

Features *of* Hearing

Applications of machine learning
to uncover the building blocks
of hearing

Lotte Weerts



IMPERIAL COLLEGE LONDON
Department of Electrical and Electronic Engineering

Features of Hearing

Applications of Machine Learning To Uncover the
Building Blocks of Hearing

Lotte Weerts

supervised by

Prof. Claudia Clopath
Dr. Dan Goodman

Submitted in part fulfilment of the requirements
for the degree of Doctor of Philosophy at
Imperial College London, July 2021

I hereby declare that this thesis was composed by me; that all the work reported herein originated from and was undertaken by me; and that information derived from the works of others has been acknowledged and referenced in the body of this work.

– Lotte Weerts (July 2021)

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence (CC BY-NC-ND). Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes, and that they do not distribute modified versions of the work. For any reuse or redistribution, researchers must make clear to others the licence terms of this work. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ABSTRACT

Recent advances in machine learning have instigated a renewed interest in using machine learning approaches to better understand human sensory processing. This line of research is particularly interesting for speech research since speech comprehension is uniquely human, which complicates obtaining detailed neural recordings. In this thesis, I explore how machine learning can be used to uncover new knowledge about the auditory system, with a focus on discovering robust auditory features. The resulting increased understanding of the noise robustness of human hearing may help to better assist those with hearing loss and improve Automatic Speech Recognition (ASR) systems.

First, I show how computational neuroscience and machine learning can be combined to generate hypotheses about auditory features. I introduce a neural feature detection model with a modest number of parameters that is compatible with auditory physiology. By testing feature detector variants in a speech classification task, I confirm the importance of both well-studied and lesser-known auditory features.

Second, I investigate whether ASR software is a good candidate model of the human auditory system. By comparing several state-of-the-art ASR systems to the results from humans on a range of psychometric experiments, I show that these ASR systems diverge markedly from humans in at least some psychometric tests. This implies that none of these systems act as a strong proxy for human speech recognition, although some may be useful when asking more narrowly defined questions.

For neuroscientists, this thesis exemplifies how machine learning can be used to generate new hypotheses about human hearing, while also highlighting the caveats of investigating systems that may work fundamentally differently from the human brain. For machine learning engineers, I point to tangible directions for improving ASR systems. To motivate the continued cross-fertilization between these fields, a toolbox that allows researchers to assess new ASR systems has been released.

It has been said that the best place to hide a leaf is in the forest, and presumably the best place to hide a voice is among other voices.

—GEORGE ARMITAGE MILLER (1947)

"THE MASKING OF SPEECH."

PSYCHOLOGICAL BULLETIN

Contents

Acknowledgements	xiii
List of Figures and Tables	xvi
Notation	xvii
1 Introduction	1
1 What are Auditory Features?	2
2 A Task-Driven Approach	4
3 Thesis Roadmap	5
4 One for the Road	7
 PART I. FOUNDATIONS	
2 Acoustic Phonetics and the Auditory System	11
1 An Introduction to Phonetics	12
2 Acoustic Correlates	17
3 Speech Robustness	31
4 Auditory Processing	33
5 Chapter Summary	37
3 Explainable Automatic Speech Recognition	39
1 A Brief History of ASR	40
2 ASR Features and Noise Robustness	41
3 An Introduction to Deep Automatic Speech Recognition	45
4 Explainable Artificial Intelligence	55
5 Chapter Summary	69

CONTENTS

PART II. NEURO-INSPIRED FEATURE DETECTION

4 Auditory Feature Hypothesis Generation	73
1 Introduction	73
2 Method	74
3 Results	80
4 Conclusions	86

PART III. FEATURES OF AUTOMATIC SPEECH RECOGNITION

5 A Primer on Comparing Human and Machine Listening	89
1 Introduction	89
2 ASR Systems	91
3 The HumanlikeHearing Toolbox	95
4 Speech Data Sets	96
5 General Methods for Psychometric Testing of ASR	97
6 Conclusions	101
6 Distortion Sensitivity of ASR Systems	103
1 Introduction	103
2 Sensitivity to Bandpass Filtering	105
3 Peak and Centre Clipping	107
4 Spectral and Temporal Modulations	110
5 Temporal Fine Structure	113
6 Target Periodicity	118
7 Conclusions	121
7 Taking ASR For A SPIN Test	125
1 Introduction	125
2 Competing Talker Background	126
3 Masker Modulations and Periodicity	137
4 Model Retraining	147
5 Conclusions	149

PART IV. CONCLUSIONS

8 Conclusions	155
1 Contributions of This Work	155
2 Limitations and Future Directions	157
3 Concluding Remarks	160
Appendices	163
A Author’s Publications	165
B Information Bottleneck Approximation	167
C Temporal Fine Structure on CRM corpus	169
D Extended Attributions for Competing Talker Experiment	171
Glossary	175
Bibliography	183

Acknowledgements

From the bottom of my heart, I would like to extend my sincerest thanks

to Dan,

for generously sharing your endless expertise and curiosity,
for your continuing belief in both me and my project,
for inspiring me to be bold;

to Claudia,

for your invaluable and constructive feedback,
for encouraging me to pursue my ideas,
for always keeping me on track;

to my examiners Patrick Naylor and Sarah Verhulst,

for reading my work with a critical eye and providing valuable feedback,
for the stimulating (and even enjoyable!) scientific discussion during my viva;

to Tobias, Eyal and Stuart,

for providing valuable feedback throughout my PhD;

to Darije,

for getting me through lockdown,
for being an amazing colleague and an even better friend;

to my Centre of Neurotechnology cohort,

for creating the intellectually stimulating environment that kickstarted my PhD,
in particular to Kate, for providing outstanding support and care,
in particular to Martin, for organising so many wonderful events,
in particular to Jeevan, for being a great friend;

to Wilten, Victor, Sara, Jacopo, Guillaume, Yann, Katha, Amadeus, Christos,

Marie, Jean, Yang, Jonny, Pamela, Nicolas, Isaac and Gabriel,

for creating a workplace filled with wisdom and kindness,

CONTENTS

for all those entertaining lunches and insightful meetings,
in particular to Wilten, for teaching me how to run,
in particular to Sara, for being a badass role model,
in particular to Christos, for helping me get my dream job;

to Hilde, Oscar, George A., Matt, Shanko, Daphne and John,
for reading and improving the many drafts of this thesis;

to Imperial College Musical Theatre and Big Band,
for keeping me sane;

**to Floor, Isaline, Kelly, Maartje, Rosalie, George B., Amrit, Angèle, Alessandra
and Emma,**
for friendships that endure separation by distance;

to Will, Sorcha, Oscar, Matt, Mary, Liv and Geraldine,
for making London feel like home,
in particular to Liv, for believing in me above and beyond (and making sure I know),
in particular to Geraldine, for being the loving partner in crime that got me through
this;

to my family,
for your love and and never-ending support,
in particular to my parents, for raising me to be the type of person that believes that
one day she can live in a castle,
in particular to Anneke, for setting an example of how to push through and follow
your dreams,
in particular to Marjolein, for showing me how its done, for counting noses and for
bottomless brunches,
and of course, to Hilde, for being the best source of informed optimism a clone could
ask for;

to Gabriel,
for being the calm of my storm,
for making me laugh, even if the penny takes a while to drop,
for always looking after me whenever I need looking after (which is perhaps more
often than I would ever admit).

List of Figures and Tables

1.1	Elephant drawings	3
2.1	Distinctive feature phone hierarchy	14
2.2	Acoustic correlates of source features	23
2.3	Acoustic correlates of vowel place features	25
2.4	Acoustic correlates of sonorant manner features	26
2.5	Acoustic correlates of sonorant place features	27
2.6	Acoustic correlates for obstruent manner features	28
2.7	Acoustic correlates for plosive place features	29
2.8	Acoustic correlates for fricative place features	30
2.9	Visualisation of the early auditory pathway	35
3.1	Hidden Markov Model (HMM) visualisation	46
3.2	Long Short-Term Memory (LSTM) visualisation	52
3.3	Time Delay Neural Network (TDNN) visualisation	53
3.4	Transformer visualisation	54
4.1	Neural auditory feature detection model overview	75
4.2	Sample responses of neural auditory feature detection models	81
4.3	Phone classification task performance of neural feature detectors	82
4.4	Feature importance of neural feature detection model	84
4.5	Noise robustness of neural auditory feature detection model	85
5.1	DNN-HMM and LSTM architecture	93
5.2	CNN-Transformer architecture	94
5.3	Sound level sensitivity of ASR systems	99
6.1	ASR performance in bandpass, and peak and centre clipping experiment	106
6.2	ASR performance in modulation filtering experiment	112
6.3	ASR performance in temporal fine structure experiment	116

LIST OF FIGURES AND TABLES

6.4	ASR performance in target periodicity	120
7.1	ASR performance in competing background talker task	133
7.2	Measurements of IM in the CNN-Transformer	136
7.3	Measurements of EM in the CNN-Transformer	137
7.4	ASR performance in modulating and periodic maskers	142
7.5	ASR accuracy versus masker modulation and periodicity benefits	143
7.6	Phone confusion matrices with periodic maskers	145
7.7	Side effects of matched condition testing	148
B.1	Visualisation of information bottleneck parameters	167
B.2	Correlation of LSTM accuracy and information bottleneck	168
C.1	Temporal fine structure experiment results on CRM data set	170
D.1	Measurements of EM for DeepLIFT and Integrated Gradients	173
D.2	Measurements of EM with amplitude, saliency and Gradient $\times$ Input	174

Notation

Unless explicitly stated otherwise, I will use the following mathematical notation throughout this thesis:

x	A lowercase letter indicates a scalar or a vector.
X	An uppercase letter indicates a matrix or a set.
$f(x)$	The function f applied to x .
$F(x)$	A filter F applied to x .
$[x, y]$	A vector comprising of the scalars x and y
x_i	The i th value of the vector x .
x_t	The value of x at time t .
$x \cdot y$	If both x and y are either vectors or matrices, this is the dot product between x and y . If x is a scalar and y a vector, matrix or scalar, it denotes scalar multiplication (i.e. every element of y is multiplied with x).
$ x $	The absolute value of x . If x is a matrix or vector, this denotes the elementwise absolute value, not the vector norm.

CHAPTER 1

Introduction

It is humbling to think about how the ticking of three tiny bones is responsible for so much of human communication. In one of evolution's most astounding feats, the human auditory system is able to transform invisible air pressure into language. This language allows us to share knowledge, plan the future and even develop a theory of mind. In this thesis, I focus on a particularly neat feat of human hearing: its robustness against distortions and noise. The importance of this ability is most evident when it is absent. For example, for those living with a hearing impairment, a crowded room can make it difficult or even impossible to have a conversation. Furthermore, as our communication becomes increasingly more digital, so grows our reliance on Automatic Speech Recognition (ASR) and, with it, our frustration when those machines fail to hear us. Despite recent advances in speech recognition technology, the disrupting soundscapes of the real world still form a significant obstacle for these systems.

Of course, the desire to understand the perception of speech in challenging listening environments is nothing new. As early as 1953, Colin Cherry described the cocktail party effect, referring to the ability of the human brain to focus its attention on one auditory stimulus while filtering out others. Over the years, many psychologists, neuroscientists and computer scientists alike have tried to untangle how human hearing achieves its remarkable robustness. This thesis is my small contribution to this large endeavour. At the forefront of my work are the basic features and feature extraction mechanisms that underlie robust speech processing. I will show how task-driven machine learning can be used as a tool to discover new auditory features as well as a way of confirming the importance of more well-known auditory features. However, I will also show through the application of rigorous auditory psychometric experiments that despite some recently reported similarities (Kell et al., 2018; Spille et al., 2018; Spille & Meyer, 2017), some of the auditory features used by ASR systems are fundamentally different from those used by humans. Through this interplay of auditory psychometrics, computational neuroscience

and machine learning, I hope to convince you that the continued cross-fertilization between these fields is a fruitful direction for improving our understanding of both human and automatic speech recognition.

But before we dive into the details, let's zoom out first. What is meant by *features of hearing*? Why would one even assume such features exist, and are important? And if they do exist, why would one assume that task-driven machine learning can be used to identify them? In the remaining part of this chapter, I will elaborate on the motivations for this work. This is followed by a thesis roadmap, which explains how the rest of this thesis is structured.

1. WHAT ARE AUDITORY FEATURES?

First, let me address the elephant in the room - what are auditory features? This question may be best answered through an analogy with vision. Imagine you were asked to draw the aforementioned elephant. You may draw its large ears, long trunk, wrinkly texture and gray skin. Depending on your drawing skills, only a small set of lines could be enough to convincingly draw an elephant (Figure 1.1). By making that drawing, you have reduced the elephant to a relatively small set of features. It is these features that set the elephant apart from, say, a golf ball:

Why are golf balls small and white?

- Because if they were big and grey they would be elephants.

Around the same time that elephant jokes such as the one above were circulating in the United States*, major steps were being taken into understanding the basic neural principles of visual processing. In their famous 1959 experiment, Hubel and Wiesel found that neurons in the primary visual cortex of a cat fired rapidly when presented with lines at particular angles. They called these neurons simple cells. Another group of neurons were referred to as complex cells, which responded best to lines in a certain angle moving in one direction.

In hearing, it is more difficult to come up with a similarly satisfying set of basic features. One could argue that the most basic component of any sound are its frequencies. After all, the first step in auditory processing involves a frequency decomposition that takes place in

* Elephant jokes are riddles or conundrums that involve an elephant. They are usually of an absurdist nature. In 1960, L.M. Becker Co of Appleton released a set of 50 trading cards titled *Elephant Jokes*. By July 1963, elephant jokes were ubiquitous and could be found in newspapers and magazines across the US. They continue to be popular on playgrounds to this date.

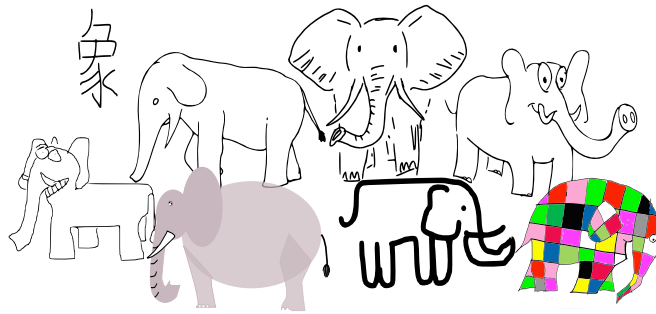


Figure 1.1. These are the artworks by some of my colleagues when prompted to ‘draw an elephant’.

the cochlea. But where do we go from there? What are the auditory equivalents of moving lines, curves and shapes? The answer to this question depends on who you ask. A musician may describe a sound by its rhythm, timbre or harmonicity. A signal processing engineer could turn to amplitude or frequency modulations. A linguist may try to find the acoustic correlates of phones, the most basic components of speech. Having identified a candidate feature, a neuroscientist may carefully poke a set of microelectrodes in a cat’s brain and measure the neural response, hoping to identify where - if anywhere - neurons are waiting for that stimulus.

The principal study from Hubel and Wiesel provides evidence for the idea that our brains build an image from simple stimuli, such as lines, into more complex representations. This idea further materialised during the connectionism movement in the 1980s (Goodfellow et al., 2016; McClelland et al., 1986). At that time, cognitive scientists were looking to explain cognitive function in a way that was somewhat grounded in neural computation. The central idea of connectionism was that a large number of computational units can achieve intelligent behaviour when networked together. These units were represented by perceptrons, which are simple linear units that represent neural computation as a weighted sum of a neurons input.

A key concept that arose during the movement is that of distributed representation. This refers to the idea that each input to a system should be represented by many features, and each feature should be involved in the representation of many possible inputs. Describing an input as a collection of potentially hierarchical features allows for the representation of more complex concepts using a limited amount of neurons. Rather than having each neuron represent one specific complex concept (sometimes referred to as ‘grandmother cells’, as one cell may represent one’s grandmother), distributed representations use patterns of activations of a group of neurons. Single neurons may even encode multiple features simul-

taneously, a concept referred to as multiplexing. Compared to sparse representations (like grandmother cells), the representational capacity of distributed representations increases combinatorically.

Although the term connectionist has gone out of fashion, many of these ideas are still present in current views on neural computation. In particular, the field of computational neuroscience is closely related to connectionism, although there are notable differences. Work in computational neuroscience is usually more concerned with biological accuracy and neurophysiological data than classic connectionist models would have been (Rogers & McClelland, 2014). The individual nodes in neural networks of computational neuroscientists are usually thought of as idealised descriptions of actual neurons, whereas connectionists tend to remain neutral about how the individual nodes map to actual neurophysiological entities. Consequently, notions such as firing rates and action potentials play an important role in computational neuroscience, but are rarely spoken of in connectionist research.

To those readers that are familiar with deep learning, the ideas behind connectionism and distributed representation may sound awfully familiar. Indeed, both concepts are central to the field of deep learning (Goodfellow et al., 2016). When popular media describes deep learning as ‘algorithms inspired by the brain’, this is the loose similarity that they are referring to. A particularly important contribution of the connectionist movement in relation to deep learning was the application of back-propagation to train deep neural networks with internal representations (Rumelhart et al. 1986a, Lecun, 1987). This algorithm is currently the dominant approach to training deep neural networks (Goodfellow et al., 2016).

2. A TASK-DRIVEN APPROACH

Considering the overlap between deep learning and computational neuroscience, it is unsurprising that both fields continue to influence one another. As a result of the dramatic performance improvements induced by deep neural networks, machine learning systems are now able to achieve near human-level performance on a range of sensory tasks such as image recognition and automatic speech recognition. Where the brain once loosely inspired the development of artificial neural networks, there now is a renewed interest in using deep neural networks as a model for human sensory processing.

The advantages of using machine learning models over real brains are obvious: it is cheaper, faster and ‘neural’ activity can be measured with as much temporal and spatial

precision as required. This line of research is of particular interest for the analysis of speech processing, since the ability to understand speech is unique to humans. This makes it difficult to explore speech processing in detail using animal models. Furthermore, due to the importance of the temporal component of speech it is often difficult to obtain non-invasive human data with the preferred temporal and spatial resolution.

As is made evident by recent works (Baby et al., 2021; Hu et al., 2020; Kell et al., 2018; Kollmeier et al., 2020; Rader et al., 2015; Spille et al., 2018; Spille & Meyer, 2017; Stolcke & Droppo, 2017), there is an increased desire to use machine learning to progress our understanding of human hearing. In this thesis, I explore both the potential and caveats of how machine learning can be used to uncover new knowledge about robust auditory features. First, I will show how a task-driven approach can be used to explore models of computational neuroscience, and can help to generate new ideas for candidate auditory features. Second, I will illustrate some of the caveats that come with using task-driven machine learning to answer questions about human hearing. Unlike models in computational neuroscience, high-performing DNNs are not designed to mimic human neural processing. I will show that in some cases, these machine learning systems may solve sensory tasks in vastly different ways than humans do.

3. THESIS ROADMAP

This thesis is organised in four parts. The first part, Part I, details the relevant theoretical foundations on which this work builds. Part II introduces an approach for generating new hypotheses about robust auditory features that leverages both computational neuroscience and machine learning. Part III focuses on the similarities and dissimilarities between the auditory features that are used in human and automatic speech recognition. The final part, Part IV, summarises the findings and limitations of this work and sets out a direction for future work. In the remainder of this section, each of the chapters will be described in more detail.

In Part I, the theoretical foundations upon which this work builds are introduced. First, relevant literature from the fields of acoustic phonetics, psychoacoustics and auditory neuroscience is reviewed in Chapter 2. This involves an introduction to acoustic phonetics, a field that investigates the acoustic cues associated with the production of speech. I then provide a short overview of the remarkable robustness to noise of the human auditory system. The chapter concludes with a review of neural mechanisms in the early auditory brainstem, in which I focus specifically on mechanisms that contribute to noise robustness.

In Chapter 3, I extend the notion of auditory feature to the basic auditory features that have been used in automatic speech recognition systems, specifically features and mechanisms that were developed to improve noise robustness. This is followed by a brief introduction to Deep Neural Networks (DNNs). These models learn to extract relevant features in the input data without manual feature engineering. In the last section of this chapter, I provide an overview of explainable artificial intelligence (XAI) techniques that may be applied to elucidate these relevant features, as well as the role that domain knowledge plays in understanding the inner workings of ASR systems.

In Part II, which consists of Chapter 4, I focus on models for feature extraction that are consistent with the biology of the early auditory system. This part of the thesis is based on Weerts et al. (2019). First, I introduce a neuro-inspired feature detection model that relies on a modest amount of parameters. This feature detection model, which is roughly based on the early auditory pathway, uses precisely timed inhibition as a mechanism for feature detection. I will first show that the small amount of parameters in the model can be tuned to detect a range of well-known features such as Amplitude Modulations (AMs) and onsets or combinations thereof. Notably, this capability is not hard-coded but arises from the basic neural mechanism the model employs. The precise type of features the model can detect depends on the chosen parameter settings, leading to a range of model variants. In order to objectively determine which variants of the model retain relevant information about a speech segment, the model variants are assessed using a classification framework. More specifically, a simple classifier is trained to the output of the feature detector models. Classification performance is then used to objectively determine which models retain task-relevant information from natural speech data. By exploring the parameter space of the model, a set of feature detectors that are important for understanding aspects of speech (such as phonetic content) are uncovered.

Part III of this thesis, which covers Chapter 5 to 7, focuses on the analysis of three ASR systems. This part of the thesis is based on Weerts et al. (2021). The first chapter in Part III, Chapter 5, is a generic method section for this part of the thesis. It introduces three open source ASR systems, as well as generic methods to assess their performance. This chapter also introduces *HumanlikeHearing*, a toolbox I developed to assess the performance of ASR systems on a barrage of psychometric experiments.

In Chapter 6, the similarities between the three ASR systems are tested on a range of basic distortions which include bandpass and modulation filtering, peak clipping as well as the importance of temporal fine structure and periodicity. This chapter shows that although some qualitative similarities exist between human and machine performance, none of the

three ASR systems use temporal fine structure and periodicity in the same way as humans.

In Chapter 7, the focus is shifted from distortions to two Speech in Noise (SPIN) tasks. The first experiment evaluates the effect of an increasing number of competing background talkers on the performance of the three ASR systems. Here, qualitative similarities between some of the ASR systems and human performance can be observed. Furthermore, Explainable Artificial Intelligence (XAI) techniques are applied to show how Informational Masking (IM) and Energetic Masking (EM) affect the performance of one of the ASR systems in this task. The second study focuses on the role of periodicity and modulations in a masker. It is shown that only one of the ASR systems treats masker periodicity and modulations in the same way as humans do. The types of errors that underlie these benefits are explored in more detail.

Finally, Part IV provides an overall summary and conclusion. This part consists of Chapter 8, in which I consolidate the results of the previous chapters and propose an outline for future work.

4. ONE FOR THE ROAD

This introductory chapter has introduced the core idea on which this thesis is based, namely that task-driven machine learning could be a useful tool for generating new hypotheses about the auditory features that underlie robust human hearing. The outcomes of the work presented in this thesis have implications for auditory neuroscientists and machine learning engineers alike.

For neuroscientists, Part II provides a case study of how a task-driven setting may be used to generate new hypotheses about human hearing. However, the results from Part III suggest that one should tread carefully when drawing conclusions from systems that may work fundamentally differently from the human brain.

For machine learning engineers, the comparison between automatic speech recognition and human speech recognition presented in Part III provides tangible directions for future research to improve ASR systems. To further motivate the cross-fertilization between fields beyond the outcomes presented in this work, an open source toolbox called *HumanlikeHearing* has been released. This toolbox allows researchers to assess new ASR systems and contribute additional psychometric tests.

Altogether, this chapter has outlined the aim, main contributions and general roadmap of this thesis. Now you know what is to come, it is time to buckle up - I sincerely hope you enjoy the ride.

Part I

Foundations

CHAPTER 2

Acoustic Phonetics and the Auditory System

Speech is both highly variable and inherently redundant. There is an infinite variety of ways in which the same word may be reflected in an acoustic waveform. This makes it difficult to relate an acoustic waveform to human speech through simple one-to-one connections. At the same time, humans can understand speech despite its highly variable nature, even under severe distortions. This is partly due to the inherent redundancy of speech: any utterance may be associated with a myriad of acoustic cues, of which only a few may be sufficient for accurate recognition. This thesis revolves around generating new insights regarding the basic auditory features the auditory system uses to unravel a speech signal. Fundamental to answering this question is a thorough understanding of both the acoustic components of a speech wave and the auditory system itself. In this chapter, I provide an overview of the acoustics of speech through the lens of acoustic phonetics, as well as an introductory description of what happens during the early stages of auditory processing in the brain.

To gain an intuition for the type of features that might be relevant, this literature review commences with an introduction to acoustic phonetics, a field that studies the physical properties of speech. This review is grounded in distinctive feature theory, a framework that describes speech through phonetic features. Phonetic features are language-independent features that can be used to precisely categorise the speech sounds found in most languages. Although the way in which speech materialises as an acoustic signal is highly variable, the way in which speech is produced (for example, through the positioning of the mouth and tongue), is relatively consistent. Understanding the acoustic correlates of phonetic features (i.e. the spectral and temporal modifications in the acoustic waveform that result from speech production) allows us to make predictions about the type of acoustic features the auditory system might be tuning into. Ultimately, however, it is not sufficient

for such acoustic features to merely exist; the auditory system must also have mechanisms in place to detect these specific features. To understand which mechanisms might underlie the detection of basic auditory features, I will also briefly review the neural mechanisms that have been studied in the mammalian early auditory pathway.

The rest of this chapter is organised as follows. Section 1 introduces distinctive feature theory. Specifically, I will describe a range of phonetic features and the way in which they are related to each other. In Section 2, the acoustic correlates associated with these phonetic features will be discussed. Although some of these acoustic correlates are sufficient for speech recognition in quiet, the auditory system may use different strategies when speech sounds are distorted. The noise robustness of the auditory system for speech recognition is discussed in Section 3. Finally, the basic components of auditory computational neuroscience are reviewed in Section 4. The focus here lies specifically on proposed neural mechanisms that may contribute to speech in noise perception.

1. AN INTRODUCTION TO PHONETICS

The first step to understanding the phonetic features described by distinctive feature theory requires a basic understanding of how speech is produced. Speech production starts in the lungs, which send air through the windpipe to the larynx, the organ that contains the vocal cords. In between the vocal cords is a space called the glottis. When the vocal cords vibrate by quickly opening and closing, the size of the glottis changes. This affects the flow of air, producing what is referred to as a glottal source. The sound then passes through the vocal tract, which consists of the throat, nose and mouth. By changing the shape of the vocal tract (for example, by changing tongue position or mouth shape) the source is modified. The new resonances and turbulence added in this process underlie the distinctive sounds we know as speech.

Phonetic features can be categorised under three general descriptors: the ‘source characteristics’, ‘articulator-free features’, which refers to the way in which the vocal tract is shaped, and ‘articulator-bound features’, which is where in the vocal tract the most significant shape change takes place (Halle & Stevens, 1991). Within each of these three groups, specific phonetic features can be defined. These are language-independent features that describe ways in which speech is produced with strong articulatory as well as acoustic correlates (Juneja, 2004). Within distinctive feature theory, each phonetic feature is considered to be binary, which means it is either present (which I will denote as *+feature*) or not (*-feature*). Often only a subset of phonetic features is required to uniquely describe

any phone, as some features are related to each other in a hierarchical fashion. An overview of the features relevant to English and how they divide up groups of phones in English is displayed in the decision tree in Figure 2.1.

In the remainder of this section, each of the three general descriptors (source characteristics, articulator-free features and articulator-bound features) and their associated phonetic features will be discussed in more detail. Note that throughout this thesis, I will use the IPA (International Phonetic Alphabet) notation for phone identification (International Phonetic Association and International Phonetic Association Staff et al., 1999). In this notation, most consonants are easily recognisable (e.g. /s/ and /t/), while vowels may be described in more unfamiliar characters (e.g. /æ/ for the *a* in *cat*). When needed, pronunciations in RP (Received Pronunciation) English* will be provided as extra guidance for readers that are unfamiliar with this notation.

1.1 Source Characteristics

The glottal source produced by the vibration of the vocal cords is a periodic complex tone with a relatively low F_0 (fundamental frequency, the lowest frequency of a periodic waveform). When a sound is associated with a glottal source, it is referred to as *+voiced*. A sound produced without vibration of the vocal cords has a more noise-like spectrum, and is referred to as unvoiced or *-voiced*. Examples of unvoiced sounds are certain consonants (such as /f/ and /s/) as well as whispered speech. By changing the shape of the vocal tract (for example, by changing tongue position or mouth shape), resonances are introduced at certain frequencies. In voiced sounds, these resonances are referred to as formants and show up as local peaks in the sound spectrum. Vowels are characterised by strong formants. In unvoiced sounds a distinct pattern of formants is generally not as pronounced, but changes in the vocal tract do induce distinct spectral shapes. Certain consonants occur in voiced/unvoiced pairs, such as /v/ and /f/. These consonants are identical from a speech production perspective, except that the voiced /v/ is produced through a glottal source, whereas the unvoiced /f/ does not involve glottal pulses, resulting in a ‘whispered’ sound with a noise-like spectrum.

* RP English is the accent traditionally viewed as the ‘standard’ for British English. It is by no means the most common British English accent, as it is used by only a very small proportion of the population of native English speakers in Britain (Trudgill, 2001). Nevertheless, since RP is one of the most commonly described and widely studied varieties of spoken English across the globe, I will stick with it for now.

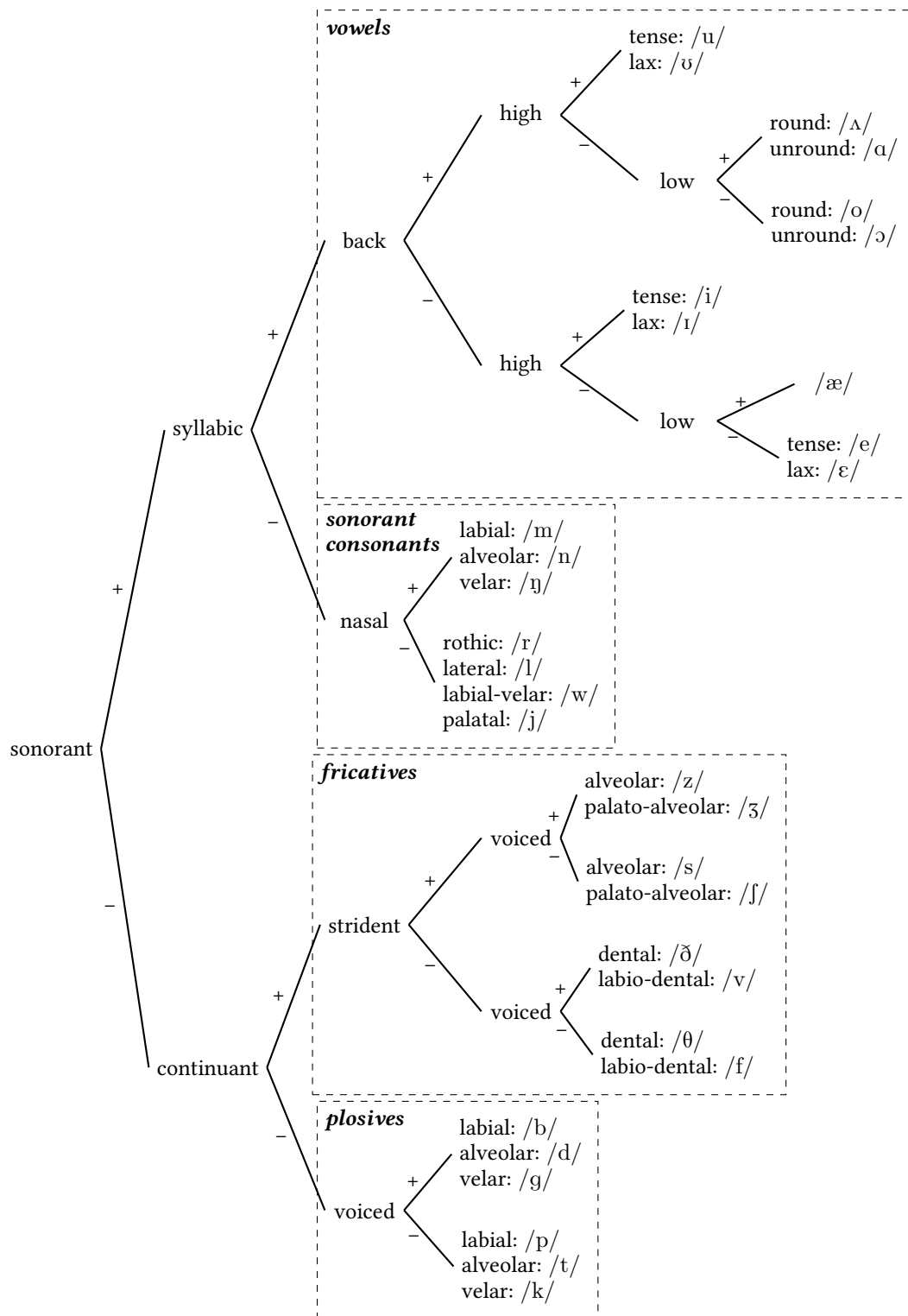


Figure 2.1. Overview of the hierarchy of phones based on their distinctive features

1.2 Articulator-Free Features: Manner of Articulation

The articulator-free features describe the way in which the source sound is modified by the vocal tract, without relating it to the articulator (e.g. mouth or tongue) that actually produces the constriction. This category is also referred to as ‘manner of articulation’ features. More specifically, articulator-free features describe how open or close the vocal tract is, how weak or strong the constriction is and whether air flows through the mouth or nasal cavity.

All sounds that are produced through an audible constriction are referred to as *+consonantal* sounds, which includes all phones but vowels and glides (/j/, /w/). Sounds in which the vocal tract constriction is not strong enough to produce a stop or turbulent noise are *+sonorant*, a category that includes vowels, nasal consonants (/m/, /n/) and glides (/j/, /w/). Vowels are distinguished from other sonorants as they are produced through an open vocal tract, which is described by the phonetic feature *+syllabic*. The non-syllabic sonorants (nasals and glides) can be further distinguished through the *+nasal* feature, which characterises phones in which air flows out through only the nasal cavities and not the mouth. Non-sonorants are usually referred to as obstruents and are characterised by the production of turbulent noise. This group includes fricatives (e.g. /f/ and /s/) and plosives (e.g. /p/ and /b/). Fricatives differ from plosives because they are *+continuant*, which is a phonetic feature that describes an incomplete constriction. Plosives, on the other hand, do involve complete constriction. Fricatives can be further divided by the degree of friction or greater turbulent noise, which is referred to as *+stridence*. These sounds are produced by placing an obstacle in the airway downstream from the constriction, which enhances turbulent noise in the front cavity. Examples of strident fricatives are /s/ and /z/, whereas /f/ and /v/ are examples of non-strident fricatives.

1.3 Articulator-Bound Features: Place of Articulation

The articulator-bound features describe the place of articulation. That is, they describe where the most significant constriction takes place or, in the case of vowels, what the shape of the tongue is. The main articulators for consonants are the lips, tongue blade, tongue body, the pharynx (throat) and larynx (voice box), each of which has an associated set of phonetic features.

A constriction at the lips is referred to as *+labial*. Consonants in this group include /b/, /p/, /w/ and /m/. When the lips are positioned against the teeth, for example in the consonant /v/, it is referred to as *+labio-dental*. The lips can be *+round*, which is typical

for the consonant /w/, but is also associated with certain vowels. For example, the vowel /ɑ/ (*ba**l**m*) is not rounded (*-round*), whereas the vowel /ɔ/ (*sto**r**y*) is produced through rounding of the lips.

When the tongue blade is the main articulator, it can be positioned in contact with or in front of the alveolar ridge, the small protuberance located just behind the upper front teeth. This is referred to as *+anterior*. Examples of consonants with this feature are /z/, /s/ and /n/. It can also be *+distributed*, which means it extends for some distance in the mouth. In English, consonants are usually either *+alveolar*, which refers to anterior and non-distributed consonants (/z/, /s/), or *+palato-alveolar*, which are distributed and non-anterior (/ʒ/, as *rou**g**e*, /ʃ/ as *wa**t**ch*). The exception to this is the /r/, which is neither anterior nor distributed. This consonant is sometimes distinguished from others as *+rhotic*, although some argue this class of consonants can not be uniquely defined by place or manner (Chabot, 2019). A further distinction between alveolar consonants can be made through the *+lateral* feature, which is formed by lateral airflow around one or both sides of the tongue (/l/).

When the tongue body is the main articulator, the sound is referred to as *+velar*. This includes consonants such as /k/ and /g/, as well as all vowels. The consonant /w/ is labial-velar, which means a constriction is made both with the lips as well as the tongue body. The tongue can be *+low*, *+high* or neutral (*-low* and *-high*). Note that *+low* and *+high* cannot be positive at the same time. In some vowels the tongue is positioned towards the *+back* (as in /a/ and /u/) whereas in others it is not (e.g. in /i/ and /e/). When the tongue body is high but not back, it is typically raised against the hard palate. This specific combination is referred to as *+palatal*.

The last articulator considered here is the pharynx (the throat). For example, by advancing the tongue root, the pharynx is widened. This results in what is called a *+tense* vowel. This can distinguish between for example /i/ (*bea**t***) and /ɪ/ (*bi**t***) as well as /e/ (*ma**y***) and /ɛ/ (*be**t***). The distinction between tense and lax vowels is rare in low vowels. The pharynx can be narrowed by retracting the tongue, which is sometimes indicated by the feature *+constricted tongue root* (K. N. Stevens, 2000), and occurs in the vowels /o/ (*ya**w**n*) and /ɔ/ (*sto**r**y*). However, in English the tense versus lax distinction are characterised by more than just the tongue root position (e.g. vowel length and movement of the tongue body throughout the course of the vowel), which is why I will use the *+tense* to encompass this phenomenon.

Lastly, some phones use the larynx as sole articulator. In English, this applies to the phone /h/. Although /h/ is historically referred to as the ‘voiceless glottal fricative’ and

usually gets assigned the *+glottal* feature, it is not always associated with frication in the glottis. From a phonetic point of view, the shape of the vocal tract is that of the succeeding and (when applicable) preceding vowels. Usually /h/ is assumed to have no other features for place or manner (Ladefoged & Maddieson, 1996). As such, it is not included in Figure 2.1. However, I will group it under fricatives when discussing its acoustic properties, as it patterns as a fricative phonologically.

2. ACOUSTIC CORRELATES

Each of the phonetic features introduced in the previous section is associated with distinct modifications of the acoustic waveform. In this section, I will first discuss four different classes of acoustic features. I then provide an overview of some of the most prominent acoustic correlates that tend to be associated with each of the phonetic features introduced in the previous section.

2.1 Types Of Acoustic Features

Generally speaking, we can distinguish between two classes of acoustic features: spectral features, which describe the frequencies present in the sound, and temporal features, which describe the way the sound changes over time. Of course, frequency and time are not orthogonal features, as changes in frequency will result in changes in the temporal pattern and vice versa. However, as will be described in more detail in Section 4, frequency and time may be represented in two somewhat distinct ways in the auditory system. Auditory nerve fibers respond to a frequency band with a preferred centre frequency and are tonotopically organised in space. This means that tones that are close to each other in terms of frequency are represented in neighbouring regions in the brain. The ‘place’ of neural activity in the auditory system thus provides information about frequency, which is also referred to as place-frequency information. However, there is also a temporal component to the neural action potentials produced in the auditory system. For example, an auditory nerve fiber does not only respond mainly to a given frequency, it also spikes in phase with its preferred frequency component of the incoming sound. This results in a temporal code that carries information in itself. Temporal codes such as these can play a significant role in hearing. In an extreme case, it has been shown that patients with a single-channel cochlear implant (thus receiving input through a single electrode that merely provides loudness and timing cues) are still able to understand spoken speech, despite the lack of

place-frequency information (Rosen, 1992). This suggests that purely temporal information can be an effective cue for speech perception.

Throughout this discussion it is important to recognise that although the auditory system may be able to use a certain cue (such as temporal information) in the absence of other cues (such as place frequency features), this merely implies the cue is sufficient, not that it is necessary. As speech signals are highly redundant, the auditory system can (and does) use a large variety of cues, sometimes interchangeably. In phonetics, this is referred to as cue trading, where the presence of one cue compensates for the removal of another. The remainder of this section outlines place-frequency features as well as three types of temporal features: the envelope, periodicity and temporal fine structure (Rosen, 1992).

Place-Frequency Features

As will be discussed in more detail in Section 4, the first step in auditory processing comprises a frequency decomposition of the incoming sound in the cochlea. Consequently, the shape of the sound spectrum (i.e. the power spectral density) is likely to be one of the primary cues for auditory perception. In signal processing, spectral representations are usually obtained through a Fourier transform, or, when frequency is considered over time, the Short Time Fourier Transform (STFT). However, when relating spectral features used by the auditory system, it is useful to note the differences between the frequency analysis performed in the cochlea and Fourier transforms. The cochlea can be roughly modelled as an array of non-linear, level-dependent, coupled band-pass filters known as auditory filters. The bandwidth of higher frequencies tends to be wider than that of low frequencies, making the auditory system less precise in its frequency selectivity for higher frequencies than it is for lower frequencies.

Formants are an example of distinctive frequency-based acoustic features. As mentioned in Section 1, formants are local maxima in the sound spectrogram that result from resonances in the vocal tract. As will be discussed in the next section, many of the acoustic correlates of phonetic features that have been studied in the past are based on formants. The formant with the lowest frequency is called F_1 , the second F_2 , etc. Experiments using synthesized speech (during which a click train that simulates the glottal pulse is modified through two bandpass filters) have shown that the first two formants are usually sufficient for identifying vowels (Hayward, 2014). When more formants are present, the auditory system may still make use of two frequency peaks in the auditory filter response. When asked to match a vowel with multiple formants to a synthetic two-formant vowel, human

subjects tend to choose a synthetic vowel with an F_1 that matches the original sound, but an F_2 that may be higher depending on the values of F_3 and F_4 (Carlson et al., 1974). This second frequency is referred to as the ‘equivalent perceptual second formant’ or F'_2 (the prime denoting that this is not the true F_2 of the sound but its perceptual equivalent). Carlson et al. (1974) showed that a peak of a similar frequency can be obtained through a model of auditory filtering.

Unfortunately, recognising vowels is not as simple as matching formants to a pre-defined set of frequencies associated with each vowel. One of the reasons for this is that formant frequencies vary per individual and depend in part on the length of the vocal tract of the speaker. Although the average F_1 for a male voice ranges from 250 Hz to 850 Hz (/i/ to /a/), and F_2 ranges from 595 Hz to 2.4 kHz (/u/ to /i/), they might be significantly higher for a female or child’s voice (Catford, 1988). Furthermore, although formants are distinctly present in vowels when spoken in isolation, they may only be present for a short period of time in natural speech. The auditory system may instead tune into formant transitions rather than individual formants. These transitions are the result of coarticulation, during which the articulators move smoothly from one position into the next. There is evidence to support this idea. For example, the ability to identify vowels can be better for vowels in consonant-vowel-consonant words than for vowels presented in isolation (Strange & Bohn, 1998). This suggests that frequency transitions and/or modulations play an important role in vowel recognition.

Spectral cues are equally important for distinguishing between consonants. For example, place features in consonants have acoustic correlates in the frequency spectrum of the initial release burst and following formant transition (Hazan & Rosen, 1991). Again, however, it has been shown that the temporal aspects of frequency transitions play an important role. For example, Strange and Bohn (1998) have shown that dynamic changes in the spectra are necessary for plosive consonant recognition.

Envelope Features

When considering temporal features, a distinction is often made between the envelope and the Temporal Fine Structure (TFS). Speech is an amplitude modulated signal, which implies it has two components: an envelope, which refers to the slower fluctuations in the amplitude of the sound, and the TFS, which refers to rapid oscillations of a carrier. The envelope is not synonymous with the low frequency component of a sound. For a single tone with a carrier frequency f_c , a sinusoidal amplitude modulation of f_m will show up as

two peaks at $f_c \pm f_m$. Envelope features are sometimes considered to be the modulations of the incoming sound wave itself, but may also refer to the modulations found in the responses of the auditory filters in the basilar membrane. Following Rosen (1992), envelope features will be considered to refer to fluctuations of up to 50 Hz in this work.

The envelope of a speech sound can act as a segmental cue. For example, because vowels are produced through an open vocal tract, they tend to be louder. As a result, speech sounds tend to modulate with the syllabic rate of speech. Furthermore, because the duration of vowels tend to vary with vowel quality, the duration of the syllable can hint at vowel identity (e.g. *heed* versus *hid*). Short silence gaps can indicate the presence of voiceless plosives or affricatives, and short transients in loudness can be used to identify the location of plosive consonants. Envelope features also indicate prosodic cues, such as word duration and general speaking rate, which can carry a distinctive meaning.

Envelope cues in speech perception are typically studied by making use of a vocoder (Dudley, 1939). Generally speaking, the term vocoder refers to a category of speech coding that synthesizes speech. Typically, the term vocoder is used to refer to a channel vocoder, which filters a speech signal into several contiguous frequency bands and processes the signal in each band in such a way that only envelope cues are preserved. This is done by first extracting the envelope in each band and by then modulating either a noise band or sinusoid centered at the same frequency band. These two different methods of modulating are referred to as noise vocoders and tone vocoders, respectively. The modified signals in each band are added together to obtain the vocoded sound wave. In its most extreme form, only one band is used. It has been shown that in quiet, speech tends to be highly intelligible with a moderate (4-16) number of bands (Loizou et al., 1999; Shannon et al., 1995).

Periodicity

Periodicity information relates to the distinction between periodic and aperiodic stimulation as well as the rate of periodicity. Generally speaking, periodicity refers to a function that repeats its values at regular intervals. In the context of speech, periodicity is usually the result of the periodic vibration of the vocal cords when voiced speech is produced. As a result of these vocal cord vibrations, periodic speech sounds typically fluctuate at rates between 50-500 Hz, whereas aperiodic sounds tend to fluctuate at rates from a few kHz up to 10 kHz. As a temporal feature, periodicity mainly refers to the ability to distinguish between periodic and aperiodic sounds. Periodic sounds are perceived with a degree of ‘buzziness’, which results from their temporal regularity as well as the presence of harmonics in the

frequency spectrum. Aperiodic sounds tend to be irregular and have a continuous spectrum, which results in a noisy quality.

Periodicity can function as a strong acoustic cue for voiced sounds, as the presence of low quasi-periodic acoustic energy reflects the presence of glottal pulses. Aperiodic sounds are usually the result of turbulence noise, which is associated with voiceless speech. The delay between the release of a consonant and the start of voicing is known as the *voice onset time* (VOT). The duration of the VOT, as well as the presence or absence of periodicity during the closure interval, can help in distinguishing voiced and unvoiced consonants. Intonation, or the variations in pitch, are also characterised by the rate of periodicity. This can serve as a prosodic component of speech, but may also have a lexical function in tonal languages such as Mandarin.

Temporal Fine Structure

Temporal Fine Structure (TFS) refers to fluctuations that typically range from 600 Hz-10 kHz that cause variations of wave shape within single periods of periodic sounds, or short time intervals of aperiodic sounds. TFS refers to the specific patterns of activation in the temporal structure, not merely the presence of specific high frequencies (although the two are related). In the auditory filters, TFS shows up as rapid oscillations with a rate close to the centre frequency of the band. Information about TFS is thus carried in the patterns of phase locking in the auditory nerve (Young & Sachs, 2008). The upper limit of phase locking in humans is unknown (Moore, 2008). For comparison, the ability to phase lock becomes weaker at frequencies above 4 kHz to 5 kHz for most mammals. However, some useful phase locking information may persist for frequencies up to at least 10 kHz (Heinz et al., 2001).

TFS information in both low and high frequency bands can be used by the auditory system to improve speech recognition in noisy environments. For example, K. Hopkins et al. (2008) defined a vocoder in which bands up to band number j remained unprocessed (intact TFS and envelope information), whereas bands of a higher centre frequency were vocoded (only intact envelope information). They showed that performance improved as the value of j increased, moving from a fully vocoded to fully intact signal. Similar effects were observed by progressively adding TFS from the highest frequency band (K. Hopkins et al., 2010).

Summary

In this section, I introduced four types of acoustic features that play a role in human speech perception. These features have been roughly divided into two categories: place-frequency features and temporal features. Place-frequency features are of a predominantly spectral nature, and include features such as formants. The transitions between formants (which are reflected in the spectrogram as Frequency Modulations (FMs)) are also considered a prominent cue for speech recognition. In this review, the temporal features have been further divided into three categories: envelope, periodicity and TFS features. These features are all Amplitude Modulations (AMs), although each of them work at different AM frequencies. Envelope features refer to slow fluctuations in overall amplitude, while TFS refers to the rapid oscillations of a carrier. Periodicity features, as defined here, lie somewhat in between: they relate to the distinction between periodic and aperiodic stimulation as well as the rate of periodic stimulation.

2.2 Acoustic Correlates of Phonetic Features

Each of the phonetic features described in Section 1 modifies the acoustic waveform in distinct ways. In this section I provide an overview of some of the cues in the acoustic signal associated with each phonetic feature. Some of these cues are invariant, meaning they are always present when a sound is modified in a particular way, while others are contextual, meaning they depend on the context (such as preceding and following phones). A compact overview of the acoustic correlates for each phonetic feature discussed in this section are displayed in Table 2.2-2.8. Most prominent cues reported in the literature relate to spectral properties (particularly formants and formant transitions), envelope features (for example for distinguishing between sonorants and non-sonorants) as well as periodicity (for example for distinguishing between voiced and unvoiced segments). TFS has traditionally not been explicitly discussed in this context. One reason for this discrepancy is that the importance of TFS is less apparent in idealised environments without additive noise or other distortions (see e.g. K. Hopkins et al., 2008). Furthermore, the ways in which TFS (or the lack thereof) may affect speech intelligibility are generally less well understood.

Voiced Speech

The glottal pulse associated with voiced speech is most pronounced in vowels, but plays a critical role in distinguishing certain consonants. Fricatives such as /f/ and /v/ have the same articulator-free and articulator-bound features and are only distinguished through

Source Features			
Features	Examples	Description	Acoustics
+voiced	Sonorants, certain fricatives (e.g. /v/, /z/), certain stops (e.g. /b/, /d/, /g/)	Vocal fold vibration	Clear presence of a low F_0 as a result of the quasi-periodic glottal pulse

Table 2.2. Acoustic correlates of source features.

their *voiced* feature. Methods for the detection of voicedness tend to be closely related to pitch detection methods, as the pitch of a speaker is generally determined by the frequency at which the vocal cords vibrate.

A range of basic acoustic cues could be used for voiced phone detection. For example, voiced sounds tend to have high average energy levels and high amplitudes. Furthermore, unvoiced speech segments tend to be dominated by noise, and typically have higher energy at high frequencies than unvoiced sounds. In signal processing, the *zero crossing rate* is sometimes used as a rudimentary way of distinguishing between voiced and unvoiced segments. This method measures the rate of sign changes across a signal, which tends to be lower for signals that have higher energy at lower frequencies.

Unfortunately, neither energy nor zero-crossing count are noise-robust methods, as they can easily be distorted by the addition of noise in certain frequency bands. The periodicity associated with voiced sounds could act as a more robust cue. In the temporal domain, periodicity can be detected using autocorrelation, as repeating temporal patterns will result in a peak at the lag associated with the fundamental frequency of a periodic sound. Furthermore, the periodic glottal source does not only result in a low-frequency F_0 , but also induces harmonics, which are spectral peaks at frequencies that are multiples of the fundamental frequency of a periodic sound. Periodicity cues can thus be detected in a way that does not rely on the ratio of high and low-frequency spectral energy, which makes these methods more robust to disruptions such as low-frequency additive noise. Models of pitch perception commonly assume that both place frequency (particularly for lower harmonics) and temporal integration (especially for higher harmonics) play a role in how the auditory system detects pitch (Ives & Patterson, 2008; Meddis & Hewitt, 1991; Moore, 2013; Patterson et al., 1987; van Noorden, 1982).

Vowels

Table 2.3 provides an overview of some prominent acoustic correlates associated with vowel place features. Formants are often considered to be a prominent cue for vowel recognition (Hayward, 2014). These resonances are largely dependent on the positioning of the tongue body, although the positioning of the lips and the pharynx also affect the quality of a vowel. Algorithmic ways of formant detection automatically include linear predictive coding (Snell & Milinazzo, 1993) as well as Hidden Markov Models (HMMs) and Kalman filters (Q. Yan et al., 2007).

If the tongue is positioned towards the back of the mouth (*+back*), F_1 tends to be low (up to 1200 Hz) and F_1 and F_2 tend to be close to each other. Delattre et al. (1952) showed using synthetic speech that a low F_1 can be sufficient to distinguish between back and front vowels, but recognition is improved when appropriate F_2 peaks are present. Carlson et al. (1974) showed that F_3 and F_4 can play a considerable role in determining the exact vowel quality that is heard for front vowels in which F_2 is very close to F_3 .

A *+low* tongue position results in a high F_1 , whereas a *+high* tongue position lowers F_1 (Pickett, 1980). For *+high* vowels, F_1 can lie as close as 100-200 Hz to the fundamental frequency. When the tongue is in a medium height position (*-low*, *-high*), F_1 lies somewhere in between the F_1 frequency of *+low* and *+high* vowels (K. N. Stevens, 2000).

The rounding of lips (*+round*) tends to result in more salient and prominent spectral peaks in vowels (K. N. Stevens, 2000). Specifically, for *+back* vowels, F_2 and to a lesser extent F_1 are lowered and sharpened. For front vowels (*-back*), the F_2 - F_3 spectral peak tends to be lower and more prominent. This is because lip rounding causes F_2 and F_3 to be closer to each other and lower overall. It also reduces the bandwidth of F_3 .

Lastly, the *+tense* feature of vowels depends on a combination of articulators. This feature occurs most frequently in *-low* vowels and is rarely applied to *+low* vowels. The acoustic effect of the tense articulators is a higher F_1 and generally lowered F_2 (K. N. Stevens, 2000). Tense vowels also tend to be shorter in English (Pickett, 1980).

Sonorant Consonants

Sonorants can typically be distinguished from obstruents (*-sonorant*) through the lack of noise and silent plosives in their spectrum. This group contains both vowels and sonorant consonants. The main acoustic cue that can help to distinguish between vowels and sonorant consonants is loudness: as vowels are produced with an open vocal tract, they tend to be louder than sonorant consonants. Most sonorant consonants are *+voiced*. An

Place Features of Vowels			
Features	Examples	Description	Acoustics
+back	/u/ (<i>boot</i>), /ʊ/ (<i>hook</i>), /a/ (<i>bath</i>), /ʌ/ (<i>gut</i>), /o/ (<i>yawn</i>), /ɔ/ (<i>not</i>)	Tongue positioned towards back of mouth	Low F_2 ; weaker amplitude in F_3 and above; F_2 tends to be very close to F_1
-back	/i/ (<i>free</i>), /ɪ/ (<i>bit</i>), /æ/ (<i>cat</i>), /e/ (<i>may</i>), /ɛ/ (<i>bed</i>),	Tongue position towards the front of the mouth	High F_2 ; stronger amplitude for F_2 - F_4 ; the maximal F_2 tends to be higher for low vowels that also have a high F_1
+low	/a/ (<i>bath</i>), /ʌ/ (<i>gut</i>), /æ/ (<i>cat</i>),	Low tongue position	High F_1
+high	/u/ (<i>boot</i>), /ʊ/ (<i>hook</i>), /i/ (<i>free</i>), /ɪ/ (<i>bit</i>)	High tongue position	Low F_1 ; generally within 100-200 Hz of the fundamental frequency
+round	/ʌ/ (<i>gut</i>), /o/ (<i>yawn</i>)	Rounding of the lips	Tendency for the major broad spectral peak to be more salient and prominent; for +back vowels, F_2 and to a lesser extent F_1 are lowered; for front vowels, the F_2 - F_3 spectral peak is lower and more prominent.
+tense	/u/ (<i>boot</i>), /i/ (<i>free</i>), /e/ (<i>may</i>), /a/ (<i>bath</i>), /ɔ/ (<i>not</i>), /o/ (<i>yawn</i>)	Tense articulators (e.g. advancing tongue root)	Tenser articulators combined with a non-low tongue position results in a higher F_1 and a generally lowered F_2 ; often shorter than lax (-tense) vowels.

Table 2.3. Several acoustic correlates for place features of vowels.

Manner Features of Sonorants			
Features	Examples	Description	Acoustics
+sonorants	Vowels, nasals, glides, liquids	No or weak constriction	Lack of stop and noise-like spectrum; tends to have more prominent formants; overall higher acoustic energy than obstruents
+syllabic	Vowels	Open vocal tract	Higher overall energy in spectrum
-syllabic	Nasals, glides, liquids	Slightly constricted vocal tract	Less energy than vowels, shorter than vowels
+nasal	Nasals (/m/, /n/, /ŋ/ (<i>sing</i>))	Mouth is closed, air flows through nasal cavities	Low frequency murmur added around 250ms; narrowband with weak midformants; typically brief formant transition of F_1
-nasal	Glides (/w/, /y/), liquids (/r/, /l/)	Mouth open	Weak constriction; slow movement of articulators

Table 2.4. Several acoustic correlates for manner features of sonorants.

overview of the acoustic correlates of manner features and place features in sonorant consonants are shown in Table 2.4 and 2.5, respectively.

The *+nasal* consonants tend to have a low frequency prominence around 250 Hz, as well as a weaker prominence in the vicinity of 1000 Hz (K. N. Stevens, 2000). Furthermore, nasals tend to be relatively narrowband, with only weak midformants (Pickett, 1980). When a nasal consonant is followed by a vowel, this is usually associated with abrupt transitions of F_2 and F_3 (K. N. Stevens, 2000). The exact transitions depend on the following vowel, but are similar to transitions induced by the release of plosive consonants that are produced by the same articulator. Similarly, the formant transitions during the implosion of a nasal consonant are similar to those induced by consonant release, but in the reverse order. The English nasal consonants /m/, /n/ and /ŋ/ (as in *sing*) are distinguished through their place features, produced through the lips (*+labial*), tongue blade (*+alveolar*) and tongue body (*+velar*), respectively. The place feature of these nasal consonants are typically described through the F_2 transition following the nasal murmur. For example, A. M. Liberman et al. (1954) showed that the upward transition in F_2 results in the perception of /m/ (*+labial*), a downward transition in /n/ (*+alveolar*) and an upward transition into a high F_2 to /ŋ/ (*+velar*).

Place Features of Sonorants			
Features	Examples	Description	Acoustics
+labial	/m/	Constriction through lips	Upward F_2 transition
+alveolar	/n/	Constriction of tongue blade against alveolar ridge	Downward F_2 transition
+velar	/ŋ/ (<i>sing</i>)	Constriction through tongue body	Upward F_2 transition to a high F_2
+rothic	/r/	Retroflex positioning of the tongue	Large F_3 transition to or from low frequency
+lateral	/l/	Airstream along the sides of the tongue	Low frequencies strong and broadband, plus moderate strength midformants; relatively wide gap between F_2 and F_3 ; abrupt changes following release
+labio-velar	/w/	Rounding of the lips as well as constriction through tongue body	Wide broadband F_2 that is far below 1 kHz
+palatal	/j/ (<i>you</i>)	Tongue body against palate	High frequency F_2 and F_3 , with F_3 close to F_4 ; large F_3 bandwidth

Table 2.5. Several acoustic correlates for place features of sonorants.

The non-nasal (*-nasal*) sonorant consonants are glides and liquids. This group is sometimes referred to as approximants. As was the case for vowels, the distinguishing factor to differentiate between approximants is the positioning of the tongue. In contrast to other consonants, approximants are typically not associated with abrupt changes in frequency. When positioned in-between vowels, approximants tend to affect the mid- and high-frequency range over a time interval of 200-250 ms.

The phones /r/ and /l/ are referred to as liquids, and can be distinguished from glides through their shorter constriction period. The liquid /r/ is characterised by a retroflex or ‘rothic’ tongue. Acoustically, this tongue position results in a large F_3 transition, either from low to high or high to low. The other main liquid in English is /l/, which is produced through a *+lateral* positioning of the tongue. This tongue position results in strong and broadband low frequencies, with additional midformants of moderate strength (Pickett, 1980). At the release of the consonant, some abrupt changes in the frequencies and amplitudes of spectral prominence occur (K. N. Stevens, 2000).

Glides are produced in a similar way as high vowels, except that the oral cavity is more constricted. As a result of the high positioning of the tongue, they typically have a relatively low F_1 frequency in the range of 250-300 Hz (K. N. Stevens, 2000). The glide /w/

Manner Features of Obstruents			
Features	Examples	Description	Acoustics
-sonorants	Fricatives, plosives	Significant constriction	Noise-like spectrum, which may have formants in certain regions depending on place of friction
-continuant	Plosives	Complete constriction of the vocal tract	Silence in the spectrogram followed by a burst; rapid and brief formant transitions
+continuant	Fricatives	Incomplete constriction	No silence in the spectrum; presence of continuous noise
+strident	/s/, /z/	High degree of friction and greater turbulent noise; typically palatal or alveolar	Stronger in total energy than non-strident fricatives
-strident	/f/, /v/	Lower degree of friction; typically labial or labio-dental	Weaker total energy than non-strident fricatives

Table 2.6. Several acoustic correlates for manner features of obstruents.

is largely similar to the vowel /u/ and has a *+high*, *+back*, *+tense* tongue body position. In addition, it has *+rounded* lips. The tongue body and lip constrictions lead to an F_2 with a wide broadband that is far below 1 kHz. On the other hand, the glide /j/ has a similar tongue position as the vowel /i/ (*+high*, *+tense*). In addition, the tongue is raised to a *+palato-alveolar* position. This constriction gives rise to a high-frequency F_2 and F_3 , with F_3 being close to F_4 . Furthermore, F_3 has a relatively large bandwidth (K. N. Stevens, 2000).

Plosive Consonants

Obstruent consonants are *-sonorant* phones. The manner features associated with obstruents are displayed in Table 2.6. Obstruents can be further divided into plosive consonants, which are discussed in this section, and fricative consonants, which will be covered in the next section.

Plosive consonants are distinguished from other obstruents through the *-continuant* feature. Like all obstruents, they are characterised by a noise-like spectrum caused by a constriction that introduces turbulent noise. In contrast to fricatives, plosive consonants involve a full constriction of an articulator, which results in a silent interval. In addition to this, they tend to be fast and introduce brief formant transitions into vowels.

We can distinguish between six plosive consonants in English through their place features (Table 2.7). The phones /p/ and /b/ are *+labial*, /t/ and /d/ are *+alveolar* and /k/ and /g/ are *+velar*. In each of these pairs, the first plosive consonant is unvoiced

Place Features of Plosive Consonants			
Features	Examples	Description	Acoustics
+labial	/b/, /p/	Constriction through lips	Most energy during burst in the low frequencies 500-1500 Hz with a diffuse spectrum that is flat or falling; slightly faster formant transitions than alveolar and velar plosives
+alveolar	/t/, /d/	Tongue blade against the alveolar ridge	Flat spectrum with energy concentration around 500 Hz as well as potential peaks in higher frequencies (>4000); overall spectrum tends to be diffuse and rising
+velar	/k/, /g/	Produced through the tongue body	Spectral peak in mid frequencies around F_2' of the vowel that follows; spectral peak is not diffuse but has narrow bandwidth 1 kHz

Table 2.7. Several acoustic correlates for place features of plosive consonants.

(-voiced) and the second +voiced. Voiceless plosive consonants are usually associated with an abrupt decrease of glottal pulses at the instant of the closure, and are resumed 5-15 ms after release of the closure. There is typically a downward movement of F_1 immediately prior to implosion, and a rise in F_1 following the release (K. N. Stevens, 2000).

Typically, the place features affect the spectral shape of the friction noise around and immediately following the consonant release (K. N. Stevens, 1980). The +labial plosives /b/ and /p/ have a primary concentration in the low frequencies 500-1500 Hz (Halle et al., 1957). They typically have slightly faster formant transitions than +alveolar and +velar plosives. The spectra for +labial consonants tend to spread out over time ('diffuse') in frequency and are flat or falling (K. N. Stevens, 1980). The +alveolar plosives /t/ and /d/ have either a flat spectrum or one in which the higher frequencies (above 4000 Hz) predominate, aside from an energy concentration in the region of 500 Hz. Like +labial plosives, their frequency spectrum tends to be diffuse, but it is rising rather than flat or falling (K. N. Stevens, 1980). As with most +velar consonants, the +velar plosive consonants /k/ and /g/ have a prominent spectral peak in the range of the second major spectral prominence in F_2' during the burst (K. N. Stevens, 2000). At the release of +velar plosive consonants, there is a prominent peak in the mid-frequency range, and the amplitude of this peak is comparable to the F_2 prominence in the vowel that follows. In contrast to

Place Features of Fricative Consonants			
Features	Examples	Description	Acoustics
+dental	/θ/ (<i>thin</i>), /ð/ (<i>this</i>)	Tongue in between lips	No major spectral peak; weak overall energy; tends to be short;
+labiodental	/f/, /v/	Lips touches teeth	No major spectral peak; weak overall energy
+alveolar	/t/, /d/	Tongue blade against the alveolar ridge	Spectral peaks in F_4 - F_5 region or even higher that exceed the energy of the peak of the adjacent vowel
+palatoalveolar	/s/, /z/	Tongue raised towards the palate and alveolar	Spectral peak in F_3 - F_4 region that exceed the peak of the adjacent vowel
+glottal	/h/	Aspiration through glottis	Noisy spectrum with vowel-like resonances

Table 2.8. Several acoustic correlates for place features of fricative consonants.

+alveolar and *+labial* plosive consonants, the spectral peak is not diffuse but shows a rather narrow bandwidth. It has been shown that the general shape of the spectrum is sufficient to correctly classify these consonants (Blumstein & Stevens, 1980).

Fricative Consonants

Fricatives are characterised by a continuous turbulent noise spectrum that results from the constriction. In contrast to plosive consonants, there is no full constriction with a subsequent release. Instead, fricatives tend to have a relatively stationary constriction. The place of articulation can be inferred from the spectra that occur during this constriction (Table 2.8). It has been shown that the location of the constriction is the most reliable predictor of the spectrum of voiceless fricatives, with a more posterior constriction resulting in a greater weighting of low frequency noise (Gordon et al., 2002). Fricative duration, on the other hand, was shown to be less useful for fricative identification.

Which resonances are introduced in the spectra of fricatives depends on the size of the cavity that is created in front of the constriction (K. N. Stevens, 2000). In the case of *+labiodental* fricatives, there is essentially no cavity, resulting in a lack of any major spectral peak. Typically, the amplitude of the noise is also much weaker than that of the following or preceding vowel. These characteristics are similar to those of *+dental* fricatives, although dental fricatives tend to be a bit shorter. The large overlap in spectra for *+dental* and *+labiodental* fricatives makes them difficult to distinguish, even for human listeners.

Typically, *+labiodental* and *+dental* fricatives are *-strident*, resulting in a weaker overall energy of the sound compared to *+alveolar* and *+palato-alveolar* consonants. In the case of *+alveolar* fricatives, the positioning of the tongue blade against the alveolar ridge introduces resonances through the front cavity, resulting in a spectral peak in F_4 - F_5 or even higher. Typically, this peak will exceed the peak of an adjacent vowel in this high frequency range. Similarly, *+palato-alveolar* consonants have peaks of greater amplitude in their spectrum, but these are typically in the F_3 - F_4 range. The *+velar* fricative consonants, which occur in some English dialects (e.g. *loch* in Scottish English) have even lower peaks (Gordon et al., 2002).

Although phonetically not a fricative, the voiced glottal ‘fricative’ /h/ (or ‘aspirate’) is included in Table 2.8. The constriction at the glottis results in a noisy turbulence excitation. Typically, /h/ has vowel-like resonances that are influenced by the following vowel (K. N. Stevens, 2000).

3. SPEECH ROBUSTNESS

As discussed in the previous section, there is a large variety of acoustic cues that can be used for speech recognition. Often, only a few of them are sufficient to correctly identify speech. Furthermore, high level language knowledge (such as syntax and semantics) can be utilised to correct mistakes made along the way. The inherent redundancy in speech helps us to overcome a range of distortions that regularly occur in the natural world. These include noisy environments, channel distortions and poor articulation. In this section, I provide a general overview of the remarkable robustness of human speech recognition.

3.1 Speech In Noise

Although current automatic speech recognition systems can accurately recognise speech in quiet as well as certain noisy conditions they were specifically trained in, the ability of humans to recognise speech in a large variety of noisy environments remains unchallenged. Accurate communication may take place when the speech level is higher than that of the noise, i.e. a positive speech-to-noise ratio (SNR) (Moore, 2013). However, speech can also be intelligible at negative SNRs. The precise degree to which noise is able to mask the target speech depends on a large range of factors, such as the spectral content as well as temporal properties. For example, intermittent noise may affect perception differently depending on the interruption rate (G. A. Miller & Licklider, 1950). Besides the relationship between the

acoustic components of the noise and the speech itself, there are also top-down processes that influence speech perception in noise. For example, Golestani et al. (2009) showed through testing pairs of words that were either semantically related or not that speech in noise perception is improved when the listener can make use of semantic information. They found that this effect was much stronger in native speakers than in non-native speakers.

3.2 Speech Distortions

The auditory system is also robust against distortions of the speech wave itself. For example, sound amplifiers may cause peak clipping, in which the peaks of the waveform are flattened or clipped. Although this acoustic manipulation results in an audible distortion, speech intelligibility remains high. Even under the most severe peak clipping, when the original complex speech wave is reshaped into a series of rectangular pulses, speech intelligibility can remain as high as 90% (J. C. R. Licklider & Pollack, 1948).

Many of the acoustic correlates mentioned in the previous section, such as formants and formant transitions, rely explicitly on frequency information. This might lead us to assume that there are particular frequency components that are crucial for speech recognition. However, it turns out that the information in speech is not confined to a very specific frequency range. For example, it has been shown that speech is equally intelligible (67% speech intelligibility score) when filtered with a high pass filter of 1800 Hz as when filtered with a low pass filter of the same frequency. Furthermore, it is possible to achieve high levels of intelligibility of speech that is confined to a narrow frequency band. For example, restricting the speech signal to a frequency band of 1 kHz to 2 kHz is sufficient for sentence intelligibility scores of around 90% (Moore, 2013). As discussed in Section 2.1, similar conclusions regarding the importance of temporal fine structure for speech in noise perception can be drawn. It has been shown that access to TFS information in both low and high frequency bands can independently improve speech recognition in noisy environment (K. Hopkins & Moore, 2009).

Although there is no critical frequency band for speech recognition, it is possible to identify specific amplitude and frequency modulations that are necessary for speech intelligibility. For example, Elliott and Theunissen (2009) showed that although most of the energy of speech lies in temporal modulations between 1-40 Hz, only temporal modulations from 1-7 Hz are critical for speech recognition. Similarly, they showed that spectral modulations < 1 cycles/kHz (which are changes in frequency that roughly correspond to vowel formant separation and vowel formant transitions in English) are critical for recognition of American English, whereas the higher spectral modulations are not as important.

3.3 Speech Versus Speech

Lastly, speech recognition often takes place in environments with competing speech. Often, the noise associated with a background of competing talkers is continuous and overlapping with the target speech, which results in a type of masking referred to as Energetic Masking (EM). However, speech recognition can also be made more difficult by interfering speech that does not overlap acoustically. This effect is referred to as Informational Masking (IM). IM is particularly strong when the interfering speech has similar characteristics as the target speech, for example if the interfering speaker speaks at a similar pitch.

It should be noted that to resolve interfering speech, spatial information can act as an important cue to separate sound sources. For example, speech intelligibility can improve when target speech and competing sounds are spatially separated (Dirks & Wilson, 1969), a phenomenon that is also referred to as spatial release from masking (SRM). This effect is commonly attributed to binaural processing (i.e. the processing of auditory input from both ears) although certain monaural cues (such as spectral notches) may also play a role (Schnupp et al., 2011). There are two commonly described ways in which binaural information may assist in sound localisation. Firstly, if a sound originates from a location that is not directly in front or behind the head, there may be differences in the time when a sound arrives at either ear (Interaural Time Difference, ITD). Furthermore, since sound strength dissipates as it travels or is blocked by the head (known as head-shadow), differences in sound level at either ear may occur (Interaural Level Differences, ILD).

4. AUDITORY PROCESSING

So far this chapter has mainly focused on aspects of speech signals. In this section I will discuss the auditory system itself. Specifically, I will review some neural mechanisms that are relevant for noise-robust speech processing in the early auditory pathway.

Many well-studied auditory features (such as pitch, loudness and timbre) are perceptual features. These features are not a distinct property of the incoming stimulus, but are perceptual attributes that result from the way in which the auditory system processes sounds. For example, although pitch relates to the F_0 of a sound, pitch can still be perceived when the F_0 is removed but other harmonics are present. A range of perceptual features have been shown to be robustly represented in the auditory cortex (Walker et al., 2011). To identify elementary components of speech processing, this section focuses on low-level mechanisms in the auditory brainstem. Figure 2.9 displays relevant components of the

early auditory pathway, each of which I will review in terms of mechanical properties as well as previous modelling approaches. Based on this review, general principles for noise robust feature extraction mechanisms in the early auditory pathway are identified.

4.1 Cochlea and Auditory Nerve Fibers

When a sound arrives at the ear it causes vibrations that are transmitted to the cochlea, a portion of the middle ear. Within the cochlea lies the basilar membrane, and as it vibrates to an incoming sound, inner hair cells on the basilar membrane provide electrical stimulation of Auditory Nerve fibres (AN fibres). A stronger vibration of a particular patch of the basilar membrane results in a higher firing rate of the corresponding AN fibers. Due to its mechanical properties, the basilar membrane effectively performs a frequency decomposition of the incoming sounds. As a result, many of the AN fibers are sensitive to tones of a preferred frequency. This decomposition is also referred to as a tonotopic arrangement and is maintained throughout the early auditory pathway. Additionally, the cochlea operates several mechanisms that help preserve the shape of the spectrum. For example, in addition to inner hair cells, there are also outer hair cells located on the basilar membrane. These outer hair cells can amplify basilar membrane motion, particularly at low input levels and at frequencies close to the characteristic of the basilar membrane where that hair cell is located (Rhode, 1971; Sellick et al., 1982). This contributes to a sharpening of the frequency response of the cochlea. Another property of the cochlea that increases the robustness of auditory processing lies in the way inner hair cells move. The temporal modulation of the wave causes the AN fibers to be stimulated in a phase-locked fashion, which means that they have a tendency to spike only at a particular phase of the incoming sound (Schnupp et al., 2011). Under moderate SNR conditions, the temporal structure of the signal causes peaks in the spike trains of the AN fibers that are more prominently visible than the average noise level (Greenberg & Ainsworth, 2004).

Various detailed mathematical and computational models of the cochlea and AN fibers have been proposed over time (see e.g. Meddis et al., 2013 and Zilany et al., 2014). However, these models typically have a high computational cost and depend on a large number of parameters. Simpler models of the cochlea have also been proposed. For example, Gammatone filter banks can provide a first-order approximation of the basilar membrane (Schnupp et al., 2011). By adding a half-wave rectification (which means that the negative part of the filterbank output is set to zero), this model can reflect the mechanical way in which neurotransmitters are released when the inner hair cells are bend in one particular direction.

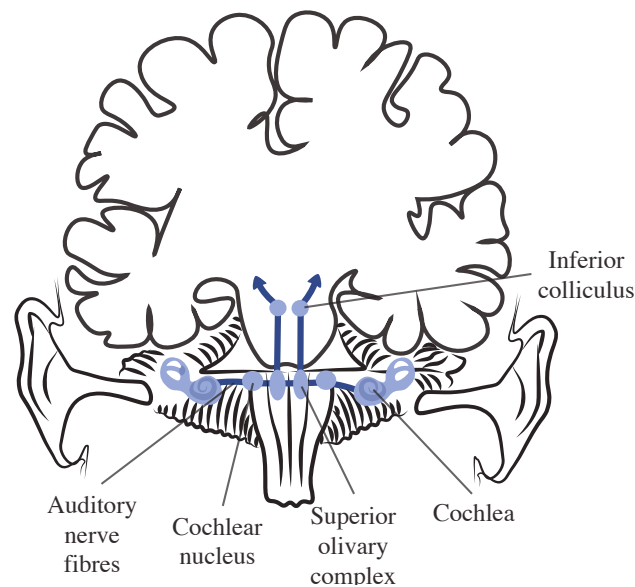


Figure 2.9. Components of the early mammalian auditory pathway. Sound is detected in the inner hair cell on the basilar membrane, which is located in the cochlea. The signal then travels through the auditory nerve fibers to the Cochlear Nucleus (CN), Superior Olivary Complex (SOC) and Inferior Colliculus (IC) until it travels further towards the thalamus and auditory cortex. Adapted from Breedlove et al. (2007).

4.2 Cochlear Nucleus

The AN fibers innervate onto the brainstem, in particular the Cochlear Nucleus (CN), which has several mechanisms in place that are involved in noise-robust auditory processing. Firstly, bushy cells in the Anterior Ventral Cochlear Nucleus (AVCN) receive strong input from multiple AN fibers, causing a sharpening of the phase-locked signal (Kandel et al., 2000). This promotes responsiveness to temporally coherent input, and reduces noise in the signal. Secondly, octopus cells in the Ventral Cochlear Nuclei (VCN) appear to be sensitive to the onset and offset of sounds. This is thought to be a result of a precisely timed inhibition and excitation that comes from stellate cells, which are responsive to a broadband frequency range rather than single centre frequencies. Determining the onsets of sounds is important for understanding speech in noisy, reverberant environments (Schnupp et al., 2011). Thirdly, the VCN is thought to project wideband inhibition to the Dorsal Cochlear Nucleus (DCN). This area of the brain is thought to analyse monaural localisation cues, such as spectral notches. Sound localisation helps in detecting a signal against a background of interfering noise (Blauert, 1997). Lastly, certain responses in the VCN are thought to be the neural correlates of the Comodulation Masking Release (CMR). CMR refers to the

paradoxical phenomenon where hearing generally deteriorates under an increase of noise bandwidth, but becomes easier when the noise is amplitude-modulated. This low-level mechanism has been explained through a simple model of narrowband excitation and wideband inhibition (Pressnitzer et al., 2001).

4.3 Superior Olivary Complex

The CN projects to the Superior Olivary Complex (SOC), which is one of the first areas of the auditory pathway that processes binaural input. The SOC plays an important role in sound localisation (Schnupp et al., 2011). More specifically, regions in the SOC assist in azimuthal sound localisation through the detection of interaural level differences (ILDs) and interaural time differences (ITDs). In particular, the Lateral Superior Olive (LSO) and Medial Superior Olive (MSO) are responsive to ILDs and ITDs, respectively. Inhibition plays an important role in both of these areas. For example, the LSO receives excitatory input from the ipsilateral side and inhibitory input from the contralateral ear. This allows encoding the location of sound based on level differences between the left and right ear (Schnupp et al., 2011). In the MSO, inhibition appears to be instrumental to the detection of interaural time differences, i.e. the time difference of a sound to arrive at the left or right ear. It has been shown that by blocking inhibitory synapses pharmacologically, the ITD tuning of neurons in the MSO changes (Brand et al., 2002). Furthermore, the relative timing of inhibition compared to the timing of excitatory inputs also plays a role in the tuning of MSO neurons (Pecka et al., 2008).

4.4 Inferior Colliculus

The Inferior Colliculus (IC) receives input from the SOC and other auditory brain areas and is thought to integrate the outputs of the preceding processing stages, such as binaural sound localisation and the pitch of a sound (Krishna & Semple, 2000). From here, the signal leaves the brainstem up to the medial geniculate body in the thalamus, which projects to the auditory cortex where further processing takes place. Similar to AN fibers and CN fibers, neurons in the IC have a characteristic frequency. In addition, some IC neurons are also tuned to a particular AM frequency (see e.g. Snyder et al., 2000) and FM sweeps (see e.g. Casseday and Covey, 1992). AM refers to the fluctuation in time of the amplitude of the signal, whereas FM refers to a fluctuation of frequencies in time. These modulations are thought to be important for speech intelligibility. For example, Elliott and Theunissen (2009) have shown that there are critical bands for both AM and FM and removing fluctuations

in these bands drastically reduces speech intelligibility. Moreover, research suggests that AM sensitivity might allow for a noise-robust neural encoding of speech, in particular for vowels (Carney et al., 2015). Additionally, neurons in the IC have been shown to respond to frequency-following responses (FFRs), which have been suggested to be a low-level neural correlate of pitch perception (Bidelman, 2013). Interestingly, many computational models of AM, FM and FFRs sensitivity share common properties. For example, it has been suggested that AM sensitivity could be a result of precisely timed inhibition (C. Kim & Stern, 2010), although other phenomena such as adaptation could also play a role (Fontaine et al., 2014). Similarly, various models have been proposed to explain FM sweep sensitivity. Such models make use of for example sideband inhibition or coincidence detection (Skorheim et al., 2014). Pitch perception has also been described through models that make use of delayed inhibition, such as the cancellation model of pitch (de Cheveigné, 1998).

5. CHAPTER SUMMARY

In this chapter, I reviewed the properties of basic auditory features through the lens of acoustic phonetics, psychometrics and auditory neuroscience.

I first provided an overview of distinctive feature theory, which is a framework that can be used to describe the way speech is produced. This framework categorises single phones based on a limited set of features: source features, the manner in which resonances are introduced by the vocal tract, and the place where the main constriction takes place. Each of these features has characteristic effects on the produced speech wave. These are referred to as the acoustic correlates or acoustic features. Four types of acoustic features have been discussed: place-frequency features, periodic features, envelope features and temporal fine structure. Most of the work in acoustic phonetics focuses on place-frequency features, such as formants and formant transients.

Although place-frequency features have been studied thoroughly, they are unlikely to be the only features the auditory system uses to make sense of speech as humans can recognise speech under a range of distortions. Some of these, such as bandpass filters or additive noise, distort the distinctive presence of formants in the spectrum. Furthermore, a large range of amplitude and frequency modulations may be removed without significantly affecting the intelligibility of speech. Nevertheless, it is important to remember that high levels of speech intelligibility under distortions may also reflect the way in which high-level knowledge about language itself is used to resolve ambiguities or uncertainty.

In the last section, a general overview of neural mechanisms in the early auditory

brainstem was presented. From this review, at least three general principles can be observed that are prevalent for noise-robust processing in the early auditory pathway. Firstly, there is a frequency decomposition and tonotopic arrangement of neurons that starts in the cochlea, which is preserved throughout the early auditory pathway and should be an integrated part of any auditory model. Secondly, auditory neurons tend to respond in a phase-locked fashion, which promotes sensitivity to temporally coherent sounds. Lastly, precisely timed inhibition and excitation can be found on many levels of processing in the auditory system. This mechanism allows for sensitivity to a range of auditory features that may be relevant for speech recognition, such as onsets, sound location, AMs and FM.

CHAPTER 3

Explainable Automatic Speech Recognition

The previous chapter was centered around auditory features, which are the focus of study in this thesis. In this chapter, I will further expand on machine learning, which has been the tool of choice in my work. Machine learning can be a powerful tool for progressing science. In this thesis, it is used in two distinct ways. In Part II machine learning techniques are applied to traditional computational neuroscience to obtain new scientific insights about neural mechanisms involved in speech recognition. In Part Part III. I investigate to what extent modern Automatic Speech Recognition (ASR) systems can be used as a model for human hearing. Furthermore, I apply explainable machine learning techniques to expose the underlying mechanisms used by these systems.

This chapter is structured as follows. I will first provide a succinct overview of the history of ASR (Section 1), which is characterized by a shift from knowledge-based template matching to black box deep learning models that require minimal feature engineering. In Section 2, I discuss the type of features that have generally been considered the standard in ASR, as well as techniques that have been applied to improve noise resistance in these features. I then briefly cover the basics of neural networks, with a specific focus on the ways in which the temporal components of speech are modelled in ASR systems (Section 3). The last section covers an overview of explainable deep learning (Section 4), in which I will list a range of methods that have been proposed over the past couple of years. This overview particularly focuses on how the incorporation of domain knowledge can improve the explanatory power of these methods.

1. A BRIEF HISTORY OF ASR

Like most early machine learning applications, the first ASR systems relied on hand-crafted input features. In ASR, many of these features were derived from acoustic phonetics, a field which aims to explain how the phonetic elements of speech (the most basic sounds of the language) are acoustically realised (Juang & Rabiner, 2005). The first ASR system is often credited to Bell Labs. In 1952, a group of engineers introduced Audrey, a system that performed isolated digit recognition for a single speaker by estimating the formant frequencies of each digit (Pieraccini & Lubensky, 2005). Formants, which are local peaks in the spectrum of a speech sound, had been studied in phonetics since the early 1900s, after the term was coined by Hermann in 1894. In the Audrey system, each digit had its own designated electrical circuit that matched the extracted formants to predefined templates.

In the late 1960s, the foundations of what would become Linear Predictive Coding (LPC) were introduced (Juang & Rabiner, 2005). LPC is based on the source-filter model of speech. It can be used to estimate responses of the vocal tract as well as extract certain speech parameters (such as formants) in an automatic way. Around the same time, the fast Fourier transform (FFT) was introduced by Cooley and Tukey (1965) and techniques such as cepstral analysis and Dynamic Time Warping were first developed.

In the 70s, one of the most successful systems was Harpy (Lowerre, 1990), developed at Carnegie Mellon University. The Harpy system was able to recognise about a 1000 words by 1976. It worked by extracting relevant parameters from the segmented input sound, which were matched to phone templates. It then searched through a finite-state graph structure that represented all legal sentences according to a given set of syntactical producten rules and word boundary rules.

The template-based approach shifted towards more rigorous statistical modelling in the 1980s. A popular method at the time was the use of Hidden Markov Models (HMMs) (Pieraccini & Lubensky, 2005). These models allowed researchers to model both the intrinsic variability of the speech signal in an acoustic model, as well as the structure of the language itself in a language model. Given an input sentence, the Viterbi algorithm could be applied to a trained HMM to find the most likely sequence of letters. Around the same time, Artificial Neural Networks (ANNs), which were first introduced in the 50s, regained popularity. Most applications of ANNs in ASR were combined with HMMs in hybrid models, as these ANNs did not cope well with temporal information (Gold et al., 2011).

The late 2000s were marked by an increased interest in ANNs for speech recognition through the development of deep learning (Deng, 2016). New improvements allowed

deeper and therefore more powerful neural networks to be trained efficiently, leading to the rebranding of ANNs as Deep Neural Networks (DNNs). Hybrid models of DNNs and HMMs became the new state of the art. These systems use relatively simple features such as spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) (see Section 2.1) as their input. The advance of Long Short-Term Memory (LSTM), a specific type of Recurrent Neural Network (RNN) that can keep track of long-term dependencies, removed the need for hybrid models. For example, great improvements in performance on clean speech recognition tasks were achieved by Graves et al. (2013), where MFCCs were used in an end-to-end LSTM. The popularity of LSTMs continues to this day, although other DNNs models such as attention networks have also gained traction. There has also been an increased interest in end-to-end systems, which use the raw audio as input to the network.

The history of ASR displays a clear shift from feature-based methods, in which domain knowledge was applied to handcraft ASR systems, towards models that require little to no feature engineering. This is desirable from an application perspective, as it removes the necessity of strong domain knowledge to design and develop ASR systems. However, as will be discussed in more detail in Section 4, the shift comes with the loss of transparency and interpretability.

2. ASR FEATURES AND NOISE ROBUSTNESS

One of the aims of the work presented in this thesis is to provide insights into auditory features that play a role in noise robust speech recognition by ASR systems. In this section, I will review the input features as well as some of the methods employed in ASR systems to improve noise-robustness.

2.1 Frequency Decompositions for Speech Recognition

Synonymous with the cochlea in the auditory system, the first step in processing of many ASR systems involves a time-frequency decomposition of the incoming sound. There are many ways to achieve such a decomposition, and I will introduce three widely used methods: the Short Time Fourier Transform (STFT), Mel-frequency spectrograms and Gammatone filterbank spectrograms. I will also discuss more recently introduced end-to-end approaches to extracting features from the raw waveform.

It should be noted that many ASR applications transform the time-frequency spectrum into what is called a cepstral representation, which is computed as the inverse Fourier

transform over the log of a spectral representation. The cepstral representations of the Mel-frequency spectrogram (see below), MFCCs, are particularly common. Cepstral representations can be useful for speech applications because (in contrast to the spectral domain) pitch and formants are additive and therefore more easily separated in the cepstral domain (Schnupp et al., 2011).

The Short Time Fourier Transform (STFT) is a popular method to obtain the time-frequency decomposition of a sound. It can be used to obtain a time-varying frequency decomposition, as it locally computes a Fourier transform over a signal (Benesty et al., 2009). The STFT is widely used in fields outside of speech recognition and includes no assumptions about the human auditory system. The STFT differs in major ways from the way the cochlea performs a frequency decomposition. For example, the frequencies in which the Fourier transform decomposes a sound are multiples of a particular fundamental frequency. In contrast, resonance frequencies in the basilar membrane are spaced logarithmically (Bregman, 1990).

The Mel-frequency spectrogram incorporates auditory knowledge into the STFT. To obtain a Mel-frequency spectrogram, the frequency axis of the amplitude spectrum of the STFT is transformed to the Mel-scale. The Mel-scale refers to a perceptual scale that spaces frequencies based on the pitches that human listeners judge to be equal in distance from one another (S. S. Stevens et al., 1937). This reflects the observation in psychoacoustics that pitches up to about 1 kHz are perceived linearly, but higher frequencies appear to be perceived on a logarithmic scale. The cepstral representation of the Mel-frequency spectrogram, the Mel-Frequency Cepstral Coefficient (MFCC), is one of the most popular auditory features in ASR system (Li et al., 2014). MFCCs are derived by first computing the Discrete Cosine Transform (DCT) of the log of the powers of the Mel-frequency spectrogram. The amplitudes of the resulting spectrogram are the MFCCs.

The Gammatone filterbank spectrogram is a different way of obtaining a frequency decomposition that does not use a Fourier transform. It uses the idea that the cochlea essentially functions as a set of mechanical filter elements that each have their own resonance frequency. Thus, instead of a Fourier transform, asymmetric Gammatone filters are applied to decompose the sound (Patterson et al., 1987). These filters are shaped as a sinusoid windowed with a gamma function. Applying the filter results in a reasonable first-order approximation of the cochlea. One disadvantage, however, is that the filters do not model the level-dependent characteristics of the auditory

filters in the cochlea (T. Irino, 1997). This can be accounted for by compressing the filter outputs, for example with a cubed root compression (Hermansky, 1990). Although Gammatone filterbanks are more computationally expensive than STFTs and Mel-frequency spectrograms, there is some evidence that they may result in better ASR performance (see e.g. Gibson et al. (2014), Adiga et al. (2013) and Zhao and Wang (2013) for comparisons).

Learnable frontends are a more recently introduced way of extracting features from acoustic data. The rise of deep learning methods has allowed learning representations from raw audio instead of engineered features. In these learned frontends, standard MFCC features are replaced with learnable neural layers, such as convolutional layers (Palaz et al., 2015; Schneider et al., 2019) or more structured filters such as Gammatone (Sainath et al., 2015) and Gabor (Noé et al., 2020; Zeghidour et al., 2021; Zeghidour et al., 2018) filters. Learnable frontends are more flexible than MFCCs as they can extract features that are most relevant for the task at hand. A disadvantage of learnable frontends compared to more classical ASR input features is it that they require training procedures that are time and resource intensive.

2.2 Feature Enhancement Strategies in ASR Systems

Neither Gammatone filter cepstra nor MFCCs are particularly robust to noise (Li et al., 2014). Therefore, ASR systems have traditionally used feature enhancement and model adaptation techniques to improve noise-robustness. Feature enhancement techniques are applied directly to the features themselves, prior to recognition. Model adaptation methods, on the other hand, update ASR model parameters to include the effects of noise (Seltzer et al., 2013). Model adaptation methods rely heavily on the type of model used for speech recognition (e.g. HMMs or DNNs). Here, I focus on generic feature enhancement techniques instead. In particular, this review comprises three types of feature enhancing methods: noise-resistant features, feature normalisation and feature compensation (Li et al., 2014).

Noise-Resistant Feature Methods

Noise-resistant feature methods try to compensate for the effect noise has on the signal. Typically, these methods do not explicitly estimate noise statistics. Noise-resistant feature methods can be further divided into methods that use some form of auditory knowledge, and methods that do not.

An example of auditory-based feature enhancement methods is RASTA-PLP, which

filters the signal to remove variations that are uncharacteristic for speech (Hermansky et al., 1992). It works under the assumption that certain noise types vary slower than the speech signal itself. This method has historically been successful at increasing performance in speech recognition tasks (Hanson & Applebaum, 1993). Other examples include zero crossing peak amplitude (ZCPA) (D.-S. Kim et al., 1999), power-normalized cepstral coefficients (PNCC) (C. Kim & Stern, 2010) and Gabor filter bank features (Moritz et al., 2013). However, because there is no universally-accepted theory about which auditory information is most important for noise-resistant speech recognition, it is difficult to determine which of these methods is best from a theoretical perspective (Li et al., 2014).

When trained accordingly, features that are based on ANNs are often inherently noise robust. Although these methods do not explicitly include noise statistics, ANNs typically generalise well and are relatively noise resistant. An early example is the TANDEM system (Hermansky et al., 2000), in which the output layer of an ANN feeds into a Gaussian Mixture Model (GMM). Another class of ANN-based features are bottleneck features (Grézl et al., 2007). Bottleneck features are generated by designing a multi-layer ANN that has a narrow middle layer (the bottleneck), which effectively forces the neural network to perform a form of dimensionality reduction. More recently, deep learning techniques have shown to be promising. For example, the CD-DNN-HMM system (Seltzer et al., 2013) used the output of a DNN as input to an HMM, achieving good performance. Deep RNNs have also been successfully applied as input for noise-robust ASR systems (Weninger et al., 2015). Furthermore, learnable frontends that can be pre-trained in an unsupervised fashion have been shown to be particularly robust against noise and distortions (Ma et al., 2020). The training procedure of DNNs can also be optimised for noise robustness. For example, Park et al. (2019) introduced SpecAugment, a simple data augmentation method that has been shown to improve noise robustness of DNN-based ASR systems significantly, particularly for smaller datasets.

Feature Moment Normalisation

In contrast to the noise-resistant feature methods, feature moment normalisation methods make an explicit estimation of the noise statistics. Noise is reduced by normalising the statistical moments of speech to reduce mismatches between training and test data. For example, channel distortions can be reduced by applying cepstral mean normalisation, where the mean of a sequence of cepstral vectors is subtracted from each vector (Atal, 1974). This can be approximated in real time using a high-pass filter with a cut-off frequency

arbitrarily close to 0 Hz (Huang et al., 2001). Higher statistical moments can also be applied to normalise the features. For example, cepstral mean and variance normalisation normalises both the mean and covariance of the cepstrum vectors (Viikki et al., 1998). Histogram equalisation methods (Molau et al., 2003) take this idea even further by normalising the entire distribution of the training and test data.

Feature Compensation

In feature compensation, the noise in the sound signal is explicitly removed from the observed speech features. For example, spectral subtraction estimates the noise spectrum during a period where no speech is present and then suppresses the incoming noise in the speech signal. This algorithm assumes that the noise and clean speech signal are additive and uncorrelated in the time domain. However, estimating the noise spectrum is not a trivial task. It relies heavily on voice activity detectors, which do not work well under low SNRs (Li et al., 2014). Another approach to feature compensation is Wiener filtering (Lim & Oppenheim, 1979), in which the clean speech signal is approximated by applying a linear filter that has been determined using a minimum mean squared error approach. Wiener filtering is an important component of the Advanced Front-End for Distributed Speech Recognition, a standardised front-end which has historically been used as a baseline comparison for noise-robust ASR systems (ES ETSI, 2002).

It is important to note that although feature enhancement techniques sometimes give way to a better performance in ASR systems, this is not always the case. For example, Seltzer et al. (2013) show that the performance of a DNN that employs feature enhancement methods sometimes decreases. They suggest this happens because the system observes fewer variations in the training data if feature enhancement techniques are being employed, making it less robust to mismatched conditions. Rather than removing noise from the input, most modern machine learning approaches aim to improve robustness by including data augmentation techniques (such as SpecAugment or additive noise for ASR applications) to increase the size and variability of the training set.

3. AN INTRODUCTION TO DEEP AUTOMATIC SPEECH RECOGNITION

In this section I provide a brief introduction to deep learning methods for ASR. I will start the review with a discussion of the more traditional ASR pipeline, which involves the use

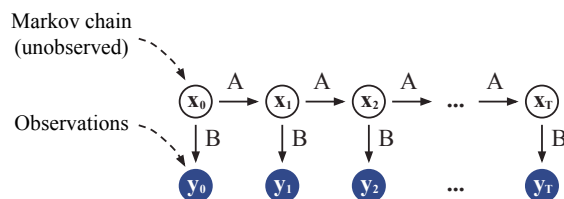


Figure 3.1. A visualisation of an HMM.

of HMMs. I then focus on the ways in which this pipeline has been altered to work well with deep learning methods. In particular, I will cover the architectures that have been developed to deal with the sequential of speech, such as RNNs.

3.1 The Traditional ASR Pipeline

The traditional ASR pipeline consists of three components: an acoustic model, a language model and a decoder (Gales & Young, 2008). The acoustic model maps speech waves to phoneme sequences, while the language model models the likelihood of word sequences. A decoder is used to combine the acoustic and language model into a final transcription.

As discussed in Section 1, HMMs were imperative in the shift from template-based speech recognition to statistical modelling. These generative models can, amongst many other applications, be used for acoustic modelling. HMMs model a process as a Markov model (Figure 3.1), in which each ‘state’ only depends on the previous state (Baum & Petrie, 1966). In the context of speech recognition, the HMM maps the relationship between speech and the most likely sequence of phonemes. Phonemes are the perceptually distinct speech in a given language that distinguish one word from another. The hidden states are the phonemes that are assumed to generate the observed sequence of acoustic features (e.g. MFCCs, more generally referred to as frames). The HMM incorporates three probability distributions: the initial state probability π , the transition probabilities A for moving from one state x_t to another and the emission probabilities b , which denote the likelihood B of the observation y_t given a hidden state x_t . The initial state probability π indicates how common certain phonemes are, while the transition probabilities A convey information about language, namely which phonemes are likely to follow one another within the language. The emission probabilities b reflect how the phonemes are acoustically realised. When HMMs are combined with DNNs, the DNN is usually trained to predict the likelihood of a certain phone given a fixed-length sequence of frames.

Another component of the traditional ASR pipeline is the language model. This is a

model that is used to ensure that the predicted output is linguistically meaningful. More specifically, a language model gives the probability of the next word (or character) in a sequence of words (or characters). One of the simplest language models are N-gram models, which predicts the next item in a sequence of $(n - 1)$ words. The output probabilities are based on the number of times the n th word occurred in a text corpus, compared to other words associated with the $(n - 1)$ -length sequence. Of course, for n larger than 2 or 3, a sequence may have never occurred in the corpus. This can be somewhat compensated for through the usage of smoothing techniques. These techniques reserve some probability mass for unseen n-grams. More recently, DNNs have also been used for language modelling (see Section 3.2).

Together, the HMM and the language model produce a large search graph of potential transcriptions for a given speech sequence: the language model produces probabilities for word sequences, while those words themselves are governed by the underlying phoneme sequences. It is the job of a decoder to come up with the most likely transcription. As the graph is too large for an exact search, methods such as greedy or beam search decoding have been developed to perform an approximate search.

3.2 Deep Automatic Speech Recognition

One disadvantage of the traditional setup is that the model is trained to predict phonemes rather than letters. This means that the training data needs to be transcribed into phonemes by experts, which is a costly affair and may be a problem for less common languages. End-to-end neural ASR systems are a more recently introduced approach to circumvent the need for phoneme transcriptions. In this section, I will first shortly introduce ANNs. I will then discuss the Connectionist Temporal Classification (CTC) criterion, a particular loss function that can be used to train sequential neural networks (such as RNNs) to tackle sequence problems without the need for phoneme alignment. This is followed by an introduction to neural sequence-to-sequence modelling.

Artificial Neural Networks

In their most basic form, ANNs consist of a collection of units called perceptrons or artificial neurons (Deng, 2016). These units are connected to each other in a directed weighted graph. Typically, neurons are organised in layers that are connected in a feedforward fashion, in which neurons of one layer connect to neurons of the next layer. The first layer is referred to as the input layer, the last layer as the output layer, and all intermediate layers are called

hidden layers. Similar to real biological neurons, artificial neurons receive inputs from other neurons, and combine the incoming signals into an output activation. The activation of a single neuron is determined by its propagation function:

$$y = f(w \cdot x + b) \quad (3.1)$$

Here, x is the input vector from the previous layer, w represents the weight vector, b is called the bias, and $f(x)$ is referred to as the activation function, a non-linear function that can take various forms. Common examples of activation functions are Rectified Linear Units (ReLUs; $\max(0, x)$), sigmoid function ($\frac{1}{1+e^{-x}}$), and $\tanh(\frac{e^x - e^{-x}}{e^x + e^{-x}})$. The output of a single neuron is thus the weighted sum of its input with an added non-linearity. The neural network as a whole is then a chain of non-linearities that together are capable of performing complex functions.

For a network to produce the desired output for a given input, the network needs to be trained. This is done by first defining a loss function which incorporates how far the model prediction is from the desired prediction. A common example of a loss function is the mean squared error:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (3.2)$$

Here, y_i is a predicted outcome and $\hat{y}_i$ the corresponding true label out of a data set of n samples. Often, a regularisation term is added to the loss function to avoid overfitting. Overfitting occurs when a network fits the training data too precisely and fails to do well on unseen data.

Given the loss function, a learning algorithm is applied to optimise the weights such that the loss function is as small as possible. This is usually done through an optimization strategy called gradient descent, an iterative algorithm that can find a local minima of a differentiable function. By computing the gradient of the function (i.e. the partial derivative of the weight with respect to the loss function) and taking a step in the negative direction proportional to that gradient, the function is minimised. For neural networks, the gradients can be computed efficiently through an algorithm called backpropagation (Rumelhart et al., 1995). At each iteration of the algorithm, the weights are updated proportionally to their gradients. The proportion is determined by the learning rate, a hyperparameter that controls the speed of learning. Certain optimisation algorithms, such as AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014), dynamically change the learning rate

for more efficient and effective learning. Usually, the gradient is not computed over the whole training set but a smaller batch of training samples, which is referred to as Stochastic Gradient Descent (SGD) (Saad, 1998).

Alignment-Free Loss Functions

Traditionally, sequential neural networks such as RNNs required pre-segmented training data. This sequence labeling was usually done using HMMs. The CTC criterion introduces a sequence level objective function that enables neural networks to be trained on a sequence of speech signals without the requirements of prior alignments (Graves et al., 2006). By removing the need for aligning speech with phonemes, the CTC criterion can be used to train neural networks to directly predict a character sequence from the speech input, thus removing the need for phoneme transcriptions and even a language model.

In the CTC setup, the neural network predicts a probability distributions over all possible characters (including a ‘blank’ output) at each time step. This can be achieved by having the same number of output neurons as characters and applying a softmax layer to obtain a probability distribution. Rather than attempting to align each character with a specific speech segment, labeling sequences are considered equivalent if they only differ in alignment. In this process, the blank outputs are ignored. The CTC objective is to maximize the sum of the probabilities of all possible alignments. This can be computed efficiently using a dynamic programming approach. At inference time, the most likely output can be extracted by simply taking the most probable character at each time step. Typically, however, the character probabilities are fed into a language model to produce a transcription with fewer errors.

The CTC criterion is not the only method used to train neural networks on sequence to sequence tasks without prior alignment. Other methods include RNN Transducer (RNN-T) (Graves, 2012) and the Lattice-Free Maximum Mutual Information (LF-MMI) (Povey et al., 2016).

Sequence to Sequence Learning

An important milestone for deep sequence modelling was the introduction of Sequence-to-Sequence (Seq2Seq) models (Cho et al., 2014; Sutskever et al., 2014), which were originally designed for machine translation tasks. A Seq2Seq model typically consists of an encoder and a decoder network. In the original design, the encoder processes the input sequence into a single vector, called the context vector. The context vector is of a fixed size that does

not depend on the length of the sequence. After processing the entire input sequence, the context is sent to the decoder, which produces the output sequence. If the encoder and decoder are both RNNs, the context vector simply reflects the number of hidden units in the encoder RNN.

Since the encoder outputs only a single context vector for the whole sequence, original Seq2Seq models had difficulty dealing with long sequences. To account for this problem, Bahdanau et al. (2015) and Luong et al. (2015) respectively introduced and refined a technique referred to as ‘attention’. This method allows the model to focus on relevant parts of the input sequence. Rather than forwarding the encoder context vector at the end of processing to all decoder time steps, the attention framework passes all encoder context vectors to the decoder. At every decoding time step, the decoder computes attention scores for each context vector. It then weighs all context vectors by this score. The scores or ‘attention weights’ thus quantify which encoder context vectors the decoder is ‘attending to’ at a given time step. There are various flavours of scoring functions such as the cosine (Graves et al., 2014) or dot product (Luong et al., 2015). Since the scoring function is applied on all inputs, the attention mechanism allows a model to draw from a state at any preceding point along the sequence, even those input that are far away.

The first Seq2Seq models were implemented using RNNs and later Convolutional Neural Networks (CNNs). In their already seminal 2017 paper, Vaswani et al. showed that attention mechanisms are powerful in themselves. They introduced Transformers, which implement attention mechanisms without using RNNs at all. Rather than processing the input sequentially, all inputs are processed at the same time and attention weights are computed between them. As all inputs are processed independently (rather than sequentially), Transformers are particularly well-suited to parallelisation.

3.3 Sequential Neural Architectures

As discussed in the previous section, various techniques have been introduced to allow DNNs to process sequential data. In this section, I will describe three such architectures. To keep this review brief, I will skip over some of the implementation details of these models. To get the full story, please see the introductory texts from Schmidhuber and Hochreiter (1997), Waibel et al. (1989) and Vaswani et al. (2017) for LSTMs, TDNNs and Transformers, respectively.

Recurrent Neural Networks and LSTMs

Traditional neural networks are feedforward, which means neurons in one layer of the network are only connected to neurons in succeeding layers. RNNs also allow for temporal connections. In its most basic form, an RNN unit will have a connection to itself. This means that the neural activation (or ‘hidden state’) at timestep t is one of the inputs to that same neuron at time step $t + 1$. When training RNNs, an adapted version of backpropagation is used called Backpropagation Through Time (BPTT) (Mozier, 1989; Robinson & Fallside, 1987; Werbos, 1988). BPTT works by unrolling all input timesteps. Spatially, each timestep of the unrolled neural network can be viewed as an additional layer that consists of a copy of the network at the previous timestep. The weights are updated after the entire input sequence has been presented to the model. As the number of timesteps increases, BPTT becomes more computationally expensive. Truncated BPTT is a modification of standard BPTT in which the network is updated after a fixed number of time steps, which reduces the computational cost significantly.

Basic RNN units can only keep track of recent events, as their internal state is easily overwritten by new inputs. To allow for long-term dependencies, Schmidhuber and Hochreiter (1997) introduced the LSTM unit. Networks that use these units are often referred to as LSTMs, and are a commonly used model for speech recognition tasks.

LSTMs have a more complicated architecture than standard RNNs (Figure 3.2). The LSTM has an internal cell state c_t . In contrast to the hidden state of standard RNNs, this internal cell state is not automatically passed onto the next neuron. Instead, an LSTM unit contains three gates that control how the information passes through the unit: the forget gate, input gate and output gate. Each gate functions as a standard artificial neuron (as per Equation 3.1) as it weights a given input by its own weight vector w and bias b and then adds a non-linearity. For LSTM units, all gates are governed by a sigmoid non-linearity, which squashes their output between 0 and 1. The input to each gate consists of both the input at the current time step x_t , as well as the output of the LSTM unit at the previous time step h_{t-1} .

The three gates interact with the cell state in the following way to produce the output h_t of the LSTM unit. First, a candidate cell state $\hat{c}_t$ is computed:

$$\hat{c}_t = \tanh(w \cdot [h_{t-1}, x_t] + b) \quad (3.3)$$

This candidate cell state can take values between -1 and 1 , as dictated by the usage of $\tanh(x)$, using a weight w and bias b specific to that cell. The actual new cell state c_t is

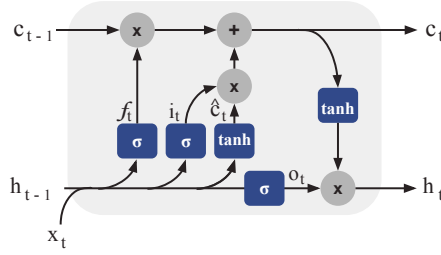


Figure 3.2. A visualisation of the internal workings of the LSTM cell.

then determined by weighting the previous cell state c_{t-1} with forget gate value f_t as well as by weighting the candidate cell state $\hat{c}_t$ with the input gate value i_t :

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \hat{c}_t \quad (3.4)$$

The forget gate f_t thus controls how information from the current input x_t and the output from the previous timestep h_{t-1} flow in the current cell, while the input gate i_t decides how much of the new information should be added to the cell state. The final output of the LSTM unit h_t is a value between -1 and 1, computed as the cell state c_t weighted by the output gate o_t :

$$h_t = o_t \cdot \tanh(c_t) \quad (3.5)$$

The output gate thus determines which proportion of the cell state c_t will become the LSTM output h_t .

The LSTM has long been the go-to RNN due to its ability to deal with longer temporal sequences. More recently, the Gated Recurrent Unit (GRU) was introduced, a simpler version of the LSTM architecture without an output gate (Cho et al., 2014).

Convolution Over Time

Another way in which temporal information can be modelled is through Time Delay Neural Networks (TDNNs; Figure 3.3; Waibel et al., 1989). These are feedforward networks designed to classify with shift-invariance, which means the classifier does not require explicit segmentation. Each neuron in a TDNN network receives input from the activation patterns over time from the neurons in the previous layer. Shift-invariance is achieved through weight sharing over time. More specifically, time-shifted copies of the network across time are created during training. Rather than updating the weights for each of these

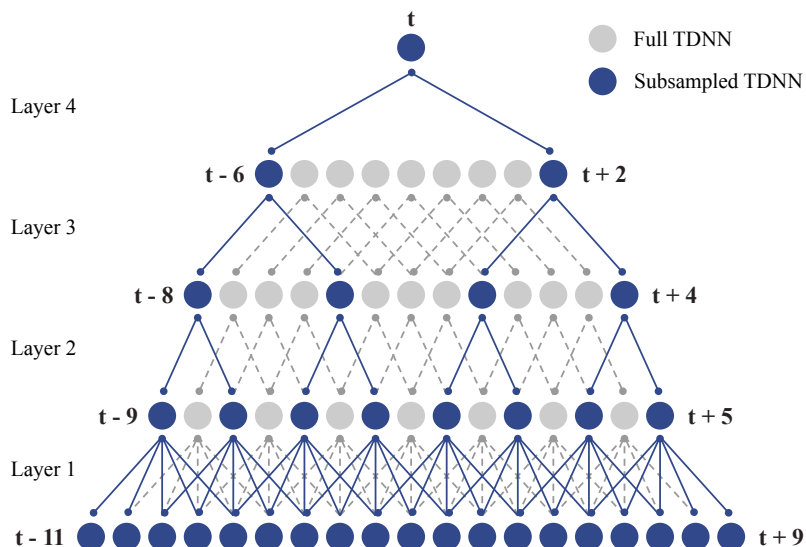


Figure 3.3. A visualisation of a subsampled TDNN. The grey connections exist in the standard TDNN, but the subsampled TDNN (which is more efficient) only uses the dark coloured weights. Note that regardless of these weights, all time steps in the input layer have some path to the final node in the last layer. This specific TDNN is asymmetric, as it scans further into past timesteps (note the $t - 11$) than in future timesteps ($t + 9$). Note that this representation only shows the output for a single time step: the same network (with the same weights) is applied to every time step, similar to how a 1D convolutional neural network has kernels that are repeated across one dimension.

copies separately, the error gradients of the shifted copies are averaged.

To those acquainted with deep learning approaches in computer vision, the concept of weight sharing will sound familiar. Indeed, TDNNs are closely related to CNNs, as they can be viewed as one-dimensional CNNs that convolve over time. A convolutional layer can be visualised as a set of filters, called convolutional kernels, that slide over the input (usually an image). Each filter is multiplied with the input and summed at each step. During training, the values of the kernels are trained. The amount of movement between applications of the filter to the input is referred to as the stride. TDNNs typically have a stride of 1. Dilated convolutions, in which the filter skips input values with a certain step size, can be used to allow the network to operate on a coarser scale and can make the network more efficient. TDNNs can be designed to require much fewer computations while still using information from each time step using a subsampling approach that is closely related to dilation (Peddinti et al., 2015).

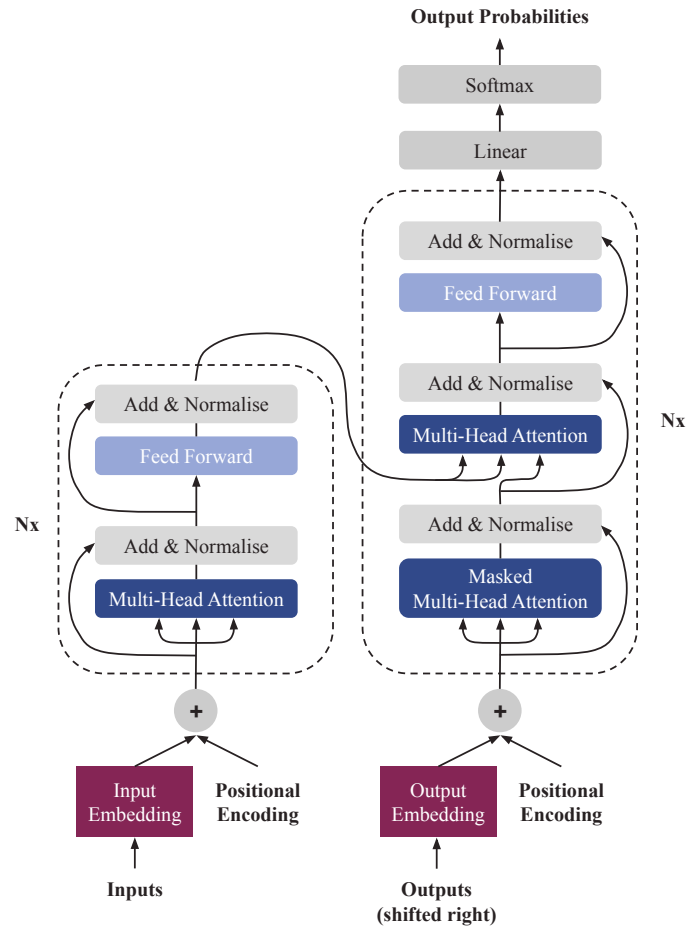


Figure 3.4. A visualisation of the Transformer architecture, which consists of an encoder (left) and decoder (right), each of which use multi-headed attention to process the sequential input.

Transformers

As discussed in Section 3.2, Transformers are a recently introduced architecture for dealing with sequence-to-sequence tasks (Figure Figure 3.4; Vaswani et al., 2017). Like its Seq2Seq predecessors, the Transformer architecture has a encoder and decoder part. A typical Transformer architecture consists of multiple stacked encoders and decoders.

Each encoder consists of a self-attention mechanism and a feed-forward neural network. The first encoder takes positional information and embeddings of the input sequence as its input, while subsequent encoders takes as input the encodings from previous encoders and weights their relevance. This is referred to as self-attention, since it relates the different positions of a single sequence to compute a representation of that same sequence, rather

than a separate output sequence. The feed-forward neural network further processes the output encodings, which are then passed to the next encoder as well as the decoder.

Like the encoder, the decoder also consists of a self-attention mechanism and a feed-forward neural network. In addition, it has an attention mechanism. This attention mechanism draws relevant information from the encodings generated by the encoders. To prevent the decoder from using current or future outputs to predict an output, the output sequence is partially mask to prevent reverse information flow.

In any sequence-to-sequence task, there are many ways in which the two sequences may be related. For example, in the context of a language task adjectives are related to the nouns that they describe, and a verb may be related to its subject and/or object. To allow for the representation of multiple of such relationships, each layer in a Transformer has multiple ‘attention heads’. Each attention head has its own set of weight matrices and can be processed in parallel with other attention heads. This architecture is referred to as multi-headed attention.

4. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Although current ASR methods based on DNNs are powerful, predictive power is only one of the desirable aspects of a machine learning system. The existence of powerful ASR systems does not necessarily expand our understanding of speech recognition, as models with the same classification performance may use completely different internal feature representations for their predictions (Arras et al., 2016, 2017; Lapuschkin et al., 2016). However, it is difficult to assess these internal representations for black box models such as DNNs, as they often lack transparency and interpretability. The scientific interest in machine learning is not constrained to its ability to correctly predict single quantities. The value of machine learning in science also lies in its potential to improve scientific understanding through the inference of causal relationships from observational data, the identification of hidden laws of nature (Schütt et al., 2017; Sturm et al., 2016) and the role in obtaining new scientific insights (Roscher et al., 2020).

In this section, I will review recent findings in Explainable Artificial Intelligence (XAI), a field that is concerned with demystifying the inner workings of machine learning methods. Some of the discussed XAI methods will be applied in future chapters, such as the application of attribution methods in Chapter 7. I will also explicitly discuss the incorporation of domain knowledge in XAI. This is particularly useful when considering machine learning applications in science, as domain knowledge can be used to improve

the input data, the model design, explain prediction outcomes as well as ensure scientific consistency. An example of the incorporation of domain knowledge can be found in Chapter 4, where I combine ideas from auditory computational neuroscience and machine learning to obtain new insights into auditory feature detection. Besides expanding scientific knowledge, XAI can also assist in achieving a range of other qualities that are usually desired in practice, such as fairness, privacy, reliability, robustness, causality, usability and trust (Doshi-Velez & Kim, 2017). Furthermore, from a computational perspective, black box models tend to be harder to debug as it is more difficult to clearly pinpoint weaknesses (Samek et al., 2018). All of these reasons warrant a focus on XAI.

The remainder of this section is structured as follows. First, relevant concepts of XAI will be defined. I will review interpretability and discuss methods that can help in achieving transparency and post-hoc interpretability. In the last section, I focus on what constitutes a scientific explanation, and provide examples of how domain knowledge has been integrated with XAI to achieve scientific outcomes and discovery.

4.1 A Taxonomy of Explainability

In contrast to the standard, objective measures that are typically used to measure the quality of a machine learning system, such as accuracy scores, XAI is a less rigid concept. Terms such as transparency, interpretability and explainability are often used somewhat interchangeably. In this review, I roughly follow the definitions used by Roscher et al. (2020), Lipton (2018) and Montavon et al. (2018). These authors define these three terms in a hierarchical fashion.

Within this hierarchy of XAI, interpretability reflects the idea that certain properties of a machine learning model are presented in human-understandable features, such as verbal descriptions or heatmaps that are atop the input (e.g. an image for a computer vision task) to indicate regions of interest. Interpretability can be achieved in two ways. Firstly, the machine learning approach itself might be transparent, that is, inherently understandable. For example, simple linear models may be considered transparent, provided they are of low enough dimension to be understood by a human. Secondly, a non-transparent model can be interpreted through what is referred to as post-hoc interpretability (Lipton, 2018) or reverse engineering (Guidotti et al., 2018). Post-hoc interpretability methods present human understandable interpretations of the model predictions and internal mechanisms.

In this review, I will not consider an interpretation to be synonymous with an explanation. A feature in the interpretable domain can usually be produced in an algorithmic fashion by using the model and the data. However, these interpretations do not necessarily

provide a self-contained explanation. To truly explain the decisions of a machine learning system, more context might be required. Following Roscher et al. (2020), I define an explanation as a set of features in the interpretable domain (e.g. a visualisation or text-description) that contributed to a decision, combined with domain knowledge and a clear analysis goal.

4.2 Interpretability

Within the machine learning community, the term interpretability does not have a universally agreed-upon meaning (Lipton, 2018). It is often defined through its relationship to human understanding. For example, Doshi-Velez and Kim (2017) define interpretability as ‘the ability to explain or to present in understandable terms to a human’, and (T. Miller, 2019) define it as ‘the degree to which a human can understand the cause of a decision’. Others have used terms such as comprehensibility (Freitas, 2014; Martens et al., 2011) and model-centric explainability (Edwards & Veale, 2017) to describe similar concepts. None of these definitions are necessarily self-contained, as it is unclear to what degree the ‘cause of a decision’ should be explained to constitute an interpretation. Instead, I will follow the definition coined by Montavon et al. (2018), who define an interpretation as the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of. Examples of this ‘interpretable’ domain are images or descriptive texts. The domain of the human-understandable features can correspond to that of the input data (Doran et al., 2018; Doshi-Velez & Kim, 2017), such as the use of heatmaps for classifiers in computer vision, or relate to aspects of the model architecture itself, such as visualisations of the distributions in the layers of a neural network. In this definition, interpretability can thus be viewed as a function of the model and the data used (Roscher et al., 2020).

There are two general forms of interpretability: transparency, which is considered in relationship to the model itself, and post-hoc interpretability, which uses both the model and the data to produce an interpretation (Lipton, 2018). Each of these will now be discussed in more detail.

4.3 Transparency

Transparency refers to the inherent interpretability of the machine learning approach itself. In some ways, the best explanation of a model is a transparent model, as it perfectly represents itself and is easy to understand (Lundberg & Lee, 2017). Often, transparency is considered to be a trade-off with predictive power, as it is assumed that transparent models are inherently less powerful (Došilović et al., 2018). Transparency can be considered

at three different levels: the level of the entire model (simulatability), the transparency of individual components of the model (decomposability) and the level of the training algorithm (algorithmic transparency) (Lipton, 2018; Roscher et al., 2020). These three levels can help us understand why DNNs are usually considered to be black box models that lack transparency.

Simulatability

Simulatability refers to the idea that a person can contemplate the entire model at once. DNNs can be defined as ‘model transparent’ in the sense that the input-output function is clearly defined mathematically (Roscher et al., 2020). However, most definitions of model transparency incorporate constraints such as model size, sparsity and monotonicity (Freitas, 2014; Lipton, 2018; Lundberg & Lee, 2017). It follows from these constraints that an interpretable model is a simple model (Lipton, 2018). Models that are thought to be inherently transparent are sometimes referred to as ‘pure transparent’ (Došilović et al., 2018). Examples of these models include decision trees (J. Wang et al., 2015), prototype selection (Bien & Tibshirani, 2011), linear models, k-Nearest-Neighbour models as well as Bayesian frameworks (Letham et al., 2015). DNNs are likely too complex to fulfill the simulatability requirement, as they tend to consist of thousands (if not millions) of parameters and behave in a highly non-linear fashion. However, it is important to note that even ‘pure transparent’ models can become difficult to grasp at once. For example, a high-dimensional linear regression model may not be considered transparent (Lipton, 2018).

Decomposability

Decomposability refers to the idea that certain components of the model may admit an intuitive explanation (Lipton, 2018). This is closely related to design transparency (Roscher et al., 2020), which refers to the idea that even if the model as a whole is not considered transparent, the choice of certain components might be. For example, design transparency can be increased by mixing transparent machine learning models with black box methods. This is sometimes referred to as hybrid transparency. By replacing parts of an interpretable model with black box models, performance may be increased while still retaining a certain level of transparency (Endriss & Leite, 2014; van Gestel et al., 2005). Furthermore, domain knowledge can help to increase the transparency of individual components of a model. For example, the choice of convolutional neural network layers for image recognition was

loosely inspired by models of neural connectivity of the animal visual cortex. On the other hand, the number of layers and neurons in DNNs are often chosen in an ad hoc fashion, thus lacking design transparency.

Some work is underway that aims to improve the inherent transparency of DNN components. For example, neural networks with attention mechanisms have shown great promise in natural language processing tasks. One of the major breakthroughs was the introduction of the Transformer neural network (Vaswani et al. (2017), see Section 3.3) which uses multiple attention layers that allow the neural network to focus on a subset of its inputs. These networks seem to be considered more transparent by some, as it is assumed that by looking at the attention vector one can directly infer which parts of the input the network uses in its prediction (see e.g. Y. Wang et al. (2016) and Lin et al. (2017)). However, caution should be taken when directly interpreting attention weights, since higher attention weights do not always correlate with greater impact on model predictions (Serrano & Smith, 2019).

Some argue that to increase transparency, the feature-oriented nature of neural networks should be replaced by a more symbolic representation. For example, object-oriented deep learning aims to replace the uninterpretable components of DNNs (tensors) by ‘objects’ or ‘symbols’ which explicitly package common properties of visual object (Liao & Poggio, 2017). Other work focuses on creating network architectures that allow for easy extraction of relationships between input features or between neurons in the network. For example, Y. Yan et al. (2019) introduce grouping layers, which can be used to identify subgroups of neurons in end-to-end models. Tsang et al. (2018) propose a feedforward network that is structured in such a way that statistical interactions between input features can be extracted from the weights of the hidden units in the network.

Algorithmic Transparency

Lastly, algorithmic transparency covers the learning algorithm itself. For DNNs, the learning algorithm is transparent to the extent that the algorithm itself can be written down. However, the choice of hyperparameters (such as batch size and learning rate) are not transparent in nature. In the case of linear models there may be an understanding of the shape of the error surface, or it can be proven that training will converge to a unique solution. As of the time of writing, such a thorough mathematical understanding of the learning algorithms used in deep learning has yet to materialize (Roscher et al., 2020).

4.4 Post-Hoc Interpretability

It is possible to achieve interpretability for non-transparent models through post-hoc interpretability, which offers a way to extract information about the decision making process of trained models without needing to explain the entire model at once. This form of interpretability is similar to how people make justifications for their own decisions without fully understanding the decision making mechanisms employed in their brain. Where transparency describes a property of the model only, post-hoc interpretability is tightly linked to both the model and the data. Often (but not always) post-hoc interpretability is concerned with explaining single predictions. By collating explanations of the predictions of a range of inputs, this can provide a general overview of how the model behaves. As it can be time consuming to investigate large numbers of individual interpretations, it can be useful to cluster or otherwise collate interpretations automatically (Lapuschkin et al., 2019). Note that when considering post-hoc interpretability, a high level of scrutiny is required to prevent systems from producing plausible but misleading explanations (Došilović et al., 2018).

In the upcoming sections I will describe a range of post-hoc interpretability methods. During this discussion, I will sometimes refer to methods as model agnostic versus model specific. Model agnostic methods can be applied to any type of machine learning algorithm, whereas model specific methods only work for a specific family of machine learning models (in the case of this review, DNNs). Furthermore, post-hoc interpretability methods can generally be classified as either local methods, which describe what individual predictions depend on locally, or global methods, which describe behaviour of the model as a whole (Doshi-Velez & Kim, 2017). As local methods do not have to consider the entire model, they may provide more accurate interpretations for specific instances. However, as they only approximate the model behaviour locally, the various local interpretations may be inconsistent with one another.

Explanations Through Examples

When the input data itself is in the interpretable domain (such as images or texts), example-based interpretations may be used. These methods select particular instances of the dataset to explain the behaviour of a model or the underlying data distribution (Molnar, 2020). Certain families of machine learning approaches are considered transparent because they are example based. For example, predictions of the k-Nearest-Neighbour algorithm are explicitly based on the average of majority label of similar instances in the training set. An

advantage of example-based methods is that they are usually easy to create, particularly when the input data is of the interpretable domain. However, this approach lays the burden of constructing a meaningful explanation on the interpreter, as it requires humans to construct mental models without providing an indication of what the most important aspects of those examples are. I will now discuss three types of examples that are used to interpret model behaviour: counterfactual examples, adversarial examples and prototypes.

A counterfactual example displays the smallest possible change to feature values of an instance that would have changed the original prediction to a predefined output (Wachter et al., 2017). For example, a counterfactual interpretation of a denied loan application system may read *‘If your income had been £45k instead of £30k, you would have been offered a loan’*. Counterfactual examples can be generated algorithmically through standard optimisation algorithms. The loss function is usually a function of two components: the distance between the desired and the predicted outcome, as well as the distance between the original instance and the counterfactual instance in feature space. Not all input features may be considered by the optimisation algorithm. For example, in the case of a loan application decision system, unchangeable variables such as age may be disregarded. It is possible for multiple counterfactual examples to exist, as there may be several ways to achieve the predefined outcome. Note that counterfactuals do not convey the internal state or logic of an algorithm, but instead describe its dependency on the external factors that led to that decision.

Similarly to counterfactual examples, adversarial examples are generated by perturbing a data instance. However, the goal here is not to provide an explanation, but to pinpoint weaknesses in the model. Adversarial examples are designed to look identical to a given baseline instance to humans, but result in a completely different (and false) prediction. The method to find adversarial examples is similar to counterfactual examples: an optimiser is used to find the smallest perturbation possible that will result in a different prediction. DNNs appear to be particularly susceptible to these types of attacks. Changing even a single pixel can be used to deceive some DNN-based image classifiers (Su et al., 2019). Model-specific adversarial techniques (also called ‘white-box attacks’) make use of knowledge about the model, for example by using the gradient to speed up the optimisation process (Goodfellow et al., 2015; Szegedy et al., 2014). Adversarial examples can also be generated through model-agnostic methods, for example by training a surrogate model based on a set of examples that approximates the decision boundaries of the model under attack (Papernot et al., 2017). Although most work on adversarial attacks focuses on the visual domain, DNN-based speech recognition systems have also been shown to be susceptible to this type of attack (Alzantot et al., 2018). It is possible to counteract certain adversarial

attacks. Notably, Dong et al. (2017) used an adversarial training scheme that also increased the interpretability of the neural network.

Prototypes are another form of example-based post-hoc interpretability. These methods aim to find a prototypical instance that is representative of a larger group of instances. For example, an adapted version of k-means algorithm uses prototypes rather than data set averages to establish classification rules (Garcia et al., 2012). A similar idea can be applied to neural networks through activation maximisation, which find an input pattern that produces the maximal response for a quantity of interest through optimisation (Simonyan et al., 2014). The method can be used to maximize the activation with respect to either a certain class prediction (thus providing prototypical examples of a specific class), or the activation of certain neurons in a neural network. This last method helps explaining which features the specific neurons respond to best (Mordvintsev et al., 2015). For the prototypes to be interpretable, it can be necessary to add a regularisation factor that ensures the prototype follows the same feature distribution as the training data (Goodfellow et al., 2014). Prototype approaches have been applied to natural language processing models by generating representative phrases that input to an LSTM model (Karpathy et al., 2015). In contrast to counterfactuals or adversarial examples, prototypes contribute to the interpretation of the model as a whole instead of single instances. They can be made more local by adding a localisation term to the loss function (Montavon et al., 2018).

Attribution Methods

Attribution methods aim to reveal components of high importance in the input and their effect as the input is propagated through the network (Ancona et al., 2017). These methods go under a large umbrella of names, such as contribution (Shrikumar et al., 2017), saliency (Simonyan et al., 2014), influence (Kindermans et al., 2016), relevance (Bach et al., 2015; Binder et al., 2016; Robnik-Šikonja & Kononenko, 2008) and feature importance (Došilović et al., 2018). This group of methods lends itself well for visualisations as the attributions are usually in the same domain as the input data, such as images or text.

One way of obtaining local attribution-based explanations are model-agnostic perturbation methods. In particular, sensitivity analysis can be used to inspect how a model output locally depends upon the different input parameters. Classically, sensitivity analysis perturbs features along one or two dimensions and records the responses. Zeiler and Fergus (2014) proposed a simple occluding method as a comparison benchmark for perturbation-based methods, in which one feature x_i at the time is replaced with a zero baseline, and

measure its attribution as the extent to which the perturbation changes the target output. Ribeiro et al. (2016) introduced the model-agnostic approach LIME (Local Interpretable Model-Agnostic Explanations). This method generates interpretations by creating a local linear proxy model in the neighborhood of a data point. This is done by perturbing the input and observing the change in predictions. In contrast to classical sensitivity methods, LIME captures small changes amongst all feature dimensions. A further generalisation is the SHapley Additive exPlanation (SHAP) approach (Lundberg & Lee, 2017). In contrast to LIME, Shapley values use all possible combinations of inputs to consider all possible predictions of an instance. More formally, SHAP computes the average marginal contribution of a feature value over all possible coalitions of features. In contrast to LIME, SHAP can offer certain guarantees such as consistency and local accuracy. However, although fast approximations exist, SHAP is more computationally expensive and can quickly become intractable when dealing with large, continuous feature spaces such as images and audio.

It is possible to determine attributions in more efficient ways by considering model-specific methods. For example, for models such as DNNs gradients can provide a more efficient way to analyse the effects of perturbations. Gradient-based attribution methods utilise the partial derivative $\frac{\partial f(x)}{\partial x_i}$ of the model function $f(x)$ with respect to a given input feature x_i . The sensitivity analysis can then be simplified by computing the partial derivative of a given sample with respect to each input feature (e.g. the pixel in an image) separately. There are a variety of ways in which this gradient can be used to compute attributions. Simonyan et al. (2014) introduced saliency maps in which the attributions are equal to absolute value of the gradient. Shrikumar et al. (2017) propose Gradient $\times$ Input, where the partial derivative is multiplied with the input itself. For Gradient $\times$ Input, the input acts as a smoothing filter that improves the sharpness of the attribution map. Sundararajan et al. (2017) proposed Integrated Gradients, an extension of Gradient $\times$ Input that does not only compute the gradient at the input, but instead computes the average gradient while the input varies along a linear path from some baseline input $\hat{x}$ (such as a black image or empty sentence) to x .

There are also DNN-specific methods that use neural activations rather than gradients. For example, Layer-wise Relevance Propagation (LRP) is a method to obtain global attributions for deep neural networks (Bach et al., 2015). Instead of using gradients, this method makes use of the actual neuron activations. The algorithm performs a backward pass through the network: it starts at the output layer and proceeds layer by layer to redistribute the prediction score according to a propagation rule. The propagation rule is designed to adhere to a conservation property, which means that the relevance received

by a neuron must be redistributed to the lower layer in equal amount. A range of rules have been proposed, see Montavon et al. (2019) for an overview. It has been shown that each step of the propagation procedure in LRP can be modeled as a Taylor decomposition performed over local quantities in the graph (Montavon et al., 2017). A related method is DeepLIFT (Shrikumar et al., 2017). Similar to LRP, DeepLIFT moves through the network in a backward fashion and uses a recursive propagation rule. Like Integrated Gradients, DeepLIFT makes use of a reference baseline input $\hat{x}$. More specifically, each unit is assigned an attribution that represents the relative effect of the unit activated at the input x compared to the activation achieved by the reference input $\hat{x}$.

Ancona et al. (2017) point out that there is a close relationship between LRP, DeepLIFT, Gradient $\times$ Input and Integrated Gradients by reformulating the former two in a modified gradient framework. They show that LRP is equivalent to Gradient $\times$ Input if only neurons with ReLU activations are used. Furthermore, they show that LRP is equivalent to DeepLIFT if the network has no biases and all nonlinearities adhere to $f(0) = 0$ (i.e. ReLUs and tanh neurons). They also note that DeepLIFT approximates the quantity computed in Integrated Gradients by replacing the gradient at each nonlinearity with the average gradient. As a result, DeepLIFT acts as an efficient approximation of Integrated Gradients as long as the gradients behave approximately linearly. However, when the gradients are highly nonlinear (for example in the case of LSTMs), DeepLIFT fails to produce meaningful results. All four methods are considered to be global methods, as they identify the marginal effect that the presence of a feature has on the output (Ancona et al., 2017). This is in contrast to sensitivity analysis, which attempts to explain how the input should be changed in order to obtain a desired variation on the output.

Methods such as Gradient $\times$ Input, Integrated Gradients, LRP and DeepLIFT capture the cross-interaction of features, rather than the effect an individual feature may have. If the goal is to find the few most important features, rather than to capture global nonlinear effects between a lot of features, a method that occludes features one-by-one (e.g. the baseline occlusion method proposed by Zeiler and Fergus (2014)) may produce more meaningful results (Ancona et al., 2017).

Global Visualisations

Example- and attribution-based methods usually provide interpretations in a similar domain as the input data. Alternatively, global visualisation techniques can be used to visualise the internal mechanisms of neural networks.

Two general methods for visualising DNNs are point-based and network-based techniques (Liu et al., 2017). Point-based techniques use scatter plots to reveal the relationships between neural network components, such as neurons or learned representations (Rauber et al., 2016; Zahavy et al., 2016). Popular techniques to visualise high-dimensional data include principal component analysis (Wold et al., 1987) and t-SNE (van der Maaten & Hinton, 2008). Network-based techniques display the internal workings of a neural network by generating and visualising a network topology. Here, the neural network is represented as a directed acyclic graph (DAG) and important information is represented through the size, color and glyphs of the nodes and edges. For example, Zhang et al. (2018) generate an explanatory graph to represent the knowledge of a pre-trained CNN.

It can be helpful to combine visualisations with generic performance measures. For example, Alsallakh et al. (2014) have developed an interactive model performance package, which consists of a confusion wheel that depicts the prediction score distribution, as well as an automated feature analysis, which depicts how well two groups of samples can be separated using certain features.

A large proportion of modern ASR systems use RNNs. These networks can be more difficult to visualise due to the added dimension of time. Karpathy et al. (2015) investigate long-term interactions of an LSTM network by comparing performance to character level language models, which revealed the existence of interpretable cells that keep track of long-range dependencies. Another notable example is LSTMVis, an interactive tool that can be used to visualise the hidden state dynamics of the activations of LSTM models (Strobelt et al., 2017).

Proxy Models

Proxy models are models from an interpretable model family that are used to approximate the predictions of a more complex approach (Gilpin et al., 2018; Guidotti et al., 2018). The goal is for the interpretable model to globally approximate the predictions of a black box model (Došilović et al., 2018). Examples of this include learning a single decision tree from a random forest (van Assche & Blockeel, 2007) or extracting decision-tree rules from a Support Vector Machine (SVM) (Martens et al., 2007). Proxy models are sometimes used to explain the model more locally. For example, Dong et al. (2017) showed that by applying Bayesian regression to DNNs, SVMs and random forests, local explanations can be produced.

Alternatively, a proxy model can be used to explain the internal representation of a

complex model. This method is common in the analysis of speech and natural language processing systems. In this context, certain linguistic properties are predicted from the activations of the neural network using a simple model. The underlying assumption is that the classification performance of the simple model is an indication of how well that linguistic property is represented in those activations. In its most basic form, the presence of a certain concept can be shown through simple correlations. For example, Nagamine et al. (2015) compute the mean response of neural units for particular phonetic features and show that certain nodes became selective to features such as place of articulation and voicing. Alternatively, simple classifiers can be used to determine how well a certain model component represents a feature of interest. Often, the components of interest are embeddings. These are mappings from the input data to vectors that are trained to represent important aspects of input data concisely. For example, multi-layer perceptrons with one or several hidden layers have been used to study properties such as speaker identity, sex, speaker rate, word contents, word order and morphology of word-, speaker- and sentence embeddings (Adi et al., 2016; Conneau et al., 2018; Ganesh et al., 2017; Qian et al., 2016; S. Wang et al., 2017). Similar work has focused on uncovering where properties such as utterance length and words are represented in RNN neurons (Chrupała et al., 2017).

Integrated Explanations

Rather than using separate post-hoc interpretability methods to interpret opaque models, some work focuses on automatically generating explanations for opaque models. For example, Hendricks et al. (2016) propose an image classifier that outputs explanatory text by using an RNN trained by descriptive labels and descriptions of images. However, caution should be taken when using a separate model to generate explanations, as there is no guarantee that the explanation model follows the reasoning of the real model. This approach may therefore lead to misleading explanations. Ultimately, the goal is to produce an explainable system that emits symbols that enable user-driven explanations of how a conclusion is reached (Doran et al., 2018). In such systems, automated reasoning would be central to the output explanations, without the need for human post processing. Some early work has been done on such inherently explanatory systems (see e.g. Besold et al. (2017)), but this is still an active area of research.

Assessing The Quality of Post-Hoc Interpretations

It can sometimes be difficult to compare post-hoc interpretability methods, as there is often no ground truth available that can be used to algorithmically assess the quality of the interpretation. There have been some proposals for algorithmic measures that test the quality of feature importance methods. Macdonald et al. (2019) introduced a formal notion of ‘relevance’. They pose that a set of input features is deemed relevant for a classification decision if the expected classifier score remains nearly constant when randomising the remaining features. They also prove that the problem of finding small sets of relevant features is NP-hard. This could be viewed as a justification for the current use of heuristic methods in practical applications (Roscher et al., 2020). Arras et al. (2017) describe an objective measure for the quality of feature importance by iteratively perturbing the relevant variables sorted by relevance. The average decline of the prediction score represents the explanation quality, where a large decline indicates that relevant input variables were truly relevant.

The lack of an objective measure of quality is sometimes circumvented by using an axiomatic approach, in which a set of preferred properties (called axioms) are set out and it is shown that the proposed method fulfills these axioms. For example, methods such as SHAP, Integrated Gradients and DeepLIFT, fulfil the axiom of ‘completeness’, which means that these attributions sum to the target output minus the baseline. It is then up to the practitioner to decide whether the provided axioms are deemed sufficient and/or necessary.

A more rigorous (and costly) approach to assess whether an interpretation is not only algorithmically sound but also provides an understandable explanation, would be to make use of human-centric experiments. Gilpin et al. (2018) propose the use of a common interpretability taxonomy for evaluating interpretations. They distinguish between three types of evaluation: application grounded, where real humans use the analysis method in real tasks; human-grounded, in which real humans use the interpretation method in a simplified task; and functionally grounded, where a proxy task is assessed analytically. Moving up this hierarchy from functional up to application grounded is equivalent to moving from algorithmic interpretations to explanations that explicitly take context and human interpretation into account.

4.5 Explainability and Domain Knowledge

Although interpretability methods can help in increasing our understanding of machine learning models, a meaningful explanation will often require the context of domain knowl-

edge and a clear analysis goal (Roscher et al., 2020). The requirement of a model to be in line with domain knowledge is sometimes referred to as justifiability (Martens et al., 2011). To check the justifiability of a model, a domain expert must be able to understand the presented interpretations. This requires an explanation to be subject-centric (Montavon et al., 2018), that is, targeted to its audience. When machine learning is applied to achieve scientific outcomes, scientific consistency is often coined as an additional requirement (Karpatne et al., 2017; Reichstein et al., 2019; Roscher et al., 2020; von Rueden et al., 2019). Scientific consistency can be applied in a post hoc fashion, where outputs are explained using additional scientific knowledge. It may also be incorporated in the design process itself, by acting as a regularizer that restricts the solution space to scientifically consistent solutions (von Rueden et al., 2019). In this section, I focus on the ways in which the incorporation of domain knowledge into XAI can and has been used to improve scientific consistency and understanding. A more comprehensive overview can be found in Roscher et al. (2020).

Firstly, domain knowledge can be used to check the model output for scientific consistency. For example, Maheswaranathan et al. (2019) use a convolutional neural network architecture to accurately model neural responses of the retina to natural images. They then apply post hoc consistency checks using domain knowledge. Despite only being trained on natural images, it was shown that the trained model simulated responses consistent with a range of observed retinal responses to highly artificial stimuli (such as bars and flashes of light).

Instead of testing for consistency after the model has been trained, it is also possible to incorporate consistency in the model design. This may result in more transparent models. For example, algorithmic transparency can be improved by including constraints to the model loss function. Stewart and Ermon (2017) and Wu et al. (2016) both use CNN architectures to predict physical parameters (such as the height or velocity of a thrown object) from videos. By explicitly introducing physical laws and constraints into the loss function of the network, scientific consistency with underlying physical processes is improved. Furthermore, model transparency can be improved by using domain knowledge in the design of particular components or the model as a whole. For example, Raissi et al. (2017) use Gaussian Processes to learn the coefficients of general parametric linear differential equations from noisy data. They incorporate domain knowledge into their model design by introducing a prior that is based on the corresponding known types of closed-form differential equations.

Post-hoc interpretability methods can be used to improve the explainability of model

predictions, to perform scientific consistency checks or to discover new relevant features. For example, Ghosal et al. (2018) train a CNN model that can predict plant stress based on images. They use interpretation tools to automatically identify important visual cues, showing a high level of agreement with visual cues that were manually marked by experts. Proxy models can also be a useful tool for interpreting models with the goal of generating new scientific understanding. For example, in follow-up work to the retina modeling work by Maheswaranathan et al. (2019) discussed earlier, Tanaka et al. (2019) reduce the complex DNN to a simple neural network with one hidden layer that can be readily related to neuroscientific processes. The number of units in the complex network was reduced by applying attribution methods to determine the subset of neurons most important for generating retinal responses for simple stimuli. The extracted computational mechanisms were shown to be consistent with prior scientific literature, but also generated a new hypothesis for a neural mechanism that may underlie vision.

5. CHAPTER SUMMARY

In this chapter, I reviewed advances in automatic speech recognition and explainable machine learning. First, I provided a brief history of ASR, a history characterised by a shift from knowledge-based speech recognition to DNN ASR systems.

Second, I reviewed the types of input features that are typically employed in ASR systems. Most modern ASR systems make use of spectral input features (e.g a spectrogram), features derived from the spectrogram (e.g. MFCCs) or learnable frontends. Features that are more consistent with the way the auditory system works (such as Gammatone filterbanks) are not very common due to their high computational cost. I also shed light on the type of techniques that have been and are being used to make the input features to ASR systems more noise robust.

This was followed by an introduction of deep learning for ASR. This section shortly introduced DNNs. I also covered some of the deep learning architectures that have been used for ASR systems, such as LSTMs and the more recently introduced Transformer models.

The rise of deep learning has come with an increased interest in using DNNs as a tool for science. Unfortunately, using deep learning for this purpose is not always straightforward, as DNNs are often considered to lack transparency and interpretability. I provided a comprehensive overview of techniques that can be applied to improve the interpretability of DNNs. In the context of DNN-based ASR or natural language processing

systems, commonly used methods are adversarial inputs, the use of attribution methods, global visualisation techniques and the use of proxy models to explain the representation learned by embedding-based networks.

Lastly, I discussed how domain knowledge can be used to provide the context required to derive scientifically meaningful explanations. In the work presented in this thesis, I will extensively use the domain knowledge presented in this and the previous chapter (Chapter 2) in the application of XAI. In Chapter 4, I use knowledge about the early auditory system to guide the design of a biologically compatible computational model for basic auditory feature detection. In Chapter 6 and 7, I use psychometric findings to guide the search for low-level features and mechanisms employed by DNNs. In Chapter 7, I also explore the application of attribution methods to understand the use of dip listening in an ASR system.

Part II

Neuro-inspired Feature Detection

CHAPTER 4

Auditory Feature Hypothesis Generation

1. INTRODUCTION

Our current understanding of hearing and speech recognition relies primarily on experimental approaches that require simple stimuli to allow for controlled experiments. However, these simple stimuli often lack the variability and complexity that are characteristic for complex sounds such as speech. In this chapter I propose an approach that combines neural modelling with data-driven machine learning to determine auditory features that are both theoretically powerful and can be extracted by networks that are compatible with the known physiology of the auditory system. The proposed approach bridges the gap between detailed neuronal models that capture specific auditory responses, and research on the statistics of real-world speech data and its relationship to speech recognition. Importantly, the proposed feature detection model can capture a wide variety of well studied features using specific parameter choices, and allows for the unification of several concepts from different areas of hearing research.

First, I introduce a feature detection model with a modest number of parameters that is compatible with auditory physiology. I show that this model is capable of detecting a range of features from raw audio signals, such as Amplitude Modulations (AMs) and onsets. In order to objectively determine relevant feature detectors within the model parameter space, I use a simple classifier that approximates the information bottleneck, a principle grounded in information theory that can be used to define which features are ‘useful’. By analysing the performance in a classification task, this framework allows for the determination of the best model parameter settings and their neurophysiological implications. Moreover, the results can be related to findings in psychoacoustic studies. I show that in the context of a phone classification task, the proposed approach rediscovers several well-known

proposed mechanisms for robust speech processing, such as spectral peak detection and AM sensitivity, as well as less-established ideas, such as the importance of spectral splatter suppression. It is also shown that some model variants are responsive to more complex features, such as combined onset and AM sensitivity. Finally, I show how the proposed approach lends itself to be extended to more complex environments, by distorting the clean speech signal with noise.

There are various potential applications of the work presented in this chapter. Firstly, it could lead to new, testable experimental hypotheses for understanding hearing. Moreover, promising features can guide the development of new acoustic front-ends for speech recognition systems.

2. METHOD

2.1 Neuronal Feature Detection Model

The proposed feature detection model is inspired by the early auditory pathway, which includes the auditory periphery and early brainstem. The model utilises precisely timed inhibition as a mechanism for feature detection. This mechanism is thought to be employed at various locations in the early auditory pathway to improve noise-robust speech processing, such as in onset and offset sensitive octopus cells in the Ventral Cochlear Nucleus (VCN), spectral notch detection through wideband inhibition in the Dorsal Cochlear Nucleus (DCN) and AM sensitive neurons in the Inferior Colliculus (IC). The model generalises previous proposals (see e.g. Carney et al. (2015), Skorheim et al. (2014) and Smith and Fraser (2004) for neuronal models for AM, Frequency Modulation (FM) and onset sensitivity, respectively) by using precisely timed inhibition in a similar way to these models, while being less restrictive in the way inhibitory and excitatory neurons influence one another. The model proposed here unifies the detection of a variety of features that tend to be treated separately in the literature, such as onsets and AMs. The model, depicted in Figure 4.1, takes raw audio data as input. The first stage of the model approximates the Auditory Nerve fiber (AN fiber) response. Each AN fiber response x^{f_c} for a signal x is estimated according to Equation 4.1.

$$x^{f_c} = h(G^{f_c}(x))^{\frac{1}{3}} \quad (4.1)$$

Here, $h(\cdot) = \max(\cdot, 0)$ denotes a half-wave rectification and $G^{f_c}(\cdot)$ denotes a Gammatone

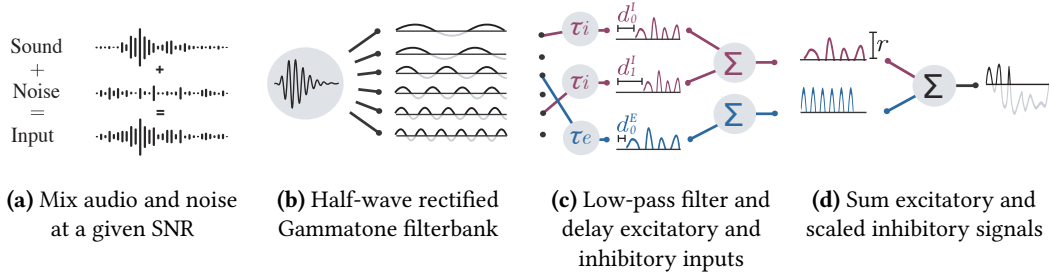


Figure 4.1. An overview of the components of the model that is used to generate auditory features. (a) The first step is to mix raw audio files of phones and noise. (b) The resulting audio file is decomposed in frequencies using a Gammatone filterbank. The output of this filterbank is halfwave rectified and compressed using cubed root compression. (c) Then, channels are selected to become inhibitory and/or excitatory. The inhibitory and excitatory signals are low-pass filtered with a cut-off frequency of τ_{i_e} and τ_{e_e} , respectively. (d) The inhibitory signals are delayed with delay d and their strength are altered through the ratio r . The auditory features are then created by summing the excitatory and inhibitory signals and applying a half-wave rectification.

filter that extracts the frequency component of x centered around the centre frequency f_c . Cubic root compression ($\cdot^{\frac{1}{3}}$) is added to account for non-linear amplification in the cochlea. In the proposed model, the bandwidths of the Gammatone filters are determined according to the Equivalent Rectangular Bandwidth (ERB) space, which defines a bandwidth for a given centre frequency similar to bandwidths found in the auditory system (Patterson et al., 1992). I used the implementation of Gammatone filter banks of the Brian Simulator (Fontaine et al., 2011), which implements a digital Gammatone filterbank through a cascade of four second-order IIR filters as per Slaney et al. (1993).

The model divides the Gammatone filter outputs into an excitatory and inhibitory stream. Given a set of centre frequencies E and a set of associated delays D^E , the excitatory stream s_t^E at time t is computed following Equation 4.2.

$$s_t^E = \sum_{e \in E} (L^{\tau_e}(x^e)_{t+d_e}) \quad (4.2)$$

Here, $L^{\tau_e}(\cdot)$ represents a low-pass Butterworth filter with a cut-off frequency of τ_e , and d_e represents the delay in D that is paired with the centre frequency $e \in E$. The excitatory stream is thus the sum of the AN fibers that respond to the centre frequencies in the set E after they have been low-pass filtered with cut-off frequency τ_E and are each delayed by a delay given by the set D^E . Similarly, the inhibitory stream s_t^I at time t depends on the set

of centre frequencies I and associated delays D^I , as shown in Equation 4.3.

$$s_t^I = \sum_{i \in I} (L^{\tau_i}(x^i)_{t+d_i}) \quad (4.3)$$

Given the excitatory and inhibitory stream s^E and s^I , the model output o_t at a given time t is computed as follows:

$$o_t = h(s_t^E - r \cdot s_t^I) \quad (4.4)$$

Here, $h(\cdot)$ once again denotes a half-wave rectification, and r is the ratio between the strength of the inhibitory and excitatory stream. The model thus relies on seven sets of parameters: E , I , D^E , D^I , r , τ_e and τ_i . In order to keep the parameter space tractable, the model variants in the results presented here were restricted to up to two excitatory centre frequencies with a wideband inhibition that was centered around the one or two frequencies in E . Under this restriction the model is still able to detect onsets and amplitude modulations within the given frequency bands, but the ability of the model to detect frequency sweeps and modulations is significantly reduced.

2.2 Data Preparation

The main source for spoken audio in this chapter is the LibriSpeech dataset, a freely available corpus of approximately 960 h of 16 kHz read English speech by 40 different speakers (Panayotov et al., 2015). The data of LibriSpeech is derived from read audio books from the LibriVox project, which contains audio books recorded by volunteers across the world. This means that there is a varying recording quality.

To distort the speech signals, the phones are mixed with natural noise recordings. These are drawn from the CHiME3 dataset (Barker et al., 2017), which contains over 8 h of noise of four types (bus, cafe, pedestrian, street). All these natural background noises have most of their energy in the lower parts of the frequency spectrum (up to 1 kHz).

For the analysis, the data was prepared in two ways: the speech inputs sound are aligned to phones and the noise and speech signals are artificially mixed with a given SNR. Each of these data preparation steps will now be discussed in more detail.

Phone Alignment with Montreal Forced Aligner

The spoken utterances in the LibriSpeech dataset are aligned with their transcription. However, for the purpose of this work, the speech data has to be aligned on phone level. This can be done automatically using applied forced alignment, a technique that takes the orthographic transcription of an audio file and generates a time-aligned version of the audio file and phone transcription. Here, I used the Montreal Forced Aligner (McAuliffe et al., 2017), an implementation based on the Kaldi ASR toolkit, to align utterances in LibriSpeech. The Montreal Forced Aligner uses both monophone and triphone models to perform the alignment, and also takes speaker characteristics into account to improve the accuracy of the overall alignment.

Mixing Noise and Sound Using FaNT

To mix the LibriSpeech utterance with the CHiME3 noise dataset at a particular SNR, it is important to correctly identify the energy in the speech and noises. For this, I used FaNT (Hirsch, 2005), an SNR calculator that was created as part of the AURORA 2 project, which is a well-known standard in ASR datasets for speech recognition in noise (Pearce, 2000). In general, the SNR is defined as follows:

$$SNR_{dB} = 20 \cdot \log_{10}\left(\frac{s_{rms}}{n_{rms}}\right) \quad (4.5)$$

Where s_{rms} and n_{rms} are the root mean square (i.e. $s_{rms} = \sqrt{\frac{1}{N} \sum_n (s_n^2)}$, and similarly for n_{rms}) of the signal s and noise n respectively.

FaNT uses both speech activity detection and frequency weighting to make the SNR computation more reliable. Speech activity detection is necessary to single out parts of the speech signal where no words are being spoken. Without speech detection, adding long pieces of silence could artificially reduce the total energy of the speech signal, thus causing the SNR to appear much lower than it actually is. Frequency weighting is applied to account for differences in perceived loudness of different frequencies. Psychoacoustic studies have shown that at the same SNR value, low-frequency noises tend to cause less distortion in speech understanding than for example white noise, which covers all frequency spectra (Hirsch, 2005). By applying frequency weighting to the signals before the SNR is computed, lower frequencies can be given less weight in their contribution to the SNR. In particular, FaNT implements a-weighting, a standard for frequency weighting (Fletcher & Munson, 1933).

2.3 Model Sensitivity

In order to compare the properties of different model variants, I developed two measures that assess AM sensitivity and onset sensitivity.

The onset strength s_o computes to what extent the model variant has the tendency to respond more strongly to the onset of a sound. It is defined as follows:

$$s_o = \frac{\mu_o - \mu_r}{\mu_o + \mu_r} \quad (4.6)$$

Where μ_o and μ_r refer to the mean response to the first o periods and last r periods of a sound that has the same frequency as the preferred frequencies of the model. The period is defined as $\frac{1}{f_c}$ in ms, where f_c is the preferred frequency. If the model variant has multiple preferred centre frequencies, o and r are defined based on the lowest frequency component. In the analysis presented here, $o = 50\text{ms}$ and $r = 200\text{ms}$ were chosen, as these values resulted in a high s_o for models that displayed clearly visible onset peaks. Note that the onset strength can be negative, which I will refer to as *onset suppressive*. Onset suppressive model variants have a delayed response to the onset of the sound.

To measure sensitivity to AM, I first determine the Modulation Transfer Function (MTF) of the auditory features. This function characterises the response of the feature detection model to sounds with various amplitude modulations. This is done by feeding the model with a signal that has a carrier frequency f_c equal to the centre frequency of the Gammatone filterbank channel and that is sinusoidally amplitude modulated with an amplitude modulation frequency f_m . The mean responses of the model to these inputs are used to determine the MTF. The MTF can be used to determine the best modulation frequency (BMF), which is defined as the modulation frequency at which the neuron has the highest response. A strong sensitivity to AM requires responses to the BMF to be significantly higher than responses to other modulation frequencies. The AM sensitivity s_{AM} is computed as follows:

$$s_{AM} = 1 - \frac{o_\mu}{o_{BMF}} \quad (4.7)$$

Where o_μ is the mean output response and o_{BMF} is the output response at the best modulation frequency. Since o_{BMF} is always higher or equal to o_μ and $o \geq 0$, s_{AM} ranges between 0 and 1, where 0 indicates no AM at all and 1 indicates strong AM sensitivity. Note that if $o_{BMF} = 0$, the AM sensitivity is undefined. In this case s_{AM} is set to zero. When $s_{AM} > 0.5$, model variants are considered AM sensitive. This threshold was determined through a

manual comparison of some MTF curves and their s_{AM} values.

2.4 Analysis Through Classification

The parameters of the model determine the type of features it responds to. To study the importance of each of these features in a speech processing context, the feature detectors are used in a phone classification task. First, the model is fed a given phone signal x to obtain the model output o . The model output is used as input to a single-layer Long Short-Term Memory (LSTM). This Recurrent Neural Network (RNN) was chosen because LSTMs are well-suited for classifying sequenced data. The LSTM is used to classify the phone label y , which is optimised using the categorical cross-entropy loss and an Adam optimiser (Kingma & Ba, 2014).

The task-driven setup proposed in this chapter is motivated by information theory. The Data Processing Inequality (DPI) tells us that the model output o can only reduce the information about label y compared to the information that was available in x . One way of defining a ‘good’ feature for the target label, is by looking for features that provide a useful compression of x and only retain the relevant information about y , while forgetting the irrelevant information present in x . This general principle is known as the information bottleneck. The advantage of the information bottleneck formulation is that it is model-independent, since it only relies on the distributions of the signal and label. However, determining the information bottleneck directly is computationally intractable for high-dimensional variables (such as sound waves, when considered over time). Instead, I apply the underlying idea of the information bottleneck by using a small, single layer LSTM with only 128 neurons. This ensures that high accuracy can only be achieved by features that have a relatively simple relationship to the label. In Appendix B, I show that in a reduced setting where the information bottleneck is more tractable, this approach finds similar model variants as the information bottleneck would.

2.5 Spectro-Temporal Receptive Field Estimation

A common way of visualising the response patterns of auditory neurons are Spectro-Temporal Receptive Fields (STRFs). This visualisation represents which type of stimuli excite or inhibit the neuron. Under the assumption of linearity, the neural response can be modelled by the convolution of a given stimulus with the STRF of the neuron. To get an intuition about the responses of some of the neural feature detectors, I estimated their STRF using the MNE toolbox (Gramfort et al., 2013). The ‘neural response’ of the feature

detection models was measured on the phone data set consisting of vowels, fricatives and plosives. This data was then fit to time-lagged inputs ranging from 0 ms to 100 ms using ridge regression. Several fits were made using different L2 regularisation parameters, and the fit with the best performance on a separate test set was selected as the final STRF.

3. RESULTS

3.1 Model Can Detect Multiplexed Features

Figure 4.2 displays three examples of the feature detectors that can be created using the proposed model. In particular, Figure 4.2a, 4.2b and 4.2c respectively exemplify the responses to the consonant /s/ of onset sensitive, AM sensitive and both onset and AM sensitive model variants, which I will refer to as ‘multiplexed’ feature detectors. Figure 4.2d-4.2f show the MTF of these same three detectors. The MTF measures the response of the model to several amplitude modulations. In this and all following experiments, the MTF curves were calculated over AMs ranging from 1 Hz to 128 Hz. Figure 4.2e and 4.2f show a clear peak in the MTF. This suggests they have a preferred AM and are thus AM sensitive, whereas the model only sensitive to onsets (4.2d) has a flat MTF. Although AM sensitive model variants could be found within all AM frequencies that were measured, the majority of AM sensitive model variants are sensitive to AMs in the mid range with the most common BMF being around 50 Hz. Note that since AMs were only measured up to 128 Hz, model variants that are sensitive to AMs higher than 128 Hz are not categorised as AM sensitive in this work.

A parameter space analysis (see next section) reveals that there is a range of parameter settings that result in either or both AM and onset sensitivity. Although qualitatively the same, these three classes of responses do vary in AM and onset sensitivity strength. Moreover, the AM sensitive neurons vary in their best modulation frequency (defined as the peak in the MTF). Note that although AM and onset sensitivity are the only features explored here, the model is expected to be extendable to other features such as FM and pitch, which have been previously implemented through similar models (see e.g. Skorheim et al. (2014) and de Cheveigné (1998) respectively).

3.2 Overall Classification Performance

A set of 2000 model variants with randomly sampled parameter settings were applied to a classification task of twelve phones, consisting of four vowels (/Λ/, /I/, /ε/, /æ/),

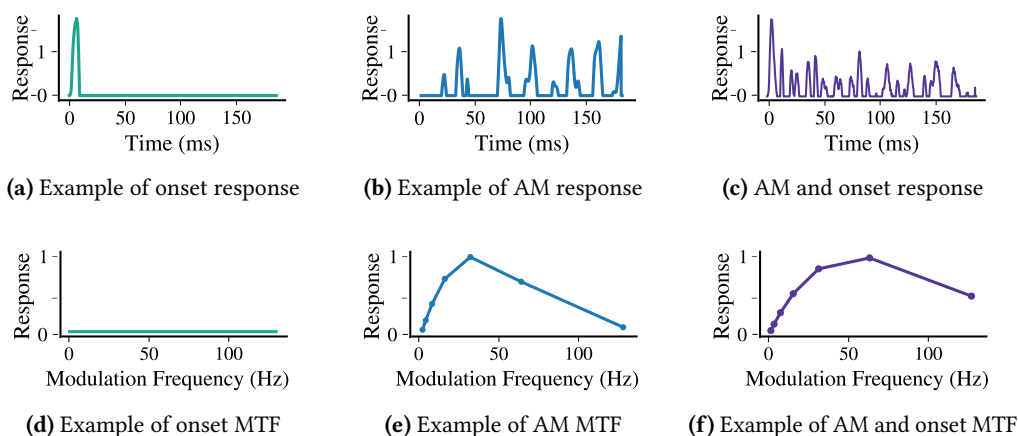


Figure 4.2. (a-c) Responses of an onset (top) and AM sensitive (middle) and combined (bottom) feature detectors to a recording of /s/, (d-f) Modulation Transfer Function of the model variants in plots a-c, measured as the normalised mean response to various AMs.

four fricative consonants (/f/, /v/, /s/, /z/) and four plosive consonants (/p/, /b/, /t/, /d/). A summation of the confusions of the 100 best-performing model variants in the three different groups shows that, as expected, most confusions are made within the three phone groups (Figure 4.3a). Overall, the average performance of the models is relatively low compared to human performance on similar tasks (see Figure 4.3c and 4.3d for human performance on consonant and vowel recognition tasks). This is expected since a single model variant filters out a lot of information (particularly frequency information, as the models in this experiment are restricted to a limited set of centre frequencies) and is unlikely to be useful for distinguishing between all phones. The analysis presented here therefore focuses on the relative performance between model variants rather than their absolute performance.

The sampled model variants tend to be better suited for fricative compared to vowel and plosive classification (Figure 4.3b). A correlation analysis of the accuracies between the three groups indicates that good fricative detection models tend to be good at detecting vowels and plosives too. However, there is a much lower correlation between the performance on a vowel versus plosive classification task. This suggests that there is an overlap between the features that can be useful for vowel and fricative recognition as well as between fricative and plosive features, but that a feature useful for vowel detection is unlikely to be helpful for plosive detection.

The 100 best model variants were applied to a more difficult 10-vowel recognition task (Figure 4.3c) and 13 consonant recognition task (Figure 4.3d). The results show that for certain parameter settings, a single feature detector can outperform accuracy obtained

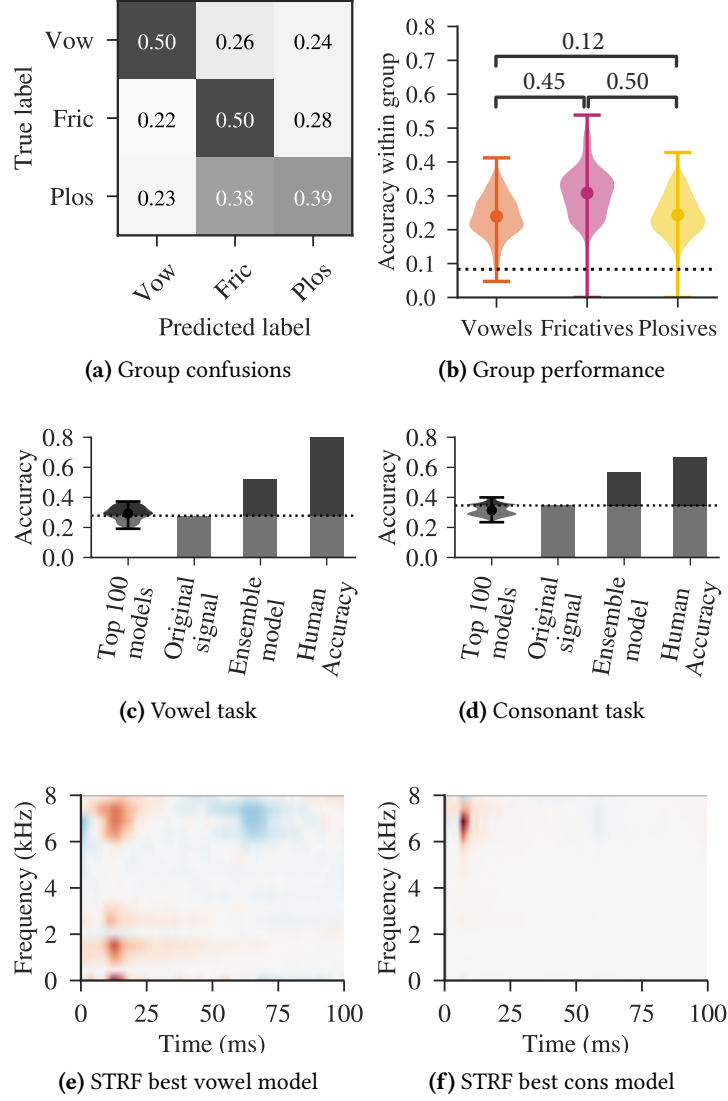


Figure 4.3. (a) Mean between-group confusion matrix for best model variants. (b) Distribution of within-group accuracies for all model variants, and between-group accuracy correlations. The dashed line refers to chance level ($\frac{1}{12}$) in this task. (c-d) Accuracy achieved with the best model variants from (a), the original filtered signal, and ensemble models on a 10-vowel (c) and 13-consonant (d) classification task. Human performance on a similar task from Meyer et al. (2010) is included. The dashed line in these figures refers to the performance on the original filtered signal, which is the baseline accuracy achieved without any feature extraction. (e-f) The spectro-temporal receptive field of the two best performing model variants for vowels and consonants.

with the original filtered signal. This indicates that the detected features allow for a more accessible representation of the relevant information in the original signal. The STRFs indicate that the best performing features for consonants (Figure 4.3f) tend to have shorter time constants in higher frequencies (>7 kHz), whereas vowel recognition is facilitated by responses to lower frequencies (2-4 kHz) and larger, slower time constants (Figure 4.3e).

The single feature performance can be improved by employing ensembles of inputs. Using the 23 features that individually perform best for each vowel and consonant as input to a newly trained LSTM increases accuracy by around 30%. This is particularly high for consonants, where performance becomes comparable to humans (Figure 4.3d and 4.3c).

3.3 AN Fibers Around Formants and Spectral Peaks Indicate Good Performance

The accuracies of the different phone groups are influenced by the centre frequencies of the model variants. The results suggest that vowel recognition is facilitated by lower centre frequencies, whereas best-performing model variants for fricatives and particularly plosives tend to have a high centre frequency (Figure 4.4a). This effect is likely closely related to the formants in vowels, which tend to be between 200-1000 Hz, and spectral peaks in consonants, which tend to be much higher. This relation can be made explicit by correlating the formants of all phones with the mean output of the model for these phones (Figure 4.4b). Models that are good at classifying vowels tend to have a negative correlation (i.e. models have a high output when the formants are low) whereas a positive correlation (i.e. model responds to high frequency formants) is typical for improved plosive recognition. Fricatives, which have high formants, benefit from either. Models outputs with a correlation close to zero (i.e. models that do not respond well to spectral peaks) perform badly in all groups.

3.4 AM Sensitivity Improves Fricative Recognition

The proposed model can be used to detect a range of features, in particular onsets, AMs or combinations of the two (Figure 4.2a-4.2c). I analysed the performance of 350 model variants that could be grouped into one of these three classes, each restricted to 1 kHz centre frequency for excitation and inhibition and only positive delays. When applied to a classification task of four fricative consonants, the results indicate that AM sensitivity is important for fricative recognition (Figure 4.4d)). Importantly, the accuracy of certain model variants is higher than the accuracy achieved by directly using the 1 kHz channel as input, indicating that the feature detector extracts useful features from the original

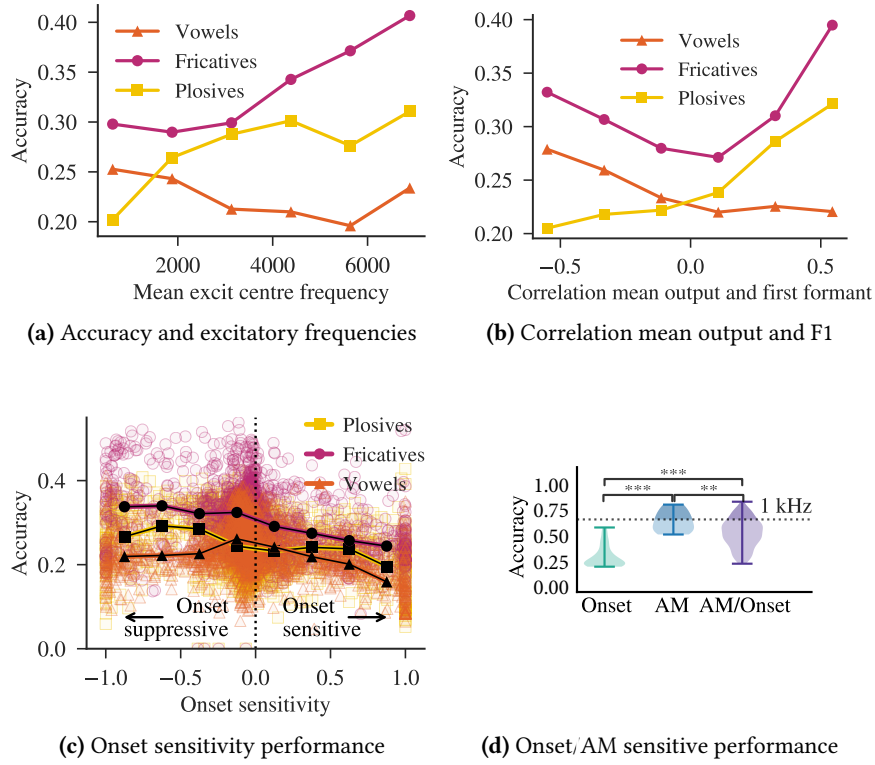


Figure 4.4. (a) Within-group accuracy in relation to the mean centre frequency for 12-phone task (b) Within-group accuracy and the correlation of the average model output and the first formant (spectral peak) in the 12-phone dataset. (c) Within-group accuracy versus onset strength. (d) Distribution of accuracies of three model types (params $E = I = 1$ kHz and $\tau_i > \tau_e$ and $d_i > d_e$.) in a fricative classification task ($p < 0.005$ and $p < 0.0005$ for ** and ***).

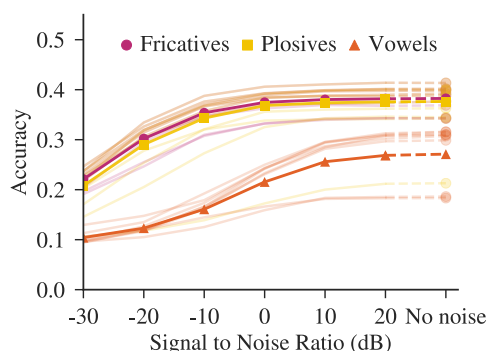


Figure 4.5. Overall accuracy under various SNRs with additive real world noise (CHiME3 data set) of the top 30 best model variants in a 12-phone recognition task (same as in Figure 1a-b). The models were trained under varying noise levels.

signal. These results are in line with results from experimental research which suggests that consonant recognition is degraded when certain AMs are removed (e.g. van der Horst et al., 1999).

3.5 Onset Suppression Facilitates Classification

The analysis is further extended by investigating the typical onset strength of a model variant, with -1 and 1 indicating a strongly onset suppressive or sensitive response, respectively (Figure 4.4c). The results indicate that strong onset sensitivity generally leads to worse performance, whereas onset suppression improves performance. This is particularly true for consonant recognition, although vowel recognition also benefits from a slight onset suppression.

3.6 Robustness Against Low-Frequency Noise

The proposed framework naturally lends itself to investigate features under varying environments, such as noise robustness. As a proof-of-principle, the SNR-curves of the 10 best performing models per phone group were computed after training them on a dataset with additive real-world noise from the CHiME3 data set (Barker et al., 2017) under varying SNRs (Figure 4.5). The results indicate that vowel recognition degrades faster than consonant recognition. This is likely because the CHiME3 noises mask vowels more than consonants, as the important formants for vowels lie in lower frequency regions and the CHiME3 noises tend have strong low-frequency components. Overall, the SNR curves start to degrade around 0-10 dB SNR, which is similar to human SNR curves (Meyer et al., 2010).

4. CONCLUSIONS

In this chapter, I introduced a neuro-inspired feature detector model and proposed an analysis methodology that can be used to identify features that are important for speech recognition. The proposed approach provides a unifying framework that can be used to confirm and explain existing concepts, such as the importance of spectral peak detection and amplitude modulations.

The real value of the proposed framework, however, is its ability to generate suggestions for future work in less well-known theories of hearing. For example, the results suggest that both onset suppression and AM sensitivity are important for phone classification. The importance of onset suppression is in line with the hypothesis that the suppression of onset noise (or ‘spectral splatter’), as observed in the mammalian auditory brainstem, can improve the clarity of a neural harmonic representation (Spencer et al., 2015). Furthermore, the importance of the onset response of the model variants suggests that phase information may play an important role in early brainstem processing. Typically, neurophysiological studies of AM sensitivity in the brainstem focus on the magnitude of the MTF, rather than the phase of the response (Joris et al., 2004). One notable exception is the work of Delgutte et al. (1998), in which both the MTF magnitude and phase of responses in the auditory nerve, Cochlear Nucleus (CN) and Inferior Colliculus (IC) were used to predict responses of these neurons to speech. In their work, Delgutte et al. (1998) stress the importance of incorporating phase information when modelling neurons in the IC. IC neurons tend to be sensitive to modulation frequencies similar to those considered in this chapter (up to 100 Hz). Their results suggest that in some IC neurons, noise bursts induce a biphasic response, with an initial delayed onset response followed by suppression. They showed that incorporating this biphasic information is important for accurate modelling of the IC response. Like the work presented in this chapter, the work of Delgutte et al. (1998) emphasise the importance of considering a variety of auditory features (e.g. both magnitude and phase) in a multiplexed fashion.

There are various potential applications of the proposed approach. Firstly, it could lead to new, testable experimental hypotheses for understanding hearing, both on the level of features as well as the underlying neurophysiological parameters. Moreover, promising features can guide the development of new acoustic front-ends for speech recognition systems.

Part III

Features of Automatic Speech Recognition

CHAPTER 5

A Primer on Comparing Human and Machine Listening

1. INTRODUCTION

In Part II, I showed how models in computational neuroscience can be used to explore the auditory features used by the human auditory system. An underlying assumption of this approach is that if the computational model somewhat resembles the auditory system, the features found using such a model may also reflect the features used by humans. In this part of the thesis, I explore the same fundamental question using a different approach: can Automatic Speech Recognition (ASR) be used as a model for human hearing, and, if so, as a tool for auditory hypothesis generation? To answer this question, a thorough analysis of three open-source ASR systems using a range of psychometric* experiments is presented in this part of the thesis. To allow researchers to assess their own ASR systems under similar conditions, the associated code has been published in *HumanlikeHearing* (Weerts et al., 2021), an new open source Python toolbox.

As a result of dramatic improvements in machine learning, particularly Deep Neural Networks (DNNs), automated systems are now able to achieve near human-level performance on a range of sensory tasks, such as image recognition and ASR. Where the brain once loosely inspired the development of Artificial Neural Networks (ANNs), there now is an increasing interest in using DNNs as a model for human sensory processing (Baby et al., 2021; Hu et al., 2020; Kell et al., 2018; Kollmeier et al., 2020; Rader et al., 2015; Schädler et al., 2015; Spille et al., 2018; Spille & Meyer, 2017; Stolcke & Droppo, 2017).

* To stress the similarity between the experiments presented here and those applied to humans, I use the word ‘psychometric’ in this part of the thesis even for the experiments are applied to ASR systems. Although this nomenclature is in line with previous work (see e.g. Spille and Meyer, 2017), the word ‘psycho’ semantically relates specifically to the human mind only, making the use of the word ‘psychometric’ in this context a slight misnomer.

For instance, a recent comparison of ASR and human speech recognition suggests ASR systems may be used to predict speech intelligibility in normal-hearing humans (Spille & Meyer, 2017). By applying Explainable Artificial Intelligence (XAI) techniques, the use of ‘dip listening’ to more accurately predict target sentences in the presence of modulated maskers was identified in an ASR model. This ASR model was trained in noisy conditions. In another example, Kell et al. (2018) showed that a task-optimized neural network was able to replicate human auditory behaviour without being explicitly trained to do so. This model was subsequently used to predict fMRI voxel responses throughout the auditory cortex substantially better than traditional spectrotemporal filter models. A more practical application of ASR systems in the auditory field is Framework for Auditory Discrimination Experiments (FADE) (Schädler et al., 2015), a framework that can be used to predict speech intelligibility in both normal and hearing impaired listeners. FADE has been shown to successfully estimate trends in speech intelligibility of older hearing-impaired listeners, even when trained on intact speech material (Fontan et al., 2020).

This line of research is of particular interest for the analysis of speech processing, since the ability to understand speech is unique to humans and therefore difficult to explore in detail using animal models. However, unlike models in computational neuroscience, high-performing DNNs are not designed to mimic human neural processing and may solve sensory tasks in vastly different ways. Particularly in the case of speech signals, which are inherently redundant, the same overall recognition performance may be achieved using different strategies and auditory cues. For an ASR system to be used as a model of human hearing, it is imperative that the system behaves as similarly as possible to human speech recognition. Are the mechanisms used by DNNs similar enough for them to function as a model for human hearing? Even if they are, how can black box models such as DNNs actually provide further insights? These questions are at the centre of this and the following chapters.

The remainder of this thesis is centered around the study of three open-source ASR systems, each of which uses a different underlying architecture. In Chapter 6, the results of a range of psychometric distortion experiments on ASR systems are presented. I will show that although certain qualitative similarities are present between human speech recognition and ASR, a variety of auditory features that are known to play an important role in robust speech processing in humans are used either very differently or not at all by the tested ASR systems. In Chapter 7, this analysis is further extended to a range of Speech in Noise (SPIN) experiments. The results from this study show that the more recent systems with better performance also tend to better match human results. Furthermore, the SPIN performance

of the best performing model is analysed using explainable machine learning techniques.

This chapter lays the groundwork for the remainder of Part III of this thesis. I will first introduce the three ASR systems that will be our guinea pigs in the next few chapters, as well as the speech data sets used in the experiments. Furthermore, this chapter covers some of the generic methodology necessary to apply psychometric tests to ASR systems, such as the need for sound level normalisation as well as how to determine Speech Reception Thresholds (SRTs) for ASR systems. Lastly, I will introduce *HumanlikeHearing*, an open source Python toolbox I wrote that allows other researchers to compare their own ASR systems with human hearing. The psychometric tests that are included in the toolbox will be described in more detail in the following chapters.

2. ASR SYSTEMS

Three open source ASR models are analysed in this work: Kaldi nnet3 chain model (trained as part of the Vosk project), DeepSpeech (trained as part of Mozilla’s DeepSpeech project) and Wav2Vec 2.0 (trained as part of Facebook’s fairseq toolbox). These systems were chosen based on two criteria. Firstly, the models are open source and trained on freely available data. Secondly, each of these models relies on a different class of commonly used ASR architectures. These models represent some of the recent history of ASR systems, with the initial application of DNNs to Hidden Markov Models (HMMs), followed by the successful application of Long Short-Term Memory (LSTM) models and finally the recent introduction of Transformer-based ASR systems. Each of the model architectures are described in more detail below. This section concludes with a comparison of the way in which each of these systems deals with both the temporal and spectral components of speech.

2.1 DNN-HMM

Vosk’s Kaldi nnet3 chain model is a DNN-HMM hybrid model, a type of model that combines DNNs with HMMs (Povey et al., 2018; Figure 5.1a). The model takes as input Mel-Frequency Cepstral Coefficient (MFCC) features and i-vectors. MFCCs are commonly used ASR input features that concisely describe the overall shape of a spectral envelope of a sound wave. The DNN-HMM model extracts 40 MFCCs from a spectrogram ranging in 20 Hz-7600 Hz, with windows of 25 ms shifted by 10 ms each time. I-vectors are latent variables that model speaker-specific characteristics of speech and can be obtained through factor analysis (Dehak et al., 2010).

The MFCC features and i-vectors are fed into a 15-layer Factorised Time Delay Neural Network (TDNN-f) (Povey et al., 2018). The layers of a TDNN-f are designed to classify patterns with shift-invariance, which means the classifier does not require explicit segmentation. Shift-invariance is achieved by averaging the backpropagation update over time-shifted copies of the network across the temporal dimension. Another feature of TDNN-f is that each layer receives not only the output of the previous layer at the current time step, but receives a contextual window of outputs from the layer below. This allows TDNN-f to model the temporal context. To efficiently model long temporal contexts, the DNN-HMM model uses a subsampling technique in which hidden activations at only few time steps are computed at each level, while ensuring that information from all time steps in the input context is processed by the network (Peddinti et al., 2015). Compared to a standard TDNN, the factorised TDNN-f applies a form of model reduction similar to singular value decomposition at every layer, which reduces model size by effectively applying a bottleneck to each layer. In the DNN-HMM model, layers of 1536 hidden neurons are reduced to a bottleneck layer of 160 nodes. The model was trained using SpecAugment (Park et al., 2019), a data augmentation technique, and an end-to-end variant of the Lattice-Free Maximum Mutual Information (LF-MMI) criterion (Povey et al., 2016).

2.2 LSTM

Mozilla's DeepSpeech model (Mozilla, 2020; Figure 5.1b) is a LSTM, which is a specific type of neural network that can model long temporal relationships in the input data. Mozilla's DeepSpeech is a variant of the original DeepSpeech model (Hannun et al., 2014), which used simpler Recurrent Neural Network (RNN) units.

In contrast to the original DeepSpeech model, which takes the speech spectrogram as input, the input to Mozilla's DeepSpeech model consists of 26 MFCCs that have been extracted from a spectrogram in a range of 20 Hz-4 kHz with a window length of 32 ms and a stride of 20 ms. One set of MFCCs reflects the spectral content over the 20 ms window.

The model architecture consists of six layers, all of which are standard feed-forward layers except for the fourth layer, which consists of recurrent LSTM units. The last layer is a softmax layer. The output of the model are lower case letters and the model is trained using the Connectionist Temporal Classification (CTC) criterion (Graves et al., 2006).

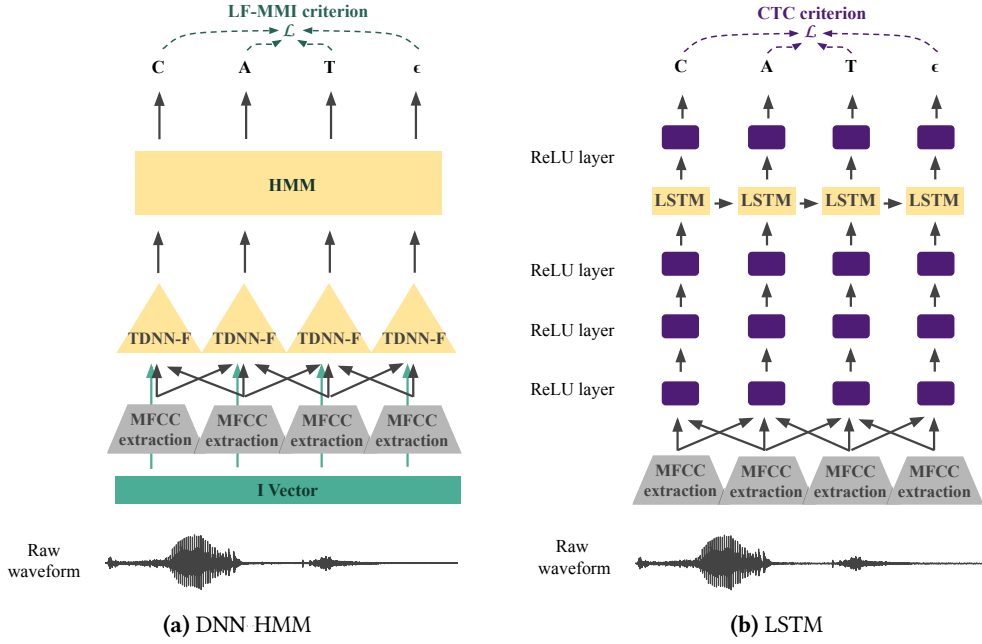


Figure 5.1. Illustration of the architecture of the DNN-HMM model (a) and LSTM model (b). Both models first extract MFCC features from the raw waveform. The DNN-HMM also extracts an i-vector of the utterance. These input features are fed into the models. The DNN-HMM model is trained using the LF-MMI criterion, while the LSTM model uses the CTC criterion. All yellow coloured components (TDNN-f, HMM and LSTM) denote the architectural components that are specifically designed to model the temporal relationships of speech.

2.3 CNN-Transformer

The last model considered is Wav2Vec 2.0, which consists of a Convolutional Neural Network (CNN) and a Transformer model (Baevski et al., 2020; Figure 5.2). In contrast to the other two ASR systems, Wav2Vec 2.0 takes the raw audio as input.

To extract relevant acoustic features from the raw audio, Wav2Vec 2.0 is initially ‘pre-trained’ in a self-supervised fashion. First, raw audio data is encoded into a latent speech representation by a seven-layer CNN that functions as a feature encoder. These representations are fed into a Transformer, which is used to build representations that capture information from the entire sequence. The model used here consists of 24 Transformer blocks with model dimension 1024, inner dimension 4096 and 16 attention heads.

To train the CNN-Transformer, the output of the feature encoder (i.e. the CNN component) is discretised. These discrete units are then used to represent targets in a self-supervised objective. During self-supervised training, some of the latent quantized

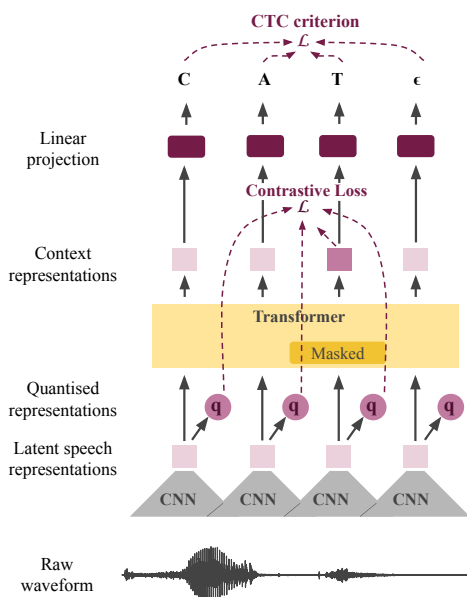


Figure 5.2. Illustration of the architecture of the CNN-Transformer model. In contrast to the DNN-HMM and LSTM models (Figure 5.1), the CNN-Transformer takes raw audio as its input. The model uses the CNN to extract discretised speech units which are fed into the Transformer, which learns contextualized speech representations. Both the discretised speech units and contextualized speech representations are learned using a pretraining procedure that uses a contrastive loss. The final transcription is learned in a finetuning step using a linear projection that is trained with a CTC loss. As in Figure 5.1, the yellow architectural component (Transformer) is responsible for modelling temporal relationships.

audio representation are masked. The self-supervised objective requires the model to identify the correct quantized latent audio representation from a set of distractors. The distractors are sampled at random from other masked latent representations in the same utterance. After pre-training, the model is finetuned for speech recognition by adding a randomly initialised linear projection on top of the context network into classes that represent the vocabulary (e.g. letters) of the task. The whole model is then optimised for speech recognition by minimizing the CTC loss.

2.4 Temporal and Spectral Limits of the ASR systems

Some basic limits on the spectral and temporal components of speech that each of these models can utilise can be derived from their inputs and architectures. For example, as all models are trained on sounds sampled at 16 kHz, frequency information of at most 8 kHz is present in the input. The LSTM model is limited to 4 kHz, as the MFCCs in that

model are derived from a spectrogram limited to that frequency. Although speech that is low-pass filtered at 4 kHz is highly intelligible to humans in quiet, the removal of high frequency information will affect the range of Frequency Modulations (FMs) this model can utilise, as well as certain high frequency cues that may be useful for speech detection in low frequency noise.

In both the DNN·HMM and LSTM architecture, spectral information is initially captured by the MFCC representation, in which much of the Temporal Fine Structure (TFS) is lost. In the CNN·Transformer, spectral information is extracted by the CNN layers, which effectively function as Finite Impulse Response (FIR) filters. The latent speech representations generated by the encoder of the CNN·Transformer have a window size and stride of approximately 25 ms and 20 ms, respectively, which is comparable to the window and stride of the MFCC features of the DNN·HMM (25 ms and 10 ms) and LSTM model (32 ms and 20 ms).

To model long temporal relationships between frames, each model uses a different strategy. In the DNN·HMM architecture, each layer at time t receives the outputs of the previous layer at times between $[t - s, t + s]$, where s denotes the ‘time stride’ of that layer. Given the time strides (1,1,1,0,3,3,3,3,3,3,3,3,3,3) of the 14 TDNN-f layers in the DNN·HMM model, the output layer has access to temporal information of 67 MFCC frames (685 ms). In the LSTM model, the input of the model at any given timestep consists of not only the MFCCs at that specific frame, but also those of the nine preceding and following frames. The model input thus covers a range of 19 MFCC frames (392 ms). Additionally, the LSTM layer can retain information over multiple timesteps, allowing the LSTM model to encode longer temporal relationships. In the CNN·Transformer model, temporal information between latent speech representations are handled by the Transformer architecture, which uses an attention mechanism to relate inputs from different timesteps. In theory, Transformer models can span temporal relationships of infinite duration, but in practice the CNN·Transformer used here was trained using sequences of at most 20 s.

3. THE HUMANLIKEHEARING TOOLBOX

As much as psychometric experiments may reveal information about the internal workings of the tested ASR systems, the results in the following chapters remain limited to the three ASR systems that were tested. To further motivate the cross-fertilization between the fields of machine learning and auditory neuroscience beyond the outcomes of these

studies, I have released the open source toolbox *HumanlikeHearing*^{*}. In contrast to FADE (Schädler et al., 2015), which can be used to train an ASR system to perform well in a specific psychometric experiment, the HumanlikeHearing toolbox allows researchers to assess externally developed ASR systems and add further psychometric tests. The ASR systems tested in the remainder of this thesis serve as baseline performance for future work. As of the time of writing, the toolbox incorporates a battery of seven tests that measure psychometric curves under a range of distortions. The tests were chosen such that each test measures a conceptually different aspect of hearing (e.g. spectral content, TFS and SPIN robustness), and so that each relies on sentence- or word-based speech recognition. Phoneme and syllable recognition tasks were excluded because many ASR systems (which are often trained to predict full sentences) perform poorly on such short nonsensical segments. The final set of tests can be roughly divided into distortion and SPIN experiments, which will be described in detail in Chapter 6 and Chapter 7, respectively. Out of the box, the toolbox supports the three ASR systems described in this chapter. Adding support for a new ASR system only requires the ASR system to be callable in Python.

4. SPEECH DATA SETS

To assess the performance of the ASR systems, a speech data set was selected that closely resembles the type of data used in psychometric testing, while also being freely available for future researchers to use. Furthermore, to prevent an overestimation of performance, it is imperative that the test data set is not incorporated in the training data of the ASR system.

Based on these criteria, the main dataset selected for the experiments described in the following chapters is the ARU speech corpus (C. Hopkins et al., 2019), a freely available dataset which consists of recordings of the IEEE (Harvard) sentences (IEEE, 1969) spoken by twelve adult native British English speakers in anechoic conditions. The IEEE sentences are commonly used in auditory tasks because they are designed to be phonetically balanced that use specific phonemes at the same frequency they appear in English. They do not contain a strong semantic signal: the sentences are grammatically correct but make little to no semantic sense. All sentences are of approximately the same length and short so that they can easily be repeated by human participants. All of the experiments measure performance

^{*} More information on installation and usage can be found on <https://github.com/neural-reckoning/humanlikehearing>

as percentage of keywords correct. To automatically extract keywords from any data set, part-of-speech (POS) tags were determined using the Natural Language Toolkit (NLTK, Loper and Bird, 2002). Then, a heuristic was applied in which groups such as nouns and verbs were included as keywords, whilst ignoring auxiliary verbs and forms of *to be*. For the IEEE dataset, this approach led to the selection of approximately five keywords per sentence.

The three ASR models themselves were trained on publicly available speech data. The CNN·Transformer model was pre-trained on the LibriSpeech dataset, a corpus of approximately 1000 h of 16 kHz read English speech derived from audiobooks of the LibriVox project (Panayotov et al., 2015). It was finetuned on a subset of 100 h labeled clean speech. In contrast to the CNN·Transformer, which was published as a state-of-the-art benchmark on the LibriSpeech dataset, the DNN·HMM and LSTM models were published as part of open source projects for real ASR applications. To improve model performance in speech recognition outside of the audiobook domain, these models were trained using the full LibriSpeech dataset together with other publicly available English training data, leading to a total of 3000 h and 3800 h for the DNN·HMM and LSTM, respectively.

To the best of my knowledge, none of these publicly available datasets incorporate IEEE sentences. Note that the toolbox also supports usage of the LibriSpeech data set. However, it should be stressed that if the ASR system was trained on LibriSpeech, it should only be tested using the test subset of the LibriSpeech data set to avoid an overestimation of the performance.

5. GENERAL METHODS FOR PSYCHOMETRIC TESTING OF ASR

In many ways, testing the performance of an ASR system on a free-form speech recognition test is not that different from testing humans. However, there are a few considerations that need to be taken into account when testing machines. In this section two of these, sound normalisation and SRT determination, will be discussed.

5.1 Sound Level Normalisation

For human testing, it is common to ensure that sound levels are at a comfortable level, usually around 65 dB for normal hearing humans. ASR systems, on the other hand, are not fed sounds of which the sound pressure levels can be measured. Instead, sound waves are transformed into a digital signal. The format in which the sound is stored can impact the

range of values the samples take on, and it is important to know which format the ASR system expects.

In the toolbox, sound files in a range of formats (such as WAV or FLAC files) are loaded as floating point values ranging mostly between -1 and 1. Under the assumption that these values represent the sound pressure level in Pa, the Root Mean Squared (RMS) and the Sound Pressure Level (SPL) in dB are computed as follows:

$$RMS = \sqrt{\frac{1}{n} \sum_t (s_t - \mu_s)^2} \quad (5.1)$$

$$SPL = 20 * \log_{10}\left(\frac{RMS}{2 \cdot 10^{-5}}\right) \text{ dB} \quad (5.2)$$

Here s_t represents the sample of a signal s at time t , n reflects the total number of samples in s and μ_s is the average value of all samples in s . The factor $2 \cdot 10^{-5}$ reflects the reference pressure of the smallest sound humans can hear in Pa.

In addition to this basic RMS-based sound level, the toolbox offers more advanced ways of estimating sound levels. Firstly, a-weighting can be applied to accommodate for the fact that the perceived loudness of a sound given a certain acoustic pressure varies depending on the frequency of the sound. Secondly, the IEEE P56 recommendation has been implemented. This implementation adjusts the estimated speech level based on a speech activity detector. This sound level estimation mitigates the fact that in the normal RMS computation segments of silence may skew the SPL to look lower than it actually is. For example, the RMS from Equation 5.1 can be made much smaller by simply appending zeros to the original sound, even though this would have no effect on the perceived loudness of the parts of the audio that contain speech.

Based on the computed sound levels, a preferred Signal to Noise Ratio (SNR) can be achieved by varying either the noise or the speech level. The level is adjusted by multiplying each sample in the sound with a gain that is computed as follows:

$$gain = 10^{(SPL_{new} - SPL_{old})/20} \quad (5.3)$$

Here, SPL_{old} is the current sound level and SPL_{new} is the target sound level.

In humans, setting the level to around 65 dB will result in a comfortable loudness. To determine the preferred loudness of the ASR system that is being tested, the toolbox includes a ‘normalisation test’ that measures ASR performance of clean speech as a function

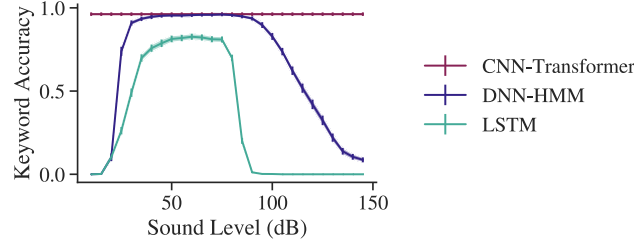


Figure 5.3. Sound level sensitivity of the three different ASR systems (CNN-Transformer, DNN-HMM and LSTM) tested on the first ten lists of the ARU dataset, measured as the keyword accuracy of sentences presented at a range of sound levels.

of sound level in dB. Applying this test to 100 sentences shows that the CNN-Transformer (which normalises input sounds) performs well regardless of sound level (Figure 5.3), whereas the LSTM and DNN-HMM only perform well within a range of 50 dB to 80 dB and 30 dB to 90 dB, respectively. Based on these results, the speech levels in all experiments in the following two chapters are held constant at 65 dB, and the noise levels are varied accordingly. Note that in all cases, sound levels are calibrated per individual sentence.

5.2 Speech Reception Threshold Determination

The performance of listeners in a speech in noise experiment are often summarised using the Speech Reception Threshold (SRT), which refers to the SNR at which a certain level of accuracy (usually 50% or 70.7%) is achieved. The SRT is determined by fitting a psychometric function (usually of sigmoidal shape) to a set of SNR values and accuracy pairs.

The Psychometric Function

The psychometric function $p = \Psi(x)$ models the relationship between the stimulus intensity x (the SNR value in our case) and the probability of a correct response p (the expected accuracy) and is assumed to have the following general form:

$$\Psi(x, \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)f(x; \alpha, \beta) \quad (5.4)$$

Here, $f(x; \alpha, \beta)$ is typically a sigmoid function such as a probit or logit that ranges between 0 and 1, where α and β denote the midpoint and slope of the function. The value of γ denotes the base rate of performance in absence of a signal. In forced-choice paradigms, γ is set to chance-level, but in free-form speech recognition it is often assumed to be 0. The upper bound of the psychometric function, i.e. the performance level for an arbitrarily large

stimulus signal, is given by $1 - \lambda$. In SRT experiments, this denotes the maximum accuracy achieved when no noise is present (i.e. when the SNR is very high). To fit the psychometric function to a given set of samples, the toolbox uses the Maximum Likelihood Estimation (MLE) as described in Wichmann and Hill (2001). As noted in Wichmann and Hill (2001), fixing the lapse rate λ to 0.0 can lead to biased parameter estimations as the MLE procedure is very sensitive to stimulus-independent errors. When estimating the psychometric functions, it is best to leave λ as a free parameter or to set a prior that encourages λ to be around the performance of the ASR system measured in quiet. Furthermore, the dataset should include at least one sample at a high probability (e.g. $p > 0.95$) Wichmann and Hill (2001), so it is important to ensure at least some of the sampled SNR/accuracy pairs are near this performance level.

Sampling Regime

To reduce the number of trials per condition in humans, data points are usually collected through an adaptive sampling procedure in which the SNRs are sampled around the threshold of interest (e.g. 50 % accuracy). For ASR systems, which unlike humans do not suffer from attention fatigue or sentence memorisation, the psychometric curves can be measured more precisely through a grid search that tests every sentence on a range of SNRs, which I will refer to as the static SRT procedure. Both the adaptive and static procedures are implemented in the toolbox.

In adaptive procedures, the SNR level is first set to a level at which performance is expected to be good. This point can be estimated adaptively by presenting the first sentence at a very low SNR and increasing the SNR until maximum accuracy is achieved. In humans, this is often assumed to be 100%. However, as it is not guaranteed that an ASR system can achieve 100% performance, the maximum accuracy is set to the accuracy in quiet in the toolbox. For subsequent sentences, the SNR is either increased (or decreased) depending on whether the performance was worse (or better) than the target threshold. The step size of the SNR changes are typically reduced after a given number of ‘reversals’. For example, in the original competing talker experiment (Rosen et al., 2013; see Chapter 7), the step size is reduced from 4 dB to 2 dB once the direction of SNR change has reversed twice. As a simpler alternative to MLE, ‘staircase’ sampling methods are sometimes used (Leek, 2001). Here, a response is assumed to be either positive or negative. To target SRT-50, a one-up one-down procedure is used, where a positive trial results in a decrease of SNR and a negative trial results in an increase. To target SRT-70.7, a two-down, one-up procedure is used, where

two positive responses are required to result in a lowering of the SNR. The SRT can then be estimated by taking the mean of all the SNR values at the last few reversals. Note that more recently, Bayesian approaches to adaptive SRT estimation have been proposed. For example, BASIE (Bayesian Adaptive Speech Intelligibility Estimation) (Gaubitch et al., 2010) is a Bayesian method that is able to achieve SRT estimates accurate to within ± 1 dB in under 30 iterations.

In the static SRT procedure, a reasonable range of SNR values can be determined through a pilot study using a small set of sentences. In contrast to the adaptive procedure, which targets SNR values around the SRT, the static procedure obtains data points over the whole SNR range and can thus be used to determine the SRT at any performance threshold. However, as it collects more data points it is much slower than the adaptive procedure. All plots in the remainder of this work are measured using the static grid-search sampling method. Provided the performance in quiet was higher than the SRT accuracy level, I found that the adaptive sampling method gave very similar results, albeit with higher variance.

6. CONCLUSIONS

In this chapter, I have laid out the basic components that form the foundations of the upcoming chapters. First, I introduced the three ASR systems that will be put to the test in the next three chapters. This included a review of the differences in temporal and spectral limits between these models, which result from their differences in input and architecture. This was followed by an introduction to HumanlikeHearing, an open source Python toolbox that allows researchers assess their own ASR systems, and a description of the speech data sets used in the experiments in the upcoming chapters. Finally, some details were provided about certain methodologies that are specific to applying psychometric tests to ASR systems. In the upcoming chapter, I will explore the robustness of the three ASR systems to a range of distortions, revealing their use (or lack thereof) of a range of well-known auditory cues.

CHAPTER 6

Distortion Sensitivity of ASR Systems

1. INTRODUCTION

As discussed in more detail in Chapter 5, there has recently been an increased interest in using Automatic Speech Recognition (ASR) systems as a model for human hearing. For an ASR system to be used as a model of human hearing, it is imperative that the system behaves as similarly as possible to human speech recognition. However, speech waves are inherently redundant. This implies that the same performance may be achieved using different strategies and auditory cues. As described in Chapter 2, any given phone or phone transition may be associated with a myriad of acoustic cues, of which only a few may be sufficient for accurate recognition. Combined with syntactic and semantic knowledge, the inherent redundancy of speech signals allows for the incredible robustness of human hearing. A common experimental approach used to assess the importance of certain specific auditory cues is to remove or distort certain features and to determine the effect their absence has on speech intelligibility (Moore, 2013). These types of experiments have shown that even under distortions so severe that poor performance may have been expected, human speech recognition often remains surprisingly accurate.

In this chapter, the qualitative similarity between human speech recognition and ASR is explored through the application of a range of distortion experiments to three open source ASR systems. The experiments included in this chapter each cover a different facet of speech, including spectral content, modulations, Temporal Fine Structure (TFS) and periodicity. I will show that although there are certain similarities between human and machine hearing, many auditory cues are used differently or not at all by the tested ASR systems. In the next chapter, this analysis will be extended to the disrupting effects of specific types of additive noise. All experiments in this and the following chapter have been made freely available as part of the HumanlikeHearing toolbox (see Section 3 in Chapter 5).

As part of this study, the following freely available ASR models have been compared:

1. The Kaldi nnet3 chain model from the Kaldi Active Grammar project (Zurow et al., 2021) uses a hybrid Deep Neural Network (DNN) and Hidden Markov Model (HMM) (**DNN·HMM**; Povey et al., 2018). It takes as input Mel-Frequency Cepstral Coefficient (MFCC) features, a classic set of features inspired by the human auditory system in which time is divided into short, overlapping ‘frames’, and the spectral envelope of the sound is computed for each frame. At each layer of the network, several frames of output from the previous layer are provided as input to give temporal context.
2. Mozilla’s DeepSpeech model (Mozilla, 2020) is a Long-Short-Term-Memory (**LSTM**) neural network, a widely used type of neural network that is designed to model long temporal relationships. It is a variant of the original DeepSpeech model that was introduced by Hannun et al. (2014). Like the DNN·HMM model, it uses MFCC features as its input. However, the MFCC features in the LSTM model are only based on spectral information up to 4 kHz, whereas the DNN·HMM uses spectral representations of up to 8 kHz.
3. Facebook’s fairseq Wav2Vec 2.0 (Baevski et al., 2020) is a **CNN·Transformer** model that, in contrast to the previous two models, takes raw audio as input. Features are extracted using a Convolutional Neural Network (CNN) that convolves the input signal over time. Temporal context is modelled using a Transformer architecture (Vaswani et al., 2017), which uses an attention mechanism to relate inputs from different timesteps.

Each of these systems uses a different approach to model the temporal component of speech. The input and architecture of each of them is described in more detail in Chapter 5.

All experiments discussed here were performed in a mismatched condition. That is, all models were trained using (mostly) clean speech, but are tested on data that is manipulated in certain acoustic dimensions. Unsurprisingly, the ASR systems tend to have poorer performance in noise compared to humans (see Chapter 7). For some of these experiments (e.g. modulations and TFS), the importance of the tested auditory cue only shows up in humans when background noise is present. Where appropriate, I corrected for poor speech in noise performance of the ASR systems by allowing higher Signal to Noise Ratios (SNRs).

The remainder of this chapter is divided into separate sections discussing five different psychometric experiments. Each section will describe the motivation and method. Furthermore, the performance of the three ASR systems will be compared with previously reported human performance on the same experimental paradigm. Finally, the results of all experiments are collated and compared in the concluding section.

2. SENSITIVITY TO BANDPASS FILTERING

Although humans can perceive frequencies from 20 Hz up to 20 kHz, even a small frequency band can be sufficient to achieve high intelligibility rates in quiet. For example, Warren et al. (2000) show that by using pass bands with a near vertical slope, high intelligibility rates are achieved when all but the 1 kHz to 2 kHz region is filtered out. Here, I investigate whether ASR systems can similarly make use of this ‘spectral redundancy’ in speech.

2.1 Method

To simulate the experiment presented in Warren et al. (2000), each sentence is filtered using a bandpass filter centered at 1500 Hz. The 2000-order Finite Impulse Response (FIR) filter used has a steep slope of 1000 dB/octave. This steep slope removes the presence of transition bands, which may otherwise leak information from other frequency regions. In Warren et al. (2000), the bandwidth varies in 1-semitone steps with values ranging from 3 semitone up to 12 semitones, but this range was extended to 80 semitones to ensure all ASR systems reached ceiling-level performance.

Instead of the CID Everyday Sentences corpus (Silverman & Hirsh, 1955) used in Warren et al. (2000), which is not freely available, the IEEE sentence lists were used here. These are of similar length as the CID sentences, and in both corpora each list of 10 sentences contains 50 keywords. However, CID sentences are more semantically meaningful than IEEE sentences, which makes it easier for human listeners to guess unintelligible words. A lack of semantic context in bandpass filtered speech has been reported to lead to an accuracy reduction of around 20% (Stickney & Assmann, 2001). When considering words in isolation, i.e. without grammatical context, performance drops a further 23%. The intelligibility of bandpass filtered IEEE sentences in humans is thus expected to be lower than that of the CID sentences. Although the language model in the ASR systems allows the models to utilise some semantic context, it is unlikely they are able to exploit this to the same extent as humans can. It is therefore expected that the ASR performance for IEEE and CID sentences would be relatively similar.

2.2 Detailed Comparison

Compared to humans, the ASR systems require a much wider spectral range to perform well (Figure 6.1a). At the width at which humans have near-ceiling accuracy (12 semitones), the average accuracy is still very poor for all ASR systems (0-10% accuracy). Note that

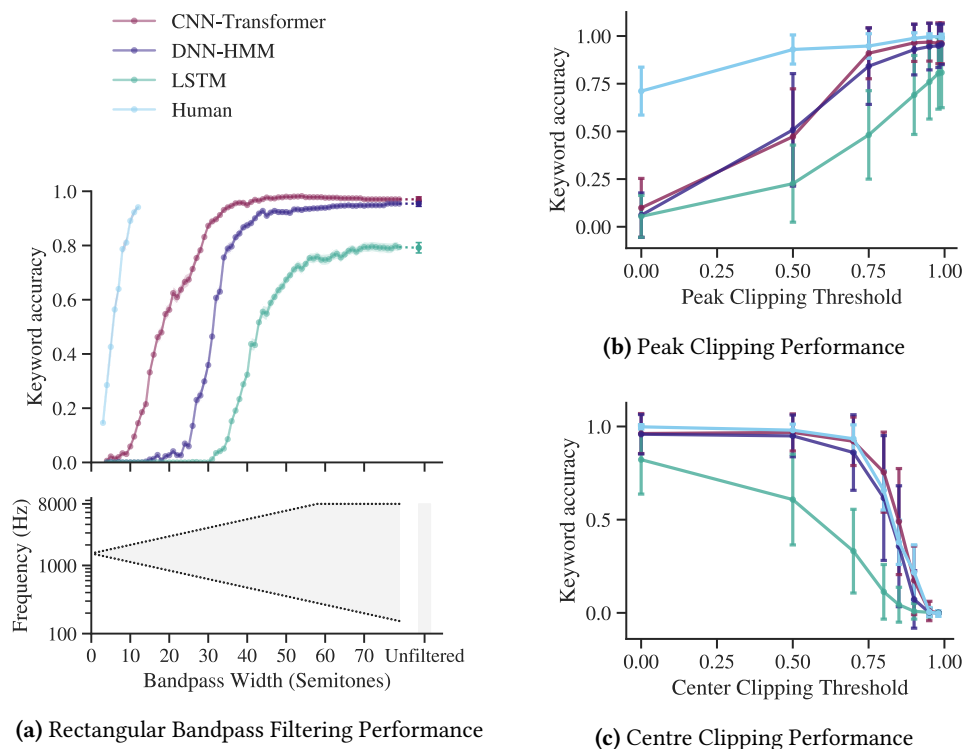


Figure 6.1. The performance of the three ASR systems compared to human performance on three types of distortions. Keyword accuracy denotes the percentage of correct keywords measured for 100 IEEE sentences (for the ASR systems). (a) Accuracy obtained when each sentence was bandpass filtered around 1500 Hz with a width that varied in semitones. Error bars denote the standard error of the mean. Below are the upper and lower cutoff frequencies of the filters. Human data from Warren et al. (2000) on the CID Everyday Sentence corpus (Silverman & Hirsh, 1955). (b) Performance in peak clipping experiment. The peak clipping threshold t denotes the percentage of the sentence samples that is not clipped (i.e. at $t = 0$ the sound is reduced to its signs). Error bars show one standard deviation. Human data from Kates and Arehart (2005). (c) As in b, but now using centre clipping, which means that all samples that are *lower* than the clipping value are set to zero. The centre clipping threshold denotes the percentage of sentence samples that have *not* been clipped, i.e. at $t = 0$ the original signal is retained. Error bars show one standard deviation. Human data from Kates and Arehart (2005) using HINT sentences (Nilsson et al., 1994).

this performance is too low to be fully explained by the usage of the slightly more difficult speech corpus.

There is a large variation between the ASR systems themselves. Despite the similar performance of the CNN-Transformer and DNN-HMM for unfiltered sounds, the CNN-Transformer is much more robust to bandpass filtering than the DNN-HMM. The LSTM, which has a lower accuracy in quiet than the two other models, is the least robust against bandpass filtering, with performance tapering off to near ceiling-level around 60 semitones

compared to approximately 35 and 45 for the CNN-Transformer and DNN-HMM, respectively. The MFCC input features of the LSTM model only retain spectral information up to 4 kHz (see Section 2 in Chapter 5), which is reached by the bandpass filter at 34 semitones. At this level of filtering, performance is still close to zero for the LSTM, and performance does not increase drastically until around 40 semitones. This suggests that low frequency information between 40-60 semitones (265 Hz to 472 Hz) is particularly important for this model. The other two models, particularly the CNN-Transformer model, reach ceiling level well before such low frequency information is available.

3. PEAK AND CENTRE CLIPPING

Communication systems such as microphones or speakers can introduce a variety of nonlinear distortions. For example, when a microphone senses very loud sounds it may saturate, resulting in the clipping of high amplitude pressure waves. This is referred to as peak clipping, a distortion that introduces additional higher harmonics as well as intermodulation products. Perhaps surprisingly, human speech recognition is robust against peak clipping distortions. Even when a speech signal is reduced to its signs (each sample replaced by +1 if positive or -1 if negative), normal hearing humans retain between 70% and 90% of intelligibility (Kates & Arehart, 2005; J. C. R. Licklider & Pollack, 1948). A related distortion is centre clipping, which refers to the reduction of low-level amplitudes. Historically, centre clipping was a technique used for noise-suppression, although it is no longer commonly applied in practice. In humans, a small amount of centre clipping is not detrimental. However, higher levels of centre clipping affect intelligibility considerably, as weak consonants (which tend to have lower amplitudes than vowels) are stripped out (Kates & Arehart, 2005; J. C. Licklider, 1946).

3.1 Methods

To study the robustness of the ASR systems to both peak and centre clipping, the experimental setup presented in Kates and Arehart (2005) was implemented. In this study, clipping distortions are parameterised by a threshold t which ranges between 0 and 1. To determine the clipping value c associated with this threshold, silence intervals at the start and end of the sentences are removed and a histogram of the magnitudes of the signal samples is computed. The clipping value c is set to the percentage t of the cumulative magnitude histogram for the sentence. Given the clipping value c , the peak clipping operation is given

by:

$$f_i = \begin{cases} c, & \text{if } x_i > c \\ x_i, & \text{if } -c \leq x_i \leq c \\ -c, & \text{if } x_i < -c \end{cases} \quad (6.1)$$

Here, x_i is the i th sample of the speech signal, and f_i the distorted output sample. The centre clipping operation is defined as follows:

$$f_i = \begin{cases} x_i, & \text{if } x_i > c \\ 0, & \text{if } -c \leq x_i \leq c \\ x_i, & \text{if } x_i < -c \end{cases} \quad (6.2)$$

The distorted sentences are then readjusted to a sound pressure level of 65 dB. As in Kates and Arehart (2005), the peak clipping thresholds were set to 0%, 50%, 75%, 90%, 95%, 98%, 99% and 100% of the cumulative magnitude histogram of each sentence. The centre clipping thresholds were set to 0%, 50%, 70%, 80%, 90%, 95%, 98%. In Kates and Arehart (2005), each of the thirteen listeners were presented with approximately 10 sentences for each threshold condition, giving a total of 130 HINT sentences (Nilsson et al., 1994) per threshold condition. In the present study, 100 IEEE sentences were used per condition. The IEEE and HINT sentences are of similar length, but the IEEE sentences provide less semantic context. This difference should not affect the results too much, as similar intelligibility rates have been measured for clipped isolated words without grammatical or semantic context (J. C. R. Licklider & Pollack, 1948).

3.2 Detailed Comparison

In both the peak and centre clipping experiments, the CNN·Transformer and DNN·HMM display similar robustness to one another, whereas the LSTM model performs much worse at all clipping thresholds (Figure 6.1b and 6.1c). This is in contrast to the bandpass filter experiment (Section 2), where the CNN·Transformer outperformed the DNN·HMM. In the peak clipping experiment, none of the ASR models perform as well as humans (Figure 6.1b). Although humans will have heard some amount of peak-clipping often throughout their lives, the more extreme levels of peak clipping are much less likely to occur in practice. This suggests that the observed large discrepancy at extreme peak clipping levels (e.g. $t = 0.0$) cannot be fully explained by previous exposure. At a peak clipping threshold of 0.5, the performance of humans is near ceiling, whereas the accuracy of the DNN·HMM

and CNN-Transformer has dropped to around 50%. At infinite peak clipping ($t = 0.0$), the performance of the latter two systems hovers around a mere 10%, compared to 72% accuracy in humans. In the centre clipping experiment, both the CNN-Transformer and DNN-HMM intelligibility rates are on par with human performance (Figure 6.1c), whereas the LSTM performance remains worse. This suggests that the CNN-Transformer and DNN-HMM, but not the LSTM, are robust against changes in the distribution of low-level amplitude parts of the speech signal.

Previous studies have reported similar detrimental effects of peak clipping on ASR performance. For example, Tachioka et al. (2014) reported 20% accuracy for a large vocabulary task (20% at $t = 0.1$), although the system in this study (a combined Gaussian Mixture Model (GMM) and HMM model) had a low ceiling performance (60% at $t = 0.9$). Malek et al. (2016) reported an increase in Word Error Rate (WER) from around 10 % for undistorted speech to up to around 50 % for speech that was distorted through nonlinear amplification and potential clipping. This effect was observed across a range of different features, including MFCCs, although ‘primitive’ Filter Bank Coefficients (FBCs) (Mel-scaled spectrograms with no cosine transforms) appeared slightly more robust.

Further research is required to understand the differences between the auditory features used by humans versus these three ASR systems that may underly the observed discrepancies in performance under peak-clipping distortions. On a rudimentary level, the poor performance may be influenced by the change in distribution of the acoustic waveform that peak clipping introduces. For the end-to-end CNN-Transformer, peak clipping affects the distribution of the input values the model observes, particularly after normalisation. For MFCCs (the input to the DNN-HMM and LSTM) peak clipping results in an increasing shift of the mean and reduction in covariance of the MFCC coefficients (Poorjam et al., 2017). Alternatively, poor performance on peak clipped audio may be the result of the specific way in which peak clipping distorts the spectral envelope, as it introduces extra energy at harmonics in the frequency spectrum. Since the ASR systems have a different response to changes in periodicity than humans do (see Section 6), other changes in harmonic structure may affect them differently too. Lastly, it should be noted that peak clipping specifically retains certain temporal information (e.g. zero-crossings). The inability of the tested ASR systems to make use of TFS (see Section 5) may therefore also contribute to their poor performance on peak clipped audio.

4. SPECTRAL AND TEMPORAL MODULATIONS

Speech is characterised by fluctuations of power in both time and frequency, which are referred to as temporal and spectral modulations. As shown in Section 2, speech can remain intelligible even after drastic degradations in spectral content, depending on which frequencies are taken out. Conversely, it has been shown that even with very coarse temporal information, speech remains intelligible in the presence of sufficient spectral information (Arai & Greenberg, 1998; Arai et al., 1999; Drullman et al., 1994).

To disentangle the importance of spectral versus temporal modulations, Elliott and Theunissen (2009) used a modulation filtering technique to remove specific spectral modulations (given in cycles/kHz) and temporal modulations (given in Hz). It was shown that in humans, most modulations can be removed without affecting speech intelligibility when presented in quiet. However, under noisy conditions temporal modulations of 1 Hz to 7 Hz and spectral modulations of 0 cycles/kHz to 1 cycle/kHz are necessary to obtain high intelligibility rates. In this section I explore to what extent ASR systems rely on modulations by following the notch filter experiment described in Elliott and Theunissen (2009).

4.1 Method

In the experiment described in Elliott and Theunissen (2009), word accuracy is measured for sounds in which specific temporal and spectral modulations have been filtered out. Here, performance was measured in white noise with an SNR of 2 dB (as in Elliott and Theunissen, 2009) as well as at the lowest SNR at which near-ceiling performance was reached without any modulation filtering (10 dB, 15 dB, 25 dB for the CNN-Transformer, DNN-HMM and LSTM respectively (I will refer to these as the high-SNR regime).

The temporal or spectral modulations of a signal can be extracted by computing the Modulation Power Spectrum (MPS), a 2D representation of a signal in which each axis represents either spectral or temporal modulations. To remove specific modulations from a target speech signal, the signal is transformed into the MPS, the desired modulations are set to zero, and the MPS is inverted back into speech. Following the procedure described in Elliott and Theunissen (2009), the MPS is derived by first computing a spectrogram using Gaussian windows on a linear frequency scale (the choice of a linear rather than log-spaced spectrogram is why results are reported in cycles/kHz rather than the octaves/kHz). The modulation power and phase spectrum are obtained from the log of the spectrogram using the 2D Fast Fourier Transform (FFT). The required modulations are set to zero in the power spectrum and randomised in the phase spectrum. To invert the filtered modulation spectrum

back into a sound, the inverse 2D FFT is applied to obtain the filtered log spectrogram. This spectrogram is exponentiated and inverted using the Griffin-Lim iterative spectrogram inversion algorithm (Griffin & Lim, 1983) to obtain the filtered speech.

To assess the importance of specific spectral and temporal regions, the notch experiment presented in Elliott and Theunissen (2009) was reproduced. In this experiment, speech filtered with a total of ten temporal and nine spectral modulations is embedded in Gaussian white noise with 2 dB SNR. In about half of the notches, all modulations outside of the ‘core’ region (set to spectral modulations up to 3.75 cycles/kHz and temporal modulations up to 7.75 Hz) are removed before removing the notch. Performance is also measured against a control condition (spectrogram inversion without modulation filtering) and the core region.

In Elliott and Theunissen (2009), each of the 17 listeners were presented with 100 sentences that were randomly assigned to a condition. Given 21 conditions (nine spectral, ten temporal notches, control and core), an average of 81 sentences total (five per listener) were presented per condition. The results presented here are based on 100 sentences per condition.

Stimuli in Elliott and Theunissen (2009) are taken from the soundtrack of the Iowa Audiovisual Speech Perception videotape (Tyler et al., 1986), which are of similar length but provide less semantic context compared to the IEEE sentences used here. Furthermore, to control for the poor robustness against white noise of the ASR systems, an additional noise condition was tested in which the SNR was set to the lowest SNR at which the ASR system achieved roughly ceiling performance (10 dB, 15 dB, 25 dB for the CNN-Transformer, DNN-HMM and LSTM, respectively). In Elliott and Theunissen (2009), SNRs were set relative to a constant noise level of 65 dB. To prevent the overall sound level to exceed the preferred ranges of the ASR systems when testing high SNR levels (see Chapter 5 Section 5.1 for details), I chose to set SNRs relative to a constant speech level of 65 dB instead.

4.2 Detailed Comparison

As with the bandpass experiment, the CNN-Transformer model is most robust against degradations, followed by the DNN-HMM model and the LSTM model (Figure 6.2). Although absolute performance levels vary depending on the ASR system and noise level, the relative importance of the spectral and temporal modulation regions are mostly consistent with the regions that are important for human speech recognition. This is particularly evident when the SNR levels are adjusted for near-ceiling performance level (the high-SNR regime, dotted lines in Figure 6.2).

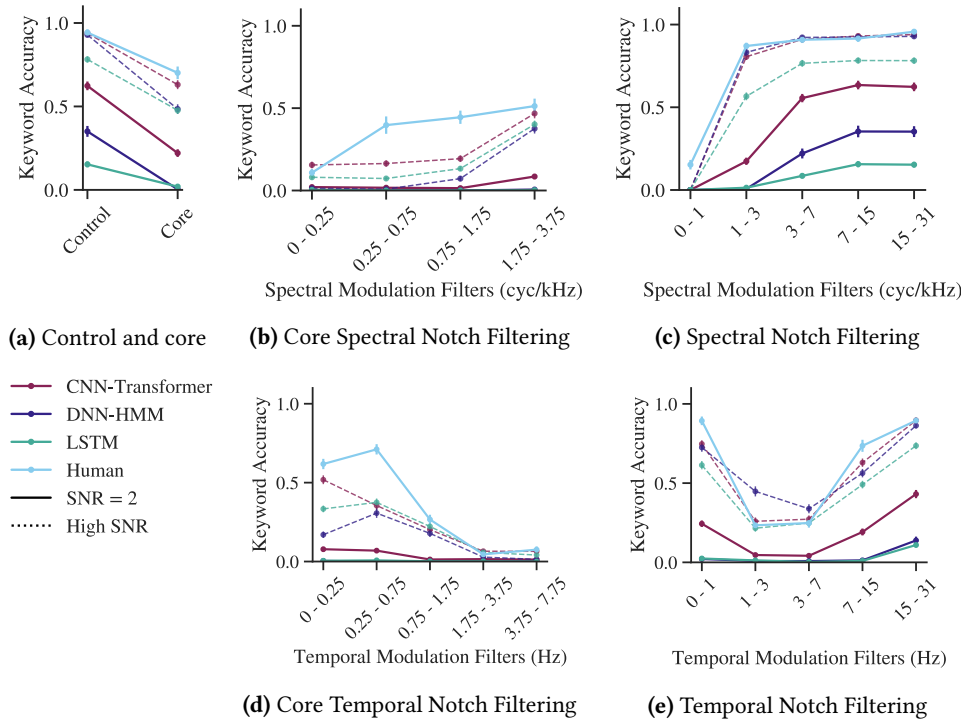


Figure 6.2. The performance of the three ASR systems in the spectral and temporal modulation filter experiment. Either spectral or temporal modulations are removed by computing the modulation power and phase spectra, setting specific regions to 0 and inverting. Performance is measured as percentage of keywords correct in 100 sentences embedded in white noise with a low (solid lines) or high (dashed lines) signal to noise ratio. Note that the maximum performance in quiet of the LSTM was only 80%. Error bars are plotted as the standard error of the mean. Human reference from Elliott and Theunissen (2009) using the Iowa Audiovisual Speech Perception videotape (Tyler et al., 1986). **(a)** Performance in the control condition (spectrogram inversion without modulation filtering) and core region condition (removal of all spectral modulations above 3.75 cycles/kHz and temporal modulations above 7.75 Hz). **(b d)** Performance when, in addition to removal of all modulations outside of the core region, either spectral modulations (b) or temporal modulations (d) in the given regions are removed. **(c, e)** Performance when spectral (c) or temporal (e) modulations in the given regions are removed.

Similar to humans, ASR performance is most disrupted by the removal of low spectral modulations in the 0-1 cycles/kHz region, after which performance monotonically increases as the notch region covers higher spectral modulations (Figure 6.2c). In humans, performance in all other regions is near ceiling, but in the ASR systems the 1-3 cycles/kHz region is more important than the higher regions. The 0-1 cycles/kHz and 1-3 cycles/kHz regions roughly correspond to vowel formant separation and formant transitions in English (Elliott & Theunissen, 2009).

In both the ASR systems and humans, temporal notches in the region of 1-7 Hz degrade intelligibility most, followed by notches in the region of 7-15 Hz (Figure 6.2e). The regions roughly correspond to syllable and phone rates, which respectively fluctuate around 2 Hz to 5 Hz (Pickett, 1999) and 15 Hz to 30 Hz (Greenberg et al., 1999; A. Liberman, 1970). In humans, the removal of both very slow (0-1 Hz) and fast (15-31 Hz) temporal modulations does not lead to a significant reduction of intelligibility compared to unfiltered speech (control). For all ASR systems, on the other hand, the 0-1 Hz range does lead to a reduction in performance. Filtering these slow temporal modulations leads to fluctuating sound levels across the sentence with a rate of up to 1 Hz. The results thus suggest that the ASR systems, unlike humans, are not able to adaptively respond to changes in sound level.

When all modulations outside of the ‘core’ modulations (defined as up to 3.75 cycles/kHz and 7.75 Hz) are removed, humans retain more than 70% accuracy under a 2 dB SNR (Figure 6.2a). At the same SNR, ASR performance drops to much lower accuracy. In the higher SNR regime, the core region performance of the ASR systems is more comparable to that of humans. The results based on additional filtering in the core region provide a more refined view of the differences in sensitivity to slow spectral and temporal modulations. This shows that the ASR systems are more sensitive to spectral modulations in the 0.25-1.75 cycles/kHz region (Figure 6.2b). Furthermore, in humans the effects of temporal modulations in the 0.25-0.75 Hz region are not significantly different from the 0.0-0.25 Hz region (Figure 6.2d). For the CNN-Transformer, there is a higher sensitivity to the 0.0-0.25 Hz range, whereas the shape of the DNN-HMM and LSTM are more comparable to that of humans.

In summary, the three ASR systems appear to be less robust against the white noise masker than humans are, but display roughly the same relative importance of spectral and temporal modulations. The similarity between human and ASR results contribute to the universality of the ‘speech modulation transfer function’ proposed in Elliott and Theunissen (2009). Furthermore, despite the use of a linear frequency scale in the computation of the MPS in Elliott and Theunissen (2009) (which differs from the approximate log scale at which both the human auditory periphery and MFCC feature extraction operate), these results further emphasise the importance of the use of ASR input features that retain (whether explicitly or implicitly) critical temporal and spectral modulations.

5. TEMPORAL FINE STRUCTURE

In the human auditory system, incoming sounds are decomposed into a set of narrow frequency bands by the cochlea. The responses in each frequency channel can be described

through two components that function at different time scales. The slower component, the temporal envelope, reflects slow fluctuations in amplitude, whereas the rapid fluctuations of the narrowband carrier are referred to as TFS. In quiet listening conditions, speech that only retains channel envelope information is highly intelligible. However, TFS is thought to play a more important role in challenging listening environments. For example, K. Hopkins and Moore (2010) showed that for normal hearing listeners TFS information contributes significantly to the perception of speech in the presence of a competing talker, particularly for channels with centre frequencies below 1 kHz. Such a benefit was not observed in hearing-impaired listeners.

As MFCC features do not retain a detailed timing structure, the use of TFS information in the DNN-HMM and LSTM is likely very limited. The CNN-Transformer model, on the other hand, would at least in principle be able to leverage TFS information. However, considering that the CNN-Transformer model was not trained in noise, it may still have learned to solely rely on envelope cues. To test these hypotheses, the two experiments described in K. Hopkins and Moore (2010) were implemented. The first experiment explores the impact of cumulatively adding TFS starting from high or low frequencies (referred to as the *TFS-low* and *TFS-high* condition, respectively). The second experiment measures the impact of removing envelope and TFS information in specific spectral regions.

5.1 Method

In the two experiments described in K. Hopkins and Moore (2010), TFS information is investigated in a cumulative as well as region-specific manner. Both experiments use a vocoding paradigm in which the speech is filtered into a range of frequency bands. Specifically, sounds are processed using a tone vocoder with 30 channels on an Equivalent Rectangular Bandwidth (ERB) scale (Glasberg & Moore, 1990) from 100 Hz to 8 kHz. The linear-phase FIR filters had a variable order, chosen such that the transition bands of each filter had similar slopes on a logarithmic frequency scale. Each filter was designed to have a response if -6 dB relative to the peak response at the frequencies at which its response intersected with the responses of two adjacent filters. The channels are divided into five regions, each spanning 6 ERB_N . To retain both TFS and envelope information, the channel is left unchanged. To remove only TFS information, a sine wave of the channel centre frequency is modulated with the channel envelope, which is extracted using a Hilbert transform. The modulated tone is then filtered again with the channel bandpass filter to ensure side bands are removed.

In the first experiment, TFS information is added successively either starting from

the low frequency region (*TFS-low*) or high frequency region (*TFS-high*). In the second experiment, all channels are tone vocoded except for one region. The channels in this region are processed in one of two ways. In the + condition, the channels remain unprocessed, thus retaining TFS. In the - condition, the channels are set to zero, thus removing both envelope and TFS information. Performance is also measured in the *allvoc* condition, in which all channels are tone vocoded. The benefit of the envelope in a specific region is computed as the difference between the Speech Reception Threshold (SRT) found in the *allvoc* and - condition, whereas the combined benefit of the envelope and TFS is the difference between the the - and + condition.

The SRT-70.1s of ASR systems were measured five times on 20 IEEE sentences (100 different sentences total) in six-speaker babble noise. A static SRT determination technique was used (see Chapter 5 Section 5.2) with SRTs ranging from -4 dB to 24 dB in 2 dB steps. As in K. Hopkins and Moore (2010), the target speech level is fixed at 65 dB.

To ensure consistency throughout this work, performance in the ASR experiments was measured against the IEEE dataset rather than the Coordinate Response Measure (CRM) data (Bolia et al., 2000) used in K. Hopkins and Moore (2009). In the CRM corpus, sentences consist of only three words that follow a common structure (*'Ready [call sign] go to [colour] [number] now.'*) and are selected from eight call signs, four colours and eight numbers. In such a closed-set task, it is much easier to achieve higher SRTs. This is particularly the case for humans, as they can be instructed to only pick words from the available set. For completeness, results on the CRM corpus are included in the appendix (see Appendix C).

Another way in which the ASR version of the experiment differs is the choice of masker. Rather than a single competing talker, a babble noise of six competing talkers was used. This was done to account for the fact that the ASR systems used here cannot attend to a specific speaker and, as will be shown in Chapter 7, are relatively much worse at recognising speech in a single competing talker scenario than humans are. The use of a different target and masker dataset is unlikely to lead to a qualitative difference in results. In a comparable experiment, SRT-50 curves for the *TFS-low* condition were qualitatively similar despite being measured against IEEE sentences in speech-shaped noise (K. Hopkins & Moore, 2009).

5.2 Detailed Comparison

In humans, an increase in TFS information (both in the *TFS-low* and *TFS-high* condition) leads to an increase in performance (reduction of SRT), particularly when TFS information is added to the low-frequency channels below 1 kHz (Figure 6.3). The overall SRT improvement



Figure 6.3. Performance in the two TFS experiments measured on 100 IEEE sentences (20 per SRT, 5 per condition) in six-speaker babble noise. In these experiments, TFS is either removed or retained in specific frequency channels using a tone vocoder. This vocoder initially filters speech in 30 frequency channels. To remove TFS from a given channel, the channel response is replaced by a tone that is modulated by the channel envelope. Human reference data, which was measured on the CRM task corpus (Bolia et al., 2000) in single-speaker babble noise, was taken from K. Hopkins and Moore (2010). Error bars are plotted at standard error of the mean. **(a-d)** In the first experiment, TFS information is added cumulatively starting either from the highest (*TFS-high*) or lowest (*TFS-low*) spectral region. The SRT-70.1s are plotted as a function of cut-off region for each of these conditions. Note that at both the highest cut-off frequency in *TFS-high* and lowest cut-off frequency in *TFS-low*, the whole sentence is vocoded. Conversely, TFS information across the whole frequency spectrum is present at the lowest *TFS-high* region and highest *TFS-low* region. **(e-h)** In the second experiment, TFS information was either removed or retained in a given spectral region, while all other regions were vocoded. A third condition was also measured in which the spectral region was completely removed. The benefit of adding envelope information was calculated as the difference between the SRT for the condition in which all regions were vocoded (i.e. TFS was removed in that region) and the SRT for the condition in which the spectral region was removed from a fully vocoded sentence. The benefit from adding envelope and TFS information together was calculated as the difference between SRT for the condition in which the spectral region was removed fully and the condition in which TFS was retained.

from speech that contains only envelope information to unprocessed channels is around 6.6 dB. The absolute SRTs observed in the ASR systems are much higher, which can partly be explained by the use of a different dataset and partly due to reduced robustness of ASR systems in noise. For all ASR systems, there is some improvement with the addition of TFS information, but the overall benefit is much smaller (around 0.5 dB, 1 dB, 2 dB for the CNN-Transformer, DNN-HMM and LSTM, respectively) and there does not appear to be a particular preference for low-frequency TFS regions. Given that the TFS benefits are so small, it is likely that the observed benefit is the result of the change in distribution of the input data more so than the utilisation of TFS cues.

In the second experiment, a comparison is made between the benefit of adding both envelope and TFS information or envelope information alone to a specific spectral region. The SRT benefit is determined by taking the difference in SRT between a vocoded sentence in which a specific spectral region is completely removed and the SRT of a sentence in which the envelope of that region is retained (*envelope benefit*) or in which both the envelope and TFS are retained (*envelope + TFS benefit*). In normal hearing listeners, the addition of TFS information leads to a benefit between 2 dB to 3 dB across all spectral regions. In contrast, the addition of TFS information does not lead to an improvement in performance in any of the ASR systems. This is consistent with the results of the first experiment, where no large difference was observed when adding TFS information from a single region. The results indicate that the spectral regions are too fine-grained to obtain an observable benefit in ASR systems.

The second experiment also exposes the differences in the relative importance of the spectral regions, in line with the results found in Section 2. In humans, the removal of envelopes in the high frequency regions (above 2 kHz) does not affect performance much. The ASR systems, on the other hand, display significant drops in performance when these regions are removed. Compared to the CNN-Transformer and the DNN-HMM, the LSTM model relies more heavily on the second, lower-frequency region (400 Hz to 1000 Hz), something that was also observed in 2. Note that the LSTM does not benefit from 4 kHz to 8 kHz since the input MFCCs to this model are based on spectral information up to 4 kHz.

In summary, in contrast to normal hearing humans, the ASR systems do not use TFS information and rely more heavily on the presence of envelope information across the entire frequency spectrum than humans do.

6. TARGET PERIODICITY

A speech sound generally consists of both periodic and aperiodic segments. Periodicity in speech is caused by periodic vibrations of the vocal cords, as can be heard in voiced speech. Aperiodicity in speech results from constrictions in the vocal tract, for example in fricatives such as /f/. Periodicity, which is closely associated with pitch, likely plays an important role in extracting certain linguistic features such as intonation. In a systematic comparison between vocoded speech with artificially varied levels of periodicity, it was shown that in quiet listening conditions, the presence of aperiodic segments is important for speech intelligibility (Steinmetzger & Rosen, 2015). In particular, noise-vocoded speech (which has no periodicity) is about as intelligible as speech with a natural balance of periodicity and aperiodicity (also called Dudley-vocoded speech), whereas fully periodic speech tends to be less intelligible. This effect is particularly noticeable when the spectral resolution is reduced due to a smaller number of vocoder channels. To further understand the use of periodicity in speech in ASR systems, the three ASR systems were tested on the same sentences that were used in the experiment of Steinmetzger and Rosen (2015).

6.1 Method

Following the procedure described in Steinmetzger and Rosen (2015), the intelligibility of three types of speech were investigated: aperiodic (noise-vocoded) speech, speech with a natural amount of source periodicity and fully periodic speech. The F_0 contours of the speech signal were extracted using PRAAT with a sampling rate of 100 Hz (Boersma, 2021). In the toolbox, the Python module Parselmouth is used to interface with PRAAT software (Jadoul et al., 2018).

To obtain the vocoded speech, the original recordings were first bandpass filtered into 6, 7, 8, 10, 12, 16 or 24 bands using a sixth order Butterworth Infinite Impulse Response (IIR) filter. Filters were spaced in a range from 100 Hz to 11 kHz following the equal basilar membrane distance (Greenwood, 1990). To extract the envelope, the output of each band was full-wave rectified and lowpass filtered at 30 Hz using a fourth-order Butterworth filter. For the noise-vocoded speech, the envelope was multiplied by a wide-band noise carrier. For Dudley-vocoded speech, which has a natural amount of source periodicity and fully periodic speech, the envelope of the voiced segments of speech were multiplied with a pulse train following the natural F_0 contour, and the segments of unvoiced speech were multiplied with a wide band noise carrier. The periodic speech was generated using pulse trains only. Here, additional F_0 contours were created by interpolation through unvoiced

sections and periods of silence using piecewise cubic Hermite interpolation in logarithmic frequency. The start and end points of each contour were anchored to the median frequency of the sentence. For all three vocoding types, the side bands were removed by filtering the modulated noise and/or pulse train using the channel bandpass filter. The Root Mean Squared (RMS) level of the channel outputs was then adjusted to match the original level in that band, after which all band filters were summed. The final waveforms were then low-pass filtered at 10 kHz.

In Steinmetzger and Rosen (2015) a second set of non-periodic, mixed periodic and fully periodic speech was generated using TANDEM-STRAIGHT (Kawahara et al., 2008). This method does not rely on bandpass filtering but produces very natural-sounding speech with a mixed source excitation that can be adapted to produce fully aperiodic or fully periodic speech. Using TANDEM-STRAIGHT, non-periodic speech was generated by keeping the default settings but fixing the F_0 to 0 Hz throughout. For Dudley-vocoded speech, to minimise the level of the aperiodic component the values of the sigmoid parameter in the source estimation routine were fixed to 1 and -40. For F_0 -vocoded speech the interpolated F_0 contours were used as input for the source extraction routine.

Since Steinmetzger and Rosen (2015) have made their material publicly available, the results reported in this section are based on their exact stimuli. The toolbox contains an implementation that leads to very similar results for the vocoded speech. However, since TANDEM-STRAIGHT is patent-protected software it is not included in the toolbox. The stimuli are recordings of ten lists of IEEE sentences (100 sentences total) spoken by a British male. A total of 220 sentences (20 sentences per 11 participants) were presented for each condition (a particular type of vocoding with a given number of channels). The ASR results are based on 100 sentences per condition (i.e. all 10 IEEE lists). In Steinmetzger and Rosen (2015), the level of the target was fixed at about 80 dB SPL, whereas here it is fixed at 65 dB SPL based on the normalisation test results of the ASR systems (see Section 5.1 in Chapter 5).

6.2 Detailed Comparison

Steinmetzger and Rosen (2015) measured the accuracy for three types of vocoded speech with up to 16 channels in quiet. It was shown that in humans, the absence of any periodicity information (and hence intonation) does not significantly affect speech intelligibility in quiet listening environments (Figure 6.4). On the other hand, additional unnatural periodicity cues as found in the periodic vocoded speech leads to substantially poorer speech intelligibility rates. In Figure 6.4, I also included estimates of the performance in quiet

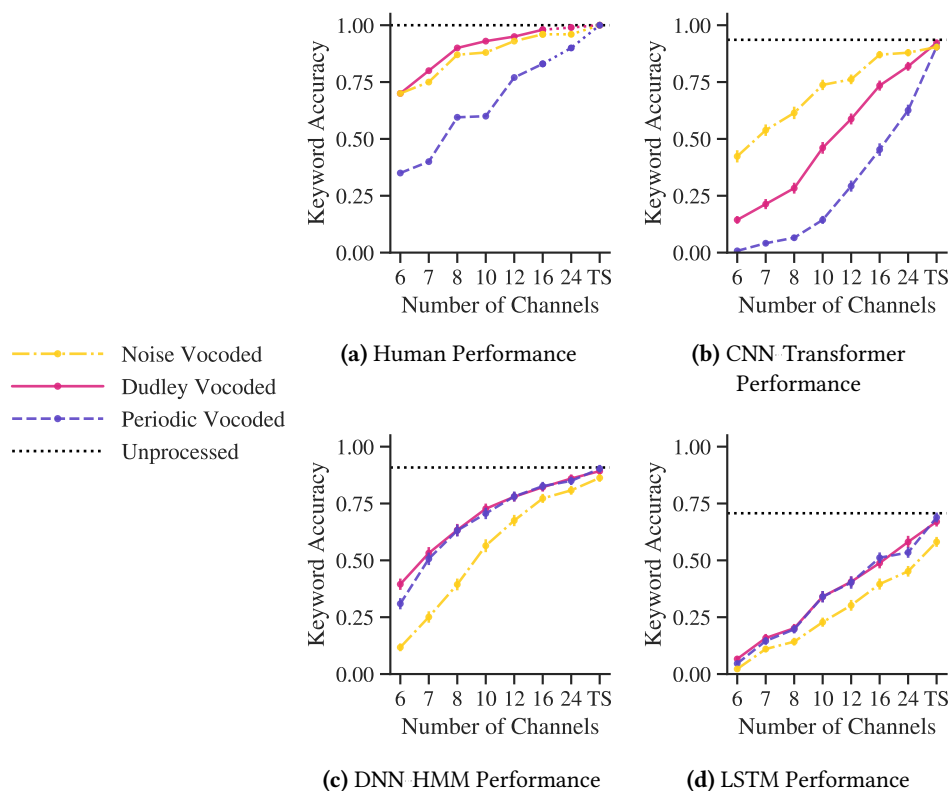


Figure 6.4. Performance in the target periodicity experiment. In this experiment, recognition accuracy is measured in three vocoding conditions that differ in periodicity: noise vocoding (aperiodic), Dudley-vocoding (mixed-periodic) and fully periodic. The TANDEM-STRAIGHT (TS) condition, in which the periodicity of the source was controlled using the TANDEM-STRAIGHT algorithm, is also included. Performance is measured as the percentage of correctly predicted keywords and is plotted as a function of the number of frequency bands for the three different vocoding conditions. The accuracy for unprocessed speech is denoted by the dashed horizontal line. All error bars show the standard error of the mean. (a) Human performance taken from Steinmetzger and Rosen (2015). Performance in the 24-channel vocoded and TS condition were estimated from psychometric function curves from a second experiment that investigated the effects of noise on speech with varying levels of periodicity. (b–d) Performance of CNN Transformer (b), DNN HMM (c) and LSTM (d) on 100 IEEE sentences presented at 65 dB. The material was identical to the material used in Steinmetzger and Rosen (2015).

based on psychometric functions for 24 channels as well as TANDEM-STRAIGHT (TS) speech from a subsequent experiment (see Section 3). The differences between the three levels of periodicity mostly disappear when using the more natural sounding TANDEM-STRAIGHT speech sounds, which were produced to be either not periodic (noise vocoded), fully periodic or mixed-periodic without reducing the spectral resolution of the sound.

None of the ASR models show the same relative sensitivity to periodicity in speech. For example, in stark contrast with the poor intelligibility for periodic speech found in

humans, the DNN-HMM and LSTM model perform equally well with respect to Dudley and fully periodic vocoded speech. For both these models, noise-vocoded speech leads to the poorest performance and remains lower than the other two speech types even for the more natural sounding TANDEM-STRAIGHT speech. Since these two models both use MFCCs as input, these may be the source of the discrepancy. Although MFCCs discard most pitch information, some voicing information is still retained (Milner & Shao, 2006). The results suggest that both the DNN-HMM and LSTM model rely on the presence of these voicing cues (which are mostly retained in the higher MFCC coefficients) in making predictions. In contrast to the DNN-HMM and LSTM, the CNN-Transformer does display a decrease in performance with respect to periodic speech compared to Dudley-vocoded speech, which suggests that the lack of periodicity is used as a cue. Surprisingly, however, the CNN-Transformer performs best when noise-vocoded speech is used. It may be the case that certain mismatches between periodic and non-periodic components that result from the imperfect automatic F_0 estimation of Dudley-vocoded speech negatively affect performance. As in humans, performance for the TANDEM-STRAIGHT speech converges to that of unfiltered speech.

7. CONCLUSIONS

In this chapter I compared the performance of ASR systems and human listeners on a range of distortion experiments. In general, ASR systems unsurprisingly performed worse than humans, particularly in the presence of noise. Where appropriate, this was corrected for by allowing a higher SNR when testing ASR systems. Even with this correction, none of the ASR systems performed overall similarly to humans in terms of which features and cues they relied on, and which distortions they were robust to. The results suggest that they are not yet ready to be used as general purpose proxies for modelling human speech recognition.

The Wav2Vec 2.0 CNN-Transformer model, which is both the most recently introduced and the biggest model of those being compared, was shown to be most robust against most distortions despite having been trained on less data. This model was most robust against bandpass filtering and spectral and temporal modulation filtering. These findings are consistent with other work that points to the robustness of pre-trained acoustic representations (such as those used in the Wav2Vec 2.0 CNN-Transformer) to mismatched train and test data domains (Ma et al., 2020). The Kaldi DNN-HMM model, which has only slightly lower accuracy levels than the CNN-Transformer in quiet and undistorted listening conditions,

tended to perform similarly or worse in most dimensions. The Mozilla's DeepSpeech LSTM, which is the smallest model of those considered here, is much less robust than the two other systems.

The first three tests discussed in this chapter assess robustness to some common distortions found in communication systems: bandpass filtering (commonly applied in phone communication), peak clipping (a common distortion when a recording device is saturated) and centre clipping (as found in simple noise-suppression systems). The results show that both the CNN-Transformer and the DNN-HMM closely follow human performance with regards to centre clipping, but none of the ASR systems come close to human performance with regards to bandpass filtering and peak clipping. Relative to the other ASR systems, the CNN-Transformer model displays the highest spectral invariance, whereas the LSTM model performs poorly in all three tasks.

The next set of tests investigate auditory features that have been commonly proposed to be important for speech processing in humans, namely spectral and temporal modulations, TFS and periodicity. The tests showed that the relative sensitivity to spectral and temporal modulations is comparable between humans and ASR systems, but the use of TFS and periodicity differs considerably. The sensitivity to spectral and temporal modulations was consistent across ASR systems, with only a slightly higher importance of slow spectral and temporal modulations compared to humans. Removing TFS affects intelligibility for normal hearing humans in noisy conditions. In contrast, the ASR systems had only a small sensitivity to TFS, an effect that completely disappeared when notches of TFS were removed that did not span the whole frequency range. This experiment also further confirmed differences in spectral sensitivity between humans and ASR systems, as the removal of high frequency regions (above 2 kHz) hardly affects intelligibility for normal hearing humans, but reduced speech in noise perception for all ASR systems. Lastly, the ASR systems appear to process periodicity information very differently from humans. Whereas humans perform poorly when vocoded speech is fully periodic (rather than mixed periodic as with natural speech), this effect was only observed in the CNN-Transformer. However, the CNN-Transformer displayed a preference for noise-vocoded speech compared to mixed periodic speech. Both the LSTM and DNN-HMM rely on the presence of periodicity and perform poorly in noise-vocoded speech. In stark contrast with humans, fully periodic vocoded speech does not negatively impact their performance compared to mixed-periodic speech.

Altogether, these results highlight the differences between ASR systems and humans. Even when they achieve similar performance in quiet listening conditions, not only are

the ASR systems much less robust to common distortions such as bandpass filtering and peak clipping, but the models also use certain auditory cues (such as periodicity and TFS) in vastly different ways than humans do. Although it is likely trivial to create ASR models that perform well under most of the distortions presented in this chapter (e.g. by including distorted data in the training set), such training would impede on the usage of the ASR systems as a model of human hearing. After all, a good model of human hearing should be predictive of unseen human behaviour without being explicitly trained to do so. Furthermore, I will show in the next chapter that such biased training does not guarantee a more humanlike system, as improvements in one auditory cue dimension may negatively impact performance in others.

Looking to the future, designing ASR models that rely on similar features of speech as humans may allow them to perform better in noisy conditions they have never been exposed to. The results presented in this chapter point to some features that may be important. For example, none of the ASR systems presented here are able to exploit TFS information. In the case of the LSTM and DNN-HMM, this is the result of the use of MFCC features as input. The end-to-end ASR systems, however, should in principle be able to utilise TFS information. Other dimensions in which the ASR systems differed significantly from humans are resistance to peak clipping distortions and the use of periodicity in the target speech.

It may also be fruitful to investigate why the CNN-Transformer model performs more similarly to humans than the other models. An immediate possibility is the unsupervised pretraining of the CNN-Transformer. In contrast to the other two models, the CNN-Transformer model does not use MFCC features but instead employs unsupervised pretraining to automatically extract acoustic features from speech data. This results in a rich speech representation that captures a wide range of phonetic information (Ma et al., 2020) as well as a surprising amount of higher level linguistic structure (Shah et al., 2021). Other pretrained acoustic representations have shown similar robustness to mismatched train and test data (Ma et al., 2020), which suggests that unsupervised, task-independent pretraining is an important step towards creating more humanlike ASR systems.

CHAPTER 7

Taking ASR For A SPIN Test

1. INTRODUCTION

One of the most challenging (and common) communication situations is that of trying to understand speech in a background of competing talkers. The robustness of human hearing shines particularly in the ability of the human brain to focus attention on one auditory stimulus while filtering out others (sometimes referred to as the cocktail party effect; Cherry, 1953). Although many recently introduced Automatic Speech Recognition (ASR) systems can reach near-human performance in quiet, speech in noise is still challenging for these systems.

In this chapter, the noise robustness of three open source ASR systems is explored in more detail. By comparing their relative performance on a variety of Speech in Noise (SPIN) tasks, more can be learned about the internal mechanisms that underlie their performance. These tasks are designed to not only test performance in a generic background of competing talkers, but to systematically compare the effects of particular masker cues (such as masker fluctuations and periodicity) that affect masker effectiveness. While none of the tested ASR systems have been explicitly trained in noise, I will show that some of them still show qualitative similarities with human speech in noise perception. Furthermore, I will show that explicitly training ASR systems in certain distorted conditions is not sufficient to make the ASR systems more humanlike, as this approach may lead to worse performance in other auditory dimensions.

Similar to Chapter 6, the following freely available ASR models are being compared in this chapter:

1. The Kaldi nnet3 chain model from the Kaldi Active Grammar project (Zurow et al., 2021) uses a hybrid Deep Neural Network (DNN) and Hidden Markov Model (HMM) (**DNN-HMM**; Povey et al., 2018). It takes as input Mel-Frequency Cepstral Coefficient (MFCC) features, a classic set of features inspired by the human auditory system in

which time is divided into short, overlapping ‘frames’, and the spectral envelope of the sound is computed for each frame. At each layer of the network, several frames of output from the previous layer are provided as input to give temporal context.

2. Mozilla’s DeepSpeech model (Mozilla, 2020) is a Long-Short-Term-Memory (**LSTM**) neural network, a widely used type of neural network that is designed to model long temporal relationships. It is a variant of the original DeepSpeech model that was introduced by Hannun et al. (2014). Like the DNN-HMM model, it uses MFCC features as its input. However, the MFCC features in the LSTM model are only based on spectral information up to 4 kHz, whereas the DNN-HMM uses spectral representations of up to 8 kHz.
3. Facebook’s fairseq Wav2Vec 2.0 (Baevski et al., 2020) is a **CNN-Transformer** model that, in contrast to the previous two models, takes raw audio as input. Features are extracted using a Convolutional Neural Network (CNN) that convolves the input signal over time. Temporal context is modelled using a Transformer architecture (Vaswani et al., 2017), which uses an attention mechanism to relate inputs from different timesteps.

Each of these systems uses a different approach to model the temporal component of speech. The input and architecture of each of them is described in more detail in Chapter 5.

The remainder of this chapter is divided into three sections. The first two sections each describe a SPIN experiment. The first experiment assesses the performance of the ASR systems in the presence of an increasing number of competing background talkers. The effects of different types of maskers are further analysed using Explainable Artificial Intelligence (XAI) methods. The second experiment compares the performance of the ASR system with respect to both periodic and modulated maskers. The best performing ASR is then used to obtain detailed confusion matrices that would be more difficult to obtain using human listeners. In the third section, the effects of training ASR systems in distorted conditions on SPIN performance is further investigated. Finally, the results are summarised and compared in the concluding section.

2. COMPETING TALKER BACKGROUND

One of the most challenging listening environments is a background of other speakers. Babble maskers affect speech intelligibility in at least two ways. Firstly, Energetic Masking

(EM) may occur, caused by the presence of masker energy in the same frequency region(s) as the energy in the target signal. As speech is a modulated signal, the amount of EM will vary over time, with short glimpses during which the Signal to Noise Ratio (SNR) is much higher. These glimpses may either be comodulated, which occurs when modulations are consistent across the spectrum, or uncomodulated, meaning the glimpses are restricted in frequency. There is evidence that some ASR systems take advantage of dips to process speech in noise (Spille & Meyer, 2017). A second type of masking is Informational Masking (IM), which occurs when the target speech and masker are difficult to disentangle despite a lack of energetic overlap, for example because the masker has a similar voice quality, or the linguistic content of the masker is distracting (lexical interference).

To investigate the relative importance of EM and IM, Rosen et al. (2013) used babble noise, noise-vocoded and envelope-modulated speech-shaped noise with an increasing number of competing talkers. Each of these maskers may introduce varying levels of EM and IM. For example, the masker effectiveness of a babble masker initially increases in going from one competing talker to two competing talkers, but slowly decreases after more than two speakers are added. The latter decrease suggests a release of IM, resulting from the individual talkers ‘blending’ together and therefore reducing intelligibility of individual speakers. This release is large enough to counteract the reduced glimpsing opportunities (i.e. increase of EM) that occur when the number of talkers increases. A similar influence of IM was not present in the noise-vocoded speech, which may be a result of the differences in sound quality and periodicity. The modulated noise masker, which can induce comodulated EM, is the least effective masker when it modulates at the rate of a single speaker. However, masker effectiveness improves as the number of speakers increases. To investigate the way background speakers affect ASR performance, the ASR systems were tested following the experimental paradigm of Rosen et al. (2013).

2.1 Method

This section describes the experimental paradigm of (Rosen et al., 2013). I will also discuss an ASR-specific analysis of IM and EM in the CNN-Transformer model. More specifically, I will describe a method for measuring IM using the Connectionist Temporal Classification (CTC), followed by a methodology that can visualise the ASR response to temporary gaps (or ‘glimpses’) in EM.

Experimental Paradigm

In the competing talker background experiments, the three ASR systems were tested using the stimuli of three masker conditions from Rosen et al. (2013). These conditions include speech babble, noise-vocoded babble and speech-shaped noise that was modulated using the envelope of the babble noise. The ‘babble’ in each of these conditions ranged from a single to up to sixteen competing talkers.

The basis for their noise maskers were sixteen male talkers from the EUROM database of English speech, a dataset in which each speaker reads five- or six passage sentences. Of these, pauses of more than 100 ms were deleted, resulting in sound files of approximately 21 s per speaker without significant pauses. In the HumanlikeHearing toolbox, segments of similar length are created by combining sentences spoken by male speakers from the IEEE ARU database. The sound files were normalised to a common Root Mean Squared (RMS).

Single speaker speech babble was created by randomly selecting one of the sixteen background talkers. As the ARU dataset only contains 6 male talkers, some talkers are sampled multiple times in the toolbox. The other talker conditions were created by randomly selecting a talker and adding it to the talker(s) already present in the previously constructed condition. Note that as the ARU dataset only contains 100 sentences, it can happen that different speakers may read the same sentence in the babble, although the chances of this occurring are low. The spectrum of each babble was equalized to the Long-Term Average Spectrum (LTAS) of the 16-talker speech babble. To create noise-vocoded babble, each of the five babbles were filtered into twelve bands using the same method described in Section 6 of Chapter 6. To create speech-shaped modulated noise, the envelope of each babble was extracted by full-wave rectification and low-pass filtering at 30 Hz. The envelope was multiplied by a broad-band noise that was equalised to the LTAS of the 16-talker babble.

The results presented here use the same stimuli material as used in the original experiment of Rosen et al. (2013), but the toolbox also incorporates an implementation that leads to similar results. The target speech consists of the first 10 lists of the IEEE dataset. In Rosen et al. (2013), Speech Reception Threshold (SRT)-50s were obtained using an adaptive approach. The results reported here were measured using the static SRT method instead, where data obtained over a range of SNRs were fitted to a logistic function using a Bayesian optimisation paradigm as described in Wichmann and Hill (2001) (see the subsection on psychometric function estimation in Section 5.2 in Chapter 5). Specifically, SRT-50s were from accuracies measured between -6 dB to 33 dB in 3 dB steps. Rosen et al. (2013) presented 320 sentences per condition (20 sentences for each of the 16 listeners). The experiments

in the present study are based on the results of all 100 IEEE sentences. The noise masker was randomly selected from the 21 s masker. In Rosen et al. (2013), noise level was fixed at 70 dB over a frequency range of 100 Hz-5 kHz. To obtain a given SNR, the target speech level was adjusted. In the experiments presented here, the target speech remained fixed at 65 dB throughout.

Measuring Informational Masking with the CTC Score

In contrast to their human counterparts, it is relatively straightforward to measure the amount of IM in ASR systems. The simplest way would be to look at the transcription of the ASR system: does it include any words or letters from the masking sentences? This idea can be taken a step further by computing the probability of the masker sentences based on the emission probabilities of the ASR system. The emission probabilities denote the probability the ASR system assigns to each letter at each time step. An approach based on emission probabilities will reveal subthreshold detection of the masker sentence, regardless of whether the letters were actually included in the final transcription. To measure the probability of the masker sentences, I will use the CTC (Graves et al., 2006). The CTC computes the likelihood of a given transcription y given the input audio x by marginalising over the set of valid alignments of that transcription:

$$p(x|y) = \sum_{a \in A_{x,y}} \prod_{t=1}^T p_t(a_t|x) \quad (7.1)$$

Here, $A_{x,y}$ is the set of all valid alignments of the transcription y for the duration of the input audio x . The value $p_t(a_t|x)$ denotes the emission probabilities for the letter a_t (i.e. the t th letter in the alignment sequence) given the input audio x of an alignment of T letters. For models that were trained using the CTC loss, in which each output neuron typically reflects the probability of one specific letter in the alphabet, the emission probabilities are simply the values of the output neurons at each time step. The CTC score can be efficiently computed using a dynamic programming algorithm (Graves et al., 2006).

Measuring Energetic Masking using Attributions

The amount of EM present in a SPIN test can be measured by computing a moving SNR. This can be done for the target and masker as a whole, to measure comodulated EM (i.e. modulations across the frequency spectrum), or for individual frequency channels, to measure uncomodulated EM. To measure the extent to which an ASR system efficiently

combats EM, attribution methods can be used. Attribution methods, discussed in more detail in Section 4.4 in Chapter 3, are designed to reveal components of high importance in the input to the final prediction of a machine learning model. These methods can be used to understand which components of the input (i.e. which segments of the input audio) contributed most to a certain prediction. For example, Spille and Meyer (2017) showed through a visualisation that a DNN-HMM trained on modulated noise appeared to have the highest attributions at moments of high SNR.

To measure which parts of the input are most important to the CNN-Transformer, I chose to use the Deep Learning Important FeaTures (DeepLIFT) attribution method (Shrikumar et al., 2017) combined with a variant of SmoothGrad (Smilkov et al., 2017), an attribution ensemble method. Ensemble methods average the attributions of the input with random additive noise to obtain less noisy attributions. The decision to use an ensemble method was motivated by a recent comparison between various attribution methods (Hooker et al., 2019), whose thorough analysis suggests that certain ensemble methods provide more robust attribution estimations than any attribution method on its own. Since ensemble methods require various attribution estimates for a single input, the attribution method they are combined with has to be relatively efficient when using limited computational resources. Ancona et al. (2017) reported that out of four popular attribution methods (Layer Relevance Propagation, DeepLIFT, Gradient $\times$ Input and Integrated Gradients, see Section 4.4 in Chapter 3), Integrated Gradients appeared most robust. Unfortunately, this method is computationally expensive, since it computes the average gradient for an input that varies along a linear path. Since DeepLIFT can be viewed as an efficient approximation of Integrated Gradients (Ancona et al., 2017), I chose to combine the ensemble method with DeepLIFT instead. Results for a range of other attribution methods can be found in Appendix D. All attribution experiments were implemented using the Pytorch Captum interpretability toolbox (Kokhlikyan et al., 2020).

The DeepLIFT method frames the question of feature importance in terms of differences with respect to a certain reference input. This reference input is usually some variation of a zero vector (e.g. a black image or silent audio), but could also be the absence of a certain feature in the input (e.g. the absence of an object). First, DeepLIFT computes a forward run of both the input x and reference input $\bar{x}$, resulting in the outputs $s_i(x)$ and $s_i(\bar{x})$ in neuron i , respectively. The difference between $s_i(x)$ and $s_i(\bar{x})$ determines the ‘relevance scores’ of all neurons in the network. More specifically, relevance scores at the output layer L are

defined as follows:

$$r_i^{(L)} = \begin{cases} s_i(x) - s_i(\bar{x}) & \text{if } i \text{ is target unit} \\ 0, & \text{otherwise} \end{cases} \quad (7.2)$$

These relevance scores are then further propagated through the network following the ‘rescale’ rule, which is defined as follows:

$$r_i^{(l)} = \sum_j \frac{z_{j,i} - \bar{z}_{j,i}}{\sum_{i'} z_{ji} - \sum_{i'} \bar{z}_{ji}} \quad (7.3)$$

Here, $\bar{z}_{ji} = w_{ji}^{(l+1,l)} \bar{x}_i^{(l)}$ is the weighted activation of the neuron i onto neuron j when the baseline $\bar{x}$ is fed into the network, while z_{ji} reflects this activation for the input x . The attribution of the i th input feature is equal to $r_i^{(1)}$, i.e. the relevance values at the output:

$$a_i = r_i^{(1)} \quad (7.4)$$

Together, the attributions a_i for each input feature i form a vector A which is of the same size as the model input. DeepLIFT satisfies the ‘completeness’ axiom, meaning that the input attributions always sum to the difference between the outputs $s_i(x)$ and $s_i(\bar{x})$. As shown in Ancona et al. (2017), DeepLIFT is equivalent to the feature-wise product of $x - \bar{x}$ and its ‘modified partial derivative’, a notion they introduce to show the similarity between DeepLIFT and attribution methods that are explicitly based on gradients. For DeepLIFT, this modified partial derivative depends on the ratio between the difference in output and the difference in input at each nonlinearity. See Ancona et al. (2017) for more details.

The ensemble method of choice is a squared variant of SmoothGrad (Smilkov et al., 2017). The original SmoothGrad formulation averages a set of j noise estimates of the feature importance (constructed by injecting a single input with Gaussian noise η independently j times). The squared SmoothGrad formulation squares each estimate before averaging:

$$e = \sum_{i=1}^j a(x + \eta)^2 \quad (7.5)$$

Where $a(x + \eta)$ denote the attributions that result from the input $x + \eta$.

Some preprocessing is required to obtain attributions of the same dimensionality as the input audio, since attributions are computed with respect to the output of a single output neuron. In the case of the CNN-Transformer, this means the attributions can only be computed for a single letter at only one time step. I will only consider the attributions of the predicted class, meaning that attributions are only calculated for the output neuron

with the largest value at each time step. The time steps at which the output neuron that denotes a ‘blank’ output had the highest value are ignored (see Section 3 Chapter 3 for more information on the role of ‘blank’ outputs in CTC-based speech recognition systems). This approach results in attribution values A_t for each valid output time step t , where A_t is a vector of the same length as the input x . Furthermore, as DeepLIFT allows negative attributions, the absolute value of the A_t is taken. Note that a negative attribution means that the input contributed to moving the output below its reference value, not that this input is ‘unimportant’. Lastly, since DeepLIFT satisfies the ‘completeness’ axiom, the attributions for a single prediction always sum to the difference between the outputs $s_i(x)$ and $s_i(\bar{x})$. To prevent particularly high $s_i(x)$ at one time step from dominating the aggregated attributions, the attributions A_t are normalised by dividing them by their mean $\hat{A}_t$. The final attribution is the mean over all T normalised attributions:

$$A = \frac{1}{T} \sum_{A_t} \frac{|A_t|}{\hat{A}_t} \quad (7.6)$$

An unfortunate reality for most XAI methods is that the ground truth is unknown. Attribution methods are no exception to this, which implies that any attribution analysis should include a verification of the method. To test whether the attributions estimated by DeepLIFT are a meaningful reflection of the features used by the CNN-Transformer, an ablation test can be used. Here, the time steps associated with the top $n\%$ attributions are set to zero, with n ranging between 0 % to 100 %. An attribution method that identifies important input features should lead to a sharp drop in accuracy. I will compare the drop induced by two other ablation orderings that are invariant of the model, namely by ordering the ablated time steps at random or by absolute amplitude of the input sound. Note that, like the attribution methods they assess, ablation tests are not infallible. It is possible for a drop in performance to be caused by a generic change in input distribution, rather than it being a true reflection of the features the model uses. An alternative method that does not suffer from this has been previously proposed (Hooker et al., 2019) but was too computationally expensive to apply, since it requires retraining the network from scratch for every ablation.

2.2 Results

This section describes the outcome of the competing talker experiment from Rosen et al. (2013) when applied to the three ASR systems. Furthermore, the effects of IM and EM on the overall best performing model (the CNN-Transformer) are further analysed.

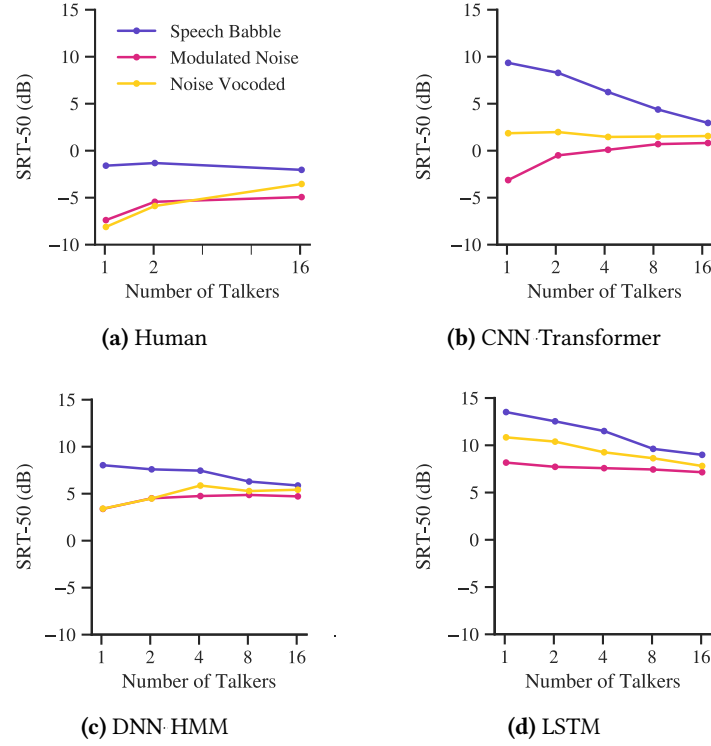


Figure 7.1. Performance in the competing background talkers experiment. SRTs at 50% correct for 100 IEEE sentences were measured in three different types of maskers: speech babble, noise-vocoded babble and modulated speech-shaped noise. These maskers were created with a varying number of talkers (1, 2, 4, 8 and 16). SRTs are plotted as a function of the number of talkers in the masker. **(a)** Human data taken from Rosen et al. (2013). **(b-d)** Performance of the CNN-Transformer **(b)**, DNN-HMM **(c)** and LSTM **(d)** ASR systems on the same material as used in Rosen et al. (2013).

Detailed Comparison of ASR Systems

Figure 7.1 shows the SRT-50s obtained for speech in the three maskers (babble, noise-vocoded babble and noise modulated with the envelope of the babble) with a varying number of competing talkers. Although the SRT-50s observed in the ASR systems are much higher than those recorded in humans, two of the three main results presented in Rosen et al. (2013) are reflected in the results of the ASR systems, particularly for the DNN-HMM.

Firstly, for all ASR systems babble was always the best masker, as it had the highest SRT-50. In humans, noise-vocoded babble and modulated noise maskers tend to closely follow one another, a result that is only clearly replicated in the DNN-HMM.

Secondly, in humans the performance for the modulated and noise-vocoded maskers tends to increase monotonically, with an initial bigger increase after which the SRT-50 plateaus. Although less pronounced, a similar effect is observed in the DNN-HMM for

both types of noise. For the CNN-Transformer, this only occurs for modulated noise, as the SRT-50s for noise-vocoded speech remain constant. As shown in Section 6 of Chapter 6, the CNN-Transformer has high intelligibility rates for noise-vocoded speech. Together these results suggest that the potential benefit of a release in EM is counteracted by IM. In contrast to the other two models, the SRT-50s of modulated noise do not change as the number of speakers increases for the LSTM. This suggests that this model is not able to make use of glimpsing opportunities.

Lastly, a major difference between these two models and the human outputs are the single babble speaker effects. Rosen et al. (2013) report that the SRT-50 for babble maskers at one and two speakers is almost identical, but report that single speaker babble noise was a less effective masker at lower SNRs than two speaker babble noise. Such a strong single speaker benefit was not observed in the ASR systems. One explanation for this is that the humans have been instructed to focus on one specific target speaker. The ASR systems, on the other hand, will likely transcribe parts of both sentences or even only the masker speaker at low SNRs.

In summary, the DNN-HMM and CNN-Transformer, but not the LSTM model, display similar interactions of IM and EM as humans do. However, in contrast to humans, performance for single speaker babble maskers is much worse. Furthermore, the CNN-Transformer shows a reduced ability to make use of glimpsing opportunities when exposed to the noise-vocoded masker, likely as a result of IM introduced by the noise-vocoder, which (as was shown in Section 6 in Chapter 6) is highly intelligible to the CNN-Transformer.

IM Interference in the CNN-Transformer

To further test the level of IM of the CNN-Transformer, the output probabilities of the system were measured against competing talker maskers at a constant SNR of 5 dB. To measure the amount of IM (particularly lexical interference) the average probability of the masker sentences as well as the probability of the target sentence were measured using the CTC criterion. In contrast to the previous section, all experiments in this section were performed on the ARU IEEE sentences. This was done to ensure the transcription of the maskers could be more easily controlled, as the maskers were not transcribed in the data from Rosen et al. (2013).

In line with the SRT experiments from the previous section, unprocessed background talkers are always the best maskers at this SNR, followed by noise vocoded speech and then modulated noise (Figure 7.2a). This is further reflected in the target speech probabilities

(Figure 7.2b). As expected based on the results from the previous section and Section 6 in Chapter 6, the noise vocoded masker sentences have an intelligibility that is relatively similar to that of the babble masker (Figure 7.2c). The exception is the unprocessed single speaker masker, which is much more intelligible than the noise-vocoded single speaker masker. As expected, the modulated masker is not intelligible at all.

Perhaps surprisingly, the target probability in the presence of the unprocessed babble masker is very low compared to the noise-vocoded masker. It does overtake the average probability of the maskers (gray dotted line in Figure 7.2c) between 8 and 16 maskers, which explains why the accuracy is still above 50 % at 16 maskers despite low target probability. The low target probability suggests that the speech babble masker induces a much stronger interference than the noise vocoded masker, despite the relative accuracy difference between these two maskers being around 20 % to 30 % (similar to that between the noise vocoded and modulated masker). The difference between the average probability of the masker sentences of unprocessed versus noise vocoded babble is much smaller than the probabilities of the target sentences. These differences could be explained through forms of IM other than lexical interference, such as voice quality or timbre.

In summary, these results provide further evidence that the CNN-Transformer experiences lexical interference from noise-vocoded maskers. Unprocessed speech maskers also induce lexical interference, but additionally seem to be affected by other forms of IM.

Evidence for Dip Listening in the CNN-Transformer

To further understand the way EM affects speech in noise perception in the CNN-Transformer, attributions were measured using DeepLIFT and squared SmoothGrad for the three different maskers at an SNR of 5 dB. To quantify the amount of dip listening, I measured the correlation coefficient between the model attributions and the moving SNR (Figure 7.3a). The moving SNR was measured every 1 ms in windows of 10 ms and then upsampled to the sound sample rate (16 kHz). If the ASR system is ‘listening in the dips’, the attributions should be high at the time segments at which the local SNR is high, since those indicate ‘glimpses’ in the masker.

For modulated noise, the correlation is consistently high, suggesting a persistent application of dip listening. The noise vocoded masker starts with a low (even slightly negative) correlation, but as the number of competing talkers increases (and the IM decreases), the usage of dip listening increases. A similar effect can be observed in the speech babble masker, although the overall use of dip listening remains much lower. This is in line with

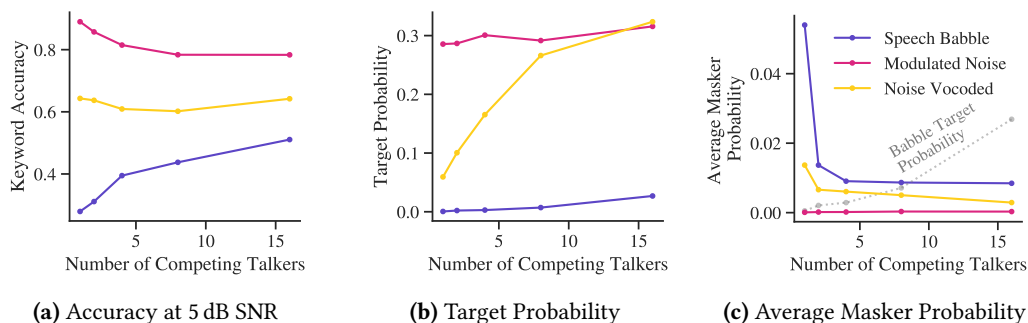


Figure 7.2. Measurements of IM in the CNN-Transformer in the competing talker experiment. All results are based on 100 IEEE sentences with a 5 dB SNR. **(a)** Keyword accuracy of the CNN-Transformer for three different maskers (speech babble, noise-vocoded babble and modulated speech-shaped noise) that were created with a varying number of talkers (1, 2, 4, 8 and 16). **(b)** The average likelihood of 100 target sentences as measured by the CTC criterion, as measured over all three noise types and different number of competing talkers. **(c)** Same as in **b**, but now the average likelihood of the masking sentences is measured. The likelihood is measured of each competing talker separately and then averaged across talkers. The dotted gray line reflects the speech babble target probability from **b** (purple line), which is shown to compare the point at which the target probability is higher than the average masker probability.

the results from the previous section, as the continued IM interference by the speech babble masker prevents efficient dip listening.

The presence of dip listening for modulated noise can be visualised by plotting the attributions against the moving SNR (Figure 7.3b). Since the use of dip listening is easier to identify when the noise fluctuates at a constant frequency, the attributions were measured for a speech shaped noise masker that fluctuates with 10 Hz at 5 dB SNR (Figure 7.3b). For visual clarity, the attributions and signal to noise ratio in this figure were lowpass filtered at 30 Hz with a 4th order Butterworth filter.

To verify whether the DeepLIFT attributions are a meaningful reflection of the features used by the CNN-Transformer, an ablation test was performed. This test shows that the performance for the ablated attributions drops much quicker than that of the random control (Figure 7.3c), providing some evidence that the DeepLIFT attributions identify temporal regions that are important for the CNN-Transformer. However, the drop in accuracy based on attribution ablations is very similar to that of the sound amplitude. This is likely because the DeepLIFT attribution method has a strong bias towards high-amplitude inputs. As discussed in Section 2.1, the DeepLIFT attributions are multiplied with $x - \bar{x}$ (where x is the input and $\bar{x}$ the baseline input). Since the baseline $\bar{x}$ is set to zero, every attribution in DeepLIFT is effectively multiplied by the input, and thus scaled by the amplitude of the incoming sound. This is not specific to DeepLIFT the same is true for many other

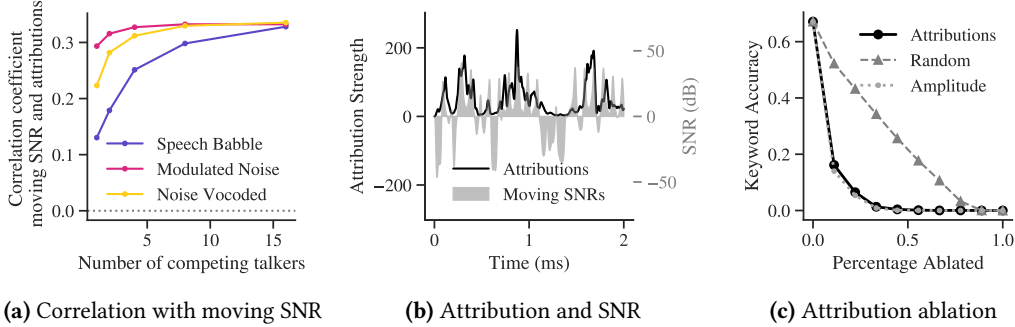


Figure 7.3. Measurements of EM in the CNN-Transformer in the competing talker experiment using DeepLIFT attributions. (a) The correlation coefficient between the moving SNR and DeepLIFT attributions for the three different maskers. A high correlation indicates that the attributions are concentrated at high local SNRs, therefore suggesting that the ASR system employs dip listening. (b) Visualisation of the measured attributions versus the moving average for a single target sentence under a masker modulating at 10 Hz. For visual clarity, the attributions and moving SNR were lowpass filtered at 30 Hz with a 4th order Butterworth filter. The attributions peak together with the moving average, suggesting the use of dip listening. (c) Accuracy in an ablation experiment. For all sentences in the competing talker experiment, performance is measured when a certain percentage of time steps is ablated (set to zero). Which time steps are ablated is either determined by their attributions (with high attributions being ablated first) or these time steps are chosen at random.

attribution methods, such as Layer-wise Relevance Propagation, Integrated Gradients and, unsurprisingly, Gradient $\times$ Input. A strong relationship between the attributions found in DeepLIFT and the absolute sound amplitude is thus unsurprising. Nevertheless, as shown in Appendix D, the absolute sound amplitudes themselves have a consistently high correlation with the moving SNR, regardless of the masker type. This suggests that DeepLIFT can uncover more than the sound amplitudes do, regardless of the results of the ablation experiment.

In summary, the attribution experiment provides further evidence for the use of ‘dip listening’ in the CNN-Transformer. This is in line with previously reported results (Spille & Meyer, 2017), although the ASR system in that study was explicitly trained with additive modulated maskers. Nevertheless, further research is required to conclusively confirm whether the attributions truly reflect segments of the input that are important to the ASR system, since the ablation experiment was inconclusive.

3. MASKER MODULATIONS AND PERIODICITY

In Section 2, the discussion mostly focused on the effects of masker modulations and masker intelligibility on speech in noise perception. However, another major difference between a

speech and noise-vocoded speech masker is the level of periodicity, which is present in a speech masker but absent in noise-vocoded speech. In Section 6 from Chapter 6, it was shown that ASR systems respond differently to periodicity in the target speaker compared to humans. Here, I compare the effects of periodicity in a masker with masker modulations following the approach described in Steinmetzger and Rosen (2015). They compared the advantage of using a periodic masker (Masker Periodicity Benefit; MPB) with the advantage from modulations in the masker (Fluctuating Masker Benefit; FMB) and found that the benefit from periodicity (MPB) tended to be higher.

3.1 Method

In this section I will first describe the experimental paradigm of the masker experiment described in (Steinmetzger & Rosen, 2015). As in the previous experiment, the CNN-Transformer is further analysed. The second part of this section describes the methodology of this analysis.

Experimental Paradigm

To investigate the effects of masker modulations and periodicity, the second and third experiments described in Steinmetzger and Rosen (2015) were applied to the ASR systems. The target speech used was the same as described in Section 6 in Chapter 6, i.e. non-periodic (noise-vocoded), mixed-periodic (Dudley-vocoded), fully periodic as well as unfiltered speech.

Two types of maskers, a speech-shaped noise masker and a harmonic complex, were used. The noise masker was a 24 s passage of white noise that was filtered to follow the LTAS of the unprocessed target speech. The LTAS was computed by taking the power spectral density of the concatenated waveforms using Welch's method (with a window size of 512 samples, 50% overlap and an FFT length of 512 samples), followed by 1 octave smoothing. The harmonic complex maskers were based on F_0 contours extracted from recordings in the EUROM database. In the toolbox, they are extracted from concatenated IEEE sentences. The contours were extracted using the PRAAT toolbox and interpolated through unvoiced and silent periods using a piecewise cubic Hermite interpolation in logarithmic frequency. The waveforms were synthesized on a period-by-period basis using the Liljencrants-Fant model (Fant et al., 1985), which approximates a typical adult male glottal pulse. In the toolbox, the glottal pulse model incorporated in the PRAAT toolbox is used instead. Using the same filtering procedure as was used for the noise maskers, the

harmonic complex maskers were matched in spectrum to the LTAS. Both the noise and harmonic complex maskers were presented either as is or were sinusoidally amplitude-modulated at a rate of 10 Hz with a modulation depth of 100%. During the experiment, the maskers were randomly selected from the 21 s segment.

As in Section 6 in Chapter 6, the results reported here are based on the material made publicly available by Steinmetzger and Rosen (2015). In their study, a total of 240 sentences (20 sentences per 11 participants) were presented for each condition (a combination of one of the three vocoded targets with a certain spectral resolution or unfiltered speech together with one of the four maskers). Here, each condition is tested using 100 sentences (all available IEEE sentences). In Steinmetzger and Rosen (2015), each vocoded target was tested under three (for noise-vocoded and Dudley-vocoded) or four degrees (fully periodic vocoded) of spectral resolution. The spectral resolution for each vocoder type was chosen to match performance in quiet to around 70%, 90% or ceiling-level. In the present study, the SRT-50 was measured for all levels of spectral resolution in which performance in quiet was above 50%. In Steinmetzger and Rosen (2015), the level of the target and masker together was fixed at about 80 dB SPL, whereas in the results presented here the target speech was fixed at 65 dB. Steinmetzger and Rosen (2015) determined SRT-50 using an adaptive procedure, whereas the results reported here are based on the static SRT method instead (see Section 5.2 in Chapter 5).

Measuring Phone Confusions

As in the previous experiment, the CNN-Transformer is further analysed. As the FMB reflects the use of dip listening, which was discussed extensively in the previous experiment, this part focuses on the MPB. A typical way of understanding the type of mistakes any classifier makes is the creation of a confusion matrix, which plots the true class against the predicted classes. For human listeners, confusion matrices may be created based on logatomes, which are artificial words of only one or two syllables. Typically, these logatomes are CVC (consonant-vowel-consonant) or VCV (vowel-consonant-vowel) sequences, and the human is tasked with recognising the middle vowel or consonant. To the best of my knowledge, no human study has compared the phone confusions that result from using harmonic versus speech shaped maskers as of writing.

Since the CNN-Transformer model predicts letters rather than phones, there is no one-to-one correspondence between the letters and nonsense words. For example, the same sound may sometimes be written as a /z/ or /s/. This makes it difficult to create

a confusion matrix. To circumvent this problem, a new CNN-Transformer model was trained on the TIMIT speech corpus (Fisher, 1986). This is a corpus of 630 speakers of eight major dialects of American English that each read ten phonetically rich sentences. The TIMIT corpus includes time-aligned phonetic transcriptions, and therefore allows the CNN-Transformer to be trained to predict phones instead of letters. Following the standard protocol, the 61 phones in the TIMIT data set were collapsed to 39 phones (Lee & Hon, 1989). The phone-specific variant of the CNN-Transformer was trained using the hyperparameter settings for the 10 h model of the original Wav2Vec 2.0 paper (Baevski et al., 2020) for 20k updates.

Phone confusions were measured using the Oldenburg Logatome Corpus (OLLO), a speech database that is designed for a comparison of recognition performance of ASR systems and human listeners (Wesker et al., 2005). OLLO contains 150 different CVCs and VCVs spoken by 40 German and 10 French speakers. Each logatome is recorded in several different speaking styles. To match the format used by the phone-specific CNN-Transformer, the OLLO logatome transcriptions (which are in SAMPA format) were transcribed to the ARPabet format used in TIMIT.

A total of around 10,000 logatome recordings were tested. The prediction of the phone-specific CNN-Transformer was recorded for each logatome in three conditions (no masker, harmonic masker and speech shaped masker) at an SNR of 5 dB. To obtain a consonant confusion matrix for each condition, a square matrix C was created in which each entry $C_{x,y}$ represents how many times the y th consonant was predicted when the x th consonant was presented. If multiple consonants were predicted for a single logatome, the matrix was updated proportionally to the number of predicted consonants. That is, $\frac{1}{n}$ was added to the respective matrix entry, with n referring to the total number of consonants. A similar matrix was created for vowels.

In addition to the confusion matrices, I also measured the ratio of the accuracy for the harmonic versus the speech-shaped maskers. This can be used to reveal which phones are relatively more affected by the use of a speech-shaped masker versus a harmonic masker. Given either the consonant or vowel confusion matrix for the harmonic masker (C_{ij}^H) or speech-shaped masker (C_{ij}^S) (representing the amount of times the j th phone was predicted when the i th phone was presented) the accuracy ratio ρ_x for the x th phone is computed as follows:

$$\rho_x = \frac{\frac{C_{x,x}^S}{\sum_y C_{x,y}^S}}{\frac{C_{x,x}^H}{\sum_y C_{x,y}^H}} \quad (7.7)$$

The more the recognition rate of a specific phone is negatively affected by a speech-shaped masker, the lower the ratio.

3.2 Results

This section describes the outcome of the modulation and periodicity masker experiments from Steinmetzger and Rosen (2015) when applied to the three ASR systems. Furthermore, phone confusions of a phone-specific version of the CNN-Transformer are analysed in more detail.

Detailed Comparison ASR Systems

The amount of MPB and FMB in humans depends mostly on the intelligibility rather than the type of target speech. For example, the MPB for near-ceiling performance is around 8 dB regardless of the type of vocoding (Figure 7.4). Similarly, the FMB hovers between 2 dB to 5 dB, with the benefit being slightly higher for noise maskers compared to harmonic complex maskers. The performance of the ASR systems similarly does not show a clear effect of periodicity in the target speech, as exemplified by the MPBs and FMBs for the TANDEM-STRAIGHT sentences (Figure 7.4). However, only the CNN-Transformer model displays a clear MPB and FMB. The LSTM and DNN-HMM, on the other hand, have a negligible MPB and a negative FMB, suggesting that masker modulations interfere with speech in noise performance.

In humans, higher intelligibility is associated with both a higher MPB and FMB (Figure 7.5). The benefit for periodic maskers is always positive, but modulating maskers interfere if the target speech intelligibility is low. Of the ASR systems, only the CNN-Transformer shows both trends. The DNN-HMM model has a very small MPB of around 1 dB, but the benefit does not improve with an increase in target speech intelligibility. Both the LSTM and DNN-HMM model show an upward trend in FMB, suggesting that the modulating masker provides less interference when intelligibility is higher. However, neither achieve a positive benefit.

In line with that observed in the speech in competing talker experiment (see Section 2), modulations in the masker provide a benefit for the CNN-Transformer. The lack of modulation benefit in the DNN-HMM is somewhat surprising, since the results of the competing talker experiment suggest a small benefit of around 1 dB for noise that is modulated by the envelope of single-speaker babble compared to noise modulated by 16-speaker babble. It may be the case that the DNN-HMM requires larger gaps than those

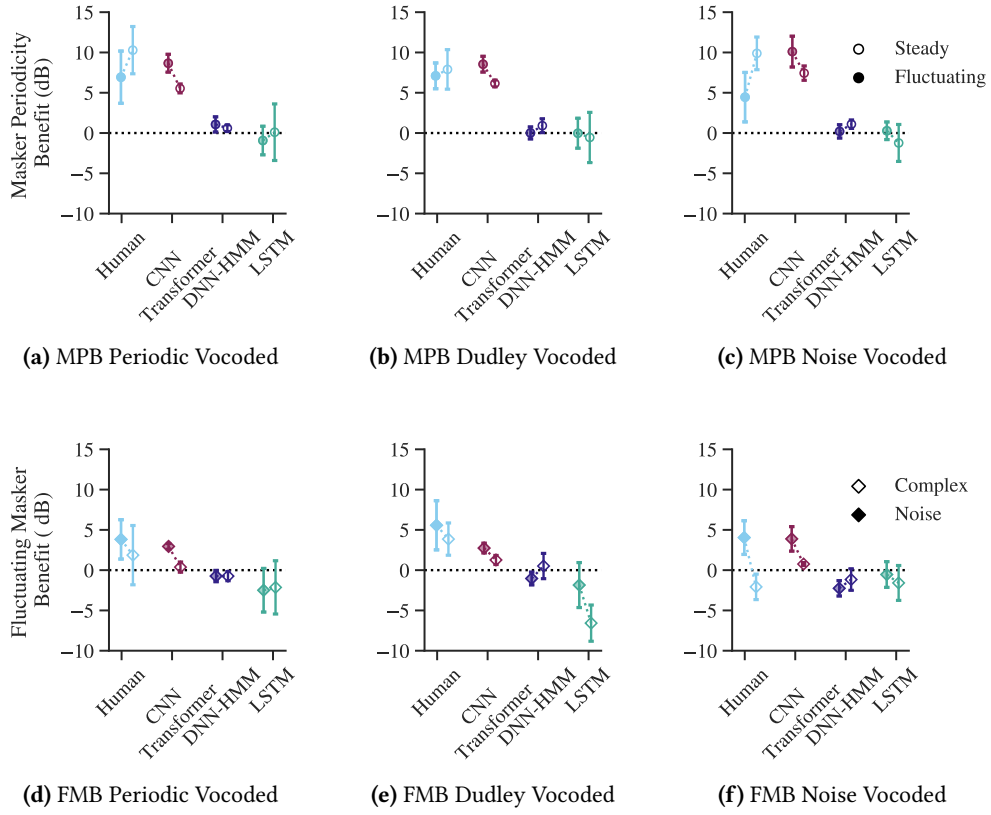


Figure 7.4. The obtained median masker periodicity benefits (MPBs) and fluctuating masker benefits (FMBs) in the masker periodicity experiment. In this experiment, SRTs were measured against both noise and harmonic complex maskers that were either modulated or not. The target speech had different levels of periodicity (see Figure 6.4), namely fully periodic (**a,d**), mixed periodic (Dudley-vocoded, **b,e**) and non-periodic (noise-vocoded, **c,f**). These were created using TANDEM-STRAIGHT, a procedure that results in natural sounding stimuli. Human data was taken from Steinmetzger and Rosen (2015), and the ASR system performance was measured on the same stimuli as human data (100 IEEE sentences, measured in trials of 20 sentences). All SRT-50s were determined using a static SRT method (see Section 5.2 in Chapter 5). Errorbars denote standard deviation. (**a-c**) MPBs are computed as the difference between the SRT-50s obtained between steady and modulated maskers. (**d-f**) FMBs denotes the difference between the and periodic (noise) and non-periodic (harmonic) masker.

introduced by the 10 Hz amplitude modulations used in this experiment. Since the average syllable rate in English ranges from 2 Hz to 5 Hz, speech-modulated noise may provide more glimpsing opportunities. Previous work has shown that a similar DNN-HMM model is able to make use of glimpsing opportunities in noise maskers that modulate with a rate of 8 Hz, although that model was trained in modulated noise (Spille & Meyer, 2017).

Steinmetzger and Rosen (2015) proposed several factors that could explain the presence

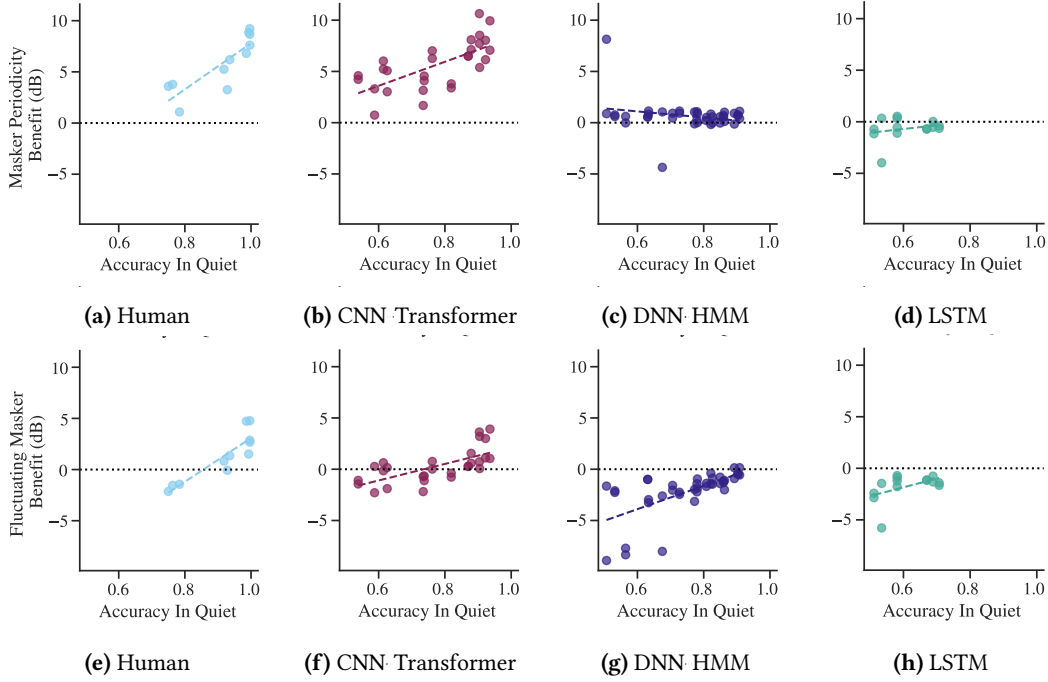


Figure 7.5. Results from the masker modulations and periodicity experiments measured over a larger range of target speech. Here, the Masker Periodicity Benefit (MPB) (a-d) and Fluctuating Masker Benefit (FMB) (e-h) are plotted as a function of the performance of the target speech in quiet for humans (a, e), the CNN-Transformer (b, f), the DNN-HMM (c, g) and the LSTM (d, h). Accuracy in quiet was measured as per target periodicity experiment (see Figure 6.4). Datapoints reflect performance for the three target vocoding types (periodic, Dudley-vocoded and noise-vocoded) with a varying number of vocoding channels, the TANDEM-STRAIGHT (TS) vocoded speech as well as unprocessed speech. The MPB and FMB were computed as described in Figure 7.4, and are based on SRT-50s pooled over all 100 IEEE sentences. Human data are based on the median values presented in Steinmetzger and Rosen (2015). For the TS and unprocessed speech, performance in quiet was estimated from the psychometric curves presented in that work.

of the periodicity benefit in humans. Firstly, harmonic complexes may allow for ‘spectral glimpsing’ in between the individual harmonics of the complex maskers. As shown in Section 2, both the DNN-HMM and the CNN-Transformer show similar patterns of sensitivity to glimpsing opportunities in noise-vocoded and modulated noise. This may account for the small periodicity benefit observed in the DNN-HMM. Secondly, periodic maskers typically hardly fluctuate outside of the F_0 modulation range itself, particularly not at the low modulation rates that are most important for speech intelligibility. However, if this were the most important factor, both the DNN-HMM and LSTM model should benefit from this too, since these two models are most sensitive to the presence of low temporal modulations (see Section 4 in Chapter 6). Lastly, harmonicity in the masker could enable

the auditory system to effectively subtract the masking sound from the signal mixture (de Cheveigné et al., 1995). In humans, the mixed target provides less MPB than the fully periodic or noise-vocoded targets. It may be that the interruptions in F_0 in the mixed target make it more difficult to form two separate auditory streams, whereas it is easier to cancel out the harmonic background in the presence of a fully aperiodic target sound or to follow two separate F_0 s. Since the DNN-HMM and LSTM models are more sensitive to the presence of periodicity at lower spectral resolutions, it may be the case that the addition of another F_0 signal is more disruptive. In contrast, the CNN-Transformer model is best at recognising noise-vocoded speech at lower spectral resolutions. The focus on unvoiced speech segments may allow this model to cancel out the harmonic complexes more easily or ignore them altogether.

Phone Confusions Matrices

As shown in the previous section, the CNN-Transformer is the only one of the tested models that displays a MPB. To further understand the underlying mechanisms that may result in an MPB, a new phone-specific CNN-Transformer model was trained. This model was tested on a logatome data set with no masker, a harmonic masker and the non-modulated speech shaped masker at a SNR of 5 dB. Confusion matrices were computed for both VCVs (Figure 7.6a-7.6c) and CVCs (Figure 7.6d-7.6f). In this section I will first briefly describe the overall shape of the confusion matrices, followed by a more detailed comparison between the phone confusions under the harmonic and speech shaped maskers. The discussion will focus on some known acoustic correlates of these phones, as discussed in more detail in Chapter 2.

Without any masking, the consonant recognition rates are high (Figure 7.6a). The most common confusion is that of the unvoiced alveolar plosive /t/ for the unvoiced alveolar fricative /s/, followed by the confusion of the /t/ with its voiced counterpart /d/. For vowels, the /ε/ is sometimes confused with an /ɪ/ (which differ in placement) while the /ɪ/ tends to be confused with the /i/ (which differ in tenseness) (Figure 7.6d).

Moving on to the harmonic maskers, many of the newly introduced confusions are exaggerations of confusions that were already present at a low level when no noise was present, such as the tendency to overpredict consonants such as /b/ and /l/ as well as the vowel /ɪ/ (Figure 7.6b and 7.6e). Notably, the harmonic complex masker tends to lead to an increased prediction of the nasals /m/ and /n/ (Figure 7.6b). Nasal consonants tend to have a low frequency prominence around 250 Hz (see Table 2.4 in Chapter 2). The ASR

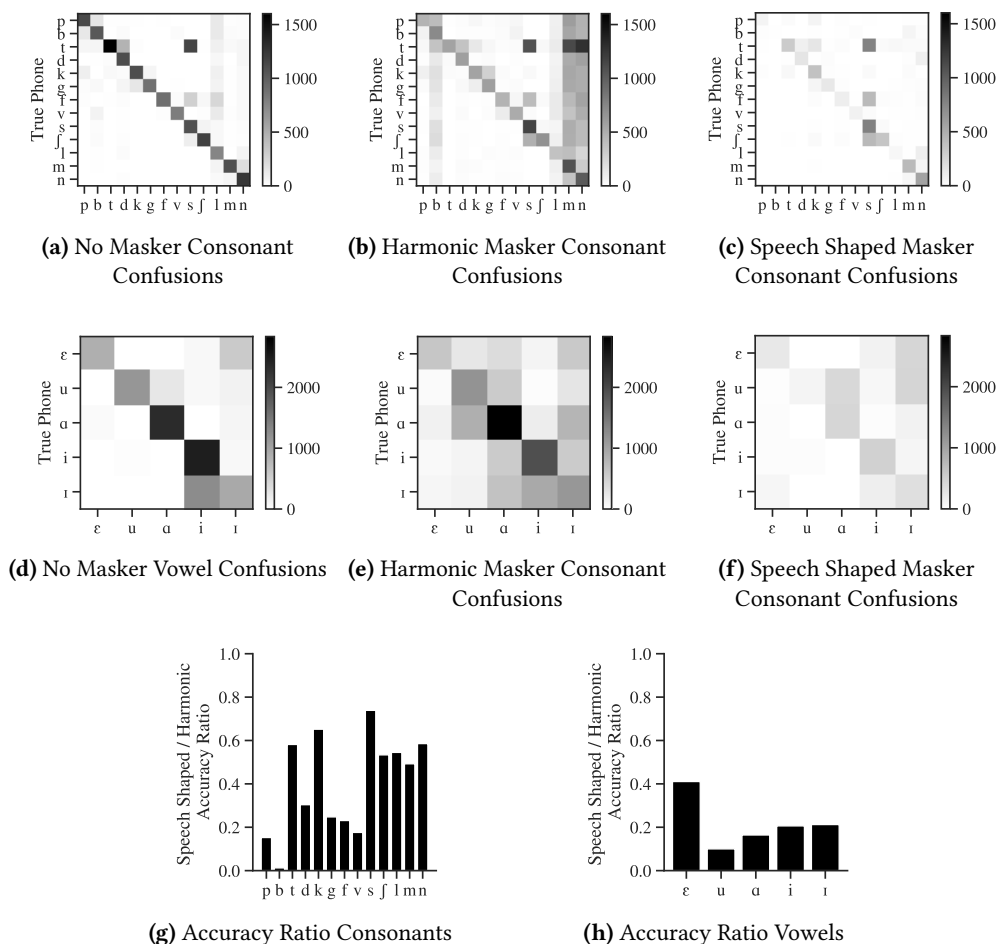


Figure 7.6. Phone confusions made by a phone-specific version of the CNN-Transformer that was trained on the TIMIT corpus and tested on the OLLO logatome corpus. (a-c) Confusion matrices for consonants measured on a set of VCVs (vowel-consonant-vowel), with either no masker (a), a harmonic complex masker (b) or a speech shaped masker (c), the latter two both at an SNR of 5 dB. Each row represents a presented phone, and each column the phone that was predicted by the ASR system. (d-f) Same as a-c but now measured for CVCs (consonant-vowel-consonant). g-h The accuracy ratio for the consonants (g) and vowels (h) between the speech shaped masker and harmonic complex masker. The accuracy for each phone under a specific masker is measured as the number of times the phone was correctly predicted (diagonal in the confusion matrix) divided by the number of times the phone was presented (the sum of each row in the confusion matrix). The accuracy ratio is then equal to the accuracies computed from the speech shaped masker confusion matrix divided by the accuracy computed from the harmonic complex masker confusion matrix.

system may mistakenly use the F_0 modulations in the harmonic complex as a cue for nasals.

As expected from the previous experiments, the performance under the speech shaped masker is much worse than under the harmonic complex masker (Figure 7.6c and 7.6f). The nasals are no longer overpredicted, and in many cases the model does not make any phone predictions at all. To more clearly understand the ways the speech shaped masker affects performance more than the harmonic complex masker, the ratio between the phone accuracies under the speech shaped and maskers are compared in Figure 7.6g and 7.6h. This shows that although overall performance is worse under the speech shaped masker in all cases, some phones are affected more than others.

For consonants, the voiced plosives (/b/, /d/ and /g/) have a lower ratio than their unvoiced counterparts (/p/, /t/ and /k/). The labial plosives (/d/ and /p/) seem particularly difficult under a speech shaped maskers. These plosives have most energy in the burst in the lower frequencies (500 Hz to 1500 Hz) and are associated with slightly faster formant transitions (see Table 2.6 Chapter 2). This suggests that the speech shaped masker interrupts the detection of consonants with an important low-frequency component, perhaps as a result of reduced spectral glimpsing opportunities. The non-strident fricatives (/f/ and /v/) also have a low accuracy ratio. As fricatives, these consonants are typically associated with a continuous noise-like spectrum. Non-strident fricatives in particular tend to have a weaker total energy than their strident counterparts. This weaker total energy may make them more susceptible to the noise-like speech shaped masker.

In the vowel experiment, the vowel /u/ has the lowest accuracy ratio. This is a high vowel, which is associated with a low F_1 (generally within 100 Hz to 200 Hz of the fundamental frequency). It is also a back vowel, which means F_2 is also low and typically very close to F_1 (see Table 2.3 in Chapter 2). The vowel with the next lowest ratio is /a/, which is a low tense back vowel. This vowel has a high F_1 (around 700 Hz to 800 Hz and an F_2 that is close to F_1 . /ɪ/ and /i/ are both high vowels and but not back vowels, resulting in a low F_1 but high F_2 . Compared to /ɪ/, /i/ has a slightly higher F_1 and lower F_2 due to its tenseness. The vowel with the highest ratio, /ε/ is a front vowel which has a low-mid tongue position. This vowel is associated with a relatively high F_2 , and a large difference between F_1 and F_2 . Altogether, it seems that vowels with a smaller formant separation are more difficult to recognise under the speech-shaped masker compared to the harmonic masker, particularly if the associated formants are low.

In summary, the results from the phone confusion experiment suggest that the harmonic complex masker produces less interference for phones that are associated with prominent low-frequency cues, such as voiced plosives or vowels with low-frequency

formants. Furthermore, fricatives with a relatively weak energy and vowels with a small formant separation also appear more difficult.

4. MODEL RETRAINING

The ASR systems tested in this and the previous chapter are ‘off the shelf’ ASR systems, meaning that they were not trained with examples of the distorted test speech. In common with most machine learning systems, ASR systems rely on the assumption that the training and evaluation data come from the same distribution. It is therefore unclear whether a reduction in performance for distorted speech is the result of the distribution shift or because the model truly relied on the feature that was distorted. By contrast, in most of the experiments described, the human listeners were given time to familiarise themselves with the distorted speech material. Unfortunately, it is not straightforward to create a perfectly fair equivalent. In contrast to humans, the tested ASR systems are not designed to be good at ‘one shot’ learning, which means their performance will only change after being presented with a large amount (hundreds or thousands) before any meaningful change in performance is measured. This makes it difficult to ‘familiarise’ these ASR systems without extensive retraining far beyond that of a human listener. To somewhat mitigate this disadvantage, distortions that humans would have been trained to listen to from a young age, such as whispered speech, were avoided in the test selection process. However, in some of the experiments humans still have a clear advantage. For example, comparing these ASR systems with speech under competing talkers is something humans have been exposed to many times before. In many ways, it is surprising that some of the ASR systems perform so similarly to humans at specific tasks without being trained to do so.

It would be tempting to think that retraining the model with examples of distorted speech would lead to a closer match. However, it is important to note that retraining to improve performance in one task can lead to worse performance at a different task. To illustrate this, I finetuned four smaller pretrained versions of the CNN-Transformer model using 100 h of the LibriSpeech dataset. These smaller pretrained models (referred to as *BASE* in the original Wav2Vec 2.0 paper of Baevski et al. (2020)) contain only 12 transformer blocks compared to the 24 transformer blocks in the large CNN-Transformer model. This smaller model size allows for faster finetuning. Like the large CNN-Transformer, the *BASE* model is pretrained using 960 h of unlabeled LibriSpeech data. For the finetuning procedure, I used the same procedure and parameter settings as those reported in Baevski et al. (2020) for finetuning of 100 h of LibriSpeech, using early stopping with at most 50k updates. For two

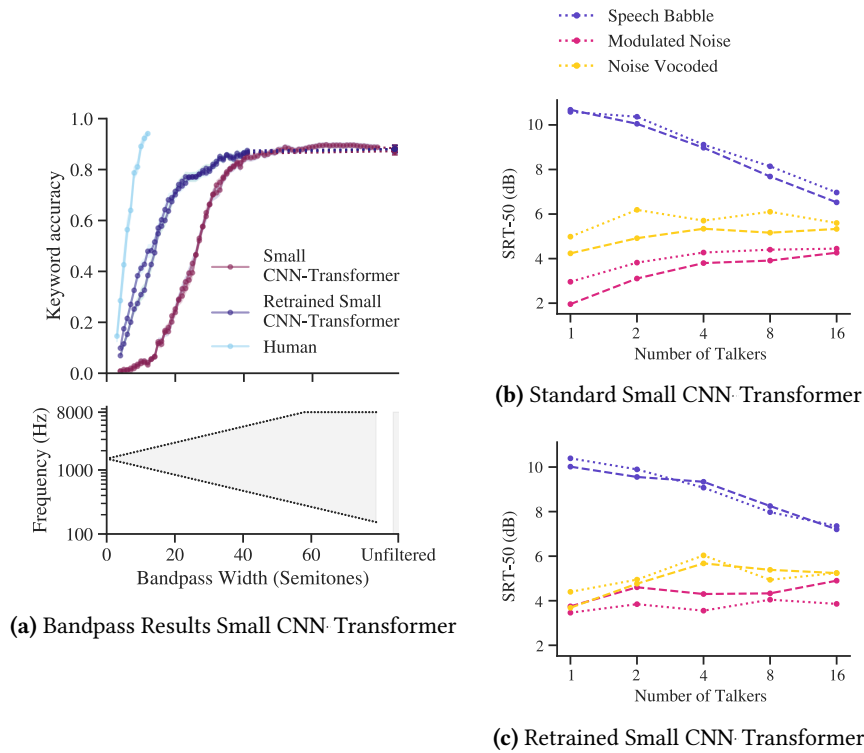


Figure 7.7. Performance of four smaller CNN-Transformer models, two of which were finetuned on bandpass filtered data, tested on the bandpass filter and masker periodicity experiment. **(a)** In the bandpass filter experiment, speech is bandpass filtered using a filter centered at 1500 Hz with varying widths (in semitones). Here the keyword accuracy is plotted as a function of the bandpass width (in semitones) at which the target speech was filtered. Same data and procedure as for Figure 6.1a, human data from Warren et al. (2000). **(b-c)** SRTs at 50% keyword accuracy in three types of noise (babble noise, noise-vocoded babble and speech-shaped noise modulated with babble envelope) with an increasing number of speakers. **b** displays the SRTs for the two small CNN-Transformers trained on unfiltered data whereas **c** shows the SRTs measured on two small CNN-Transformers that were trained on a mixture of normal and bandpass filtered data. SRT-50s were measured in the same way as in Figure 7.1.

of these models, half of the training set was bandpass filtered with a 2000 order bandpass FIR filter following the same procedure as Section 2 in Chapter 6. For each sentence, the centre frequency was randomly selected between 1 kHz to 4 kHz, and the width of each bandpass filter was randomly selected between 6 to 8 semitones.

As expected, the models trained on bandpass filtered data performed better in the bandpass filter test from Chapter 6 than the models trained on unfiltered speech data (Figure 7.7a). These models even outperform the bandpass filter performance of the much larger CNN-Transformer model that was trained on unfiltered speech (Figure 6.1a in Chapter 6).

The model accuracy is comparable to that of humans tasked with recognising bandpass filtered isolated words (Stickney & Assmann, 2001). However, further testing shows that while the spectral invariance improved, noise robustness in the retrained model is decreased (Figure 7.7b and 7.7c). In particular, the models trained on bandpass filtered speech have consistently higher SRT-50s (i.e. perform worse) for speech-modulated noise maskers (see Section 2 in Chapter 6) compared to the models trained on unfiltered data.

When one is merely interested in creating an ASR system that performs well in a specific target task, increased performance in a single dimension is likely to be sufficient. A model can also be trained on data that is distorted in multiple dimensions to further increase robustness in a multi-task setting. However, the results from this section suggest that such training may have unintended side effects. In particular, training on specific distortions may affect performance in currently unknown but nevertheless important dimensions. When the goal is to use ASR systems as a proxy for human listeners, these problems may be evaded by only training the model in realistic natural settings, such as a background of competing talkers. In the same way that human hearing is robust against certain distortions without much training, robustness against unnatural distortions may emerge in a candidate ASR system as a result of the way it has learned to account for natural noise. ASR systems that are trained using multi-condition training (see e.g. Ge et al. (2017) and Lippmann et al. (1987)), data-augmentation (see e.g. Cui et al. (2015) and Park et al. (2019)) or continual learning (which uses training data from a new application domain to gradually improve performance in old domains, see e.g. Sadhu and Hermansky (2020)) with various types of natural noise are likely good candidates for such an approach. A test battery such as the one presented in this thesis can be a good starting point for finding an ASR system that displays consistent humanlike behaviour.

5. CONCLUSIONS

In this chapter, the performances of three ASR systems were compared with that of human listeners on two SPIN tests. This set of tests teases out the factors involved in understanding speech in masking noise, including babble, modulated noise and periodic maskers. Overall, all ASR systems obtained much higher SRTs (i.e. performed worse), but certain qualitative similarities were observed between some of the ASR systems and humans. Although the results from Chapter 6 suggest that none of the ASR systems use auditory cues in the same way as humans do, the results from this chapter suggest that they may nevertheless be useful to ask more narrowly defined questions.

When comparing the three different ASR systems, it was shown that the CNN-Transformer model is most similar to humans and can make use of both glimpsing opportunities and masker periodicity to the same extent as humans. However, it does suffer more from informational masking from noise-vocoded speech. The DNN-HMM model is able to make use of glimpsing opportunities in speech-shaped modulated noise, but this ability does not extend to 10 Hz noise modulations. The LSTM model shows no benefit for masker periodicity or modulations. No benefit of having only a single competing talker was observed in any of the ASR systems, although this is likely the result of the fact that the ASR systems could not be instructed to attend to a particular speaker, whereas the human listeners were. It would be interesting to see how the performance in a competing talker experiment changes for models that are able to focus on a specific speaker.

Despite the remaining differences with robustness against certain distortions, the CNN-Transformer model tested here may already be similar enough to humans to be used to generate some narrow hypotheses about human hearing. As a proof-of-principle, the observed FMB and MPB in both experiments were further analysed using a combination of XAI methods and domain knowledge. In line with previous work (Spille & Meyer, 2017), XAI attributions suggest that the observed FMB is the result of ‘dip listening’. Furthermore, an analysis of the likelihood of various maskers confirms that noise-vocoded maskers induce lexical IM. This analysis also suggested that non-lexical informational interference (e.g. timbre or voice quality) plays a relatively large role in the amount of IM induced by (unprocessed) competing talkers. The MPB was explored further using an adapted CNN-Transformer that was trained to predict phones. Using this model, phone confusion matrices were obtained that suggest that speech-shaped noise maskers interfere more with phones that have prominent low-frequency cues than harmonic complex maskers do.

Because the CNN-Transformer is an end-to-end model it is more difficult to confirm the importance of certain spectral regions using attribution-based methods. Steinmetzger and Rosen (2015) suggested a range of mechanisms and features such as harmonic cancellation, modulations and spectral glimpsing may underlie the periodicity benefit observed in humans. The results from the phone confusion experiments cannot confirm which of these mechanisms is most important for the CNN-Transformer. By applying more extensive XAI techniques to the CNN-Transformer model, the relative contribution of these different mechanisms may be better understood, or other mechanisms could be identified.

As none of these models were trained in noisy conditions, the absolute differences in SRT levels between the ASR systems and humans are not surprising. However, in the last section of this chapter, I showed that fine tuning models to certain distortions may

negatively affect performance in other dimensions. The fact that the CNN-Transformer model is the most robust against distortions despite being trained on the least varied data suggests that it is not sufficient to simply increase the types of data it is trained on, but that further developments of the model and training procedure are required. The results from both this and the previous chapter point to tangible directions for future research into improving ASR systems.

Part IV

Conclusions

CHAPTER 8

Conclusions

Even after decades of demystifying research, the human auditory system continues to astound with its incredible robustness to distortions and noise. Understanding how the human brain accomplishes this feat may help to improve our ability to better assist those with hearing loss, as well as improve the performance of Automatic Speech Recognition (ASR) systems. The work presented in this thesis is my small contribution to this much larger endeavour. I have explored two distinct ways of using machine learning to generate new hypotheses about the basic features that may underlie human hearing. This chapter summarises the findings of these explorations. I will also discuss some limitations and outline directions for future work.

1. CONTRIBUTIONS OF THIS WORK

In Part II, which consists of Chapter 4, I showed how computational neuroscience and machine learning can be combined into a framework for auditory feature hypothesis generation. This is an example of how domain knowledge can be incorporated in a machine learning approach to strengthen the scientific consistency of the obtained results. The key to the proposed approach is simplicity. By using a simple neural model of feature detection in the early auditory brainstem, each of the model parameters can be understood in relation to their physiological correlates. By using a simple machine learning classification model, only those feature detector variants that extract useful features perform well. This simplicity is not synonymous with inadequacy. I showed that the small amount of parameters in the model can be tuned to detect a range of well-known features such as Amplitude Modulations (AMs) and onsets or combinations thereof. The simple classifier was then used to identify which feature detectors retain the relevant information present in natural speech data. With this approach I confirmed existing ideas, such as the importance of spectral peak detection and AMs. Moreover, I have shown that this framework can be used to generate ideas for

future work in lesser-known theories of hearing, such as the importance of spectral splatter suppression.

In Part III of this thesis, which covers Chapter 5 to 7, I focused on the use of ASR systems as a model for human hearing. In Chapter 6 and 7 specifically, I analysed the performance of three open source ASR system on a barrage of psychometric tests. By applying a range of distortion experiments as well as speech in noise tests, the features used by the ASR systems were compared with those used by human listeners.

The distortions tests from Chapter 6 showed that none of the ASR systems performed similarly to humans in terms of which features and cues they relied on, and which distortions they were robust to. In particular, none of the ASR systems were robust to common distortions such as bandpass filtering (commonly applied in phone communication) and peak clipping (a common distortion when a recording device is saturated). The tests showed that the relative sensitivity to spectral and temporal modulations is comparable between humans and ASR systems, but the use of Temporal Fine Structure (TFS) and periodicity differs considerably. TFS in particular is not utilised at all by any of the ASR systems, while periodicity in target speech is used in very different ways. These results suggests that these systems are not yet ready to be used as general purpose proxies for modelling human speech recognition.

In Chapter 7, I compared the speech in noise performance of the same three open source ASR systems. Although none of the ASR systems were shown to rely on the same features as humans do in Chapter 6, the speech in noise experiments from Chapter 7 suggest that there are similarities between some of the ASR systems and human listeners in a speech in noise context. In particular, I showed that the Wav2Vec 2.0 CNN-Transformer model (Baevski et al., 2020) is most similar to humans and can make use of both glimpsing opportunities and masker periodicity to the same extent as humans. However, since this model uses periodicity in target speech differently from humans, it does suffer more from informational masking from noise-vocoded speech maskers. Furthermore, no benefit of having only a single competing talker was observed in any of the ASR systems, although this is likely the result of the fact that the ASR systems could not be instructed to attend to a particular speaker. Although the results from Chapter 6 indicate that none of the ASR systems use auditory cues in exactly the same way as humans do, the results from Chapter 7 suggest that they may nevertheless be useful to ask more narrowly defined questions about speech in noise perception.

As a proof of principle of asking such more narrowly defined questions, I applied techniques from Explainable Artificial Intelligence (XAI) to explore the use of dip listening

in the Wav2vec 2.0 CNN-Transformer in Chapter 7. The similarities between ASR and human speech in noise perception has been noted in previous work that used ASR systems as a model of human hearing (Kell et al., 2018; Spille & Meyer, 2017). Similar to the results reported in Spille and Meyer (2017), the use of dip listening was confirmed through the application of attribution methods. Chapter 7 also explored the benefit of periodic maskers in the CNN-Transformer model, a phenomenon also observed in humans, in more detail through the analysis of phone-level confusion matrices.

2. LIMITATIONS AND FUTURE DIRECTIONS

While the approaches proposed in this thesis can generate new insights into the features of human hearing, they also generate a range of future research topics. In this section I will discuss some limitations of the proposed approaches as well as interesting directions for future work.

2.1 On Neuro-Inspired Feature Detector Selection

In Part II of this thesis, I applied task-driven machine learning to a simple neural feature extraction model that used precisely timed inhibition as a feature extraction mechanism. The obtained results are only based on that single model. It would be interesting to see how the outcome would change if different computational models would be applied, such as previously proposed models of AM (Carney et al., 2015), Frequency Modulation (FM) (Skorheim et al., 2014) and onset detection (Smith & Fraser, 2004).

Another limitation of this study is that the results are only based on a single classifier. This means that the results are dependant on the ability of the Long Short-Term Memory (LSTM) specifically to make use of certain auditory features. It would be interesting to see a comparison between different types of classifiers, as this would provide a less model-dependent view of which types of features and feature extraction mechanisms may be useful for a given classification task. These alternative classifiers could include other relatively complex Artificial Neural Network (ANN) approaches for time series classification, such as Convolutional Neural Networks (CNNs), but more scalable approaches are also available. For example, Dempster et al. (2020) recently introduced RandOm Convolutional Kernel Transform (ROCKET), a classifier that can achieve accurate time series classification by combining random convolutional kernels with simple linear classifiers. Methods such as ROCKET have advantages over more complex classifiers, as training is fast, the model itself

is relatively interpretable all while still achieving high performance.

A faster classifier is not the only way by which the proposed framework can be scaled up. In the current approach, each classifier is trained individually on each neural feature extraction model. An interesting direction to further explore is to feed multiple neural models as input to a classifier and impose a sparsity constraint on these inputs. This way a subset of good feature extractor models could be selected automatically, without having to train a classifier on each individual model. One step further in this direction would be to make the neural feature extraction model itself differentiable. This would allow for the parameters of the neural model to be tuned with standard gradient descent optimisation algorithms. However, I found in an exploratory study that it is difficult to optimise a neural feature detection model that incorporates delays, as these lead to a periodic loss function with many local minima in the error space. Furthermore, such an approach would be more difficult to extend to other neural models, therefore making the framework less generally applicable.

2.2 On Using ASR as a Model for Human Hearing

Looking to the future, designing ASR models that rely on similar features of speech as humans may allow them to perform better in noisy conditions they have never been exposed to. The results presented in Chapter 6 point to some features that may be important. For example, none of the tested ASR systems were able to exploit TFS information. Although TFS information is mostly lost in MFCC input features, end-to-end ASR systems should in principle be able to utilise TFS information. Other dimensions in which the ASR systems differed significantly from humans are resistance to peak clipping distortions and the use of periodicity in the target speech.

In Chapter 6 and 7, I applied a range of psychometric tests to three open source ASR systems. One potential limitation of these types of experiments is that they violate (to a certain extent) the underlying assumption of many machine learning systems that the training and test data come from the same distribution. Some work in XAI suggests that these types of problems may be overcome by retraining a model on the distorted data set (Hooker et al., 2019). However, as I showed in Chapter 7, fine tuning models to certain distortions may negatively affect performance in other dimensions. This means that such approaches may effectively change the features the machine learning model uses under the hood. An interesting direction for future work would be to develop ASR systems in which the resistance to certain distortions is an emergent property rather than something it has to be explicitly trained to do.

Building further on this idea, designing ASR models that rely on similar features of speech as humans may allow them to perform better in noisy conditions they have never been exposed to. Based on the results in this thesis, one avenue for future work involves mitigating for the lack of use of TFS as well as the differences in use of periodicity. The mitigation strategy in these cases is not as straightforward as finetuning a model to a certain distortion. In the case of TFS and periodicity the feature information is already present in the input (particularly for end-to-end systems), but is simply not used or not used correctly. Future work could focus on encouraging ASR models to use these types of features, for example by incorporating their usage through a form of regularisation. It would be particularly interesting to see whether learnable frontends (such as the contrastive loss pretraining used in Baevski et al. (2020)) can be optimised to encourage the use of these features.

Lastly, despite the remaining differences with robustness against certain distortions, the Wav2Vec 2.0 CNN-Transformer model tested here may already be similar enough to humans to be used to generate some narrow hypotheses about human hearing. Although benefits of modulated noise have been previously observed in ASR systems (Spille & Meyer, 2017), the benefit of masker periodicity in ASR systems has not been extensively reported on. This masker periodicity benefit is not yet well understood in humans. Steinmetzger and Rosen (2015) suggested a range of mechanisms and features such as harmonic cancellation, modulations and spectral glimpsing may underlie the periodicity benefit observed in humans. By applying XAI techniques to understand the mechanisms that underlie the masker periodicity benefit in the CNN-Transformer model, the relative contribution of these different mechanisms may be better understood, or other mechanisms could be identified.

2.3 On the Application of XAI to ASR Systems

In the last chapter of this thesis (Chapter 7), I used attribution methods to understand how the tested CNN-Transformer model was able to achieve a masker modulation benefit. Similar to previous work (Spille & Meyer, 2017), the attribution method pointed to the use of dip listening. It should be noted, however, that the computed attributions lead to a similar drop in accuracy in an ablation experiment as the absolute sound amplitude did. This either indicates that the attributions do not accurately represent the features that are considered important to the model, or that the ablation experiment was unsuccessful because the change in input distribution threw the model off. These two options are difficult to disentangle, since the ground truth is unknown. The lack of ground truth is a common conundrum when applying XAI methods, and an important limitation of their applicability.

This difficulty further stresses the importance of incorporating domain knowledge to verify XAI techniques rather than assuming that their proposed interpretations are meaningful.

An unfortunate side effect of using an end-to-end model is that it is much more difficult to get insights into the spectral regions of interest using attribution methods. As a result, a similarly satisfying explanation of the masker periodicity benefit that was observed in the Wav2Vec 2.0 CNN-Transformer model (Baevski et al., 2020) is more difficult to achieve using such XAI methods. Previous work has shown that by applying attribution methods to the hidden layers of a CNN, it is possible to create an interpretable proxy model for certain CNN models (Tanaka et al., 2019). This approach may be combined with attention-based XAI methods for the Transformer layers of the end-to-end model, although it should be noted that attention weights themselves are at best noisy predictors of the components' importance (Serrano & Smith, 2019). Following an approach similar to the one described in (Tanaka et al., 2019), it may be possible to create a reduced version of the CNN-Transformer model that only predicts certain phones (e.g. /b/, since the benefit of a periodic masker is most clearly present in this phone), while still displaying the masker periodicity benefit. The internal mechanisms of a small proxy model would be easier to analyse and understand. Using these or other XAI techniques may provide further insights into the results found in the psychometric experiments described in Chapter 6 and 7.

More generally, the importance of explainability in machine learning surpasses scientific interests. As machine learning systems are increasingly applied to decision-making processes in the real world, it is important to understand how these systems get to their decisions. Understanding ASR systems may provide some insights into human hearing, but may also improve their performance in the real world. For example, ASR systems may be biased to particular accents and voices (Feng et al., 2021). Understanding their internal mechanisms may provide one puzzle piece to mitigating these type of problems, alongside with improved data selection and other fairness-related mitigation strategies.

3. CONCLUDING REMARKS

There remains so much more to be uncovered about human hearing. In this thesis, I have merely scratched the surface of what is already known about the intricate complexity of the auditory system, let alone what there is still left to learn.

Looking ahead, I believe that the use of task-driven machine learning is a useful tool for generating new hypotheses about the auditory features that underlie robust human hearing. The work I presented exemplifies how machine learning can be used as a tool to

accelerate the discovery of auditory features of interest, but also points to potential pitfalls associated with using machine hearing as a model for human hearing. Furthermore, the results point to the tangible ways in which ASR systems may be improved to overcome these pitfalls. This includes the release of an open source toolbox that can be used to compare the likeness of an ASR system to human hearing.

After four years of dedicating myself to understand the human auditory system, I have to humbly admit I now have many more questions than I started with. If anything, I hope that works like mine inspire machine learning engineers and auditory neuroscientists alike to pursue the continued cross-fertilization between their fields. In the words of Alan Turing (1950), which have beamed on my screensaver throughout these past years, ‘*we can only see a short distance ahead, but we can see plenty there that needs to be done*’.

Appendices

APPENDIX A

Author's Publications

This appendix contains a list of the publications that were made during the course of this PhD, as well as the thesis chapters based on these publications.

Weerts, L., Clopath, C., & Goodman, D. F. M. (2019). A unifying framework for neuro-inspired, data-driven detection of low-level auditory features. *Conference on Cognitive Computational Neuroscience, Berlin, Germany*, PS-2B.52. <https://doi.org/10.32470/CCN.2019.1245-0>

This article introduces a framework for auditory feature hypothesis generation. This approach combines neural modelling with machine learning to determine relevant low-level auditory features. Chapter 4 is based on this article.

Weerts, L., Rosen, S., Clopath, C., & Goodman, D. F. M. (2021). The psychometrics of automatic speech recognition. *bioRxiv*

In this article, the performance of three state-of-the-art ASR systems are compared with results from humans on a barrage of standard psychometric experiments. The software is released as an open source toolbox to allow researchers to assess future ASR systems or add additional psychometric measures. Chapter 5, 6 and 7 are based on this article.

APPENDIX B

Information Bottleneck Approximation

In Chapter 4, it was mentioned that the proposed approach is inspired by the Information Bottleneck (IB) principle (Tishby et al., 2001). Generally speaking, the IB principle can be used to measure how well a variable Y (e.g. a class label) is predicted from a particular variable X (e.g. an input signal) compared to its compressed representation X^c (e.g. the model output). The relationship between the three variables Y , X and X_c is visualised in Figure B.1.

The IB principle dictates that a good compression minimises the following:

$$I(X; X^c) - \beta I(X^c; Y) \quad (\text{B.1})$$

where β defines the compression rate and I is the mutual information, which quantifies the amount of information about one variable obtained through the other. The mutual information is calculated as follows:

$$I(X, Y) = \sum_x \sum_y p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (\text{B.2})$$

Despite its strong theoretical foundations, the IB is difficult to compute because it is difficult to reliably estimate the mutual information for high dimensional variables like sounds. Instead, I approximate the IB principle in Chapter 4 by feeding the model

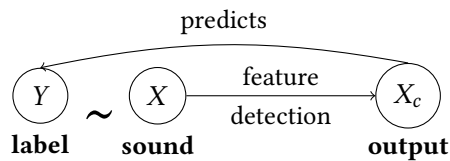


Figure B.1. Visualisation of the relation between label Y , input sound X and the model output X_c .

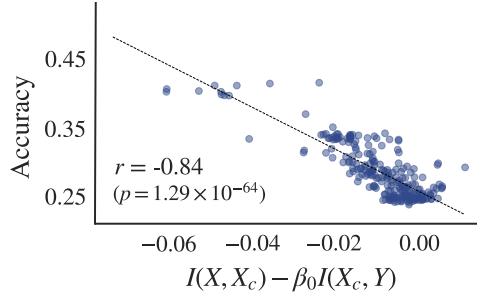


Figure B.2. LSTM accuracy versus information bottleneck for the PCA compressed outputs of 375 model variants that were restricted to a 1 kHz channel. β_0 refers to the compression ratio of the unaltered output of the 1 kHz channel which captures the maximum $I(X_c; Y)$ and $I(X; X_c)$ achievable for this restricted model.

output into a simple classifier to predict a label Y . The underlying assumption is that a simple classifier performs well if information about the label is not buried within irrelevant information (i.e. $I(X^c; Y)$ is low) while enough information about the label is retained (i.e. $I(X; X^c)$ is high). Note that a complex classifier would perhaps be able to perform well even if $I(X^c; Y)$ is high, since it can perform the relevant feature extraction itself.

To verify the extent to which the proposed approach is related to the IB, I applied it to a reduced problem. The IB can be made tractable by compressing the signal X and model output X^c to 10 dimensions using Principal Component Analysis (PCA). The mutual information values from Equation B.1 were calculated for each out of 375 neural model variants using the Java Information Dynamics Toolkit (Lizier, 2014). Following the framework described in Chapter 4, an LSTM classifier was trained for each neural model variant. The results indicate a strong correlation ($r = -0.84$, $p \ll 0.05$) between the classification accuracy and the IB formulation (Figure B.2).

APPENDIX C

Temporal Fine Structure on CRM corpus

As discussed in Chapter 6 Section 5, the CRM corpus used in K. Hopkins and Moore (2010) is a closed-set corpus in which sentences have a common structure (*"Ready [call sign] go to [colour] [number] now."*). The human participants in this experiment were instructed to pick one out of eight call signs, one out of four colours and one out of eight numbers. This makes the task much easier than recognising sentences from the IEEE dataset, which contains five keywords per sentence that are not selected from a predefined subset and are not strongly semantically related. To ensure the results are not task-dependent, the experiment was repeated on the CRM corpus. To mimic the closed-set task in a model-agnostic manner, each predicted transcription was replaced with the most similar, valid CRM sentence. This most similar sentence was determined by computing the character error rate (measured as the character-level Levenshtein distance) between the predicted transcriptions and all 256 (8 call signs, 4 colors, 8 numbers) possible sentences and selecting the sentence with the lowest error rate. As in K. Hopkins and Moore (2010), the accuracy is then reduced to a binary correct/incorrect that indicates whether or not the correct sentence was selected. To estimate the speech reception threshold the static SRT estimation procedure was used (see Chapter 5 Section 5.2).

As shown in Figure C.1, the use of the simpler task results in higher SRTs, but leads to a qualitatively similar performance. In particular, the cumulative TFS benefit is still not as high as it is in humans, and no significant TFS benefit is observed in the second notch experiment. However, some small differences in the different spectral regions in the second experiment exist. In particular, the 4th spectral region appears much less important for the LSTM tested on the CRM corpus than it was tested on IEEE data, although there is also a much higher standard error of the mean.

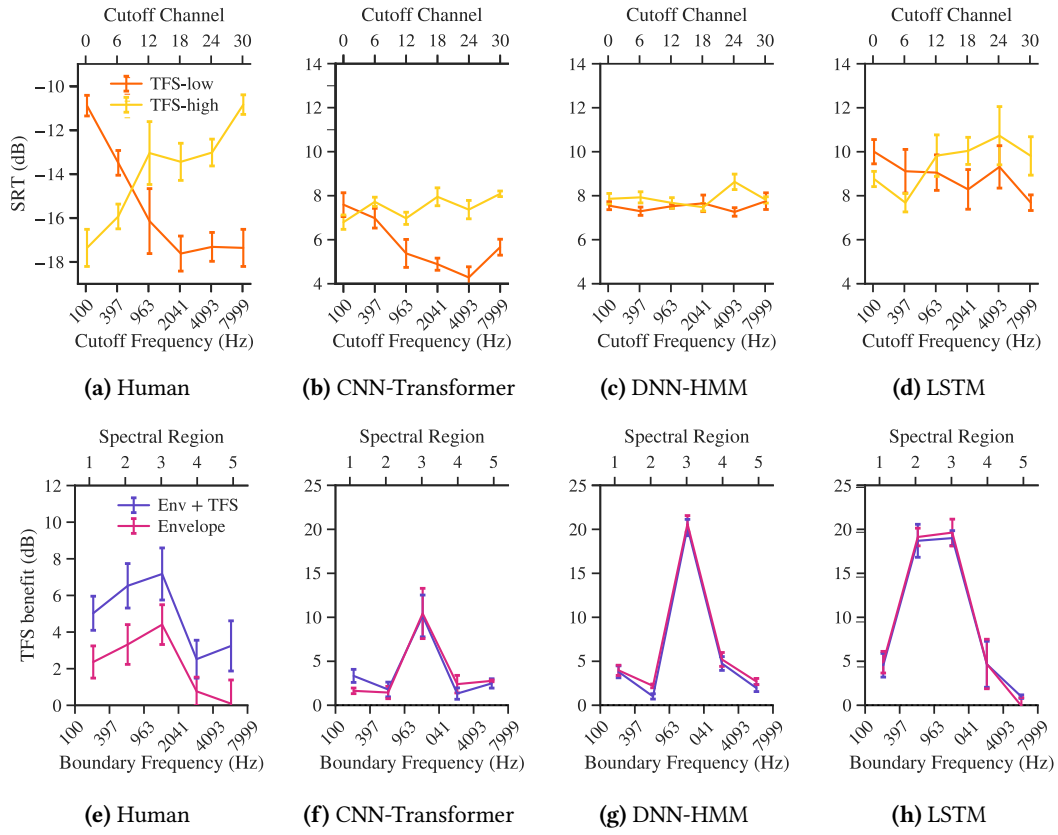


Figure C.1. Same as in Figure 6.3 in Chapter 6 but here performance is measured against 100 sentences from the CRM dataset (single male talker) rather than IEEE sentences.

APPENDIX D

Extended Attributions for Competing Talker Experiment

In Chapter 7, the performance of the CNN-Transformer model in the competing talker experiment was further investigated using the DeepLIFT attribution approach. As discussed in Section 4 in Chapter 3, there are many other proposed attribution methods. In this Appendix, the same experiment is repeated but for the saliency, Gradient $\times$ Input and Integrated Gradients attribution methods. The absolute amplitude of the sound is also included as a baseline measure. In the remainder of this Appendix, I will first briefly describe how each method works. I will then show that all attribution methods, despite some small differences, point to the same outcome: that the CNN-Transformer does indeed use dip listening, although informational masking affects its ability to do so.

1. METHOD

In addition to DeepLIFT, three more attribution methods were explored. Note that, as described in Chapter 7, each of these methods was measured using the squared variant of SmoothGrad. All methods were implemented using Captum, an interpretability library for PyTorch (Kokhlikyan et al., 2020). As a baseline, the absolute sound amplitude is also included.

Absolute Sound Amplitude Baseline As was shown in Chapter 7, the ablation test suggests that the DeepLIFT attributions performed as well as the absolute sound amplitude in terms of affecting the accuracy. It is included as a comparison baseline in this review. Of course, this is not an attribution method, since the absolute sound amplitude is the same regardless of the model.

Saliency Saliency is the most straightforward gradient attribution method out there. It

is equivalent to the absolute value of gradients with respect to the network input (Simonyan et al., 2014).

Gradient $\times$ Input As the name suggests, Gradient $\times$ Input is a variant of saliency that multiplies the model input with the gradient with respect to the network input (Shrikumar et al., 2017). This method was introduced to emphasise the attributions of the parts of the input that are actually expressed (i.e. have a high value).

Integrated Gradients Integrated Gradients is an extension of Gradient $\times$ Input that does not simply compute the gradient at the input, but instead computes the average gradient while the input varies along a linear path from some baseline input $\bar{x}$ (such as a black image or empty sentence) to the input x . In theory, Integrated Gradients is ‘complete’, meaning that all attributions add up to the output value. However, this is only true if enough samples are taken across the linear path, which can be excessively high for highly non-linear systems. DeepLIFT approximates the quantity computed in Integrated Gradients by replacing the gradient at each nonlinearity with the average gradient. As a result, DeepLIFT acts as an efficient approximation of Integrated Gradients as long as the gradients behave approximately linearly (Ancona et al., 2017). Because of the high computational complexity of Integrated Gradients, these attributions were only computed for a single sentence (rather than 100 sentences).

2. RESULTS

The results found using Integrated Gradients (Figure D.1d) are very similar to those found with DeepLIFT (D.1a). Note that there is some overlap between the attributions when they are visualised (Figure D.1f and D.1c for Integrated Gradients and DeepLIFT, respectively), as DeepLIFT can be viewed as an approximation of the Integrated Gradients method.

As discussed in Chapter 7, the removal of inputs with the highest absolute amplitude leads to a similar drop in accuracy as for most attribution methods (Figure D.2a). This could imply that Deep Learning Important FeaTures (DeepLIFT) does not give any more information besides the overall sound amplitude provides. However, the correlation coefficients of the moving SNR and the absolute sound amplitude show that, in contrast to what was found for all attribution methods, this correlation is mostly independent from the type of noise masker.

The differences between the correlations found using DeepLIFT and the absolute

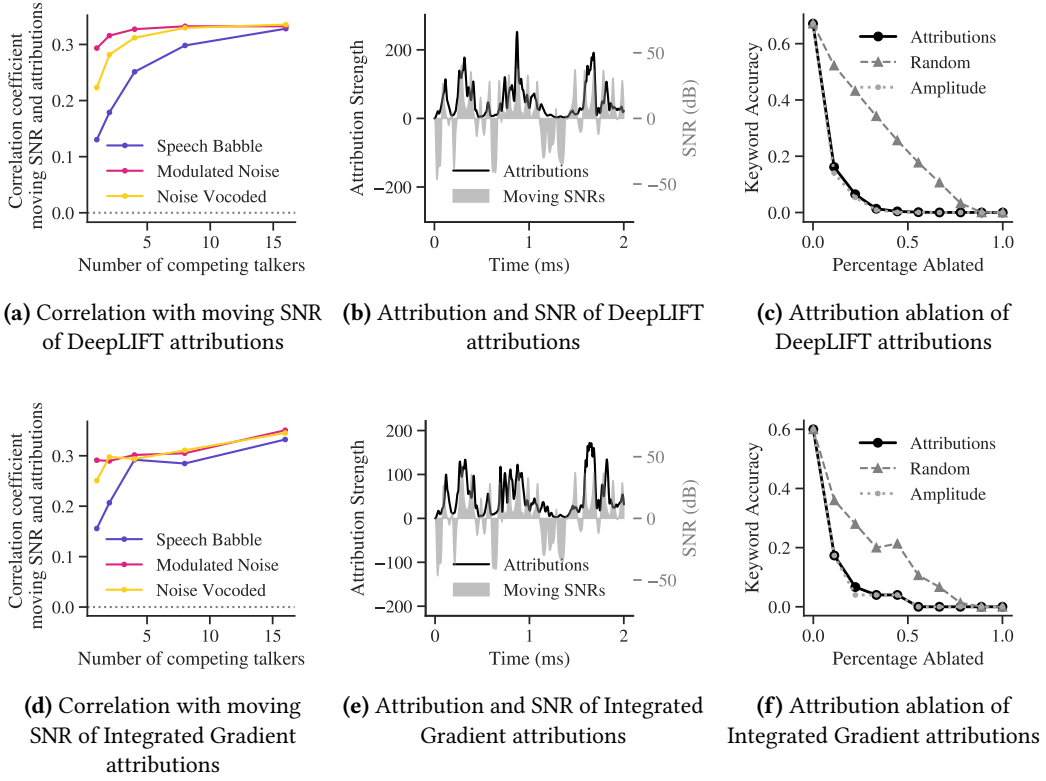


Figure D.1. Measurements of EM in the CNN-Transformer in the competing talker experiment (as per Figure 7.3 in Chapter 7) for DeepLIFT (a-c) and Integrated Gradients (d-f). Note that while all other attribution methods were measured against 100 sentences, Integrated Gradients was only measured against one sentence due to its high computational cost. (a, d) The correlation coefficient between the moving SNR and attributions for the three different maskers. (b, e) Visualisation of the measured attributions versus the moving average for a single target sentence under a masker modulating at 10 Hz. (c, f) Accuracy in an ablation experiment, where inputs are removed in order of attribution size. A random order and ordering based on absolute amplitude are included.

amplitude can likely be attributed to the gradients of the model with respect to the input. This is evidenced by the correlation of the SNR with the saliency attribution method, which only measures this gradient. Saliency attributions display the clearest divide between the three different noise types (Figure D.2d). However, the saliency method is slightly less effective at dropping the accuracy than the other attribution methods (Figure D.2f).

Perhaps unsurprisingly, the Gradient $\times$ Input method leads to correlations that are in between those found using sound amplitudes and the saliency method (D.2i). The results found with Gradient $\times$ Input method are highly similar to those found using DeepLIFT.

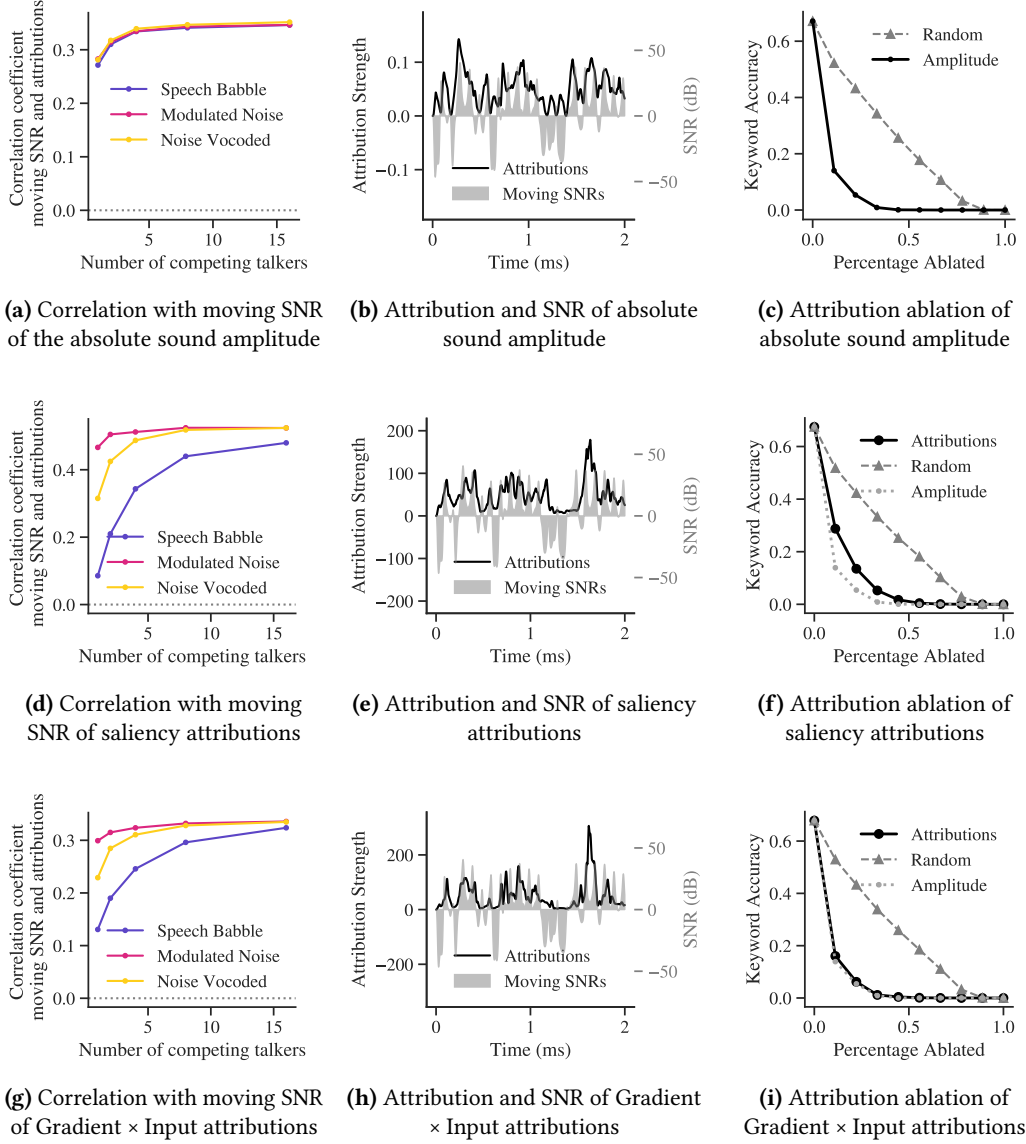


Figure D.2. Measurements of EM in the CNN-Transformer in the competing talker experiment (as per Figure 7.3 in Chapter 7) using the absolute sound amplitude (**a-c**), saliency (**d-f**) or Gradient $\times$ Input (**g-i**) as attribution method. (**a,d,g**) The correlation coefficient between the moving SNR and attributions for the three different maskers. (**b,e,h**) Visualisation of the measured attributions versus the moving average for a single target sentence under a masker modulating at 10 Hz. (**c,f,i**) Accuracy in an ablation experiment, where inputs are removed in order of attribution size. A random order and ordering based on absolute amplitude are included.

Glossary

Acoustic phonetics A field that studies the physical properties of speech. 11

Amplitude Modulation (AM) Fluctuations in the amplitude (signal strength) of a carrier wave. 6, 22, 36-38, 73, 74, 78, 80, 81, 83-86, 139, 155, 157, 178

Artificial Neural Network (ANN) A machine learning algorithm (very) loosely inspired by biological neural networks, consisting of units (called 'neurons' or sometimes perceptrons) that can send their respective outputs to one another through weighted connections. The connections are typically organised in layers, with all neurons of the preceding layer feeding into the next. Each neuron computes a non-linear function of the weighted sum of its inputs, and outputs a real number. 40, 44, 47, 89, 157, 176, 177, 179

Attribution method An Explainable Artificial Intelligence (XAI) technique that aims to reveal components of high importance in the input of a machine learning algorithm, typically visualised as a feature importance map. 55, 62, 69, 159

Auditory filters The cochlea is sometimes described as an array of overlapping bandpass filters, which are referred to as auditory filters. 18

Auditory Nerve fibres (AN fibres) Nerve fibres that carry auditory sensory information from the cochlea to the brain. 34-36, 74, 75

Automatic Speech Recognition (ASR) The transcription of spoken language into text by computers. 1, 6, 7, 39-41, 43-47, 55, 65, 69, 70, 89-91, 93, 95-101, 103-113, 115, 117-123, 125-129, 132-135, 137, 138, 140-142, 144, 145, 147, 149-151, 155-161, 178

Basilar membrane A stiff structure in the cochlea that moves up and down in response to incoming sound waves and plays an important role in the frequency decomposition of incoming sounds in the auditory pathway. 34, 35

- Cochlea** Spiral structure in the inner ear that converts sound pressure patterns from the outer ear into electrochemical impulses. These impulses are sent to the brain through the auditory nerves. 34, 35, 113, 175
- Cochlear Nucleus (CN)** Nucleus in the brainstem on which the auditory nerve fibres innervate 35, 36, 86
- Connectionist Temporal Classification (CTC) criterion** A sequence level objective function that enables neural networks to be trained on a sequence of speech signals without the requirements of prior alignments. 47, 49, 92-94, 127, 129, 132, 134, 136
- Convolutional Neural Network (CNN)** A class of Deep Neural Network (DNN) with a shared-weight architecture of convolutional filters that slide along the input features. This type of neural network is commonly used in computer vision tasks, but can also be applied to other domains such as audio. 50, 53, 65, 69, 93-95, 97, 99, 104, 106-111, 113, 114, 116, 117, 120-123, 126, 132-141, 143-145, 147, 148, 150, 151, 157, 160
- Deep Learning Important FeaTures (DeepLIFT)** A method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. 130-132, 135-137, 172
- Deep Neural Network (DNN)** A rebranding of the Artificial Neural Network (ANN), with ‘deep’ referring to the increased number of layers. 6, 41, 43, 44, 46, 47, 50, 55, 58-61, 63, 65, 69, 70, 89-95, 97, 99, 104, 106-111, 113, 114, 116, 117, 120-123, 125, 126, 130, 133, 134, 141-144, 150, 176
- Distinctive feature theory** A framework that describes speech through phonetic features. 11, 12, 37
- Energetic Masking (EM)** Masking of target speech as a result of the presence of masker energy in the same frequency region(s) as the energy in the target signal. 7, 33, 126, 127, 129, 130, 132, 134, 135
- Envelope** The slower fluctuations in the amplitude of a sound, as contrasted with temporal fine structure. 19, 22, 104, 109, 114-118
- Explainable Artificial Intelligence (XAI)** Forms of artificial intelligence in which the results of the algorithm and/or the way it got to those results can be understood by people. 7, 55, 56, 68, 70, 90, 126, 132, 150, 156, 158-160, 175

- Factorised Time Delay Neural Network (TDNN-f)** A TDNN is an Artificial Neural Network (ANN) architecture that can classify temporal patterns with shift-invariance through 1D convolution over time). The factorised TDNN is structurally the same as a TDNN whose layers have been compressed via a Singular Value Decomposition (SVD). 92, 93, 95
- Fluctuating Masker Benefit (FMB)** The difference between the Speech Reception Threshold (SRT) of a fluctuating masker versus a steady masker in dB. 138, 139, 141–143, 150
- Formants** Resonances introduced by changing the shape of the vocal tract. These show up as local peaks in the sound spectrum. 13, 18, 24, 32, 40, 83, 85, 112
- Frequency Modulation (FM)** Fluctuations in the frequency of a carrier wave. 22, 36–38, 74, 80, 95, 157
- Fundamental frequency F_0** The lowest frequency of a periodic waveform. 13, 23, 33, 118, 119, 121, 138, 143, 144, 146
- Gaussian Mixture Model (GMM)** A probabilistic model that models the data as a mixture of a finite number of Gaussian distributions with unknown parameters. 44, 109
- Harmonics** Spectral peaks at frequencies that are multiples of the fundamental frequency of a periodic sound. 20, 23, 107, 109, 138, 139, 141–144, 150
- Hidden Markov Model (HMM)** A statistical model in which the system that is being modeled is assumed to be a Markov process with unobservable 'hidden' states. Each hidden state only depends on the previous state. The HMM is commonly used to model temporal sequences such as speech. 24, 40, 41, 43, 44, 46, 47, 49, 91–95, 97, 99, 104, 106–111, 113, 114, 116, 117, 120–123, 125, 126, 130, 133, 134, 141–144, 150
- Inferior Colliculus (IC)** The principal midbrain nucleus of the auditory pathway thought to integrate the outputs of preceding auditory processing stages. 35, 36, 74, 86
- Informational Masking (IM)** Type of masking which occurs when the target speech and masker are difficult to disentangle despite a lack of energetic overlap, for example because the masker has a similar voice quality, or the linguistic content of the masker is distracting (lexical interference). 7, 33, 127, 129, 132, 134–136, 150

Interpretability The extent to which certain properties of a machine learning model are presented in human-understandable features (such as verbal descriptions or heatmaps). This can be achieved through transparency and post-hoc interpretability. 56, 57, 60

Lattice-Free Maximum Mutual Information (LF-MMI) A criterion that enables sequence-level HMM state posteriors to be estimated using DNN-based acoustic models. 49, 92, 93

Long Short-Term Memory (LSTM) A type of Recurrent Neural Network (RNN) designed to model long-term temporal dependencies through the use of a internal cell state and three gates that regulate the flow of information in and out of the cell. 41, 50–52, 62, 64, 65, 69, 79, 91–95, 97, 99, 104, 106–114, 116, 117, 120–123, 126, 133, 134, 141, 143, 144, 150, 157

Long-Term Average Spectrum (LTAS) The power spectrum of the frequencies of a speech sample as estimated through a fast Fourier transform. 128, 138, 139

Masker Periodicity Benefit (MPB) The difference between the Speech Reception Threshold (SRT) of a periodic masker versus a non-periodic masker in dB. 138, 139, 141–144, 150

Maximum Likelihood Estimation (MLE) A statistical method for estimating the parameters of a probability distribution by maximising the likelihood of the data, that is, the parameters for which the observed data is most probable under the assumed statistical model. 100

Mel-Frequency Cepstral Coefficients (MFCCs) A cepstral representation of the short-term power spectrum of a sound that is commonly used as input features to Automatic Speech Recognition (ASR) systems. MFCCs are based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. 41–43, 46, 69, 91–95, 104, 107, 109, 113, 114, 117, 121, 123, 125, 126

Modulation Power Spectrum (MPS) The time-frequency power in modulation across both spectral and temporal dimensions, i.e. the spectrogram of a spectrogram. 110, 113

Modulation Transfer Function (MTF) A function that characterises the response of the feature to various Amplitude Modulations (AMs). 78, 80, 81

Peak clipping A common distortion in which the peaks of a waveform are flattened or clipped. 32, 106-109, 122, 123

Periodicity Periodicity refers to the repetition of signal waveforms. In relation to speech, periodicity is the result of the periodic vibration of the vocal cords when voiced speech is produced. It determines the fundamental frequency and is reflected in the spectrum by the appearance of harmonically spaced spectral components. 20, 22, 23, 103, 109, 118-123, 125-127, 138, 141-144, 148, 150, 156, 159

Phonetic features Language-independent features that can be used to precisely categorise most speech sounds found in most languages. 11, 12

Post-hoc interpretability Methods to extract information about the decision making process of trained models without needing to explain the entire model at once. 56, 57, 60, 62, 66-68, 178

Recurrent Neural Network (RNN) A class of ANNs in which outputs from previous time steps are allowed to feed into nodes in the current time step. This allows for the modelling of temporal sequences. 41, 46, 47, 49-52, 65, 66, 79, 92, 178

Root Mean Squared (RMS) The root of the mean of the squared differences between two values. In the context of audio processing, it usually refers to the RMS of difference between the audio samples and the mean audio sample value and can be used to estimate the Sound Pressure Level (SPL). 98, 119, 128

Short Time Fourier Transform (STFT) A transform that can determine the sinusoidal frequency and phase content of segments of a signal as it changes over time. 18, 41-43

Signal to Noise Ratio (SNR) A measure that compares the level of a target signal to the level of background noise. It is defined as the ratio of signal power to the noise power (each of which is typically given in decibels). 98-101, 104, 110, 111, 113, 121, 127-130, 134-137, 140, 144, 145

Sound Pressure Level (SPL) The pressure level of a sound (or more precisely, the local pressure deviation from the ambient atmospheric pressure), measured in decibels. 98, 179, 180

Spectro-Temporal Receptive Field (STRF) A representation of which types of stimuli excite or inhibit an auditory neuron, represented as a graph with a temporal and spectral axis. 79, 80, 83

Speech in Noise (SPIN) Test A speech recognition test where the recognition rate is measured as a function of the level of additive noise. These types of tests are often used in clinical assessments of hearing loss. 7, 90, 96, 125, 126, 129, 149, 180

Speech Reception Threshold (SRT) The Sound Pressure Level (SPL) at which a listener achieves a certain level of accuracy (usually 50% or 70.7%) in a Speech in Noise (SPIN) task. 91, 97, 99-101, 115-117, 128, 133, 134, 139, 142, 143, 148-150, 177, 178

Superior Olivary Complex (SOC) Collection of nuclei in the brainstem that play an important role in multiple aspects of hearing, such as sound localisation. 35, 36

Support Vector Machine (SVM) A supervised machine learning model that maps examples in space so as to maximise the width of the gap (also called the support) between two categories of training examples. 65

Temporal Fine Structure (TFS) Rapid oscillations that range from 600-10 kHz that cause variations of wave shape within single periods of periodic sounds, or short time intervals of aperiodic sounds, as contrasted with the slower fluctuations of the sound envelope. 19, 21, 22, 32, 95, 96, 103, 104, 109, 114-117, 122, 123, 156, 158, 159

Transformer A sequence-to-sequence neural network which uses an attention mechanism. The attention mechanism weighs the influence of different parts of the input data, allowing the model to ‘attend’ to various parts of the input sequence. 50, 69, 93-95, 97, 99, 104, 106-111, 113, 114, 116, 117, 120-123, 126, 132-141, 143-145, 147, 148, 150, 151

Transparency A machine learning system is transparent if it is inherently understandable because the model itself, its components or the training algorithm are interpretable. 56, 57, 60, 178

Vocoder A category of speech coding that synthesizes speech. Typically, the term vocoder is used to refer to a channel vocoder, which filters a speech signal into several contiguous frequency bands and processes the signal in each band in such a way that only envelope cues are preserved. 20, 114-122, 127, 128, 133, 134, 136, 138, 139, 142-144, 148, 150

Word Error Rate (WER) A common metric of the performance of a speech recognition system based on the minimum number of substitutions, insertions and deletions required to obtain the target transcription from the predicted transcription, averaged by the total number of words in the target sentence. 109

Bibliography

- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.
- Adiga, A., Magimai, M., & Seelamantula, C. S. (2013). Gammatone wavelet cepstral coefficients for robust speech recognition. *TENCON 2013-2013 IEEE Region 10 Conference (31194)*, 1–4. <https://doi.org/10.1109/TENCON.2013.6718948>
- Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., & Rauber, A. (2014). Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics*, 20(12), 1703–1712. <https://doi.org/10.1109/TVCG.2014.2346660>
- Alzantot, M., Balaji, B., & Srivastava, M. (2018). Did you hear that? Adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*.
- Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.
- Arai, T., & Greenberg, S. (1998). Speech intelligibility in the presence of cross-channel spectral asynchrony. *1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2, 933–936. <https://doi.org/10.1109/ICASSP.1998.675419>
- Arai, T., Pavel, M., Hermansky, H., & Avendano, C. (1999). Syllable intelligibility for temporally filtered LPC cepstral trajectories. *The Journal of the acoustical society of America*, 105(5), 2783–2791. <https://doi.org/10.1121/1.426895>
- Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2016). Explaining predictions of non-linear classifiers in NLP. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 1–7. <https://doi.org/10.18653/v1/W16-1601>
- Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8), e0181142. <https://doi.org/10.1371/journal.pone.0181142>
- Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6), 1304–1312. <https://doi.org/10.1121/1.1914702>

BIBLIOGRAPHY

- Baby, D., Van Den Broucke, A., & Verhulst, S. (2021). A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nature Machine Intelligence*, 3(2), 134–143.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 12449–12460).
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015*.
- Barker, J., Marxer, R., Vincent, E., & Watanabe, S. (2017). The third ‘CHiME’ speech separation and recognition challenge: Analysis and outcomes. *Computer Speech and Language*, 46, 605–626. <https://doi.org/10.1016/j.csl.2016.10.005>
- Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6), 1554–1563.
- Benesty, J., Sondhi, M. M., Huang, Y., & Greenberg, S. (2009). *Springer handbook of speech processing*. (Vol. 126). Springer-Verlag. <https://doi.org/10.1121/1.3203918>
- Besold, T. R., d’Avila Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K.-U., Lamb, L. C., Lowd, D., Lima, P. M. V., et al. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv: 1711.03902*.
- Bidelman, G. M. (2013). The role of the auditory brainstem in processing musically relevant pitch. *Frontiers in psychology*, 4, 264. <https://doi.org/10.3389/fpsyg.2013.00264>
- Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 2403–2424. <https://doi.org/10.1109/TPAMI.2011.142>
- Binder, A., Bach, S., Montavon, G., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. *Information science and applications (icisa) 2016* (pp. 913–922). Springer. https://doi.org/10.1007/978-981-10-0557-2_87
- Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization*. MIT press. <https://doi.org/10.7551/mitpress/6391.001.0001>

- Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *The Journal of the Acoustical Society of America*, 67(2), 648–662. <https://doi.org/10.1121/1.383890>
- Boersma, P. (2021). Praat: Doing phonetics by computer. <http://www.praat.org/>.
- Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107(2), 1065–1066.
- Brand, A., Behrend, O., Marquardt, T., McAlpine, D., & Grothe, B. (2002). Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, 417(6888), 543–547. <https://doi.org/https://doi.org/10.1038/417543a>
- Breedlove, S. M., Watson, N. V., & Rosenzweig, M. R. (2007). *Biological psychology: An introduction to behavioral, cognitive, and clinical neuroscience (5th ed.)*. Sinauer Associates.
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. <https://doi.org/https://doi.org/10.7551/mitpress/1486.001.0001>
- Carlson, R., Fant, G., & Granström, B. (1974). Two-formant models, pitch, and vowel perception. *Acta Acustica united with Acustica*, 31(6), 360–362. <https://doi.org/10.1016/B978-0-12-248550-3.50008-8>
- Carney, L. H., Li, T., & McDonough, J. M. (2015). Speech Coding in the Brain: Representation of Vowel Formants by Midbrain Neurons Tuned to Sound Fluctuations. *eNeuro*, 2(4), 1–12. <https://doi.org/10.1523/ENEURO.0004-15.2015>
- Casseday, J., & Covey, E. (1992). Frequency tuning properties of neurons in the inferior colliculus of an FM bat. *Journal of Comparative Neurology*, 319(1), 34–50. <https://doi.org/10.1002/cne.903190106>
- Catford, J. C. (1988). *A practical introduction to phonetics*. Clarendon Press Oxford.
- Chabot, A. (2019). What’s wrong with being a rhotic? *Glossa: a journal of general linguistics*, 4(1). <https://doi.org/http://doi.org/10.5334/gjgl.618>
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5), 975–979. <https://doi.org/10.1121/1.1907229>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. <https://doi.org/10.3115/v1/D14-1179>

BIBLIOGRAPHY

- Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 613–622. <https://doi.org/10.18653/v1/P17-1057>
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136. <https://doi.org/10.18653/v1/P18-1198>
- Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90), 297–301. <https://doi.org/10.1090/S0025-5718-1965-0178586-1>
- Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9), 1469–1477.
- de Cheveigné, A. (1998). Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, 103(3), 1261–1271. <https://doi.org/10.1121/1.423232>
- de Cheveigné, A., McAdams, S., Laroche, J., & Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *The Journal of the Acoustical Society of America*, 97(6), 3736–3748. <https://doi.org/10.1121/1.412389>
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798. <https://doi.org/10.1109/TASL.2010.2064307>
- Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word*, 8(3), 195–210. <https://doi.org/10.1080/00437956.1952.11659431>
- Delgutte, B., Hammond, B., & Cariani, P. (1998). Neural coding of the temporal envelope of speech: Relation to modulation transfer functions. *Psychophysical and physiological advances in hearing*, 595–603.
- Dempster, A., Petitjean, F., & Webb, G. I. (2020). Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5), 1454–1495. <https://doi.org/10.1007/s10618-020-00701-z>
- Deng, L. (2016). Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, 5. <https://doi.org/10.1017/ATSIP.2015.22>

- Dirks, D. D., & Wilson, R. H. (1969). The effect of spatially separated sound sources on speech intelligibility. *Journal of speech and hearing research*, 12(1), 5–38.
- Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*.
- Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable ai really mean? a new conceptualization of perspectives. *Proceedings of the International Workshop on Comprehensibility and Explainability in Artificial Intelligence and Machine Learning*, 2071.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 210–215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2), 1053–1064. <https://doi.org/10.1121/1.408467>
- Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7). <https://doi.org/10.5555/1953048.2021068>
- Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, 11(2), 169–177. <https://doi.org/10.1121/1.1916020>
- Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16, 18. <https://doi.org/10.2139/ssrn.2972855>
- Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS comput biol*, 5(3), e1000302. <https://doi.org/10.1371/journal.pcbi.1000302>
- Endriss, U., & Leite, J. (2014). Multi-objective learning of accurate and comprehensible classifiers—a case study. *STAIRS 2014: Proceedings of the 7th European Starting AI Researcher Symposium*, 264, 220. <https://doi.org/10.3233/978-1-61499-421-3-220>
- ES ETSI. (2002). 202 050 v1. 1.3: Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. *ETSI standard*.

BIBLIOGRAPHY

- Fant, G., Liljencrants, J., & Lin, Q.-g. (1985). A four-parameter model of glottal flow. *STL-QPSR*, 4(1985), 1–13.
- Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.
- Fisher, W. M. (1986). The darpa speech recognition research database: Specifications and status. *Proc. DARPA Workshop on Speech Recognition, Feb. 1986*, 93–99.
- Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4), 377–430.
- Fontaine, B., Goodman, D. F., Benichoux, V., & Brette, R. (2011). Brian hears: Online auditory processing using vectorization over channels. *Frontiers in Neuroinformatics*, 5, 9. <https://doi.org/10.3389/fninf.2011.00009>
- Fontaine, B., MacLeod, K. M., Lubejko, S. T., Steinberg, L. J., Köppl, C., & Peña, J. L. (2014). Emergence of band-pass filtering through adaptive spiking in the owl's cochlear nucleus. *Journal of neurophysiology*, 112(2), 430–445. <https://doi.org/10.1152/jn.00132.2014>
- Fontan, L., Cretin-Maitenaz, T., & Füllgrabe, C. (2020). Predicting speech perception in older listeners with sensorineural hearing loss using automatic speech recognition. *Trends in hearing*, 24, 2331216520914769.
- Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1), 1–10. <https://doi.org/10.1145/2594473.2594475>
- Gales, M., & Young, S. (2008). The application of hidden markov models in speech recognition.
- Ganesh, J., Gupta, M., & Varma, V. (2017). Interpretation of semantic tweet representations. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 95–102. <https://doi.org/10.1145/3110025.3110083>
- Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE transactions on pattern analysis and machine intelligence*, 34(3), 417–435. <https://doi.org/10.1214/11-AOAS495>
- Gaubitch, N. D., Brookes, M., Naylor, P. A., & Sharma, D. (2010). Bayesian adaptive method for estimating speech intelligibility in noise. *Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges*.
- Ge, F., Li, K., Wu, B., Siniscalchi, S. M., Yan, Y., & Lee, C.-H. (2017). Joint training of multi-channel-condition dereverberation and acoustic modeling of microphone array speech for robust distant speech recognition. *INTERSPEECH-2017*, 3847–3851.

- Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18), 4613–4618. <https://doi.org/10.1073/pnas.1716999115>
- Gibson, J., van Segbroeck, M., & Narayanan, S. (2014). Comparing time-frequency representations for directional derivative features. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH-2014*, 612–615.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. <https://doi.org/10.1109/DSAA.2018.00018>
- Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2), 103–138. [https://doi.org/10.1016/0378-5955\(90\)90170-T](https://doi.org/10.1016/0378-5955(90)90170-T)
- Gold, B., Morgan, N., & Ellis, D. (2011). Brief History of Automatic Speech Recognition. <https://doi.org/10.1002/9781118142882.ch4>
- Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism (Cambridge, England)*, 12(3), 385. <https://doi.org/10.1017/S1366728909990150>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 2672–2680. <https://doi.org/10.5555/2969033.2969125>
- Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*.
- Gordon, M., Barthaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, 141–174. <https://doi.org/10.1017/S0025100302001020>
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267), 1–13. <https://doi.org/10.3389/fnins.2013.00267>
- Graves, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

BIBLIOGRAPHY

- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine learning*, 369–376. <https://doi.org/10.1145/1143844.1143891>
- Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3, 6645–6649. <https://doi.org/10.1109/ICASSP.2013.6638947>
- Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.
- Greenberg, S., & Ainsworth, W. A. (2004). Speech Processing in the Auditory System: An Overview. *Speech processing in the auditory system* (pp. 1–62). Springer-Verlag. https://doi.org/10.1007/0-387-21575-1_1
- Greenberg, S., Arai, T., Kingsbury, B., Morgan, N., Shire, M., Silipo, R., & Wu, S.-L. (1999). Syllable-based speech recognition using auditory like features. *Journal of the Acoustical Society of America*, 105, 1157–1158.
- Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, 87(6), 2592–2605.
- Grézl, F., Karafiát, M., Kontár, S., & Cernocky, J. (2007). Probabilistic and bottle-neck features for LVCSR of meetings. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4, IV–757. <https://doi.org/10.1109/ICASSP.2007.367023>
- Griffin, D., & Lim, J. (1983). Signal estimation from modified short-time fourier transform. *1983 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 32(2), 236–243. <https://doi.org/10.1109/ICASSP.1983.1172092>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1–42. <https://doi.org/10.1145/3236009>
- Halle, M., Hughes, G. W., & Radley, J.-P. (1957). Acoustic properties of stop consonants. *The Journal of the Acoustical Society of America*, 29(1), 107–116. <https://doi.org/10.1121/1.1908634>
- Halle, M., & Stevens, K. N. (1991). Knowledge of language and the sounds of speech. *Music, language, speech and brain* (pp. 1–19). Springer. https://doi.org/10.1007/978-1-349-12670-5_1
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deepspeech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

- Hanson, B. A., & Applebaum, T. H. (1993). Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech. *1993 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2, 79–82. <https://doi.org/10.1109/ICASSP.1993.319235>
- Hayward, K. (2014). *Experimental phonetics: An introduction*. Routledge.
- Hazan, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Perception & Psychophysics*, 49(2), 187–200. <https://doi.org/10.3758/BF03205038>
- Heinz, M. G., Colburn, H. S., & Carney, L. H. (2001). Evaluating auditory performance limits: I. one-parameter discrimination using a computational model for the auditory nerve. *Neural computation*, 13(10), 2273–2316. <https://doi.org/10.1162/089976601750541804>
- Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. *European Conference on Computer Vision*, 3–19. https://doi.org/10.1007/978-3-319-46493-0_1
- Hermann, L. (1894). Nachtrag zur Untersuchung der Vocalcurven (“Addendum to the investigation of vowel lines”). *Arch. Ges. Physiol*, 58, 264–279.
- Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4), 1738–1752. <https://doi.org/10.1121/1.399423>
- Hermansky, H., Ellis, D. P. W., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. *2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3, 1635–1638. <https://doi.org/10.1109/ICASSP.2000.862024>
- Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1992). RASTA-PLP speech analysis technique. *1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1, 121–124. <https://doi.org/10.1109/ICASSP.1992.225957>
- Hirsch, H. G. (2005). FaNT: filtering and noise adding tool. *Niederrhein University of Applied Sciences*.
- Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32.
- Hopkins, C., Graetzer, S., & Seiffert, G. (2019). ARU speech corpus (University of Liverpool). <http://datacat.liverpool.ac.uk/681/>
- Hopkins, K., & Moore, B. C. J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *The Journal of the Acoustical Society of America*, 125(1), 442–446. <https://doi.org/10.1121/1.3037233>

BIBLIOGRAPHY

- Hopkins, K., & Moore, B. C. J. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, 127(3), 1595–1608. <https://doi.org/10.1121/1.3293003>
- Hopkins, K., Moore, B. C. J., & Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, 123(2), 1140–1153. <https://doi.org/10.1121/1.2824018>
- Hopkins, K., Moore, B. C. J., & Stone, M. A. (2010). The effects of the addition of low-level, low-noise noise on the intelligibility of sentences processed to remove temporal envelope information. *The Journal of the Acoustical Society of America*, 128(4), 2150–2161. <https://doi.org/10.1121/1.3478773>
- Hu, G., Determan, S. C., Dong, Y., Beeve, A. T., Collins, J. E., & Gai, Y. (2020). Spectral and temporal envelope cues for human and automatic speech recognition in noise. *Journal of the Association for Research in Otolaryngology*, 21(1), 73–87. <https://doi.org/10.1007/s10162-019-00737-z>
- Huang, X., Acero, A., Hon, H.-W., & Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3), 574–591.
- IEEE. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3), 225–246. <https://doi.org/10.1109/TAU.1969.1162058>
- International Phonetic Association and International Phonetic Association Staff et al. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- Ives, D. T., & Patterson, R. D. (2008). Pitch strength decreases as F0 and harmonic resolution increase in complex tones composed exclusively of high harmonics. *The Journal of the Acoustical Society of America*, 123(5), 2670–2679. <https://doi.org/10.1121/1.2890737>
- Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1–15. <https://doi.org/10.1016/j.wocn.2018.07.001>
- Joris, P., Schreiner, C., & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological reviews*, 84(2), 541–577.

- Juang, B.-H., & Rabiner, L. R. (2005). Automatic speech recognition - a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1, 67.
- Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks* (Doctoral dissertation).
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2000). *Principles of neural science* (Vol. 4). McGraw-hill New York.
- Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.
- Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10), 2318–2331. <https://doi.org/10.1109/TKDE.2017.2720168>
- Kates, J. M., & Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The journal of the acoustical society of America*, 117(4), 2224–2237. <https://doi.org/10.1121/1.1862575>
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3933–3936. <https://doi.org/10.1109/ICASSP.2008.4518514>
- Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3), 630–644. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kim, C., & Stern, R. M. (2010). Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4574–4577. <https://doi.org/10.1109/ICASSP.2010.5495570>
- Kim, D.-S., Lee, S.-Y., & Kil, R. M. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on speech and audio processing*, 7(1), 55–69.
- Kindermans, P.-J., Schütt, K., Müller, K.-R., & Dähne, S. (2016). Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

BIBLIOGRAPHY

- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch.
- Kollmeier, B., Spille, C., Martínez, A. M. C., Ewert, S. D., & Meyer, B. T. (2020). Modelling human speech recognition in challenging noise maskers using machine learning. *Acoustical Science and Technology*, 41(1), 94–98. <https://doi.org/10.1250/ast.41.94>
- Krishna, B. S., & Semple, M. N. (2000). Auditory temporal processing: Responses to sinusoidally amplitude-modulated tones in the inferior colliculus. *Journal of neurophysiology*, 84(1), 255–273. <https://doi.org/10.1152/jn.2000.84.1.255>
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages* (Vol. 1012). Blackwell Oxford.
- Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2912–2920. <https://doi.org/10.1109/CVPR.2016.318>
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1), 1–8. <https://doi.org/10.1038/s41467-019-08987-4>
- Lee, K.-F., & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11), 1641–1648. <https://doi.org/10.1109/29.46546>
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & psychophysics*, 63(8), 1279–1292. <https://doi.org/10.3758/BF03194543>
- Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3), 1350–1371. <https://doi.org/10.1214/15-AOAS848>
- Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. <https://doi.org/10.1109/TASLP.2014.2304637>
- Liao, Q., & Poggio, T. (2017). *Object-oriented deep learning* (tech. rep.). Center for Brains, Minds and Machines (CBMM).
- Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8), 1. <https://doi.org/10.1037/h0093673>

- Liberman, A. (1970). Some characteristics of perception in the speech mode. *Perception and its Disorders*, 48, 238–254.
- Licklider, J. C. R., & Pollack, I. (1948). Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *The Journal of the Acoustical Society of America*, 20(1), 42–51. <https://doi.org/10.1121/1.1906346>
- Licklider, J. C. (1946). Effects of amplitude distortion upon the intelligibility of speech. *The Journal of the Acoustical Society of America*, 18(2), 429–434. <https://doi.org/10.1121/1.1902443>
- Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12), 1586–1604. <https://doi.org/10.1109/PROC.1979.11540>
- Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
- Lippmann, R., Martin, E., & Paul, D. (1987). Multi-style training for robust isolated-word speech recognition. *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 12, 705–708.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1), 48–56. <https://doi.org/10.1016/j.visinf.2017.01.006>
- Lizier, J. T. (2014). Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1, 11. <https://doi.org/10.3389/frobt.2014.00011>
- Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, 106(4), 2097–2103. <https://doi.org/10.1121/1.427954>
- Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *arXiv preprint cs/0205028*. <https://doi.org/10.3115/1225403.1225421>
- Lowerre, B. (1990). The HARPY speech understanding system. *Readings in speech recognition*, 576–586. <https://doi.org/10.1016/B978-0-08-051584-7.50053-X>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 4765–4774.

BIBLIOGRAPHY

- Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. <https://doi.org/10.18653/v1/D15-1166>
- Ma, D., Ryant, N., & Liberman, M. (2020). Probing acoustic representations for phonetic properties. *arXiv preprint arXiv:2010.13007*.
- Macdonald, J., Wäldchen, S., Hauch, S., & Kutyniok, G. (2019). A rate-distortion framework for explaining neural network decisions. *arXiv preprint arXiv:1905.11092*.
- Maheswaranathan, N., McIntosh, L. T., Tanaka, H., Grant, S., Kastner, D. B., Melander, J. B., Nayebi, A., Brezovec, L., Wang, J., Ganguli, S., & Baccus, S. A. (2019). The dynamic neural code of the retina for natural scenes. *bioRxiv*. <https://doi.org/10.1101/340943>
- Malek, J., Cerva, P., Seps, L., & Nouza, J. (2016). Study on the use and adaptation of bottleneck features for robust speech recognition of nonlinearly distorted speech. *SIGMAP*, 65–71. <https://doi.org/10.5220/0005955500650071>
- Martens, D., Baesens, B., van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3), 1466–1476. <https://doi.org/10.2139/ssrn.878283>
- Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, 51(4), 782–793. <https://doi.org/10.1016/j.dss.2011.01.013>
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *INTERSPEECH-2017*. <https://doi.org/10.21437/Interspeech.2017-1386>
- McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. *MIT Press, Cambridge MA*, 3–44.
- Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, 89(6), 2866–2882. <https://doi.org/10.1121/1.400725>
- Meddis, R., Lecluyse, W., Clark, N. R., Jürgens, T., Tan, C. M., Panda, M. R., & Brown, G. J. (2013). A computer model of the auditory periphery and its application to the study of hearing. *Advances in Experimental Medicine and Biology*, 787, 11–20. <https://doi.org/10.1007/978-1-4614-1590-9-2>
- Meyer, B. T., Jürgens, T., Wesker, T., Brand, T., & Kollmeier, B. (2010). Human phoneme recognition depending on speech-intrinsic variability. *The Journal of the Acoustical Society of America*. <https://doi.org/10.1121/1.3493450>

- Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 22(2), 167–173. <https://doi.org/10.1121/1.1916995>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Milner, B., & Shao, X. (2006). Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction. *IEEE transactions on audio, speech, and language processing*, 15(1), 24–33. <https://doi.org/10.1109/TASL.2006.876880>
- Molau, S., Hilger, F., & Ney, H. (2003). Feature space normalization in adverse acoustic conditions. *2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1, I–I. <https://doi.org/10.1109/ICASSP.2003.1198866>
- Molnar, C. (2020). *Interpretable Machine Learning*. Lulu. com.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 193–209). Springer. https://doi.org/10.1007/978-3-030-28954-6_10
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining non-linear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65, 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>
- Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, 9(4), 399–406. <https://doi.org/10.1007/s10162-008-0143-x>
- Moore, B. C. J. (2013). An introduction to the psychology of hearing (6th ed.). *Brill, Boston*, 11.
- Mordvintsev, A., Olah, C., & Tyka, M. (2015). Inceptionism: Going deeper into neural networks. <https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
- Moritz, N., Schädler, M. R., Adiloglu, K., Meyer, B. T., Jürgens, T., Gerkmann, T., Kollmeier, B., Doclo, S., & Goetze, S. (2013). Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction. *Proc. of CHiME*, 1–6.

BIBLIOGRAPHY

- Mozer, M. C. (1989). A focused back-propagation algorithm for temporal pattern recognition. *Complex systems*, 3(4), 349–381.
- Mozilla. (2020). Deepspeech. <https://github.com/mozilla/DeepSpeech>
- Nagamine, T., Seltzer, M. L., & Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. *INTERSPEECH-2015*, 1912–1916.
- Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2), 1085–1099.
- Noé, P.-G., Parcollet, T., & Morchid, M. (2020). Cgcnn: Complex gabor convolutional neural network on raw speech. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7724–7728. <https://doi.org/10.1109/ICASSP40776.2020.9054220>
- Palaz, D., Doss, M. M., & Collobert, R. (2015). Convolutional neural networks-based continuous speech recognition using raw speech signal. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4295–4299. <https://doi.org/10.1109/ICASSP.2015.7178781>
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519. <https://doi.org/10.1145/3052973.3053009>
- Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. *INTERSPEECH-2019*, 2613–2617. <https://doi.org/10.21437/Interspeech.2019-2680>
- Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 2.
- Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Simulation*, 83, 429–446. <https://doi.org/10.1016/B978-0-08-041847-6.50054-X>

- Pearce, D. (2000). Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition front-ends. *AVIOS 2000: The Speech Applications Conference, 2000*, 261–264.
- Pecka, M., Brand, A., Behrend, O., & Grothe, B. (2008). Interaural time difference processing in the mammalian medial superior olive: The role of glycinergic inhibition. *Journal of Neuroscience*, 28(27), 6914–6925. <https://doi.org/10.1523/JNEUROSCI.1660-08.2008>
- Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *INTERSPEECH-2015*, 3214–3218.
- Pickett, J. M. (1980). *The sounds of speech communication: A primer of acoustic phonetics and speech perception*. Univ Park Press.
- Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Allyn; Bacon Boston. <https://doi.org/10.1121/1.1289701>
- Pieraccini, R., & Lubensky, D. (2005). Spoken language communication with machines: the long and winding road from research to business. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 6–15. https://doi.org/10.1007/11504894_2
- Poorjam, A., Jensen, J., Little, M., & Christensen, M. (2017). Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis. *INTERSPEECH-2017, 2017-August*, 289–293. <https://doi.org/10.21437/Interspeech.2017-378>
- Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. *INTERSPEECH-2018*, 3743–3747. <https://doi.org/10.21437/Interspeech.2018-1417>
- Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., & Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. *INTERSPEECH-2016*, 2751–2755. <https://doi.org/10.21437/Interspeech.2016-595>
- Pressnitzer, D., Meddis, R., Delahaye, R., & Winter, I. M. (2001). Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus. *Journal of Neuroscience*, 21(16), 6377–6386. <https://doi.org/10.1523/JNEUROSCI.21-16-06377.2001>
- Qian, P., Qiu, X., & Huang, X.-J. (2016). Investigating language universal and specific properties in word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1478–1488. <https://doi.org/10.18653/v1/P16-1140>

BIBLIOGRAPHY

- Rader, T., Adel, Y., Fastl, H., & Baumann, U. (2015). Speech perception with combined electric-acoustic stimulation: A simulation and model comparison. *Ear and hearing*, 36(6), e314–e325. <https://doi.org/10.1097/AUD.000000000000178>
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348, 683–693. <https://doi.org/10.1016/j.jcp.2017.07.050>
- Rauber, P. E., Fadel, S. G., Falcao, A. X., & Telea, A. C. (2016). Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, 23(1), 101–110. <https://doi.org/10.1109/TVCG.2016.2598838>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the mössbauer technique. *The Journal of the Acoustical Society of America*, 49(4B), 1218–1231.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Robinson, A., & Fallside, F. (1987). *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering Cambridge.
- Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5), 589–600. <https://doi.org/10.1109/TKDE.2007.190734>
- Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive science*, 38(6), 1024–1077. <https://doi.org/10.1111/cogs.12148>
- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216. <https://doi.org/10.1109/ACCESS.2020.2976199>
- Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1278), 367–373.
- Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, 133(4), 2431–2443. <https://doi.org/10.1121/1.4794379>

- Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. *Backpropagation: Theory, architectures and applications*, 1–34.
- Saad, D. (1998). Online algorithms and stochastic approximations. *Online Learning*, 5, 6–3.
- Sadhu, S., & Hermansky, H. (2020). Continual learning in automatic speech recognition. *INTERSPEECH-2020*, 1246–1250.
- Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., & Vinyals, O. (2015). Learning the speech front-end with raw waveform cldnns. *INTERSPEECH-2015*, 1–5.
- Samek, W., Wiegand, T., & Müller, K.-R. (2018). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1(1), 39–48. <https://doi.org/11.1002/pub/8129fdff-en>
- Schädler, M. R., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, 54(sup2), 100–107.
- Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. *INTERSPEECH-2019*, 3465–3469. <https://doi.org/10.21437/Interspeech.2019-1873>
- Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. MIT Press. <https://doi.org/10.7551/mitpress/7942.001.0001>
- Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1), 1–8. <https://doi.org/10.1038/ncomms13890>
- Sellick, P., Patuzzi, R., & Johnstone, B. (1982). Measurement of basilar membrane motion in the guinea pig using the mössbauer technique. *The journal of the acoustical society of America*, 72(1), 131–141.
- Seltzer, M. L., Yu, D., & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7398–7402. <https://doi.org/10.1109/ICASSP.2013.6639100>
- Serrano, S., & Smith, N. A. (2019). Is attention interpretable? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. <https://doi.org/10.18653/v1/P19-1282>

BIBLIOGRAPHY

- Shah, J., Singla, Y. K., Chen, C., & Shah, R. R. (2021). What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. *arXiv preprint arXiv:2101.00387*.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234), 303–304. <https://doi.org/10.1126/science.270.5234.303>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153.
- Silverman, S. R., & Hirsh, I. J. (1955). Cx problems related to the use of speech in clinical audiometry. *Annals of Otology, Rhinology & Laryngology*, 64(4), 1234–1244.
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *Workshop at International Conference on Learning Representations*.
- Skorheim, S., Razak, K., & Bazhenov, M. (2014). Network models of frequency modulated sweep detection. *PLoS ONE*, 9(12). <https://doi.org/10.1371/journal.pone.0115196>
- Slaney, M. et al. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35(8).
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Smith, L. S., & Fraser, D. S. (2004). Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses. *IEEE Transactions on Neural Networks*, 15(5), 1125–1134. <https://doi.org/10.1109/TNN.2004.832831>
- Snell, R. C., & Milinazzo, F. (1993). Formant location from LPC analysis data. *IEEE transactions on Speech and Audio Processing*, 1(2), 129–134. <https://doi.org/10.1109/89.222882>
- Snyder, R. L., Vollmer, M., Moore, C. M., Rebscher, S. J., Leake, P. a., & Beitel, R. E. (2000). Responses of inferior colliculus neurons to amplitude-modulated intracochlear electrical pulses in deaf cats. *Journal of neurophysiology*. <https://doi.org/10.1152/jn.2000.84.1.166>
- Spencer et al. (2015). Broadband onset inhibition can suppress spectral splatter in the auditory brainstem. *PloS one*, 10(5), e0126500. <https://doi.org/10.1371/journal.pone.0126500>
- Spille, C., Kollmeier, B., & Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, 52, 123–140. <https://doi.org/10.1016/j.csl.2018.04.003>

- Spille, C., & Meyer, B. T. (2017). Listening in the dips: Comparing relevant features for speech recognition in humans and machines. *INTERSPEECH-2017*, 2968–2972. <https://doi.org/10.21437/Interspeech.2017-1168>
- Steinmetzger, K., & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *The Journal of the Acoustical Society of America*, 138(6), 3586–3599. <https://doi.org/10.1121/1.4936945>
- Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *The Journal of the Acoustical Society of America*, 68(3), 836–842. <https://doi.org/10.1121/1.384823>
- Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press. <https://doi.org/10.1121/1.1327577>
- Stevens, S. S., Volkman, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3), 185–190. <https://doi.org/10.1121/1.1901999>
- Stewart, R., & Ermon, S. (2017). Label-free supervision of neural networks with physics and domain knowledge. *Thirty-First AAAI Conference on Artificial Intelligence*.
- Stickney, G. S., & Assmann, P. F. (2001). Acoustic and linguistic factors in the perception of bandpass-filtered speech. *The Journal of the Acoustical Society of America*, 109(3), 1157–1165. <https://doi.org/10.1121/1.1340643>
- Stolcke, A., & Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. *INTERSPEECH-2017*, 137–141. <https://doi.org/10.21437/Interspeech.2017-1544>
- Strange, W., & Bohn, O.-S. (1998). Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies. *The Journal of the Acoustical Society of America*, 104(1), 488–504. <https://doi.org/10.1121/1.423299>
- Strobelt, H., Gehrmann, S., Pfister, H., & Rush, A. M. (2017). Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1), 667–676. <https://doi.org/10.1109/TVCG.2017.2744158>
- Sturm, I., Lapuschkin, S., Samek, W., & Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of neuroscience methods*, 274, 141–145. <https://doi.org/10.1016/j.jneumeth.2016.10.008>
- Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5), 828–841. <https://doi.org/10.1109/TEVC.2019.2890858>

BIBLIOGRAPHY

- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 3319–3328.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.
- T. Irino, R. D. P. (1997). A Time Domain, Level-Dependent Auditory Filter: The Gammachirp. *The Journal of the Acoustical Society of America*, 101(1), 412–419. <https://doi.org/10.1121/1.417975>
- Tachioka, Y., Narita, T., & Ishii, J. (2014). Speech recognition performance estimation for clipped speech based on objective measures. *Acoustical Science and Technology*, 35(6), 324–326. <https://doi.org/10.1250/ast.35.324>
- Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S., & Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Advances in Neural Information Processing Systems*, 8537–8547.
- Tishby, N., Pereira, F., & Bialek, W. (2001). The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49.
- Trudgill, P. (2001). The sociolinguistics of modern rp. *Sociolinguistic Variation and Change*. Edinburgh: Edinburgh Univ. Press., 176–=178.
- Tsang, M., Cheng, D., & Liu, Y. (2018). Detecting statistical interactions from neural network weights. *6th International Conference on Learning Representations*.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460. https://doi.org/10.1007/978-1-4020-6710-5_3
- Tyler, R., Preece, J., & Lowder, M. (1986). The iowa audiovisual speech perception laser videodisc. laser videodisc and laboratory report. *Iowa City, IA: Department of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics*.
- van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees. *International Conference on Inductive Logic Programming*, 269–279. https://doi.org/10.1007/978-3-540-78469-2_26
- van der Horst, R., Leeuw, A. R., & Dreschler, W. A. (1999). Importance of temporal-envelope cues in consonant recognition. *The Journal of the Acoustical Society of America*, 105(3), 1801–1809. <https://doi.org/10.1121/1.426718>

- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579–2605.
- van Gestel, T., Baesens, B., van Dijke, P., Suykens, J., Garcia, J., & Alderweireld, T. (2005). Linear and nonlinear credit scoring by combining logistic regression and support vector machines. *Journal of credit Risk*, 1(4).
- van Noorden, L. (1982). Two channel pitch perception. *Music, mind, and brain* (pp. 251–269). Springer. https://doi.org/10.1007/978-1-4684-8917-0_13
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.
- Viikki, O., Bye, D., & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. *1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2, 733–736. <https://doi.org/10.1109/ICASSP.1998.675369>
- von Rueden, L., Mayer, S., Garcke, J., Bauckhage, C., & Schuecker, J. (2019). Informed machine learning—towards a taxonomy of explicit integration of knowledge into machine learning. *Learning*, 18, 19–20.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, 841. <https://doi.org/10.2139/ssrn.3063289>
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3), 328–339. <https://doi.org/10.1016/B978-0-08-051584-7.50037-1>
- Walker, K. M., Bizley, J. K., King, A. J., & Schnupp, J. W. (2011). Multiplexed and robust representations of sound features in auditory cortex. *Journal of Neuroscience*, 31(41), 14565–14576. <https://doi.org/10.1523/JNEUROSCI.2074-11.2011>
- Wang, J., Fujimaki, R., & Motohashi, Y. (2015). Trading interpretability for accuracy: Oblique treed sparse additive models. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254. <https://doi.org/10.1145/2783258.2783407>
- Wang, S., Qian, Y., & Yu, K. (2017). What does the speaker embedding encode? *INTERSPEECH-2017*, 1497–1501. <https://doi.org/10.21437/Interspeech.2017-1125>
- Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. *Proceedings of the 2016 conference on empirical methods in natural language processing*, 606–615. <https://doi.org/10.18653/v1/D16-1058>

BIBLIOGRAPHY

- Warren, R. M., Bashford Jr, J. A., & Lenz, P. W. (2000). Intelligibility of bandpass speech: Effects of truncation or removal of transition bands. *The Journal of the Acoustical Society of America*, 108(3), 1264–1268. <https://doi.org/10.1121/1.1287710>
- Weerts, L., Clopath, C., & Goodman, D. F. M. (2019). A unifying framework for neuro-inspired, data-driven detection of low-level auditory features. *Conference on Cognitive Computational Neuroscience, Berlin, Germany*, PS–2B.52. <https://doi.org/10.32470/CCN.2019.1245-0>
- Weerts, L., Rosen, S., Clopath, C., & Goodman, D. F. M. (2021). The psychometrics of automatic speech recognition. *bioRxiv*.
- Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. *International Conference on Latent Variable Analysis and Signal Separation*, 91–99. https://doi.org/10.1007/978-3-319-22482-4_11
- Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4), 339–356.
- Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., & Kollmeier, B. (2005). Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. *INTERSPEECH-2005*.
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8), 1293–1313. <https://doi.org/10.3758/BF03194544>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37–52. <https://doi.org/10.1016/B978-0-12-409547-2.14892-9>
- Wu, J., Lim, J. J., Zhang, H., Tenenbaum, J. B., & Freeman, W. T. (2016). Physics 101: Learning Physical Object Properties from Unlabeled Videos. *BMVC*, 2(6), 7. <https://doi.org/10.5244/C.30.39>
- Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., & Andrianakis, I. (2007). Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing. *Computer Speech & Language*, 21(3), 543–561. <https://doi.org/10.1016/j.csl.2006.11.001>
- Yan, Y., Zhu, J., Duda, M., Solarz, E., Sripada, C., & Koutra, D. (2019). Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 772–782. <https://doi.org/10.1145/3292500.3330921>

- Young, E. D., & Sachs, M. B. (2008). Auditory nerve inputs to cochlear nucleus neurons studied with cross-correlation. *Neuroscience*, 154(1), 127–138. <https://doi.org/10.1016/j.neuroscience.2008.01.036>
- Zahavy, T., Ben-Zrihem, N., & Mannor, S. (2016). Graying the black box: Understanding dqns. *International Conference on Machine Learning*, 1899–1908.
- Zeghidour, N., Teboul, O., Quitry, F. d. C., & Tagliasacchi, M. (2021). LEAF: A learnable frontend for audio classification. *arXiv preprint arXiv:2101.08596*.
- Zeghidour, N., Usunier, N., Kokkinos, I., Schaiz, T., Synnaeve, G., & Dupoux, E. (2018). Learning filterbanks from raw speech for phone recognition. *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*, 5509–5513. <https://doi.org/10.1109/ICASSP.2018.8462015>
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53
- Zhang, Q., Cao, R., Shi, F., Wu, Y. N., & Zhu, S.-C. (2018). Interpreting CNN knowledge via an explanatory graph. *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7204–7208. <https://doi.org/10.1109/ICASSP.2013.6639061>
- Zilany, M. S. a., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *The Journal of the Acoustical Society of America*, 135(1), 283–286. <https://doi.org/10.1121/1.4837815>
- Zurow, D., Makarov, D., Allison Rose, R., & Drábek, V. (2021). Daanzu/kaldi-active-grammar: V2.1.0. *Zenodo*. <https://doi.org/10.5281/zenodo.4671179>