al., 1993) may provide a good starting point because it contains 450 sentences where each sentence is spoken by 7 out of 630 talkers.

Another research topic could be to improve and extend SIIB. One way that SIIB may be improved would be to consider non-parametric mutual information estimators other than the KNN mutual information estimator that is used in this thesis. SIIB could also be extended to consider binaural signals or hearing impairments. As discussed in Section 3.1.2, one approach to account for hearing impairments is to reintroduce the concept of *interpretation noise* (Kleijn and Hendriks, 2015).

Recall that one of the conclusions from the evaluation in Chapter 5 was that intelligibility metrics tend to perform poorly on data sets that were not considered during their development. SIIB and SIIB$^{\text{Gauss}}$ are theoretically motivated and were not tuned specifically for the data sets considered in Chapter 5. Even so, an independent evaluation of the accuracy of SIIB and SIIB$^{\text{Gauss}}$ would be worth considering.

7.3 Conclusion

This thesis approached speech communication from an information theoretical perspective. New methods for estimating the information rate of speech communication that rely on a chorus of talkers were proposed. When applied to real-world data, an estimate for the acoustic information rate of speech of about 180 b/s is obtained. This is not as low as estimates obtained using language models, which are around 60 b/s, but is considerably lower than previous attempts that use acoustic models. It can thus be concluded that accounting for talker variability reduces the gap between estimates that rely on language models and estimates that rely on acoustic models. An even lower information rate may be obtained by using a more powerful transform than the KLT to reduce statistical dependencies between the time-frequency units of speech signals.

Using the communication model developed in this thesis, a novel intelligibility metric called SIIB, and a variant called SIIB$^{\text{Gauss}}$, were proposed. An evaluation of SIIB and SIIB$^{\text{Gauss}}$ showed that both intelligibility metrics have state-of-the performance. This suggests that other speech-based technologies may also benefit from a more information theoretical approach to modelling speech communication.

Finally, this thesis showed that many intelligibility metrics do not generalise well to new data sets and listening environments. For this reason, when collecting listening test data for the purpose of developing new intelligibility metrics, it may be more beneficial to gather data from a wide range of listening environments, than to gather a lot of data for a single listening environment. Furthermore, this thesis showed that the accuracy of intelligibility metrics can often be improved by accounting for the statistical dependencies between the time-frequency units of acoustic speech signals.

Bibliography

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Trans. Audio, Speech, Language Process.*, 2(4):567–577.

Allen, J. B. (2005a). Articulation and intelligibility. *Synthesis Lectures on Speech and Audio Processing*, 1(1):1–124.

Allen, J. B. (2005b). The Articulation Index is a Shannon channel capacity. In *Auditory Signal Processing*, pages 313–319. Springer.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proc. Int. Conf. on Machine Learn.*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, USA. PMLR.

Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2017). A non-intrusive short-time objective intelligibility measure. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, pages 5085–5089. IEEE.

Anglin, J. M., Miller, G. A., and Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, 58(10):i–186.

ANSI (1969). *American National Standard Methods for the Calculation of the Articulation Index*. S3.5. Acoustical Society of America, New York, USA.

ANSI (1989). *American National Standard Method for Measuring the Intelligibility of Speech Over Communication Systems*. S3.2. Acoustical Society of America, New York, USA.

ANSI (1997a). *American National Standard Methods for Calculation of the Speech Intelligibility Index*. S3.5. Acoustical Society of America, New York, USA.

ANSI (1997b). *American National Standard Methods for Calculation of the Speech Intelligibility Index*. S3.5. Acoustical Society of America, New York.

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

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. (2018). Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Commun.*, 49(10-11):763–786.

Bertsekas, D. P. and Tsitsiklis, J. N. (2002). *Introduction to probability*. Athena Scientific, New Hampshire, USA, first edition. Section 4.7, `http://www.athenasc.com/Bivariate-Normal.pdf`.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Audio, Speech, Language Process.*, 27(2):113–120.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press, Cambridge, UK.

Braithwaite, D. T. and Kleijn, W. B. (2018). Bounded information rate variational autoencoders. In *Proc. Deep Learning Day*.

Bridle, J., Brown, M., and Chamberlain, R. (1983). Continuous connected word recognition using whole word templates. *Rad. Elec. Eng.*, 53(4):167–175.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In *Adv. in Neural Info. Proc. Systems*, pages 737–744.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, 120(6):4007–4018.

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

Chen, F. and Loizou, P. C. (2011). Predicting the intelligibility of vocoded and wideband mandarin chinese. *J. Acoust. Soc. Am.*, 129(5):3281–3290.

Chen, F., Wong, L. L. N., and Hu, Y. (2013). A hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech. *sc*, 55(10):1011–1020.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, volume 1, pages 539–546. IEEE.

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

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119(3):1562–1573.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: the Hurricane Challenge. In *Proc. Interspeech*, pages 3552–3556.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons, New York, USA.

Crystal, T. H. and House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *J. Acoust. Soc. Am.*, 83(4):1553–1573.

Cummins, F., Grimaldi, M., Leonard, T., and Simko, J. (2006). The chains corpus: Characterizing individual speakers. In *Proc. Int. Conf. on Speech and Comput.*, volume 6, pages 431–435. Citeseer.

Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the effective signal processing in the auditory system. i. model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622.

Davies, M. and Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. Routledge, Abingdon, UK.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.*, 28(4):357–366.

Denes, P. B. (1963). On the statistics of spoken english. *J. Acoust. Soc. Am.*, 35(6):892–904.

Doquire, G. and Verleysen, M. (2012). A comparison of multivariate mutual information estimators for feature selection. In *Proc. Int. Conf. on Pattern Recognition Applications and Methods*, pages 176–185.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). Icra noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *Audiology*, 40(3):148–157.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Am. Stat. Assoc.*, 82(397):171–185.

Elliott, T. M. and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLOS Comput. Biol.*, 5(3):e1000302.

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

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

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

Fano, R. M. (1950). The information theory point of view in speech communication. *The Journal of the Acoustical Society of America*, 22(6):691–696.

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends Cogn. Sci.*, 4(7):258–267.

Flanagan, J. L., Allen, J. B., and Hasegawa-Johnson, M. A. (2008). *Speech Analysis, Synthesis, and Perception*. Citeseer, third edition.

Flege, J. E., Schirru, C., and MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Commun.*, 40(4):467–491.

Fletcher, H. and Galt, R. H. (1950). The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22(2):89–151.

Fletcher, R. (1975). An ideal penalty function for constrained optimization. *IMA J. Appl. Math.*, 15(3):319–342.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119.

Gallois, C., Ogay, T., and Giles, H. (2005). Communication accommodation theory: A look back and a look ahead. In *Theorizing about intercultural communication*, pages 121–148. Sage, California, USA.

Gao, S., Ver Steeg, G., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Proc. Int. Conf. on Artificial Intelligence and Statistics*, pages 277–286.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

Gerkmann, T. and Martin, R. (2009). On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling. *IEEE Trans. Signal Process.*, 57(11):4165–4174.

Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children*. PhD thesis, Massachusetts Institute of Technology, Massachusetts, USA.

Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.*, 116(6):3679–3689.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, London, England. `http://www.deeplearningbook.org`.

Gordon-Salant, S. and Fitzgibbons, P. J. (1995). Comparing recognition of distorted speech using an equivalent signal-to-noise ratio index. *J. Speech Lang. Hear. Res.*, 38(3):706–713.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, pages 6645–6649. IEEE.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Adv. in Neural Info. Proc. Systems*, pages 513–520.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scand. Audiol.*, 11(2):79–87.

Hendriks, R. C., Crespo, J. B., Jensen, J., and Taal, C. H. (2015). Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 23(5):851–862.

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

Houben, R., Koopman, J., Luts, H., Wagener, K. C., Van Wieringen, A., Verschuure, H., and Dreschler, W. A. (2014). Development of a dutch matrix sentence test to assess speech intelligibility in noise. *Int. J. Audiol.*, 53(10):760–763.

Hu, Y. and Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *J. Acoust. Soc. Am.*, 122(3):1777–1786.

Huang, C., Chen, T., Li, S., Chang, E., and Zhou, J. (2001). Analysis of speaker variability. In *Proc. Eurospeech*.

Jensen, J., Batina, I., Hendriks, R. C., and Heusdens, R. (2005). A study of the distribution of time-domain speech samples and discrete fourier coefficients. In *Proc. SPS-DARTS*, volume 1, pages 155–158.

Jensen, J. and Hendriks, R. C. (2012). Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. *IEEE Trans. Audio, Speech, Language Process.*, 20(1):92–102.

Jensen, J. and Taal, C. H. (2014). Speech intelligibility prediction based on mutual information. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(2):430–440.

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

Jin, I.-K. (2014). *Development of the Speech Intelligibility Index (SII) for Korean*. PhD thesis, University of Colorado at Boulder.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.*, 130(3):1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *J. Acoust. Soc. Am.*, 134(1):436–446.

Kallenberg, O. (2006). *Foundations of modern probability*. Springer Science, New York, USA, 2 edition.

Karhunen, K. (1947). *Über lineare Methoden in der Wahrscheinlichkeitsrechnung (On linear methods in probability and statistics)*. Universitat Helsinki, Helsinki.

Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *J. Acoust. Soc. Am.*, 117(4):2224–2237.

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

Kates, J. M. and Arehart, K. H. (2015). Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality. *J. Acoust. Soc. Am.*, 138(4):2470–2482.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93.

Khademi, S., Hendriks, R., and Kleijn, W. B. (2017). Intelligibility enhancement based on mutual information. *IEEE/ACM Trans. Audio, Speech, Language Process.*

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

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am.*, 126(3):1415–1426.

Kleijn, W. B. and Hendriks, R. C. (2015). A simple model of speech communication and its application to intelligibility enhancement. *IEEE Signal Process. Lett.*, 22(3):303–307.

Kleijn, W. B., Lim, F. S., Luebs, A., Skoglund, J., Stimberg, F., Wang, Q., and Walters, T. C. (2018). Wavenet based low rate speech coding. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.* IEEE.

Kleijn, W. B. and Paliwal, K. K. (1995). *Speech coding and synthesis*. Elsevier Science Inc., New York, USA.

Koch, R. (1992). *Auditory sound analysis for the prediction and improvement of speech intelligibility*. PhD thesis, University of Goettingen, Goettingen.

Kolchinsky, A. and Tracey, B. D. (2017). Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. (2017). Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, 69(6):066138.

Krause, J. C. and Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.*, 112(5):2165–2172.

Kryter, K. D. (1962a). Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1689–1697.

Kryter, K. D. (1962b). Validation of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1698–1702.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

Levelt, W. J. M. (1999). Models of word production. *Trends Cogn. Sci.*, 3(6):223–232.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.*, 74(6):431.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.*, 54(5):358.

Loizou, P. C. (2013). *Speech enhancement: theory and practice*. CRC press, Boca Raton.

Loizou, P. C. and Ma, J. (2011). Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms. *J. Acoust. Soc. Am.*, 130(2):986–995.

Lombard, E. (1911). Le signe de l'elevation de la voix (the sign of the rise in the voice). *Annales des maladies de l'oreille, du larynx, du nez et du pharynx*, pages 101–119.

Lyon, R. F. (2017). *Human and machine hearing*. Cambridge University Press, New York, USA.

Ma, J., Hu, Y., and Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.*, 125(5):3387–3405.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press, Cambridge, UK.

Massey, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *J. Am. Stat. Assoc*, 46(253):68–78.

Miller, G. A., Heise, G. A., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.*, 41(5):329.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27(2):338–352.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Navidi, W. C. (2008). *Statistics for engineers and scientists*. McGraw-Hill Higher Education, New York, USA.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95(2):1085–1099.

Nygaard, L. C. and Pisoni, D. B. (1995). Speech perception: New directions in research and theory. In *Speech, Language, and Communication*, pages 63 – 96. Academic Press, San Diego, USA.

Oxenham, A. J. (2001). Forward masking: Adaptation or integration? *J. Acoust. Soc. Am.*, 109(2):732–741.

Pavlovic, C. V. (1987). Derivation of primary parameters and procedures for use in speech intelligibility predictions. *J. Acoust. Soc. Am.*, 82(2):413–422.

Phatak, S. A., Lovitt, A., and Allen, J. B. (2008). Consonant confusions in white noise. *J. Acoust. Soc. Am.*, 124(2):1220–1233.

Piantadosi, S. T. (2014). Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.

Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall, New Jersey, USA.

Rao, K. R. and Yip, P. (1990). *Discrete cosine transform: algorithms, advantages, applications*. Academic press, San Diego, USA.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (2016). Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. *J. Acoust. Soc. Am.*, 140(4):2670–2679.

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

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

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246.

Santos, J. F., Senoussaoui, M., and Falk, T. H. (2014). An improved non-intrusive intelligibility metric for noisy and reverberant speech. In *Proc. IEEE. Int. Workshop on Acoust. Speech Enhancement*, pages 55–59. IEEE.

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Siegler, M. A. and Stern, R. M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, volume 1, pages 612–615. IEEE.

Slaney, M. (1993). An efficient implementaion of the Patterson-Holdsworth auditory filter bank. *Apple Comp. Tech. Report 35*.

Smith, A. E. and Coit, D. W. (1997). Penalty functions. *Handbook of Evolutionary Computation*, 5:1–6.

Stevens, K. N. (2000). *Acoustic Phonetics*. MIT press, Massachusetts, USA.

Stoel-Gammon, C. (1989). Prespeech and early speech development of two late talkers. *First Lang.*, 9(6):207–223.

Taal, C. H., Hendriks, R. C., and Heusdens, R. (2014). Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Comput. Speech Lang.*, 28(4):858–872.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). On predicting the difference in intelligibility before and after single-channel noise reduction. In *Proc. IEEE. Int. Workshop on Acoust. Speech Enhancement*.

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

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011b). An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *J. Acoust. Soc. Am.*, 130(5):3013–3027.

Taghia, J. and Martin, R. (2014). Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(1):6–16.

Tang, Y. and Cooke, M. (2016). Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In *Proc. Interspeech*, pages 2488–2492.

Tiffany, W. R. (1980). The effects of syllable structure on diadochokinetic and reading rates. *J. Speech. Lang. Hear. Res.*, 23(4):894–908.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proc. 37th Allerton Conf. Commun., Control, Comput.*, page 368377.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *Proc. Speech Synth. Workshop*, page 135.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2016). An intelligibility metric based on a simple model of speech communication. In *Proc. IEEE. Int. Workshop on Acoust. Speech Enhancement*, pages 1–5. IEEE.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2017). On the information rate of speech communication. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, pages 5625–5629.

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2018a). An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Trans. Audio, Speech, Language Process.*

Van Kuyk, S., Kleijn, W. B., and Hendriks, R. C. (2018b). An instrumental intelligibility metric based on information theory. *IEEE Signal Process. Lett.*, 25(1):115–119.

Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition: Ii. noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.*, 12(3):247–251.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (2003). Design, optimization and evaluation of a danish sentence test in noise. *Int. J. Audiol.*, 42(1):10–17.

Wegel, R. and Lane, C. (1924). The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Phys. Rev.*, 23(2):266.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, 15(2):70–73.

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series*, volume 7. MIT press, Cambridge.

Wong, L. L. N., Ho, A. H. S., Chua, E. W. W., and Soli, S. D. (2007). Development of the cantonese speech intelligibility index. *J. Acoust. Soc. Am.*, 121(4):2350–2361.

Xia, R., Li, J., Akagi, M., and Yan, Y. (2012). Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin. In *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, pages 4465–4468. IEEE.

Zhao, S., Song, J., and Ermon, S. (2018). Infovae: Balancing learning and inference in variational autoencoders. *arXiv preprint arXiv:1706.02262*.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Addison-Wesley Press, Oxford, England.

Zorila, T.-C., Kandia, V., and Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Thirteenth Annual Conf. of the Int. Speech Comm. Assoc.*

Appendices

Appendix A

Mathematical Derivations

A.1 Mutual Information of Two Univariate Gaussians

Suppose that x and y are jointly Gaussian univariate random variables with a covariance matrix given by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{\mathrm{x}}^2 & \sigma_{\mathrm{x}}\sigma_{\mathrm{y}}\rho_{\mathrm{xy}} \\ \sigma_{\mathrm{x}}\sigma_{\mathrm{y}}\rho_{\mathrm{xy}} & \sigma_{\mathrm{y}}^2 \end{bmatrix},$$

where $\mathrm{var}(\mathrm{x}) = \sigma_x^2$, $\mathrm{var}(\mathrm{y}) = \sigma_y^2$, and ρ_{xy} is the correlation coefficient.

Substituting (2.19) and (2.21) into (2.13), the mutual information of x and y is

$$\begin{aligned} I(\mathrm{x};\mathrm{y}) &= H(\mathrm{x}) + H(\mathrm{y}) - H(\mathrm{x},\mathrm{y}) && \text{(A.1)} \\ &= \frac{1}{2}\log 2\pi e\sigma_x^2 + \frac{1}{2}\log 2\pi e\sigma_y^2 - \frac{1}{2}\log\det(2\pi e\mathbf{\Sigma}) && \text{(A.2)} \\ &= \frac{1}{2}\log 2\pi e\sigma_x^2 + \frac{1}{2}\log 2\pi e\sigma_y^2 - \frac{1}{2}\log(2\pi e)^2(\sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2 - \sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2\rho_{\mathrm{xy}}^2) && \text{(A.3)} \\ &= -\frac{1}{2}\log(1-\rho_{\mathrm{xy}}^2) && \text{(A.4)} \end{aligned}$$

A.2 Mutual Information for a Gaussian Markov Chain

Let x, y, and z be jointly Gaussian, univariate random variables that satisfy the Markov condition $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$. Let $\mathrm{var}(\mathrm{x}) = \sigma_{\mathrm{x}}^2$, $\mathrm{var}(\mathrm{y}) = \sigma_{\mathrm{y}}^2$, and $\mathrm{var}(\mathrm{z}) = \sigma_{\mathrm{z}}^2$. Without loss of generality, let $\mathbb{E}[\mathrm{x}] = \mathbb{E}[\mathrm{y}] = \mathbb{E}[\mathrm{z}] = 0$.

The correlation coefficient between x and z can be written as

$$\begin{aligned}
\rho_{\mathrm{xz}} &= \frac{\mathbb{E}[\mathrm{xz}]}{\sigma_{\mathrm{x}}\sigma_{\mathrm{z}}} && \text{(A.5)}\\
&= \frac{\mathbb{E}[\mathbb{E}[\mathrm{xz}|\mathrm{y}]]}{\sigma_{\mathrm{x}}\sigma_{\mathrm{z}}} && \text{(A.6)}\\
&= \frac{\mathbb{E}[\mathbb{E}[\mathrm{x}|\mathrm{y}]\mathbb{E}[\mathrm{z}|\mathrm{y}]]}{\sigma_{\mathrm{x}}\sigma_{\mathrm{z}}} && \text{(A.7)}\\
&= \frac{\mathbb{E}[\frac{\sigma_{\mathrm{x}}}{\sigma_{\mathrm{y}}}\rho_{\mathrm{xy}}\mathrm{y}\frac{\sigma_{\mathrm{z}}}{\sigma_{\mathrm{y}}}\rho_{\mathrm{yz}}\mathrm{y}]}{\sigma_{\mathrm{x}}\sigma_{\mathrm{z}}} && \text{(A.8)}\\
&= \frac{\sigma_{\mathrm{x}}\sigma_{\mathrm{z}}\rho_{\mathrm{xy}}\rho_{\mathrm{yz}}\mathbb{E}[\mathrm{yy}]}{\sigma_{x}\sigma_{\mathrm{z}}\sigma_{\mathrm{y}}^2} && \text{(A.9)}\\
&= \rho_{\mathrm{xy}}\rho_{\mathrm{yz}}, && \text{(A.10)}
\end{aligned}$$

where (A.6) follows from the law of total expectation, (A.7) follows from the conditional independence of x and z given y, and (A.8) uses the expression for the conditional expectation of two jointly Gaussian random variables (Bertsekas and Tsitsiklis, 2002):

$$\mathbb{E}[\mathrm{x}|\mathrm{y}] = \mathbb{E}[\mathrm{x}] + \rho_{\mathrm{xy}}\frac{\sigma_{\mathrm{x}}}{\sigma_{\mathrm{y}}}(\mathrm{y} - \mathbb{E}[\mathrm{y}]). \tag{A.11}$$

The mutual information between x and z is then obtained by substituting (A.10) into (A.4):

$$I(\mathrm{x};\mathrm{z}) = -\frac{1}{2}\log(1 - \rho_{\mathrm{xy}}^2\rho_{\mathrm{yz}}^2). \tag{A.12}$$

A.3 Correlation Coefficient for an Additive Communication Channel

Suppose that

$$\mathrm{y} = \mathrm{x} + \mathrm{n}, \tag{A.13}$$

where x and n are statistically independent univariate random variables. Without loss of generality, let $\mathbb{E}[\mathrm{y}] = \mathbb{E}[\mathrm{x}] = \mathbb{E}[\mathrm{n}] = 0$. The correlation coefficient of x and y can be written as

$$\begin{aligned}
\rho_{\mathrm{xy}} &= \frac{\mathbb{E}[\mathrm{xy}]}{\sqrt{\mathbb{E}[\mathrm{x}^2]\mathbb{E}[\mathrm{y}^2]}} && \text{(A.14)}\\
&= \frac{\mathbb{E}[\mathrm{x}(\mathrm{x}+\mathrm{n})]}{\sqrt{\mathbb{E}[\mathrm{x}^2]\mathbb{E}[(\mathrm{x}+\mathrm{n})^2]}} && \text{(A.15)}\\
&= \frac{\mathbb{E}[\mathrm{x}^2]}{\sqrt{\mathbb{E}[\mathrm{x}^2]\mathbb{E}[\mathrm{x}^2+\mathrm{n}^2]}} && \text{(A.16)}\\
&= \sqrt{\frac{\mathbb{E}[\mathrm{x}^2]}{\mathbb{E}[\mathrm{x}^2]+\mathbb{E}[\mathrm{n}^2]}} && \text{(A.17)}\\
&= \sqrt{\frac{\mathbb{E}[\mathrm{x}^2]/\mathbb{E}[\mathrm{n}^2]}{1+\mathbb{E}[\mathrm{x}^2]/\mathbb{E}[\mathrm{n}^2]}} && \text{(A.18)}\\
&= \sqrt{\frac{\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})}{1+\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})}}, && \text{(A.19)}
\end{aligned}$$

where $\mathrm{var}(\mathrm{x})/\mathrm{var}(\mathrm{n})$ is interpreted as the signal-to-noise ratio, and (A.16) follows because x and n are statistically independent.

By manipulating (A.17), the correlation coefficient of x and y can also be written as

$$\begin{aligned}
\rho_{\mathrm{xy}} &= \sqrt{\frac{\mathbb{E}[\mathrm{x}^2]}{\mathbb{E}[\mathrm{x}^2]+\mathbb{E}[\mathrm{n}^2]}} && \text{(A.20)}\\
&= \sqrt{\frac{\mathbb{E}[\mathrm{y}^2]-\mathbb{E}[\mathrm{n}^2]}{\mathbb{E}[\mathrm{y}^2]}}, && \text{(A.21)}
\end{aligned}$$

therefore,

$$\rho_{\mathrm{xy}}^2 = \frac{\mathbb{E}[\mathrm{y}^2] - \mathbb{E}[\mathrm{n}^2]}{\mathbb{E}[\mathrm{y}^2]} \tag{A.22}$$

$$= \frac{\mathrm{var}(\mathrm{y}) - \mathrm{var}(\mathrm{n})}{\mathrm{var}(\mathrm{y})}. \tag{A.23}$$

A.4 Mutual Information for an Additive Gaussian Vector Channel

Suppose that

$$\mathbf{x} = \boldsymbol{E}\mathbf{m} + \mathbf{p}, \tag{A.24}$$

where $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is a vector-valued random variable with c dimensions, $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}_{\mathrm{p}})$ is a vector-valued random variable with d dimensions, and $\boldsymbol{E} \in \mathbb{R}^{d\times c}$ is a full rank matrix. Furthermore, let $\mathbf{m}$ and $\mathbf{p}$ be statistically independent.

For the above model, the covariance matrix of $\mathbf{x}$ is

$$\boldsymbol{R}_{\mathrm{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^*] \tag{A.25}$$

$$= \mathbb{E}[(\boldsymbol{E}\mathbf{m} + \mathbf{p})(\boldsymbol{E}\mathbf{m} + \mathbf{p})^*] \tag{A.26}$$

$$= \mathbb{E}[\boldsymbol{E}\mathbf{m}\mathbf{m}^*\boldsymbol{E}^* + \mathbf{p}\mathbf{p}^*] \tag{A.27}$$

$$= \boldsymbol{E}\mathbb{E}[\mathbf{m}\mathbf{m}^*]\boldsymbol{E}^* + \mathbb{E}[\mathbf{p}\mathbf{p}^*] \tag{A.28}$$

$$= \boldsymbol{E}\boldsymbol{E}^* + \boldsymbol{R}_{\mathrm{p}}. \tag{A.29}$$

(A.27) follows because $\mathbf{m}$ and $\mathbf{p}$ are statistically independent.

The mutual information between $\mathbf{m}$ and $\mathbf{x}$ is

$$\begin{aligned} I(\mathbf{m};\mathbf{x}) &= H(\mathbf{x}) - H(\mathbf{x}|\mathbf{m}) && \text{(A.30)} \\ &= H(\mathbf{x}) - H(\mathbf{p}) && \text{(A.31)} \\ &= \frac{1}{2}\log\det(2\pi e \boldsymbol{R}_{\mathrm{x}}) - \frac{1}{2}\log\det(2\pi e \boldsymbol{R}_{\mathrm{p}}) && \text{(A.32)} \\ &= \frac{1}{2}\log\frac{\det(\boldsymbol{R}_{\mathrm{x}})}{\det(\boldsymbol{R}_{\mathrm{p}})} && \text{(A.33)} \\ &= \frac{1}{2}\log\frac{\det(\boldsymbol{E}\boldsymbol{E}^* + \boldsymbol{R}_{\mathrm{p}})}{\det(\boldsymbol{R}_{\mathrm{p}})}, && \text{(A.34)} \end{aligned}$$

where (A.31) follows because $\mathbf{m}$ and $\mathbf{p}$ are statistically independent, (A.32) uses (2.19), and (A.34) follows from (A.29).

For the case where $\boldsymbol{R}_{\mathrm{p}} = \boldsymbol{I}$, (A.34) reduces to

$$I(\mathbf{m};\mathbf{x}) = \frac{1}{2}\log\det(\boldsymbol{E}\boldsymbol{E}^* + \boldsymbol{I}). \tag{A.35}$$

For the case where $\boldsymbol{R}_{\mathrm{p}} = \phi^2\boldsymbol{I}$, where ϕ is a positive real-valued scalar, (A.34) reduces to

$$I(\mathbf{m};\mathbf{x}) = \frac{1}{2}\log\frac{\det(\boldsymbol{E}\boldsymbol{E}^* + \phi^2\boldsymbol{I})}{(\phi^2)^d}. \tag{A.36}$$

Lastly, suppose that

$$\mathbf{x} = g(\boldsymbol{E}\mathbf{m} + \mathbf{p}), \tag{A.37}$$

where g is a deterministic invertible function. Using (2.17), the introduction of g does not effect mutual information, thus (A.34) is also valid when the communication model in (A.24) is replaced by (A.37).

Appendix B

Rationale for STFT parameters

The STFT analysis parameters in Section 3.2 were selected for the following reasons:

- A sampling rate of $f_s = 16$ kHz was selected because it corresponds to a Nyquist frequency of 8 kHz, and frequencies above 8 kHz have a negligible effect on intelligibility.

- A frame length of 25 ms and a frame rate of $R = 80$ frames/s were selected because the human vocal tract tends to change shape at a rate less than 40 Hz. Thus, the frequency content of each speech sound produced by the vocal tract can be observed.

- Using N = 400 with f_s $= 16$ kHz corresponds to a frequency bin resolution of 40 Hz, which is less than the frequency resolution of the human ear (apart from at very low frequencies, however, these low frequencies do not contribute to speech intelligibility).

Thus, the values for the STFT analysis parameters in Section 3.2 are such that the transformation does not discard relevant information and allows the spectral content of each speech sound to be analysed.