ABSTRACT

Title of Dissertation: MODELING ADAPTABILITY MECHANISMS OF SPEECH PERCEPTION
Nika Jurov
Doctor of Philosophy, 2024

Dissertation Directed by: Professor Naomi H. Feldman and
Professor William Idsardi
Department of Linguistics

Speech is a complex, redundant and variable signal happening in a noisy and ever changing world. How do listeners navigate these complex auditory scenes and continuously and effortlessly understand most of the speakers around them? Studies show that listeners can quickly adapt to new situations, accents and even to distorted speech. Although prior research has established that listeners rely more on some speech cues (or also called features or dimensions) than others, it is yet not understood how listeners weight them flexibly on a moment-to-moment basis when the input might deviate from the standard speech.

This thesis computationally explores flexible cue re-weighting as an adaptation mechanism using real speech corpora. The computational framework it relies on is rate distortion theory. This framework models a channel that is optimized on a trade off between distortion and rate: on the one hand, the input signal should be reconstructed with minimal error after it goes through the channel. On the other hand, the channel needs to extract parsimonious information from the

incoming data. This channel can be implemented as a neural network with a beta variational auto-encoder.

We use this model to show that two mechanistic components are needed for adaptation: focus and switch. We firstly show that focus on a cue mimics humans better than cue weights that simply depend on long term statistics as has been largely assumed in the prior research. And second, we show a new model that can quickly adapt and switch weighting the features depending on the input of a particular moment. This model's flexibility comes from implementing a cognitive mechanism that has been called "selective attention" with multiple encoders. Each encoder serves as a focus on a different part of the signal. We can then choose how much to rely on each focus depending on the moment.

Finally, we ask whether cue weighting is informed by being able to separate the noise from speech. To this end we adapt a feature disentanglement adversarial training from vision to disentangle speech (noise) features from noise (speech) labels. We show that although this does not give us human-like cue weighting behavior, there is an effect of disentanglement of weighting spectral information slightly more than temporal information compared to the baselines.

Overall, this thesis explores adaptation computationally and offers a possible mechanistic explanation for "selective attention" with focus and switch mechanisms, based on rate distortion theory. It also argues that cue weighting cannot be determined solely on speech carefully articulated in laboratories or in quiet. Lastly, it explores a way to inform speech models from a cognitive angle to make the models more flexible and robust, like human speech perception is.

MODELING ADAPTABILITY MECHANISMS OF SPEECH PERCEPTION

by

Nika Jurov

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Naomi H. Feldman, Chair, Advisor
Professor William Idsardi, Co-Chair, Co-Advisor
Assistant Professor Thomas Schatz
Associate Professor Ellen Lau
Professor Jonathan Simon, Dean's Representative

© Copyright by
Nika Jurov
2024

Foreword

Chapter 5 in this thesis is joint work with Grayson Wolf and chapters 5, 6 and 7 are joint work with William Idsardi and Naomi Feldman. This work was funded by an NSF grant BCS-2120834.

Special thanks to: Charles Wu and Lori Holt for sharing their experimental data, to both Christopher Bates and to Priyatham Kattakinda for sharing their model code and for helpful advice, and to Thomas Schatz, Soheil Feizi, Philip Resnik, Ellen Lau and Jonathan Simon for helpful comments and discussion.

Dedication

I dedicate this thesis to my mother and my father, and to their mothers and fathers and so on. A long line of people who stubbornly persevered many years thinking with their own heads while being suppressed by systemic functions of the ”don’t think for yourself” society.

Acknowledgments

"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

"But it was the Great Question! The Ultimate Question of Life, the Universe and Everything," howled Loonquawl.

"Yes," said Deep Thought with the air of one who suffers fools gladly, "but what actually is it?"

A slow stupefied silence crept over the men as they stared at the computer and then at each other.

"Well, you know, it's just Everything ... everything ..." offered Phouchg weakly.

"Exactly!" said Deep Thought. "So once you do know what the question actually is, you'll know what the answer means."

– Douglas Adams, The Hitchhiker's Guide to the Galaxy

Anyone who has ever attempted to do science knows how daunting this is, especially at the beginning. But luckily no science is done in a vacuum – the science that I was taught was through people and with people. To give them the credit they deserve would require me to write at least a novel. A word ''thank you'' seems almost empty and senseless compared to a profound positive change of my life that these people are responsible for. However, for a lack of a better word, thanks are in order.

First, I thank the entire department of Linguistics. I chose this department because of all the photos of cakes on the website, and my experience here did not disappoint. The fun that you have doing science is simply contagious.

Then, importantly, my advisors, Naomi Feldman and Bill Idsardi, and then Thomas Schatz, who I have consulted in parallel for five years. I still cannot believe you took me as a student and let me talk to you about speech and models, even before I understood what they were. The guidance you gave, through questions, direct advice, discussions or simply stern looks at some of my comments will never be lost.

Naomi, thank you for hours and hours of extreme attention to detail in mechanistic understandings of human psychology that you shared with me. For the fact that you had to decipher my ”mystery novels” again and again. For the hours of math sessions and the hours of discussing the workings of cognition that taught me so much about combining different methods and scientific view points. For being able to ask you for advice on anything and for getting a sincere answer. This includes the subject of the tedious (and to my utmost surprise, successful) job hunt. Finally, I thank you as well as in always fighting for inclusion and diversity and supporting junior scientists (in the making) that are talented regardless of anything else. Your efforts changed many lives, including mine.

Bill, thank you for numerous free flowing discussions on anything that touches on speech, phonology, perception, cognitive mechanisms, models and this list goes on and on. I will always be fascinated by the depth and width of your knowledge and I admit, have been coming for advice to you sometimes thinking I am consulting a human encyclopedia. I also thank you for your somewhat cynical outlook on pretty much any scientific paper that always makes me pause in reading a paper and ask myself, what would Bill think about this? And I always end up thinking

that you would find weak points of that research with one glance, bringing depth and extreme understanding of the matter at hand. Thank you for letting me consult you and for helping me build my research more carefully than I ever could.

Thomas, first thank you for talking with me about anything that has to do with mechanisms and models for years after you left to be a professor in France. Thank you for being so generous with your time and for helping me with understanding computation in depth, I have learned so much from you. From coding, to math, to understanding and implementing mechanisms – without you I could not be a computational scientist. Thank you for taking me seriously even though I repeatedly called hippocampus a hippopotamus (some days were harder than others, I have to admit). And thank you for continuously developing scientific ideas with me, which we will hopefully continue to do in the future.

Furthermore, I thank Ellen and Philip for discussing speech with me (which I know, is a little bit of a pain...) and for numerous discussions or classes that you led. Your advice and work shaped my ideas, and made me look at the research I was doing from a higher view point. I also thank Jonathan Simon and Shihab Shamma and Dan Butts for similarly helping me understand neural mechanisms on a much deeper level than I could have hoped for. Your findings and ideas and classes were crucial to my own research.

I thank Ewan Dunbar that was the first professor I ever encountered that believed in me and gave me a serious chance at science. You were the first to show me that one can think in a different dimension when it comes to speech and you were also responsible I applied for American PhD programs. Without you, I would definitely not be where I am, nor who I am.

I thank Craig Thorburn, Joselyn Rodriguez and Leslie Famularo for randomly initiated discussions about anything speech (and occasionally about best bars, music venues or bubble tea

stores). You constituted a small, but strong speech community that always had a very perceptive insight into a variety of topics. Thanks for letting me discuss all these with you and for hearing my practice talks until the point when it was physically painful. I hope we will continue talking in the future and creating science together. I also thank Craig for being a year-above-me-experienced-and-knowledgeable guide about pretty much everything in my first years and for making some of the best cocktails I ever had.

I thank Masha Polinsky on serious support especially during my job search. You guided and supported me in times I was very uncertain and helped me without really needing to. Besides, your advice on good restaurants proved to be well taken very seriously.

Additionally, I thank ”the lunch community” (that is, anyone that was in the lunch room) for making my midday break a fun experience. I especially thank Alexander Williams for his attention to students and for his immeasurable assortment of random conversational topics, that ranged about events, places and people around the world. Your stories always made my day funnier and more positive.

I thank Kim Kwok for being the best administrator I have ever seen – you always have students at the front of your mind and I have to say I could not be more thankful. You helped with so many things and never made me feel like I am bothering you amidst your already busy day. You made things happen that seemed impossible, so sometimes I think maybe you have some magical powers.

When it comes to other fellow students, I have to thank everyone for the fun atmosphere, a lot of conversations, interactions, throwing ideas around and more. I must especially thank my cohort without whom I would not made it through my very first year: Maša Bešlin, Alex Krauska, Polina Pleshak and Jéssica Mendes (and at the time, Justin Malčić). I especially thank Maša for

wanting to exercise habits from the Balkans and sit around in coffee shops on the weekends (what atrocity, to walk around with your coffee!). I thank Luisa Seguin for talking, at length, about how to bake a good bread, that was very therapeutic to me. For cooking pies, risottos and for making my 888 cake, some of the finest I have tasted. I thank you both for willing to talk openly about problems that all young scientists have and for much more than I am able to write about here – you became friends outside academia. Additionally, I thank Alex Krauska, Zara Harmon, Alex Chabot, Katherine Howitt and Sathvik Nair for a lot of fun interactions, conversations and more that happened during our common time at UMD. I especially thank Alex Krauska for always being ready to share life or work outside campus and for making an incredible cake at my dissertation. And I thanks Alex Chabot for sharing his Virginia, cooking and memories with me.

Finally, I thank everyone close to me that continuously reminded me, just by their presence, about the fact that life is a lot bigger than whatever I am working on. First, my family that is everything to me, that supported me in so many ways, that taught me how to think critically – hvala! I thank my dear friend Marko for being the first to encourage me to pursue the PhD. I also thank Luka for housing me during the pandemic when I was stranded at home, for companionship and for elvish music. I also thank many many friends I have in the US, in Slovenia and across Europe who simply make my existence full of joy and life, which powers absolutely everything else I do in life: Pablo, Hugo, Giulia, Jonathan, Andreas, Doyna, Leire, Neus, Matic, Matic, Mateja, Maja, Nejc, Ota, Jerca, Urška, Tanja, Žiga, Nina, Rémi, Hao, Joe, David, Yao and the list goes on. You know what you mean to me and that transcends this document.

Finally, I thank Chris for sharing every up and down of my PhD and much much much more. You decided to sit with me in the front row of this roller coaster and hold my hand which

you really did not deserve, because you are one of the nicest human beings I have ever met.

Hvala.

Table of Contents

Preface ii

Foreword ii

Dedication iii

Acknowledgements iv

Table of Contents x

List of Tables xiii

List of Figures xiv

List of Abbreviations xvii

Chapter 1: Introduction 1

Chapter 2: Acoustic-Phonetic Feature/Cue/Dimension Weights 7
- 2.1 Introduction . . . 8
- 2.2 Acoustic-Phonetic Features . . . 9
- 2.3 Cue Weighting As Dependent On Long Term Statistics . . . 11
  - 2.3.1 Toscano's and McMurray's Cue Weighting Model . . . 12
- 2.4 Multiple Models Of Cue Weighting . . . 15
  - 2.4.1 Kleinschmidt's and Jaeger's Ideal Adapter Framework . . . 17
- 2.5 Conclusion . . . 20

Chapter 3: Selective Attention 22
- 3.1 Introduction . . . 23
- 3.2 Short Term Cue Weighting . . . 25
  - 3.2.1 Accented Speech . . . 26
  - 3.2.2 Noisy Speech . . . 33
- 3.3 Distorted Speech . . . 35
- 3.4 Selective Attention . . . 38
  - 3.4.1 Auditory Objects, Streaming And Attention . . . 40
  - 3.4.2 Spectro-Temporal Receptive Fields . . . 42
  - 3.4.3 Changes In Spectro-Temporal Receptive Fields . . . 47

3.5 Conclusion 49

Chapter 4: Rate Distortion Theory 51
4.1 Introduction To RDT 52
4.2 Theoretical framework of Rate Distortion Theory 55
4.3 Beta Variational Auto-Encoder: Mathematical Implementation of RDT 58
4.4 Cognitive Models Using Rate Distortion Theory 61
4.5 Conclusion 63

Chapter 5: Focus: Computational Evidence That Cue Weighting Requires More Than Long Term Statistics 64
5.1 Introduction 65
5.2 The Simulated Phenomenon 67
5.3 Model 69
5.4 Long Term Distributions: Corpora 71
5.4.1 Input Data Pre-Processing 73
5.4.2 Evaluation Stimuli 75
5.5 Evaluation: Cue Weights 76
5.6 Simulations 77
5.6.1 Acoustic Input 77
5.6.2 Articulatory Input 79
5.6.3 Low Dimension Input 81
5.6.4 Multiple Vowel Classi cation 82
5.7 Primary Cue Bias With A Cue Weighting Parameter 84
5.8 Conclusions 87

Chapter 6: Switch: Simulating Flexible Feature Weighting 89
6.1 Introduction 89
6.2 A model based on rate distortion theory 91
6.3 Model training and testing 95
6.4 Experiments 99
6.4.1 VAE 99
6.4.2 -ME-VAE 100
6.5 Discussion 101
6.6 Conclusion 103

Chapter 7: Feature Disentanglement 105
7.1 Wav2Vec2.0 106
7.1.1 Disentanglement Model 108
7.1.2 Cue Weights 111
7.1.3 Noisy TIMIT 113
7.2 Results 113
7.2.1 Consonants 114
7.2.2 Vowels 115
7.2.3 Discussion 116

7.3 Conclusion 119

Chapter 8: Conclusions 121

List of Tables

7.1 Consonants - Cue Weight Differences– Secondary cue weight subtracted from primary cue weight within each transformer layer, then averaged across layers. Numbers closer to 1 signify bigger weight on primary cue, numbers closer to -1 denote bigger weight on secondary cue. 118
7.2 Vowels - Cue Weight Differences– Secondary cue weight subtracted from primary cue weight within each transformer layer, then averaged across layers. Numbers closer to 1 signify bigger weight on primary cue, numbers closer to -1 denote bigger weight on secondary cue. 118

List of Figures

3.1 Experimental 7 step continuum- Each dimension represents a different covarying cue (either for vowels or consonants). Each square is a different stimulus with a different combination of cue values. Colored lower and upper squares denote where typical values of a cue denote a speci c category. 27

3.2 Miscorrelated Features- The green squares denote the exposure stimuli, which are now combinations of non typically correlated cue values. On the one dimension, listener may want to categorize the stimulus with category 1, whereas the second dimension will be indicative to category 2. 28

3.3 Con ict - How does a listener decide? If both features had equal weight, this would be a hard choice, since for this particular stimulus with cue 1 value 6 and with cue 2 of value 2 each denote a different category. The arrows show what the listener might think given each dimension values. For cue 2 with value 2, this would likely be category 1. It would likely be category 2 if we only looked at cue 1 with value 6. 29

3.4 Feature 1 is Upweighted- If the listener only relied on feature 1, the perceptual decision boundary would completely depend on it. The stimulus in question (with feature 1 of value 6 and feature 2 of value 2) would now be categorized as category 2. 30

3.5 Feature 2 is Upweighted- If the listener only relied on feature 2, the perceptual decision boundary would completely depend on it. The stimulus in question (with feature 1 of value 6 and feature 2 of value 2) would now be categorized as category 1. 31

4.1 RDT (illustrative, after Sims, 2016) - RDT channel has an encoder and a decoder. It is optimized both on rate (how informative the latent information is) and distortion (how many perceptual errors there is between input and output). Internal noise comes from information constrained capacity. 55

5.1 Cue Reliability Based On Statistics (Illustrative) - If means are further apart or the variances are smaller, there is more separability on that dimension (as in A and B). Inversely, means closer together or more variance introduce more confusion on which distribution the black dots belongs to (C and D). 66

5.2 Assumed Cue Distributions (Illustrative) - Here we see assumed cue distributions that cause cue 2 to be more reliable than cue 1. As we can see, cue 2 shows good separability as indicated with the green dotted line (that is, if all of the decision would be based on the dimension of cue 2. Meanwhile, cue 1 shows overlap between categories, and as such becomes more uncertain as a perceptual boundary. We can see this with the illustrative blue dotted line: parts of category 1 and 2 would be categorized as the opposite category if the decision were based entirely on cue 1. 67
5.3 Human Cue Weights (Illustrative) - Here we see reliance of cue illustratively reproduced from (Wu & Holt, 2022). Listners rely more on spectral information than on duration when identifying [ɛ] and [æ]. 68
5.4 TIMIT: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker). 75
5.5 Buckeye: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker). 76
5.6 Wall Street Journal: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker). . . . 77
5.7 Cue Weights (MFCCs)- Cue weights of [ɛ] and [æ] extracted from each corpus and then transformed into MFCCs. Cue weighting with these featural representations does not seem do mimic humans, but rather shows reverse cue weights for each of the three corpora. 78
5.8 Cue Weights (Filterbanks) - Cue weights of [ɛ] and [æ] extracted from each corpus and then transformed into lterbanks. Cue weighting with these featural representations does not seem do mimic humans, but rather shows reverse cue weights for each of the three corpora. 79
5.9 Cue Weights (Filterbanks Vs. Articulatory) - Cue weights of [ɛ] and [æ] extracted from TIMIT and then transformed into lterbanks (called Acoustic on the gure) and articulatory features extracted from (Siriwardena et al., 2022). 80
5.10 Cue Weights (extracted feature values) - Cue weights of [ɛ] and [æ], when rst formant and duration are extracted from each corpus and used as model input. Cue weighting based solely on input data matches humans only for a model traines with WSJ corpus. 81
5.11 Multiple Vowels: extracted feature values- Cue weights of [ɛ] and [æ], when we use all vowels of TIMIT during training. The input is simply the extracted and normalized rst formant and duration for each vowel. 82
5.12 Multiple Vowels: MFCCs) - Cue weights of [ɛ] and [æ], when we use all vowels of TIMIT during training. The input is all vowels of TIMIT transformed into MFCCs. 83
5.13 Cue Weights (extracted feature values) - Cue weights of [ɛ] and [æ], when rst formant and duration are extracted from each corpus and used as model input. Cue weighting based solely on input data does not show the same trend as humans who largely upweight the rst formant for TIMIT nor Buckeye. When we add the cue weighting parameter, we see that cue weighting mimics humans. . 85

6.1 Multiple encoder -VAE training - Schematic architecture of training of multiple encoder -VAE with an added categorization model. The number of encoders can be extended. Each encoder encodes the same input. A sample from each encoder E_i is decoded separately in training. The category model receives a random proportion of data from each encoder. 94
6.2 Feature weighting (single -VAE) and training data - A: Modeled decision boundary of a single channel -VAE that is forced to reconstruct VOT better than F0. It cannot adapt to the moment-to-moment statistics unless retrained. B: Visualization of training data used for all simulations. For categorization, data in the lower left corner was labeled [b], and data in the upper right corner was labeled [p]. 95
6.3 Flexible adaptation in perception - Modeled decision boundaries due to different combinations of encoders. Encoder 1 is focused on VOT, encoder 2 is focused on F0. Depending on how much weight we give to each encoder, we arrive to different decision boundaries. A: all of the decision is based on F0. B: all of the decision is based on VOT. C: 80% of the information is coming from encoder 2, 20% from encoder 1. D: 80% of the information is coming from encoder 1, 20% from encoder 2. 96

7.1 Training stimuli - We use corner 3x3 stimuli from (Wu & Holt, 2022) to train a linear classi er on representations from the models to see how they weight cues important to humans. 112
7.2 Base Model: Cue Weights - [p], [b] - Cue weights across all layers of baseline wav2vec2.0 netuned on text labels. 114
7.3 Finetuned Model: Cue Weights - [p], [b] - Cue weights across all layers of baseline wav2vec2.0 netuned on noisy TIMIT to predict phonetic labels. 114
7.4 Disentangled Model: Cue Weights - [p], [b] - Cue weights across all layers of the disentangled wav2vec2.0 trained to prevent feature correlations in labeling foreground (phonetic labels) and background (noise labels). 115
7.5 Base Model: Cue Weights - [æ], [ɛ] - Cue weights across all layers of base wav2vec2.0. 115
7.6 Finetuned Model: Cue Weights - [æ], [ɛ] - Cue weights across all layers of netuned wav2vec2.0 on TIMIT to predict phonetic labels. 116
7.7 Disentangled Model: Cue Weights - [æ], [ɛ] - Cue weights across all layers of the disentangled wav2vec2.0 trained to prevent feature correlations in labeling foreground (phonetic labels) and background (noise labels). 116
7.8 Difference In Cue Weights - Consonants - Cue weight for VOT minus cue weight for pitch across different layers. We can see slight changes across the three models, where in all models VOT is greater (upper right corner), with the disentangled model having most change in cue weight for pitch. 117
7.9 Difference In Cue Weights - Vowels - Cue weight for spectral information minus cue weight for duration across different layers. We can see slight changes across the three models, where in all models duration is greater (lower left corner), with the disentangled model having most change in cue weight for spectral information. 117

List of Abbreviations

A1	Primary auditory cortex
-VAE	Beta variational auto-encoder
BCE	Binary cross-entropy
CE	Cross entropy
CER	Character error rate
CTC	Connectionist Temporal Classi cation loss
D	Distortion
ECoG	Electrocorticography
F0	Fundamental frequency/pitch
F1	First formant
F2	Second formant
F3	Third formant
H	Entropy
LDA	Linear discriminant analysis
MFCCs	Mel-frequency cepstral coef cients
ME- -VAE	Multiple encoder beta variational auto-encoder
MSE	Mean square error
R	Rate
STG	Superior temporal gyrus
Q	Channel
RDT	Rate distortion theory
STRF	Spectro-temporal receptive elds
WSJ	Wall Street Journal
VAE	Variational auto-encoder
VOT	Vowel onset time

Chapter 1: Introduction

The speech signal is anything but invariant – it changes from speaker to speaker and from situation to situation. Listeners navigate this chaotic acoustic scenery with incredible exibility and are able to robustly translate the variable signals to invariant perceptual events and units. Much evidence shows that listeners can adapt to a variety of listening situations (Adank & Janse, 2009; Clarke & Garrett, 2004; Davis et al., 2005; Elliott & Theunissen, 2009; Francis et al., 2008; Guediche et al., 2016; Holdgraf et al., 2016; Idemaru & Holt, 2014; Li & Allen, 2009; Liu & Holt, 2015; Nittrouer, 2002, 2004; Nittrouer & Miller, 1997; Wu & Holt, 2022; X. Zhang et al., 2021; Zhou et al., 2024).

In particular, listeners map speech to representations with the help of so called cues or features or dimensions that represent different aspects of speech. These cues enable the listener to map speech to more abstract representations (phonetic categories). For example, the vowels [ɛ] and [æ] of standard American English can be differentiated by two cues, spectral energy and duration.

Some cues seem to be more relied on than others (i.e. Francis et al., 2008; Idemaru and Holt, 2014; Wu and Holt, 2022). For example, spectral information is more important than duration to listeners differentiating between [ɛ] and [æ]. Reliability is estimated by the listeners and has been previously thought to depend on long term statistics of the speech the listener hears

in their life (Toscano & McMurray, 2010).

However, reliability can change in the moment. This happens if some cues have confusing values or are inaccessible. In those situations, listeners can use other cues that seem more reliable. For example, if spectral information is not accessible, listener can use duration (Wu & Holt, 2022). Which cue is more relied on than another is then a changeable process. This adaptive exibility of cue weighting has been quite hard to capture in computational models, because they either presuppose xed cue weighting (Toscano & McMurray, 2010) or they only show how a listener may adapt their perceptual space for a single cue and do not explain how cues are weighted exibly (Kleinschmidt & Jaeger, 2015).

In this thesis, I explore the exibility of cue weighting with computational simulations based on an information theoretic framework and based on a cognitive mechanism that is hypothesized to perform exible cue weighting. To be able to do this, I bring together information theory framework called rate distortion theory as the optimal perceiver and a cognitive mechanism that has been observed in many experiments (Francis et al., 2008; Idemaru & Holt, 2014; Liu & Holt, 2015; Nittrouer, 2002, 2004; Nittrouer & Miller, 1997; Wu & Holt, 2022; X. Zhang et al., 2021; Zhou et al., 2024). This mechanism has been adapted from neuro-biology and called "selective attention" by Holt et al., 2018.

First, I provide a possible mechanistic explanation for cue weighting exibility. To do so, I argue that two main mechanistic components are needed, one I call focus and one I call switching of the focus. I argue that focus comes from internal upweighting of a cue that can be independent of the input data. This focus is due to a selective view of the signal that preferentially views a single cue.

Additionally, I argue that the switch between focus on one cue to another cue is what is

necessary to model exible adaptation or “selective attention”. This mechanism has not been computationally implemented before. This switch can be performed because the input data is received with multiple views, where each view provides a different focus on the signal. Importantly, we can choose also to switch partially among the different views, so that the decision is dependent on some proportion of view (or focus) on a speci c cue in combination with another cue.

Selective attention is a term for a hypothesized mechanism that enables switching focus between features of the attended auditory object. However, this process may not be consciously done as is assumed for processes of typical auditory attention (i.e. choosing which speaker to attend to in a situation of two competing speakers). There is little direct evidence about how and whether attention affects cues. However, Holt et al., 2018 adapted this term because ndings in neuro-biology show changes as measured in different ways on a neuronal level in auditory cortex of animals and humans (Fritz et al., 2003; Holdgraf et al., 2016; Schwartz & David, 2018). There will need to be more experiments done to establish how cues are weighted on a neural level. However, in this thesis, we assume a term selective attention for a possible top down mechanism that enables quick re-weighting of cue information coming from bottom up (because perhaps we have stopped trusting speaker's production of a certain feature, or we have learned how to adjust to very degraded speech).

To implement the mechanism of selective attention, I use models based on rate distortion theory. This is a framework coming from information theory that expresses a channel that is constrained in its information capacity: it is trained on a trade off between distortion (amount of reconstruction errors) and rate (whether the representations it extracted are parsimonious). This model has been used previously in cognitive science, for example to model attention in vision

(Bates & Jacobs, 2021), visual memory (Bates et al., 2019; Jakob & Gershman, 2023; Sims, 2016) and pragmatic reasoning (Zaslavsky et al., 2020), to name a few.

I use rate distortion theory as an optimal perceiver model. As such, it serves as a model that either completely depends on the speech data it has been trained with, or has an added parameter that implements selective view or focus. Overall, the rate distortion theory framework is good to adopt because given speci c data, it represents an optimal channel that can compress the input information in a goal directed way, while this compression and any potential perceptual errors are explained by information constraints imposed on the channel.

Additionally, I also argue that realistic speech input is an important component of cognitive models of speech perception. I train models on big speech corpora that represent a more realistic scenario than carefully articulated speech in typical experimental setups. This allows us to see how listeners would behave if they only depended on realistic input data.

Finally, I also ask whether cue re-weighting happens because of the listener's ability to separate (or disentangle) other sounds (or noise) from speech signal, thus being informed of where the noise may mask speech cues. If the listener knew what the noise is masking, they may rely more on whatever the noise is not masking. To test this, we perform feature disentanglement on a bigger speech-to-text model (wav2vec2.0), with the goal of disentangling speech features from noise labels, and noise features from speech labels.

The structure of this thesis is as follows. First, to motivate my computational simulations, I discuss prior research. In Chapter 2 I present what is typically meant by cues and I discuss prior mechanistic accounts on cue weighting and its shortcomings. In Chapter 3 I discuss momentary cue reweightings as observed in psycholinguistic experiments on speech in quiet with an accent, on speech in noise and on speech that has been distorted and no longer has typical speech cues

preserved. In addition, I discuss ndings in neuro-biology on attention that could underlie cue weighting mechanism that has been called “selective attention” by Holt et al., 2018. Further, I present rate distortion theory and motivate its use as a cognitive model in Chapter 4, as it serves as a framework of our cue weighting modeling.

Then, in Chapter 5 I ask whether cue weighting can be based only on long term statistics that the listener has learned in their lives. This is done by simply training rate distortion theory model on a speech corpus and extracting cue weights from the trained model. Based on simulations in this thesis, relying on data cannot be the entire story: rst, the cue weighting dependent on the long term statistics does not mimic humans very well, and second, this type of cue weighting simulation gives xed cue weighting. In other words, such a model cannot change its cue reliance in different listening situations.

In Chapter 6 I present a model that exhibits quick and exible cue weighting depending on any moment the listener has found themselves in. This model is a beta auto-encoder model, based on rate distortion theory, that has multiple encoders. Each encoder represents a single view of the signal that primarily detects a single feature. Each encoder then provides a single focus on a speci c cue. The information coming from each encoder can then be weighted exibly according to the listening situation at the second stage. This represents a model unlike Toscano and McMurray, 2010 which only relies on long term statistics.

Finally, in Chapter 7 I perform feature disentanglement and ask whether that can explain adaptation seen in humans. This is done with a model that tries to disentangle foreground and background features of noisy speech by penalizing the mutual information between foreground features and background labels (and vice versa). This model is adapted from a model developed for a similar purpose in vision by Kattakinda et al., 2022. To see whether disentanglement drives

adaptation, I extract cue weights from this disentangled model and from two baselines. All three models are based on a relatively state of the art speech to text model, wav2vec2.0 (Baevski et al., 2020). I use this model as an example of the recent development of deep neural networks that has shown an incredible performance in mapping speech to text in quiet, but has poorer performance in noise. To examine cue weights I use experimental stimuli from Wu and Holt, 2022. I show that in such a model there is less cue weighting variability for speech in quiet versus speech in noise, unlike humans. However, the disentangled model shows more weight on spectral information than its baselines, so there is perhaps some effect of disentanglement on cue weights.

Overall, this thesis computationally explores cue weighting and adaptability mechanisms that have been called selective attention. This is done with a rate distortion theory framework and two mechanisms: focus and switch between focuses. I show models for focus can mimic humans better than models that purely depend on data. I also show that a model with multiple independent focuses and a exible reliance on these focuses implements “selective attention” – exible cue re-weighting. Additionally, I explore whether cue weighting is informed by speech and noise feature disentanglement in a bigger speech-to-text model (wav2vec2.0) that can learn better features and show that the disentanglement goes in the direction of human behavior, but it does not fully achieve it. Overall, an emphasis in this thesis is put on developing cognitive models with real speech and with closer connection to cognitive mechanisms than what is seen in prior computational models.

Chapter 2: Acoustic-Phonetic Feature/Cue/Dimension Weights

This chapter serves as a review of the relevant literature on cue/feature weighting that the work of this thesis is built upon. Acoustic-phonetic features or cues or dimensions[1] are an important part of the speech perception process. They offer a window into how human listeners map speech to abstract representations.

This chapter discusses the nature of cues and numerous experimental results that uncovered the weighting or trading relations between them as observed in human behavior in identifying speech sounds. Importantly, this chapter serves to summarize general cue weighting thought to be possibly learned from long term statistics and to describe prior mechanistic accounts of cue weighting. The goals of this chapter are also to explain the link between cue weights and the perceptual representations as has been posited by prior research. This will allow us to computationally explore cue weighting later in chapters 5 (exploring whether cue weighting can be based solely on long term statistics), in 6 (proposing a more flexible computational model of cue weighting that is built on selective attention mechanism) and in 7 (exploring whether feature disentanglement in artificial cognition drives human like cue weighting).

[1]There is a multitude of terms that has been used for the same phenomenon: cues, features, dimensions. Features may be confusing because there are also phonological features which are separate percept from acoustic-phonetic feature. Cue could also be confusing because it may be used as a prompt to something and has indeed been used as such in other research outside audition (i.e. vision, memory, attention etc.). In this thesis, acoustic-phonetic features or cues or dimensions are synonymous and are used interchangeably to denote the phonetically important statistics that the listener can extract from the signal, without abstracting from it.

2.1 Introduction

Listeners employ speech perception mechanisms to form internal representations of speech sounds. Speech is a complex and variable signal, therefore this mechanism has to deal with redundancy, variability, noise and uncertainty. This suggest a highly exible mechanism that is able to map the signal to linguistic representations for a wide variety of situations that the listener is in and is likely using redundancies to its advantage.

Acoustics signi cant for linguistic information found in speech has been described as a cue (i.e. Holt and Lotto, 2006; Liberman, 1957). A cue enables the listener to map the signal to some phonological category. In speech, the cues denote acoustic-phonetic features that inform the listener about some crucial aspect of the speech sound. For example, one such feature is voice onset time (VOT), or the difference between the burst and the start of phonation. VOT linguistically cues voicing, a phonological feature: is the sound voiced (like in bat) or unvoiced (like in pat)? Other features also denote voicing, and as many as 16 have been identi ed in intervocalic position (Lisker, 1986).

It is important to note that phonological features are different from acoustic-phonetic features. Acoustic-phonetic features are measurable and are found in the actual signal. The phonological features are higher abstract representations belonging to the listener's grammar and can have as correlates multiple acoustic-phonetic features. This suggests hierarchical perceptual processing of features, rst the acoustic-phonetic, then phonological features that represent the phonological category. In this thesis, feature will denote acoustic-phonetic features/cues. Additionally, the exact nature of the representation of the categories (whether they are discrete or continuous, for example Apfelbaum et al., 2022) is beyond the scope of this work.

The following sections will give an overview of acoustic-phonetic features and of their reliability. In other words, we will see how and why listeners rely on cues as explained by prior research.

2.2 Acoustic-Phonetic Features

Acoustic-phonetic features or dimensions are properties of the speech signal and will be different depending on the perceiver's grammar. For example, voice onset time (VOT) will be a cue for voicing in several languages, but its values will be different for English speakers than for speakers of other languages (such as some Slavic languages, i.e. Slovenian, Serbo-Croatian, Russian etc.). In English, VOT is typically positive, with shorter values denoting voiced sounds. In Russian, voiced sounds are denoted with a negative VOT (with prevoicing) (Kazanina et al., 2006). There are other examples, such as nasality in English vowels is not contrasting (that is, nasalised English vowel V is considered an allophone of the vowel V and both are versions of a single category), while it is in French vowels (V and nasalized V are different vowels/categories in French grammar).

Acoustic-phonetic features then are linguistically informative parts of the acoustic signal. There are usually several acoustic-phonetic features pointing to the same phonological feature (or the contrastive quality between categories) (Lisker, 1986). One of the well researched examples in English word initial stops is voicing signaled by both the pitch of the voice (F0) and the VOT (i.e., “pier” vs. “beer”) (Idemaru & Holt, 2014). This signi es redundancy of information present in the signal.

However, speaker's variability and the listening situation uctuate so much that no cue

so far has been found to be invariant (Holt & Lotto, 2010). This suggests that the listener has to be open to several cues, which may partially explain the redundancy of cues in the signal. Nonetheless, given that cues are different at different points in time, this makes speech perception a more complex process than it would seem.

The fact that the listener's grammar is an important factor for cues is also evident in second language perception. For example, Japanese speakers usually have a hard time distinguishing American English [l] (as in “low”) from [ɹ] (as in “row”). The primary cue for American English listeners is often determined to be the third formant (F3) (Yamada & Tohkura, 1992), whereas Japanese listeners also use F2. This is because in Japanese, there is only a single lateral ap (Ladefoged & Maddieson, 1996) and variability in F3 dimension is not important and contrastive for liquids in Japanese. Thus, Japanese listeners may not perceive F3 as a cue as important as F2 and this may drive their confusion.

This “importance” of a cue is a signi cant phenomenon observed in many studies. In other words, it reveals cue reliability that the listener has about it: a so called primary cue/feature. That is, one cue seems more important in perception than another (or several others). For example, whereas voicing can be signaled by both VOT and F0, VOT is considered a primary cue. This primacy is re ected in experiments that test decisions alongside two cues that contrast (Holt & Lotto, 2006; Idemaru & Holt, 2014, 2020; Liu & Holt, 2015; Wu & Holt, 2022), where participants on average rely more on one cue than another. Cue reliability has also been thought of as either dependent on long term statistics or as (un)certainty about a cue. The former will be discussed in section 2.3 and the latter in section 2.4.

2.3 Cue Weighting As Dependent On Long Term Statistics

In speech perception cues are not equally relied on (Idemaru & Holt, 2014; Liu & Holt, 2015; Schertz et al., 2015; Wu & Holt, 2022). Relying on cues disproportionally to map speech to a category implies a mechanism that can prioritize one cue over others. The prioritized cue is often called a primary cue. For example, VOT is often termed a primary cue in identifying word initial stops (Idemaru & Holt, 2014), with F0 as the secondary cue. In contrast, Korean uses F0 in addition to VOT as a primary cue (Kwon, 2015; Schertz et al., 2015). Therefore, there may be multiple primary cues, depending on the contrasting categories alongside a dimension (English has a two way contrast, Korean has a three way stop contrast: lenis, fortis and aspirated).

Primacy necessarily implies that there is a hierarchical relationship between the cues. Primary cues are thought to be more relied on than the secondary. In fact, secondary cues are thought by some to only be recruited when a primary cue is not available (i.e. Francis et al., 2008; Harmon et al., 2019). Why a cue is primary is not clear, especially since primacy changes depending on the context and the contrast. For example, VOT although primary in word initial stops, is not informative for a stop in a mid-word position. The relation between cues also depends on the situation: in noise, VOT is less informative than F0 (Wu & Holt, 2022).

As such, a mechanism must be at play that weights the cues: more important primary cues are weighted more heavily. In other words, they are more relied on. What is important can be thought of as coming from the way cues are distributed in the input: listeners prefer to rely on more easily separable distributions. This idea was computationally modeled in Toscano and McMurray, 2010, described in the following subsection.

2.3.1 Toscano's and McMurray's Cue Weighting Model

The essential assumption of the model by Toscano and McMurray, 2010 is that the cue weights can simply be deduced from the statistical distribution of the categories the listener has learned in their life in an unsupervised fashion.

Toscano's and McMurray's model is thought of as a mechanism that integrates multiple sources of information, just like in other domains of cognition, i.e. combining visual and haptic information for estimating properties of an object (Ernst & Banks, 2002). In addition to combining multiple pieces of information, they frame speech cue weighting as a weighting-by-reliability mechanism. This means that some sources of information are inherently more reliable than others. For Toscano and McMurray, the reliability comes from the statistics of speech the listener observes in their lifetime, i.e. as learned through development (Maye et al., 2002; McMurray et al., 2009; Saffran et al., 1996).

This is tied to the fact that cues help to map continuous speech to a perceptual space that enables to interpret the signal as a discrete category. Thus, categorical information importantly contributes to reliability. In other words, if the variability (denoted by it variance ^{2}) of a cue is big, this will also make the distinction of the categories harder, which is expressed as the cue weight w:

$$w = \frac{1}{^2} \tag{2.1}$$

Furthermore, categories that are closer together will be harder to distinguish because they will be more easily confused. For example, if production of some cue varies a lot for a particular

category, that big variability will reduce the reliability of the aforementioned cue. In short, cue reliability is then a function of the variability and the distance between the categories. Therefore, a cue weight of a dimension is mathematically expressed as:

$$w = \frac{(\ _1 \quad _2)^2}{\ _1 \ _2} \tag{2.2}$$

where $_1$ and $_2$ denote the means of cue values (i.e. VOT values) typical for category 1 and 2 (i.e. /p/ and /b/ categories), while $_1$ and $_2$ are their standard deviations. Categories are thought of as Gaussian for simplicity, with the two parameters (and). The further apart the means are and the smaller the variances are, the bigger the cue weight.

Their model learns from the statistical distribution of the production data obtained by Allen and Miller, 1999. The model is a mixture of Gaussians model that explicitly represents categories, from which it is able to represent their parameters and thus their weights.

This model learns cue weights in a multiple categories case as follows:

$$w_i = \sum_{n}^{K} \sum_{m}^{K} \frac{\ _m \ _n (\ _m \quad _n)}{\ _m \ _n} = 2 \tag{2.3}$$

where w_i represents the weight of i-th cue (i.e. VOT) and and again denote the parameters of a speci c category (i.e. [p]) denoted by this cue, while controls the relative weight of a particular unused Gaussian for this cue: if a Gaussian did not contribute to the weight, will be near or at 0.

While their base model learned categories, it did not learn cue weights appropriately, because the model entirely depended on the primary cue. To mitigate that, they added another mixture of Gaussians to hierarchically represent a phonological dimension. This model learns

two categories that are further apart than their veridical means are (as obtained from the data), and gives cue weights that largely prioritizes primary cue, with secondary cue weight being lower than observed in humans.

Starting the simulations with $K = 2$ (i.e. not allowing the model to deduce the number of the categories itself), they experiment with several values of the means. Only if the means were further apart than the real means did the model succeed. Here we see that in fact, the model does not successfully derive cue weights from the actual speech statistics, but needs to have an exaggerated distance between the categories.

This failure to re ect the actual statistics is a main shortcoming of this model. While it is possible that the categories are perceptually represented further apart than they are in the observed statistics, this undermines the argument that people can learn cue weights perfectly from the data. It would then be more appropriate to say that this model learned cue weights from a combination of the data and the perceptual ampli cation of the differences between the learned categories, which the authors acknowledge themselves. How and why this ampli cation happens is left to further research.

Another problem that arises from presupposing that cue weights come simply from the learned lifelong statistics is that they are not xed, but that people seem to be adaptable (Clarke & Garrett, 2004; Holt et al., 2018; Wu & Holt, 2022). That is, what is a primary cue at one moment may be a secondary cue at another moment.

Lastly, this model has been t on actual production data. However, that consisted of only 4 talkers that produced words with initial stops in a laboratory setting (Allen & Miller, 1999). This is another problem since the actual data statistics have much less separable dimensions between categories, as will be shown in later chapters, but has also been described many times in prior

literature, such as Hillenbrand et al., 1995; Pierrehumbert, 2003.

I test whether cue weights come from the observed long term statistics in chapter 5. In addition, I use bigger speech corpora with more speakers in more naturalistic setting than Toscano and McMurray, 2010 to extract cue weights. With this I show that the long term distributions of cues are unlikely to be solely responsible for cue reliability as proposed by Toscano and McMurray, 2010.

2.4 Multiple Models Of Cue Weighting

Perceiving cues can be also thought of as inference under uncertainty. If the goal is to perceive the category (and after that, extract the meaning of what was said), but the cues are multiple and may be changing (and thus are also contributing to the uncertainty), this is not a trivial task. Importantly, there are typically several cues to a particular category. It has been assumed that some seem to offer less uncertainty to the listener (primary cues): these cues may be combined disproportionally to arrive to category mapping.

In speech perception, inference under uncertainty expresses the fact that listener never has actual access to what the speaker produced, but instead has to infer what the speaker intended. Inference under uncertainty also incorporates the idea that the category is not inferred merely through the bottom up processing, but by comparing any possible high-level explanation for it. These may change with time, for example when perceiving continuous speech one does not get all of the information instantaneously (McMurray et al., 2008). Indeed, prior research showed that that listeners maintain this uncertainty for some time when perceiving continuous speech (Bard et al., 1988; Connine et al., 1991; Dahan, 2010; Gwilliams et al., 2022).

Furthermore, inference under uncertainty is a phenomenon known in cognition beyond speech perception (Chater et al., 2006; Clark, 2013; T. L. Grif ths et al., 2010; Knill & Pouget, 2004; Rigoli et al., 2019). Cues can also be thought as different sources of information and different domains of cognitive science have described combining multiple sources of information as probabilistic and/or Bayesian – both within a domain, such vision (Jacobs, 1999, 2002; Landy & Kojima, 2001), audition (Kleinschmidt & Jaeger, 2015; Toscano & McMurray, 2010), as well as across domains, such as a combination of haptic and visual sources of information (Ernst & Banks, 2002), information from both visual and vestibular systems (Fetsch et al., 2009), combination of auditory and visual information (Battaglia et al., 2003) and information from proprio-perception and vision systems (van Beers et al., 2002) to name a few.

It is thought that combining data from multiple measurements reduces uncertainty (Fetsch et al., 2013), which an organism undoubtedly faces while internally representing the world. However, integrating multiple sources of information across domains seems to be something that newborns are not yet able to do (Stein et al., 2014) and as such it is likely to be learned with experience. Cues are highly likely to be learned across development also in speech perception, as demonstrated by a number of statistical learning experiments (i.e. Maye et al., 2002; Saffran et al., 1996) and the proposed learning mechanisms that drive the learning. (i.e. McMurray et al., 2009; Schatz et al., 2021).

Inference under uncertainty has been offered as an explanation to a variety of effects of speech perception, such as perceptual magnet effect (Feldman et al., 2009), integration of auditory and visual cues (Battaglia et al., 2003; Bejjanki et al., 2011), and also of phonetic adaptation (Kleinschmidt & Jaeger, 2015) that I describe in more detail in subsection 2.4.1.

2.4.1 Kleinschmidt's and Jaeger's Ideal Adapter Framework

Kleinschmidt's and Jaeger's model of adaptation is based on acknowledging two important facts: the listening situations being different at different moments and the lack of invariance in speech. This lack of invariance means that acoustically measurable values of speech may signify two different phonological categories for two different speakers. For example, someone's VOTs may be much shorter in general for both voiced and unvoiced sounds, such that their production of [p] and [b] in isolation may be considered as exemplars of [b] for another speaker's articulation properties.

They propose that listeners are able to overcome these issues by having prior beliefs about the statistics of the relevant categories, by recognizing the statistics of whatever is observed and by their observation that there is a need to adapt to the situation. This is re ected in their “ideal adapter framework”, which is developed to explain the adaptive power of the listener. Their claim is that although situations change constantly, there are statistical regularities that our cognitive system can exploit.

The ideal adapter framework presupposes that listeners do not have direct access to the true distribution of the data, but that they have uncertain beliefs about it. The acquisition of these beliefs is driven by the lifelong implicit learning about the speech properties and also about the talkers and groups of talkers. This framework is built upon the work that thinks of speech perception as statistical inference under uncertainty of the talker's intentions. The category c_i is inferred after observing cue v and can be computed with Bayes rule as a product of the likelihood of the cue given category c_i and the prior probability of the cue c_i:

$$P(C = c_i|x) = \frac{p(x|C = c_i)p(C = c_i)}{\sum_j p(x|C = c_j)p(C = c_j)} \propto p(x|C = c_i)p(C = c_i) \tag{2.4}$$

This incorporates the idea that category is not inferred only through the bottom up processing, but also by comparing any possible high-level explanation for it. For example, ideal adapter has predictions about the upcoming category and has beliefs about the cue distribution. The cue distribution may be inferred from the observations of the world but are likely not identical to it according to (Kleinschmidt & Jaeger, 2015), since listeners have limited observations of the world.

To go from one listening situation to another, listeners can change the likelihood function. For example, if the speaker is producing cue values that do not match the known speakers, changing likelihood function to accommodate for that can solve miscommunication problems. One speaker may articulate big VOT differences between voiced and unvoiced sounds, while another will produce much smaller differences. Therefore it is crucial for the listener to use the right likelihood function. If the talker is known, Kleinschmidt and Jaeger, 2015 assume that the likelihood function may be summoned from memory. If the listener encounters a similar speaker to the known speaker, they can help themselves with prior knowledge. However, if they do not know the speaker at all, they need to update their beliefs.

In particular, they offer a mechanistic account for two phenomena where listeners exhibit changed behavior about categorization: phonetic recalibration (Kraljic & Samuel, 2005; Norris et al., 2003) and selective adaptation (Eimas & Corbit, 1973; Samuel, 1986). Phonetic recalibration or perceptual learning is a term to describe a phenomenon where the listener gets an ambiguous

token from a continuum between two categories paired with a disambiguating label (i.e., visual or lexical). It is thought that the representation is changed to t the confusing pair of signal and the label. On the other hand, selective adaptation is a phenomenon that is thought to re ect less categorization of the tokens of an exposed category member, possibly due to fatigue “of the feature detector” (Eimas & Corbit, 1973), where it is unclear what that feature detector is.

In Kleinschmidt and Jaeger, 2015, the listeners can adapt by shifting the cue's mean or by expanding the variance of the cue (depending on which is more uncertain) to update their beliefs about a particular talker's productions. This happens incrementally as they encounder deviations from the expected. They can update their beliefs incrementally either for a speci c talker, or for a group of talkers. These updates do not interfere with each other, because they represent the beliefs about a generative model of each speci c talker (or a group).

This framework requires a different model for each speaker (or speaker group) they ever encounter in their lives, which means storing parameters in memory for each listening situation. However, the downside is that the same talker may sound a lot different in noise, i.e. talking from a different room, over the phone, on a busy street etc. This means that not only does a listener have to store one generative model per talker, but also per situation. And while they show that a listener can group similar talkers and situations together, it is unclear what that means for personal deviations from the standard (i.e. a particular speaker's properties deviating from a certain dialect the listener is familiar with).

Furthermore, adapting the parameters of the cue implies that the listener's perceptual space is full of different cue representations, since they differ for each speaker. Relatedly, the ideal adapter framework needs to continuously add and adapt these representations.

Finally, Kleinschmidt and Jaeger, 2015 offer an explanation about adaptation in a single

dimension only – that means, that in their approach, it is unknown how one cue becomes more reliable than another.

I offer an alternative computational model that can adapt without representing each speaker or group as a separate model: a single model that can adapt on a moment-to-moment basis. This eliminates the need to model the speaker differently for each different listening situation one is in. Rather, we simply model exible adjustment from one situation to the other that can happen quickly should the listener encounter a completely new type of noise masking the speaker (that the listener may already know).

2.5 Conclusion

In this chapter, we have seen that phonetic-acoustic features or cues are statistics of a part of the signal that the listener receives. They are not random, but rather carry linguistic information. Cues are what helps the listener distinguish between phonetic categories – signaling the acoustic dimension alongside which there is an evident contrast between at least two categories. They point to a higher level phonological feature or abstraction and as such there are many cues pointing towards the same phonological feature. The one that the listeners of a speci c language group tend to rely on (or weight) most on average is typically called the primary feature. The weighting mechanism has been implemented before, some of which I have described in more detail (Kleinschmidt & Jaeger, 2015; Toscano & McMurray, 2010). They each paved a way into better understanding of how listeners rely on features and how they are able to adapt, but they also have shortcomings. These shortcomings (relying completely on long term statistics, or having a multitude of models to be able to adapt) will be discussed also later in chapters 5 and 6 where I

present a different solution to cue weighting: one based on rate distortion theory, an information theoretic framework, and on bigger speech corpora than what has been used previously in the literature.

Chapter 3: Selective Attention

In the previous chapter we have seen what cues/features are and how they may be treated and weighted by the listener. Here, we will rst review in more detail experiments that test listeners with ambiguous cues, especially with short term changes to cue distributions. In addition, we will look into noisy and distorted speech which admittedly lack some of the cues listeners usually rely on. We will see that listeners adapt rapidly in a way that cannot easily be accounted by prior models (since cue weighting now no longer responds to long term statistics). Importantly, we will look into an alternative hypothesis that can explain this plasticity: a mechanism that has been called “selective attention” that could be responsible for quick adaptation to new talkers and listening situations on a moment-to-moment basis. We will also review research from neural measurements that may offer support for the selective attention mechanism. This will serve as the data that the computational modeling in this thesis is built to account for: exible cue weighting as seen in humans may be due to quick re-allocation of attention to cues that serve the listener best in the speci c moment. These data will serve us to propose computational simulations of cue weighting, rst based only on input data, and second with adjusted model that can re-weight the cues on a moment-to-moment basis.

3.1 Introduction

Speech cues are an important part of speech perception. However, they are constantly changing depending on the speaker and the listening situation. In other words, there is a notable lack of invariance. As we have seen in the previous chapter, some of the prior literature focuses on lifelong cue weighting, trying to elucidate primary cues, such as modeled by Toscano and McMurray, 2010. Other literature focuses on short term uctuations and the listener's responses to those (Kleinschmidt & Jaeger, 2015). The cues may actually change depending on the environment the listener is in, as well as on the speaker's current properties, that may also change with time.

Therefore, which cue is reliable is a changeable phenomenon. How do the listeners decide which cue is reliable? Prior research consistently nds a so called primary and secondary cues that are identical across experiments, such as primary VOT and secondary F0 for voicing for English American speakers (i.e. Francis et al., 2008; Idemaru and Holt, 2014; Whalen et al., 1988). But a lot of experiments also happened in a laboratory setting with speech that has been carefully manipulated, with a single voice, without any other noises. Thus, the primary cue is often established within a listening situation considered as a baseline: clear speech, with no disturbances.

However, we know that context and listening situations change which cues are available and also how they are interpreted, i.e. because of the range of stimuli presented (Brady & Darwin, 1978), because of presence of other phonemes (Brady & Darwin, 1978; Mann, 1980) or because of rate of speech (Miller & Liberman, 1979) (for more extensive review, see Holt and Lotto, 2006). This shows that cues important for a particular category change across listening situations

and make speech perception less trivial than it may seem.

Finally, it is also important to note individual differences in cue weighting (Kapnoula et al., 2017; Kong & Edwards, 2016; Ou et al., 2023; Schertz et al., 2015; Wu & Holt, 2022). Although often experimental results averaged over all of the participants give an impression that the listeners uniformly behave in a certain way, each individual actually exhibits slightly different cue weighting.

This changeability of cues requires the listener to always be adapting to the situation and to the speaker. Section 3.2 describes a range of experiments that exhibit quick adaptations by listeners exposed to miscorrelated cues (i.e., cues that give con icting information with a combination of non typical values). First, we will look into short term cue re-weighting due to miscorrelations, then into cue weighting when speech is presented in noise, and then into adaptations to corrupted signal, like distorted speech.

These experimental results lead us to examine a new mechanistic proposal in the literature, termed selective attention (Holt et al., 2018). Although it is unknown whether attention selectively upweights speci c features (and differently given different distortions present in the signal), it is known that it can enhance all of the representations of features of the attended speech in the brain (Shinn-Cunningham, 2008).

Likewise there is some evidence that attention provokes neural tuning shifts of the auditory cortex neurons in ferrets (Fritz et al., 2003). These speculations are centered around neural experiments, some around measured neuronal responses sensitive preferentially to different parts of auditory input, including linguistically important dimensions. It is unclear whether neurons in auditory cortices are “detectors” or “integrators” or whether there are more of “detectors” in some regions than others. It is also unclear whether “detectors” are found on the lower level and

“integrators” on higher level. However, we could hypothesize that because of the hierarchical nature of linguistic processing, feature “detector” neurons are integrated by feature “integrator” neurons that can represent more complex and abstract information.

I use the term selective attention for a mechanism responsible for exible cue weighting, although it may be a separate mechanism from actual attention that acts as a spotlight: what I mean is a mechanism that can treat information coming from each feature disproportionally (as is typically meant by cue weighting). In addition, I assume it is a mechanism that can easily shift reliance from one feature to another to facilitate intelligibility. It seems to be as a process that is outside of conscious control, but can be provoked by input, as has been described in a series of experiments described in section 3.2.

Selective attention or rather, exible and quick cue weighting based on a moment to moment basis will be implemented in chapter 6. However, before we can do that, we must rst present experimental data that gives evidence for quick cue weighting. In addition, we must also present modeling simulations done on models that do not employ selective attention to see how far they can take us.

3.2 Short Term Cue Weighting

A number of experiments (Idemaru & Holt, 2014, 2020; Lehet & Holt, 2020; Liu & Holt, 2015; A. Symons et al., 2023; Wu & Holt, 2022) have tested listeners in a very similar fashion to elucidate short term cue weighting, therefore I will summarize their structure and their results together in this section, generalizing over all of them. These experiments are important because they explicitly show cue weighting changes that seem to follow the exposure of short-term cue

miscorrelations.

In particular, these experiments show that if a listener is exposed to miscorrelated features, they will stop relying on the secondary (or the less reliable) feature, even when it becomes informative again. Second, they show that if speech is presented in noise, the primary cue will be different: whichever was the secondary cue for speech in quiet now becomes the primary cue for speech in noise.

These experiments show incredible fast adaptability of human listeners to use whatever is available to them in the moment. This research is also used for simulations in the present thesis and has been some of the main driving force to propose and test models of exible and quick cue weighting a moment to moment basis.

3.2.1 Accented Speech

In a range of experiments (Idemaru & Holt, 2014, 2020; Lehet & Holt, 2020; Liu & Holt, 2015; Wu & Holt, 2022), adult listeners of American English were exposed to speech with particular properties. Speech that they heard was natural, but it was a particular combination of two cues/features. These two cues were already established by prior research to be important for mapping signal to category. Speci cally, Idemaru and Holt, 2014, 2020 and Wu and Holt, 2022 tested listeners for word initial stops [p],[b],[t],[d] with words pier, beer, tear, deer. The cues used in this experiment were VOT and F0, which have been already determined by prior research to be covarying primary (VOT) and secondary (F0) cues (Francis et al., 2008; Lisker, 1986; Whalen et al., 1988). Lehet and Holt, 2020; Liu and Holt, 2015 and Wu and Holt, 2022 used American English [æ] and [ɛ] vowels in words sat and set (or satch and setch), with covarying primary cue

as spectral information (rst ve formants) and secondary cue as duration.

One of the main goals of the experimenters was to see how the listeners' cue weighting changes with combinations of different cue values. In particular, typical standard English usually has a short VOT and lower F0 for voiced sounds ([b], [d]). It is inverse for unvoiced [p] and [t]: their VOTs are typically longer and F0 are higher. When it comes to vowels [æ] and [ɛ], the formants (speci cally mostly f1) are typically in the lower range for [ɛ] than for [æ], while duration is typically shorter for [ɛ] than for [æ].

These cues were used in the experiments with different steps from a single voice production, in particular, with a 7-step continuum. This continuum was an interpolation between two good exemplars of each category of a single speaker, so that the further down the continuum we are closer and closer to a different vowel category. We can represent this schematically as a square, partitioned in 7x7 steps – see gure 3.1. Each step is a different value of a cue, with each corner of the grid denoting typical values associated with a category.

Figure 3.1:Experimental 7 step continuum- Each dimension represents a different covarying cue (either for vowels or consonants). Each square is a different stimulus with a different combination of cue values. Colored lower and upper squares denote where typical values of a cue denote a speci c category.

First, a baseline cue weighting is established by letting the participant identify each pair of

cue values out of the 7x7 continua. This line of research uniformely nds that VOT is a primary cue for stops and spectral information is a primary cue for vowels, with F0 and duration being secondary for consonants and vowels respectively.

Second, listeners are exposed to miscorrelated values (see gure 3.2), such as feature 1 pointing to one of the categories, and feature 2 pointing to the opposite category. This represents a con ict for the listener if they were weighting the features equally. However, from established baseline we know that this is not so – the listeners on average relied more on VOT (or spectral information), which were dubbed their primary cues.

Figure 3.2:Miscorrelated Features- The green squares denote the exposure stimuli, which are now combinations of non typically correlated cue values. On the one dimension, listener may want to categorize the stimulus with category 1, whereas the second dimension will be indicative to category 2.

This is the crucial point of the experiment, because how listeners behave reveals something very important about cues and short term reactions to them. If there was no primary feature, the listener would face a con ict they would need to resolve (for a speci c example of the dilemma and of a stimulus with miscorrelated features, see gure 3.3). Each feature points to a different category. Of course, this is assuming that the listener's perceptual space has categorized category 1 and 2 each in lower left and upper right corner of the cue values.

Figure 3.3:Con ict - How does a listener decide? If both features had equal weight, this would be a hard choice, since for this particular stimulus with cue 1 value 6 and with cue 2 of value 2 each denote a different category. The arrows show what the listener might think given each dimension values. For cue 2 with value 2, this would likely be category 1. It would likely be category 2 if we only looked at cue 1 with value 6.

We know that listeners rely more on VOT for word initial stops and more on spectral information when identifying vowels when the speech is clear, but that does not require that listener use the same feature in this situation. There are several possibilities on how to weight the features, but for simplicity let us illustrate two extreme cases: both where the listener completely ignores (downweights) one of the features. One with feature 1 being relied on completely (see gure 3.4), and the opposite case, where the feature 2 is being relied on completely (see gure 3.5).

At test time, the listeners are given stimuli that have ambiguous primary feature (such that it cannot be relied upon), and a secondary feature that is highly indicative of one of the categories. The listeners again simply had to identify the word they were hearing.

In all of the results of the aforementioned studies, it has been found that whatever was a

Figure 3.4: Feature 1 is Upweighted- If the listener only relied on feature 1, the perceptual decision boundary would completely depend on it. The stimulus in question (with feature 1 of value 6 and feature 2 of value 2) would now be categorized as category 2.

primary feature at the beginning of the experiment (that is, before the listener was exposed to the miscorrelated cues), was also the feature the listener relied on more when presented with miscorrelated features. This could be because, simply, the listener's choice depends on primary cue as learned from the long term statistics, such as would be explained by Toscano and McMurray, 2010. In other words, the listener learned in their lifetime that the primary feature is more reliable overall and decides to now discard the information coming from the secondary feature in favor of the feature they trust more. Another option is that the listener recognized that the talker now has an accent and must adapt to them. Therefore they must guess and model which of the features are unreliable. They may decide that without other disambiguating information, the primary feature should prevail as with the talker at the beginning of experiment, thereby extrapolating on a group of talkers (Kleinschmidt & Jaeger, 2015).

If the listeners accepted this talker to be producing a short VOT with a high F0, maybe now the listener should accept that for this talker, the cues are simply reversed. Since they

Figure 3.5: Feature 2 is Upweighted- If the listener only relied on feature 2, the perceptual decision boundary would completely depend on it. The stimulus in question (with feature 1 of value 6 and feature 2 of value 2) would now be categorized as category 1.

now correlate differently, the listener could adapt – however, this never happens in experimental results, not even after about 4.5h exposure of miscorrelated features over 5 days. In addition, as soon as the features are typically correlated again, the listeners “bounce back” and use the ignored (secondary) feature again.

However, even though in most experiments the listeners never learned to reverse the correlation between the features, as if completely responding to the short term statistics, H. Zhang et al., 2022 nds reversal. This reversal happens when listeners are tested with male/female sounding voices that overlap in their F0: female low F0 is male high F0, therefore the stimuli for those cues are identical. Interestingly, for female miscorrelated cues, listeners downweight F0 (as in other experiments), but for male miscorrelated cues, they temporarily reverse the cues: stimuli with high F0 are suddenly more likely to be categorized as [b]. This only happened when male and female voices produced overlapping stimuli during exposure, and did not happen if listeners heard only male voice across all experimental blocks (with identical stimuli across all

experiments for the male voice).

Interestingly, listeners do not generalize from one contrast (i.e. [p]/[b]) to another (i.e. [t]/[d]) (Idemaru & Holt, 2014) or to a new context (from bier to bear, or from bier to bill) (Idemaru & Holt, 2020). The effect of downweighting is also smaller when the speaker has changed (Liu & Holt, 2015). Liu and Holt, 2015 did nd that listeners generalize from words setch/satch to set/sat. However, all of the four words had the parts “se”/“sa” acoustically identical across words (that is, the beginning of se-t and se-tch are physically identical). It is therefore suggestive that adaptation happened to a particular voice (speaker). This is unlike a more ecological scenario in Bradlow and Bent, 2008 where listeners have been shown to generalize their accent adaptation to novel speakers.

In X. Zhang et al., 2021, listeners were given an ambiguous VOT paired with a low or high F0 and a disambiguating lexical information. This lexical information makes listeners rely on F0 even when they have been exposed to acoustic miscorrelations paired with lexically disambiguating words. That is, the exposure was a miscorrelated F0 and VOT stimulus and a lexical item that helps the listener decide whether it's a [p] or a [b] (peace vs. beef), while the test is composed of a single lexical item (peak), ambiguous VOT and indicative F0. In other words, lexical information with paired miscorrelated features seems to prevent cue adjustment. Listeners may rely on (or remember prior) lexical information rather than rely on the acoustics alone. This makes sense from the viewpoint that speech perception does serve mapping signal to categories, but ultimately that this is done to extract the message out of the signal.

Finally, when listeners learned to downweight duration (secondary, less relied on cue) in Lehet and Holt, 2020, they still used duration for subsequent categorization of consonants later in the word (“seppa” vs. “sebba”). This indicates that downweighting of a cue is not general or sus-

tained across time and phonemes. It is unclear how the ideal adapter framework (Kleinschmidt & Jaeger, 2015) would explain that, because cues in Kleinschmidt and Jaeger, 2015 are represented in one dimension and it is unclear how cues are weighted between each other. It could explain that only the cues related to the vowel are modi ed each in its own dimension. However, as soon as the listeners get back the typically correlated input, they show the behavior they started with and no downweighting of secondary cue is happening. The ideal adapter framework would need to incrementally adapt the mean and the variance of the cues of the vowel between conditions on the same speaker or create multiple models for the vowel of the same speaker.

Overall, these experiments suggest that listeners quickly and effortlessly adapt to temporary statistics from a single speaker. They do not seem to generalize that to other speakers, but perhaps this is also the case because they were never exposed to a group of talkers with miscorrelated features. They also never learn the reverse cue correlation, except in one case (H. Zhang et al., 2022), which had overlapping stimuli for two voices (male/female).

3.2.2 Noisy Speech

In addition to the experimental design described in section 3.2.1, Wu and Holt, 2022 added a new experimental parameter: noise. Whereas most prior literature on cue weighting focuses on clear undisturbed speech, Wu and Holt, 2022 tested listeners with identical experimental paradigm in noise: listeners were tested on words with word initial stops (pier/beer) and on vowels (set/sat). It is worth mentioning that this noise type was speech shaped white noise l-tered with a spectrum of an interview segment. This particular noise may have been particulary ef cient in masking primary cues (VOT and spectral information). There were no other reports

or other noises, that may cover other parts of the signal.

Interestingly, listeners' primary cue for [p]/[b] was VOT (as observed in prior studies), with secondary being F0. However, when noise was added, F0 was primary and VOT was secondary. The same happened with vowels [ɛ] and [æ]: in clear speech, spectral information was a primary cue, where duration was secondary. In noise, the reverse was true (duration was primary and spectral information was secondary).

Furthermore, the more the person used the primary cue at baseline (tested after hearing typical correlations), the more they downweighted the secondary cue after being exposed to mis-correlations. This was true for clear speech and for speech in noise. Thus, relying on one of the cues more from the beginning likely facilitated the listener to upweight that cue more and ignore the secondary cue.

In a similar fashion A. Symons et al., 2023 tested listeners with informational masking in a dichotic listening task: they received one speaker in each ear. Their reliance on F0 (as a secondary cue on average) was tested. Those listeners that relied more on F0 relied on it even more in dichotic task and all of the listener increased their reliance on F0. This is not surprising since F0 carries vital information about the speaker and can help disambiguate informational masking.

This challenges some important assumptions by prior literature – that is, that the primary cue is likely always primary in a certain context, since it has been found in so many experiments, and that it likely comes from the observed long term statistics. Indeed, maybe humans learn separate long term statistics for each different listening situation, as Kleinschmidt and Jaeger, 2015 would predict. Cue weighting based on long term reliability (Toscano & McMurray, 2010) would likely fail to predict this behavior, unless it somehow assumed that indeed long term statistics is

different in noise.

However, this particular noise type the participants heard in the experiment is likely unknown to them. It may be similar to noises they have heard before in their lives, maybe enough so to extrapolate upon, but likely this particular noise is new. The speaker is also new and therefore the ideal adapter framework (Kleinschmidt & Jaeger, 2015) would need to create several models (one for clear speech, one for noisy speech) to account for that.

Finally, this research shows that listeners can very quickly adapt to temporary cue miscorrelations. It is unclear, however, how they actually do it, as none of the prior models can satisfactorily explain the rapid adaptation. It is also unclear exactly why acoustic miscorrelations or noise clearly lead to adaptation, but do not seem to generalize well over contexts or speakers, nor when there are lexically disambiguating words present.

3.3 Distorted Speech

In the previous section we have seen that listeners have an incredible possibility of quick adjustment of cue weights – listeners seem to adapt to unusual input in a speci c way. This seems to be to rely on one of the two con icting cues, where the cue that is relied upon seems to depend on listening situation (is the speech presented in quiet or is it in noise?). However, there are many more listening situations that listeners have shown to be able to adapt to relatively quickly and some of these will be summarized in this section. This unusual input may be arguably even harder to adapt to than the manipulations seen above (most notably for example in Wu and Holt, 2022). The quick adaptation to very unusual input shows the robust exibility of human speech perception that we model in chapter 6.

In particular, listeners can adapt even to speech that is unnatural, either because it is ltered (Li & Allen, 2009) or distorted in some way (Adank & Janse, 2009; Davis et al., 2005; Elliott & Theunissen, 2009; Guediche et al., 2016; Holdgraf et al., 2016; Khalighinejad et al., 2019; Khayr et al., 2024). Of course, the degradation can be so bad that it completely prevents intelligibility, however, listeners do show rapid plasticity when they receive either an external disambiguating cue or when they are able to learn with repetition. I summarize some of the studies below.

One out of many studies on synthetic speech, Remez et al., 1981 tested listeners' intelligibility of three tone sinusoidal signal replicating natural utterances, without the traditional acoustic cues. Participants were able to identify the arti cial stimuli and the authors claim that it is because of time varying properties that match the natural acoustics. In other words, only the formant center frequencies were heard in the utterance by combining three tone patterns, which seemed to be suf cient for intelligibility.

Li and Allen, 2009 tested listeners with speech that was partially or completely masked by noise by low or high band lters. With that they tested whether these bands are treated independently in perception, however this did not hold true for consonants. This is likely because of important cues covering more than one band. Nevertheless, they showed that some amount of degradation does not prevent intelligibility (until it does).

In Davis et al., 2005, stimuli were distorted by ltering them into six bands, out of each a smoothed amplitude pro le was extracted and then noise was modulated in each band. These noise vocoded stimuli somewhat imitate cochlear implant input and served to ask whether intelligibility improved for the listeners simply over time (with more presentations), and whether receiving information in written or oral form would facilitate learning. Participants did improve over time even without disambiguating information. They also improved when they heard the

distorted utterance followed by its clear version (acoustically or the written form). In addition, when participants were presented with a distorted utterance, followed by its clear version, and then again by its distorted version, listeners suddenly experienced a pop-out effect.

In Guediche et al., 2016 listeners were exposed to severly distorted monosyllabic words. These words were vocoded and spectrally shifted so that when listeners were asked to identify the words, they showed very low intelligibility (recognition accuracy at 11%). When the listeners were presented the with written form of the word after their response, their intelligibility would improve rapidly over the course of only a few repetitions. Similar rapid adaptation was also found when listeners were presented with a gradual degradation of the stimuli (from low to high distortion over the course of the experiment). Therefore, listeners can adapt even when several cues are missing in heavily distorted speech.

A neural study (Holdgraf et al., 2016) found similar results when it came to distorted speech, and they also found neural changes that correlated with participants' improved intelligibility. Speci cally, an ECoG study was performed on passively listening participants, hearing triplets of degraded-clear-degraded utterance, somewhat similar to Davis et al., 2005. Listener's intelligibility at the rst presentation of the degraded sentence was quite poor, but after the presentation of its clear version the intelligibility rose to about 77% of words.

Overall, it is striking how adaptable speech perception is. Even heavily distorted speech can be intelligible. And, if it is not from the beginning, there is a possibility of adaptation to achieve intellegibility, either with more exposure or with disambiguating repetition of the stimulus (written or oral). In fact, actual users with cochlear implants were tested on a perceptual learning task with speech in quiet and in noise and also exhibited learning for time compressed speech (Khayr et al., 2024). These studies show that cue weighting has to operate also on the signals

that possibly lack some, if not many of the usual cues found in normal speech. The mechanism driving adaptation may be always active, since many moments in our lives have slight deviations from the regular acoustic scenes we are used to. Current models simply cannot account for such rapid plasticity that listener exhibit, which brings us to introduce an alternative hypothesis.

3.4 Selective Attention

In the previous section we have seen that listeners adapt rapidly to short term statistics and that they also adapt to many types of distortions of speech: from accent adaptation (Bradlow & Bent, 2008; Clarke & Garrett, 2004), to rapid adjustment of cues after short term miscorrelated speech exposure (Idemaru & Holt, 2014; Lehet & Holt, 2020; Liu & Holt, 2015; Wu & Holt, 2022; X. Zhang et al., 2021) to various types of noises or disturbances (Davis et al., 2005; Guediche et al., 2016; Holdgraf et al., 2016; Wu & Holt, 2022). Prominent prior cue weighting models are either based on long term statistics (Toscano & McMurray, 2010) or they posit that the listener should create a model for each situation they are in (Kleinschmidt & Jaeger, 2015). As such, neither of this models is exible and quick enough to account for a myriad of data that shows quick adaptation.

An alternative hypothesis for rapid cue re-weighting has been put forth in the literature: selective attention (Francis et al., 2008; Holt et al., 2018; A. Symons et al., 2023; A. E. Symons et al., 2021). What is selective attention, and speci cally what is it in the context of cue weighting/reliance? This section summarizes the mechanism of selective attention and argues that we should consider it as a viable hypothesis of rapid and effortless moment-to-moment cue weighting readjustments.

In Holt et al., 2018, the authors offer a term “non spatial dimension based auditory attention” to mechanistically explain rapid cue weighting that is adaptive to the listening situation. They posit that listener can adjust their attention to one cue or another, depending on the speaker and the environment. This means that attentional in uence goes all the way down to adjust, or better lter features that are represented in the brain, particularly in the auditory areas.

However, auditory selective attention is connected to a multitude of phenomena, and as such is a term loaded with many meanings for various types of research. Selective attention in audition is most often described as choosing what to listen to in a mixture of sounds (as in for example cocktail problem (Cherry, 1953)). There is currently no direct evidence that this is or is not the same mechanism that distributes weights to cues depending on the situation. In fact, research in neuro-biology seems to have shown a lot that resembles potential cue weighting, but as of yet, this is all hypothetical.

Below I will summarize some of the research that shows effects of auditory selective attention (and that has been described with such terminology) that may be indicative of a process that allocates weights cues on a moment-to-moment basis, but it remains open whether these scienti c ndings come from a single mechanism or more mechanisms, as has also been stated by Holt et al., 2018 who proposed to name dynamic and plastic cue weighting as selective attention. I will also discuss why it is interesting to entertain this hypothesis, and later in chapters 5 and 6 show how it can be implemented in a computational model.

First we will look brie y into how sounds are streamed and grouped perceptually, then more closely into how sounds may be encoded in primary auditory areas. Finally, we will look into how neural encodings may change due to attention and why this may be similar to cue weighting.

3.4.1 Auditory Objects, Streaming And Attention

Auditory objects form an important part of research on auditory perception and their formation may be amenable to attentional processes. What exactly is an auditory object depends on de nitions and use of particular research (T. D. Grif ths & Warren, 2004), but largely it seems to denote a perceptual event, that is separated by perceptual boundaries from other events and that is an abstraction of information independent of the speci c sensory representation. T. D. Grif ths and Warren, 2004 also argue that they can have temporal dimension.

Bizley and Cohen, 2013 further argue that an auditory object is a perceptual construct that can be assigned to a particular source, that it spans events over time, that it has spectro-temporal properties that separate it from other objects, since it is a combination of source speci c features. Importantly, an auditory object is still an object even though it can incur various changes to its spectro-temporal properties. For example, rate of speech changes should not prevent us from recognizing a speaker or a word.

An auditory object is supposed to be formed from acoustic features into perceptual features. Hawkins, 2010 claims that linguistic units can be understood as auditory objects. The parallels are easy to see: for example, Chomsky and Halle, 1968 posits that distinctive phonological features form a phoneme, which can be understood as a perceptual representation or an auditory object that is an abstraction from the actual sensory signal.

Although attention is not needed for auditory object formation, it can enhance the perceptual awareness of it (Bizley & Cohen, 2013). For example, Ding and Simon, 2012 show that when listening two competing speakers, individual neural representations can be extracted for each that correspond to representing them each as an individual auditory object. That is true

also when speakers are of the same gender and share more spectro-temporal properties and thus harder to disambiguate.

Golumbic et al., 2013 further found that in attending a single speaker in a cocktail party problem, the attended speaker is enhanced in its neural representation, while the ignored speech still remains represented in low-level auditory cortices (in and near superior temporal gyrus). Higher order regions (that do language processing) show more of selectivity for only attended speech.

Other studies showed neural changes due to attention. Chait et al., 2010 showed that listeners can be primed to attend certain moments while ignoring others, which showed in MEG analysis as auditory cortex attenuating responses to signals that the listeners were trying to ignore and enhancing the responses of the attended part of the stimulus. Elhilali et al., 2009 nds with an MEG analysis that the when participants attended the target masked with the background masker, target rhythm at 4 Hz emerges in the neural signal of the participant. Meanwhile, when the participants attended the background (or masker) in the inverse task, the 4 Hz cortical response is suppressed. They take this difference as a modulatory effect of task-dependent attention on the neural representation itself.

Mesgarani and Chang, 2012 reported that a speaker in a mixture of two speakers was better reconstructed from the brain recordings when the listener attended to that speaker. Relatedly, Lu et al., 2024 describes that in ferrets, attention to a sound sequence binds this sound representation to other temporally coherent sounds (as in when one attends to a single speaker in a mixture of sounds), with an enhanced representation in auditory cortex, while at the same time suppressed incoherent sequences.

Thus, attention is supposed to be operating on the level of auditory objects. However, if

attention is focused on a particular feature, there is enhanced sensitivity to other features of that source (Bizley & Cohen, 2013) – i.e. if the listener focuses on the tone frequency, the location feature is also enhanced (Zatorre et al., 1999).

Furthermore, Shinn-Cunningham, 2008 argues that auditory object formation and selective attention cannot be fully separated. In fact, attention both participates at object formation and enhances its features. First, grouping of spectro-temporal structure happens, after which several auditory objects arise with higher order features. One of these objects is enhanced by attention, but so are its features, higher and lower level as they are coming in in time. Fritz et al., 2007 also describes attention as a mechanism that can selectively lter the features that are important to the organism, with rapid reshaping of receptive elds in auditory cortex.

Similarly, Shinn-Cunningham and Best, 2008 also argue that when the listener knows that the object they want to attend has a speci c feature, they can bias their selective attention to prioritize objects with that feature. This, however, is particularly hard for individuals with hearing loss that may not have encoded the stimuli well in the auditory periphery.

These results do not offer a clear understanding on whether (and how) attention modulates the feature representations. However, studies done directly on groups of neurons or single neurons that may act as feature detectors or integrators in humans and animals may clarify this slightly.

3.4.2 Spectro-Temporal Receptive Fields

This section discusses neurons in auditory cortex of both animals and humans to explain brie y their purpose. It is important to elucidate brain structures responsible for treating auditory

inputs since that is where cues may be represented. Moreover, it is important to understand neural processing at this smaller scale to understand how speech cues/features are encoded by the brain – at least to the extent that is currently understood. This can help us understand how a cue weighting mechanism that is closer to biological reality may work.

The idea of feature detector cells as encoding perceptually signi cant events has been born as early as with Barlow, 1953, and concurrently with Hubel, 1959. Likely, there were antecedents for that even before by y Cajal Santiago, 1892 that suggested “psychic cells” (Martin, 1994). In linguistics, likely inspired by discoveries in neural encodings, feature detectors were proposed by Eimas and Corbit, 1973. Decades later, and many many experiments more, we could hypothesize that there are feature detectors also in auditory areas (Mesgarani et al., 2008). However, we do not understand well how these feature encoders work, what all they encode, what they represent and how bigger perceptual events are encoded. Importantly we do not know whether these neurons are actually detectors or whether they are integrators and whether they are context sensitive. We also do not know well what processes guide the interactions between bottom up sound information detection/integration and top down in uences from higher processes (i.e. attention, memory, other grammatical processes such as semantics, syntax etc.). There are however, many studies that try to shed light on that.

Research has found neurons in both animal and human primary/secondary auditory cortex having spectro-temporal receptive elds (STRFs). The STRFs denote spectrotemporal stimulus characteristics in time and spectral domain that drives each neuron the most. That means we can see whichever frequency the neuron was most (and least) responsive to in some window of time, depending on the stimulus that the organism heard. Neurons' responses also have different rates and depend on the stimulus. However, for simplicity we will look at general ndings about

neurons in auditory cortex with speci c interest in understanding what their responses are given some speech stimulus, how they differ and whether they change.

Mesgarani et al., 2008 showed that it can be possible to extract phoneme like STRFs from the neural recordings of ferrets' primary auditory cortex. That means that speech is neurally represented almost as a phoneme-like unit in an animal that likely does not care or understand for human speech. Later, in an ECoG study Mesgarani et al., 2014 showed that distinct speech features can also be extracted from neuronal representations from superior temporal gyrus (STG), which is a human auditory area. In fact, the speech features were represented with STRFs that resembled the spectrograms of particular acoustic properties – for example, of nasality, or of plosiveness, higher phonological features, as well as more acoustic-phonetic features such as VOT. Recently, Leonard et al., 2024 explored 685 single neurons in STG and similarly found speech cue like sensitivity in recorded neurons.

In a review of computations in non primary auditory area, superior temporal gyrus, Bhaya-Grossman and Chang, 2022 report that STG is more responsive to speech sounds than other sounds, which is not like the primary auditory cortex – it is equally sensitive to all sounds. It may be that STG encodes phonological units and ner acoustic-phonetic features on a different scale (Bhaya-Grossman & Chang, 2022).

For example, Oganian et al., 2023 explored sensitivity of neurons to formants and vowels in human STG. They did an ECoG study with several electrodes, out of which some were specifically selectively responsive to formants. None of the electrodes were speci cally selective for an entire vowel, however, some of the electrodes were selective to formants themselves. Furthermore, when they grouped electrodes in subset, i.e. high F1, low F2, they found that these receptive elds could discriminate between vowels (i.e. between [a], [e] and [i]). A combination

of two electrode pairs (most sensitive to low F1, high F2, and another sensitive to high F1, low F2) could discriminate between all vowels tested (cardinal [a], [e], [i], [o], [u]). These ndings con rm neuronal encoding of higher phonological features (i.e. high vs. low vowel).

Neural preferential response to speci c acoustic aspects of speech have been found in adults and some indications for them have also been found in a few month old infants. Gennari et al., 2021 claim they found invariant orthogonal neural codes for the dimensions of manner and place of articulation in 3 month old infants. Infants were tested with 256-channel electroencephalographic system on syllables that differ in manner and place of articulation. Researchers claim that the invariant representations for manner and place of articulation are then integrated secondarily. This would indicate that infants are able to represent the acoustic-phonetic feature and then integrate them to get the higher representations.

Using naturalistic speech, Menn et al., 2023 show an emergence of stable neural activity for phonological feature processing only in 14 month old infants. They posit that the later age may be due to the fact that they used naturalistic stimuli as opposed to more carefully curated syllables in Gennari et al., 2021. They also show that categorical feature processing is native speci c, much like behavioral ndings of phoneme acquisition, i.e. Werker and Tees, 1984, and that infants rst acquire phonemes that extend over longer time intervals.

Overall, these studies show that humans (and animals) may have neurons or groups of neurons that are particularly sensitive to linguistic feature information. Indeed, that is likely true from soon after birth as some research demonstrates (Gennari et al., 2021; Menn et al., 2023). Interestingly, some research shows that some neurons preferentially respond to linguistically relevant dimensions, such as acoustic-phonetic features (i.e. VOT) (Mesgarani et al., 2008) and higher more abstract (phonological) features such as +/-low in vowels, plosiveness in consonants

etc. (Oganian et al., 2023). Some of these neurons may act more as detectors (i.e. those more sensitive to VOT information) and others as integrators (i.e. those more sensitive to phonological features, such as +/- low), however this is still unclear.

How these neuronal encodings are composed into higher representations is unclear. Sounds are processed hierarchically in more and more complex perceptual units. For example, Bizley and Cohen, 2013 posit that there may be an increase in information abstraction in neural representations, going hierarchically further up from primary auditory cortex. Moreover, Brodbeck et al., 2018 show neural responses of incremental processing from acoustics to higher language processing. I assume that cues are part of this hierarchical neural processing and may be treated differently at different moments, depending on how reliable they are in the moment.

Furthermore, Gwilliams et al., 2018 show that the acoustic information is maintained in the auditory cortex while nding out the identity of the lexical item being heard. This uncertainty about the phoneme given its features in natural speech comprehension may be indicative of the plasticity of cue weighting that can be readjusted given the lexical knowledge. Similarly, Gwilliams and Davis, 2022 discuss how auditory responses to speech sounds are in uenced by morpheme and lexical information.

To conclude, we have seen a more “zoomed in” view of neural encodings of various features important for speech perception. Attention and further language processing may in uence neural speech sound representations. But how deep those changes go and whether they are similar to cue weighting is addressed by the following section.

3.4.3 Changes In Spectro-Temporal Receptive Fields

In the previous section we have seen that auditory areas represent acoustic-phonetic features. However, these representations may not be stable over time, which may be a direct effect of attentional processing. In fact, research has observed changes in neural encodings due to attention.

A particularly interesting example comes from Fritz et al., 2003 where ferrets were trained to detect a target tone of any frequency. The neural representations were examined as STRFs. When ferrets learned to attend to a speci c target frequency, facilitative shape changes were observed in STRFs (as seen in comparisons of STRFs for before and after training). In particular, sensitivity for target frequency seemed to be enhanced in the neural representation, even when the target was changing. These changes were not observed if the ferret was not trained, therefore behavior seems an imporant factor in the STRF changes.

In addition, Elhilali et al., 2007 did a similar study and veri ed that STRFs appear stable over a course of time (from 30 minutes to 2 hours). Furthermore, they veri ed that attentional plastic changes are not result of inherent variability of STRFs, but rather of goal directed behavior. Similar has been observed also for detection of a target tone in noisy background (Atiani et al., 2006; Fritz et al., 2007). However, plasticity is only observed for a feature that is attended to (Keuroghlian & Knudsen, 2007). Kauraäki et al., 2007 found that selective attention also enhances auditory cortex frequency selectivity in humans.

However, attention can also reduce neural spike rate: Schwartz and David, 2018 trained ferrets to perform a tone in noise detection task. Ferrets had to switch attention between several target frequencies within the same stimuli. In addition, target salience was manipulated so that

attention was better observed. Unexpectedly, they observed suppression of neuronal responses (spiking rate) not of all distractor frequencies, but mostly near the target frequency. According to the authors, this may be to suppress non informative noise around the target frequency.

Interesting changes at neuronal ensemble level have also been observed in humans. Holdgraf et al., 2016 found STRF changes of neural ensembles that followed improved intelligibility in an ECoG study with distorted speech. Stimuli were created by ltering out a portion of the modulation power spectrum of each utterance. Listeners received distorted, clear and again distorted version of the same utterance. STRF changes are observed between the rst and the second presentation of the degraded utterance. As this was coupled with a behavioral task, the researchers could con rm that STRF changes correlated with intelligibility. It seems that STRFs may become more responsive to the features of the clear version of the utterance. As such, this is either a direct consequence of memory or of successful cue reweighting (or both).

Further evidence for neural changes due to attention were also found by Khalighinejad et al., 2019. The authors recorded auditory cortex of human listeners who listened to speech with changing background noises. They report that adaptation to noise is independent of attention (the listeners were either doing a visual task or not), but that acoustic features of noise are suppressed neurally, which results in enhanced representation of acoustic features of speech.

The changes in neuronal encodings are suggestive of the fact that top down processes, such as selective attention, can in uence the way speech is ltered or processed: some studies show enhancement, other suppression, and both seem to be re ecting the facilitating effect in the animal or human doing the task. Much more remains to be explored, especially in humans, and especially with experiments that can reveal something about the general plasticity of the neurons in auditory cortex.

3.5 Conclusion

In this chapter, we have seen some evidence from behavioral research about quick cue reweighing and neural research that attention can modulate neural representations. This is linked into a working hypothesis rst set forth by Holt et al., 2018 that exible cue weighting is facilitated by selective attention.

First, we have seen that psycholinguistic experiments show quick cue re-weighting after short term exposure to non typically correlated acoustic-phonetic features. This is puzzling since humans show different cue weighting at different moments in time, and prior literature offering computational models do not have a good explanation as to how this is done.

Additionally, noise may inform listeners so that they weight they cues differently: a primary cue in quiet that may be masked by noise is now secondary, and a secondary cue now becomes primary. This quick and incredible plasticity is also discussed in the context of distorted speech. Speech that is corrupted in some way can become intelligible with either more exposure to it. However, perceptual adaptation is even faster with disambiguating information, presented either acoustically right after the distorted speech, or visually in a written form. Lexical knowledge and/or memory of the clear speech utterance then can help listeners adjust to speech cues that are completely different than in standard speech perception. In fact, some may be missing or be corrupted to a degree to be completely unreliable.

To explain these phenomena, Holt et al., 2018 proposed a cognitive mechanism called selective attention. This process may enable rapid cue readjustments seen in a lot of literature. However, it is not clear whether selective attention can operate on the level of feature representations in the brain. To this end, studies on auditory objects have been presented, which elucidate

how perceptual (auditory) events are represented in the brain. Furthermore, we have seen how the representation of the auditory object and all of its features can be neurally enhanced by attention. It remains to be tested whether attention (or another mechanism) actually performs cue weighting. However, in this thesis, we use this mechanism as the driving force of exible cue weighting that can change on a moment to moment basis.

We also examined neural recordings that can be expressed as spectro-temporal receptive elds in animals and humans. We have seen that there is some evidence for acoustic-phonetic and for phonological features. Additionally, we have seen that neurons responding most to speci c features can change in their tuning depending on the task and that there is correlational behavioral facilitation that happens with it (i.e., humans hearing clear version of degraded speech gained intelligibility).

To conclude, these ndings in neuro-biology are suggestive of plasticity that we see in human behavior. It remains to be explained how exactly this plasticity is manifested on neuronal level in humans and what that means for general speech perception. In the following chapters (5, 6), we will see a computational level (Marr, 1982) explanation that is inspired by both behavioral and neural ndings described in this chapter: a computational exploration of cue weighting done with speech corpora, based on information theoretic framework.

Chapter 4: Rate Distortion Theory

So far, we have seen evidence that people weight cues in speech perception, and that they are capable of quick adaptation in form of cue reweighting, depending on the listening situation. Prior models (Kleinschmidt & Jaeger, 2015; Toscano & McMurray, 2010) cannot seem to account for this, therefore in this chapter, we will see a framework (rate distortion theory) in which a new computational model will be proposed in chapters 5 and 6. This is a framework that comes from information theory rst introduced by Shannon, 1959 and then developed by Berger, 1971 and has been applied to various domains of cognitive science, i.e. Sims, 2016. It can be seen as an extension of ef cient coding (Barlow, 1961), which has been very in uential in understanding neural sensory encoding.

The unique power of rate distortion theory lies in acting as a channel (with an encoder and a decoder) that can compress the input information in a goal directed way, while this compression and any potential perceptual errors are characterized by informational constraints imposed on the channel. As such, the internal noise of the channel is simply a consequence of reducing information size of the input. Nonetheless, when the channel is optimized towards some goal, it can be successfully trained to balance extracting general properties of the world, while minimizing perceptual errors.

This chapter describes the workings of rate distortion theory framework that is part of

information theory. It explains the mechanisms of rate distortion theory in mathematical detail, as well as of a variational auto-encoder, its possible implementation. In addition, this chapter talks about why it is important to consider rate distortion theory as a model of speech perception and summarizes prior models of other cognitive mechanisms built within this framework.

Finally and importantly, this chapter serves as a description of a framework for computational modeling of cue weighting used in this thesis: a neural network (beta variational auto-encoder) implementing rate distortion theory with two adjustments: rst, an added decision model which allows the listener to map speech to a category and is implemented in all of our cue weighting models. With this base model we will ask whether optimal cue weighting depends purely on long term statistics (simulated in chapter 5). Second, an added parameter that can bias the model to be more "attentive" towards a speci c cue that can mimic humans more accurately than cue weighting based solely on long term statistics, as also described in in chapter 5. Finally, this model will serve as a base to a new exible cue weighting model that can change cue reliance exibly with selective attention mechanism, once we expand the number of encoders as simulated in chapter 6.

4.1 Introduction To RDT

Rate distortion theory (RDT) (Berger, 1971; Shannon, 1959; Sims, 2016) can represent an optimal perception system given speci c data and a speci c goal. In speech perception, the goal in question is to map speech sounds to some abstract representations. These serve the listener as an underlying representations for later language processing and language comprehension. In chapter 2 we have seen potential models of optimal speech perception, speci cally of models that

use speech cues/features as part of their mechanism. The goal or prior computational modeling (broadly speaking) is to show how listener can map signal to category via cues (for details, see section 2.3). Here, we will look at how cue weighting can be implemented with RDT, and why it is interesting to consider RDT as a cognitive model.

RDT comes from information theory and represents a channel that is optimized with a trade-off: on the one hand, it deals with mapping continuous signals to a representation with reduced information rate, therefore some loss of information is inevitable. On the other hand, it must extract and capture the general properties of the world. RDT has been used for various purposes but here, we will consider it an optimal model of speech perception. The incoming signal is speech and RDT's extracted representations will be useful statistics about properties of the speech.

An RDT channel is completely probabilistic, it shares some commonalities with probabilistic inference models and Bayesian models (Sims, 2018). However, the RDT framework gives speci c interpretation of the limited cognitive processing (Sims, 2018): RDT explains “internal noise” as due to sensory processing limitations. The internal noise is a term often used in various ways in the wide literature of cognitive science, neuroscience, psychology etc., and seems to often denote the discrepancies in measured biological (and also sometimes called inef- cient) processing that can have multiple sources (Faisal et al., 2008). In RDT, internal noise is explained as the discrepancies that arise due to the processing under constraints, which could be of various origins, such as not having unlimited time, energy or other resources necessary. For us, internal noise will mean the inevitable loss of some of the information in processing speech, while keeping the information about cues that help in mapping speech to category.

Additionally, RDT can be thought as an extension of the theory of ef cient coding that

comes from neural information processing (Barlow, 1961) which describes that a goal of neural processing is to extract ef cient codes of the highly redundant sensory input that is of particular signi cance to the organism. Organisms may want to solve perceptual problems which are beyond simplifying redundancy and RDT allows for a variety of cost functions that can model a variety of objectives. As such, RDT appears to be a versatile framework for modeling a wide range of mechanistic underpinnings speci c to a particular cognitive process.

Finally, RDT can also explain Shepard's universal law of generalization (Sims, 2018). This law states that stimuli are represented in a psychological space where they are separated by some distance. The generalization is easier made on stimuli that are closer together than further apart. RDT can account for that by describing a mechanism geared toward a speci c task while operating under constraints. This, as well as application of RDT in attentional allocation in vision (Bates & Jacobs, 2021), as well as in memory (Bates et al., 2019) give a good reason to apply RDT framework also on speech perception mechanisms. The constraints in speech perception are multiple, and this thesis will give an account on a speci c one, which is the variability of the signal to which the listener must constantly be adapting.

As such, mapping signal to perceptual representations is far from being a simple task, since speech sounds mapped to the same learned representation can vary to a large extent, due to the speaker's properties, and/or the environmental noises. To address this particular problem, we will rst look at the theory of rate distorion theory and then expand upon it in chapters 5 and 6.

Figure 4.1:RDT (illustrative, after Sims, 2016) - RDT channel has an encoder and a decoder. It is optimized both on rate (how informative the latent information is) and distortion (how many perceptual errors there is between input and output). Internal noise comes from information constrained capacity.

4.2 Theoretical framework of Rate Distortion Theory

RDT (Berger, 1971; Shannon, 1959; Sims, 2016) is a theory about an optimal solution to the trade-off of transmitting information with as little error as possible, while obtaining good generalization about the input. In speech perception, we can afford to lose some information when receiving the speech if that does not interfere with achieving the goal of mapping speech to representations. This goal can be achieved if the channel extracts informative representations about the general statistics of the data, while making as few perceptual errors as possible. RDT achieves that with a mathematical framework that operates under information processing constraints, while maximizing information transmission that optimizes the pre-de ned goal.

The channel optimization is mathematically formalized as (Sims, 2016):

$$Q = \text{argmin}_Q(D_Q) \quad \text{subject to} \quad I_Q(x;y) \quad C \tag{4.1}$$

where Q denotes the optimal channel (built out of an encoder and decoder), which minimizes

the channel distortion denoted as D. This is subject to the fact that channel's information capacity is at most C, which stands for information processing constraint. It is measured by the mutual information I_Q, which is a measure of the entropy (or the uncertainty) between the channel's (Q) output y and the entropy of the channel's input x. The entropy will be low if the channel's output is informative, i.e. if the channel has learned anything useful about the input.

Mutual information is defined as:

$$I(x;y) = H(x) - H(x|y) = \sum_i \sum_j p(y_j|x_i)p(x_i)\log\frac{p(y_j|x_i)}{p(y_j)} \tag{4.2}$$

where $H(x)$ is defined as the entropy of input x and represents the average surprise or uncertainty of each possible value of x:

$$H(x) = \sum_i p(x_i)\log p(x_i) \tag{4.3}$$

And $H(x|y)$ is a conditional entropy:

$$H(x|y) = \sum_i \sum_j p(y_j)p(x_i|y_j)\log p(x_i|y_j) \tag{4.4}$$

Thus, mutual information is a measure of mutual dependence between the source (or input) x and the channel produced outcome y. It tells us the amount of information obtained about y by observing x. The conditional entropy $H(x|y)$ expresses the uncertainty about x after y is known and will be small if our channel successfully learned relationship between y and x. Inversely, $H(x|y)$ will be high if there is still more information needed to explain x after knowing y.

Distortion D is defined as a product between the channel's output given some input, the

probability of this input and some pre-de ned loss function:

$$D = \sum_i \sum_j p(y_j | x_i) p(x_i) L(x_i; y_j) \tag{4.5}$$

This means that the distortion will be low if a particular channel $p(y_j | x_i)$ nds as output a y_j such that it highly likely given a speci c x_i. This is weighted by how likely is a speci c input x_i to occur (i.e. is or is not a rare event). If it is likely, then the distortion will be higher, which leads the channel to better represent more likely (or frequent) events.

The difference between the output and input is measured with the loss $L(x, y)$. This can be chosen depending on the goal and can just be a squared difference between the input and the output for our purposes:

$$L(x_i; y_j) = (x_i - y_j)^2 \tag{4.6}$$

Such a loss compares input x and output y and measures the errors depending on how far y is from x.

A model built like this is often computationally intractable, especially for raw, higher dimensional data, such as speech. Fortunately, there exists a mathematical approximation of RDT, called a beta variational auto-encoder (Alemi et al., 2018; Kingma & Welling, 2013). This model can be built using neural networks, which allow for treating corpus sized real data.

4.3 Beta Variational Auto-Encoder: Mathematical Implementation of RDT

Beta variational auto-encoder-VAE) (Kingma & Welling, 2013) is a probabilistic, generative model that is capable of unsupervised representation learning, with information-theoretic properties. It approximates RDT (Alemi et al., 2018), while it is able to treat large data. This is especially attractive for cognitive modeling, since with this kind of model, we can use raw, corpus sized amount of data. Below, I provide an overview on mathematical framework of VAE and explanations on how it is connected to RDT.

-VAE is a channel, built out of an encoder $e(z|x)$ and a decoder $d(y|z)$. The encoder and decoder represent the optimal channel Q in RDT terms described in previous section 4.2.

The encoder $e(z|x)$ receives input x and converts it into a hidden (or latent) representation z. The decoder takes this z and tries to reconstruct the x. To differentiate between the input and the output, we denote the reconstructed as y.

The measure about how much one variable contains about the other is given again by the mutual information. This is computed in the same way as in RDT equation 4.2 of mutual information in section 4.2. For simplicity, this equation is repeated here with mathematical equivalent used in beta-VAE:

$$
\begin{aligned}
I(x|y) &= H(x) - H(x|y) \\
&= \sum_i \sum_j p(y_j|x_i)p(x_i)\log\frac{p(y_j|x_i)}{p(y_j)} \\
&= \sum_i \sum_j p(x_i;y_j)\log\frac{p(x_i;y_j)}{p(x_i)p(y_j)}
\end{aligned}
\tag{4.7}
$$

However, we now compute mutual information between the input x and the latent repre-

sentation z, whereas RDT equation of mutual information was given simply between input x and output y.

This gives mutual information in continuous space beta-VAE (Alemi et al., 2018) in equation 4.8:

$$I(X;Z) = \int\int dxdz\, p(x;z) \log \frac{p(x;z)}{p(x)p(z)} \tag{4.8}$$

Where p depends on the choice of the encoder's parameters, and p depends on the data distribution of the input, namely $p(x)$ denotes true input data probability density with parameters . The true $p(x)$ is not available to us, therefore we use stochastic approximation to estimate the marginal likelihood $p(x) = \int p(z)p(x|z)$.

We want Z (the latent representations) to be a meaningful representation of the X input. However, if the encoder is merely an identity function, then $X = Z$ and the mutual information is the entropy of the data, $H(X)$. If the X and Z are completely independent, then the mutual information is zero, since X and Z have no relationship. What we want is that X and Z have some relationship, but not such that we obtain an identity function.

Alemi et al., 2018 have provided mathematical proofs for the variational lower and upper bounds of the mutual information:

$$H - D \leq I(X;Z) \leq R \tag{4.9}$$

Here, H denotes the input data entropy, D is the distortion (reconstruction accuracy) of the channel and R is the rate (compression accuracy). H depends on the data but is not changeable, so we can treat it as a constant. D in VAE can be computed as as a reconstruction error, while R

depends only on how well the encoder approximates the variational marginal distribution $p(z)$.

The rate R comes from the encoder $e(z|x)$ trying to approach $p(z)$, which is measured as an averaged KL divergence (or a "distance" metric about how well encoding distribution probability distribution $e(z|x)$ approximates the marginal distribution):

$$R = D_{KL}(e(z|x)||p(z)) \tag{4.10}$$

The distortion D can be computed with a variety of cost functions and for our purposes, a mean squared error is suf cient:

$$D = (x - y)^2 \tag{4.11}$$

If we have zero distortion, $D = 0$, then the channel will perfectly reconstruct the input, however the rate will at minimum be equal to H, the entropy of the input data. If we have zero rate, $R = 0$, then the encoding distribution will just be identical to the marginal $e(z|x) = p(z)$, which means that the encoding distribution is independent of x. In this case, the latent information is not encoding anything useful about the input.

The β parameter will enable to go more towards one or the other extreme: one where the distortion is low but the rate is high ($\beta < 1$), i.e. latent representations retaining lots of details of data, and the other, where the distortion is high but the rate is low ($\beta > 1$) (Alemi et al., 2018), i.e. more compressed or abstract representations.

The complete objective for training β-VAE is:

$$L(; ; x) = R + D = D_{KL} e (zjx)jjp (z) + (x y)^2 \quad (4.12)$$

4.4 Cognitive Models Using Rate Distortion Theory

The present thesis is not the rst nor the only implementation of RDT in cognition. Although Bayesian models used for predictive models in inference under uncertainty have been more common, RDT has been used in several domains to model goal directed ef cient coding: in visual attention (Bates & Jacobs, 2021), memory (Bates et al., 2019; Jakob & Gershman, 2023; Sims, 2016), pragmatic reasoning (Zaslavsky et al., 2020), action-centric representations (De Llanza Varona et al., 2023) and melody learning (Zhou et al., 2024). In fact, if we considered ef cient coding (Barlow, 1961) as well, there are many studies using it to model compressed neuronal encoding of sensory signals, such as Botvinick et al., 2015; Louie and Glimcher, 2012; Lu et al., 2019; Parker and Newsome, 1998; Urai et al., 2022. Thus, the idea of encoding sensory signals in an economic way (and potentially in a goal directed way) is not a foreign concept to cognitive science.

The work of this thesis is inspired by two particular RDT implementations, those of Bates et al., 2019 and Bates and Jacobs, 2021. First, Bates et al., 2019 added a decision model using latent variable as the input. This will be useful also for us to model mapping speech to category space.

Bates et al., 2019 did this by modeling ef cient memory compression with RDT. They add a separate decision model to the base VAE, with the latent code z coming from -VAE as an input to the decision model. The objective function of this model is just a sum of -VAE loss

with the decision model's loss (which in their case, was just a mean squared error). We adopt this approach to be able to model mapping extracted latent information to a category. That is, the decision model for us will serve to model adult speakers who can map speech to a category through cues.

Another important adjustment that has been made relevant to the work in this thesis is an added parameter that can simulate attention in prompted visual search (Bates & Jacobs, 2021). This parameter, which Bates and Jacobs, 2021 called λ, scales the distortion part of the loss:

$$L(x; y; C) = \sum_i \lambda_i (x_i - y_i)^2 \tag{4.13}$$

The C represented the prompt, which they call cue, but this is not to be confused with the way speech cue is described here. In visual search, a cue is a prompt the participant receives during the experiment. This prompt had 80% chance of being reliable and it indicates where on the screen the target could appear (left or right side). As such, it is a prompt for visual attention. The participant then simply has to respond whether there is a target present shortly after on the screen (which is a whiter blob in a noisy visual scene, with varying degrees of signal to noise ratio). This attention parameter enabled Bates and Jacobs, 2021 to train the model to represent a participant in this situation – the model was penalized more for poorly reconstructing the prompted side of the screen that should be better attended.

In the present work, shown in chapters 5 and 6, we borrow this parameter for another purpose: cue weighting. It will allow us to penalize the model more for poorly reconstructing one cue over the other cue and to experimentally look at how that mimics human listeners. This will allow us to bias the model to be more “attentive” to that cue.

4.5 Conclusion

In this chapter, rate distortion theory framework was presented and described mathematically. It is a framework coming from information theory that describes a channel that is trained on balancing a trade off objective: optimizing distortion (reconstruction error) and rate (information content). This framework can be implemented with a neural network called variational auto-encoder. The objective of this network also contains a rate and distortion terms, but this model can treat raw data with full dimensionality.

Other implementations of this framework within cognitive science relevant to this thesis can be summarized with two important extensions to the base architecture of VAE: an additional model that takes as input the latent VAE representation and does something with it, as well as an "attention" parameter λ that lets the experimenter control what the VAE reconstructs better or worse, with the goal being on controlling the behavior of the model.

These two adjustments will be used in the present thesis to model two important aspects of cue weighting: one, added decision model will simulate adult speakers mapping speech to a category. Two, a cue weighting parameter λ will allow us to penalize the model more for poorly reconstructing one of the cues, thus biasing it towards being more "attentive" to that cue. This is going to serve as a base cue weighting model with which we test whether cue weighting depends solely on the input data (without the added parameter λ) or on biased cue weighting (with the added parameter λ) (in the following chapter 5). I also use this model as a base to an expanded model that simulates selective attention process for quick cue reweighting in chapter 6.

Chapter 5: Focus: Computational Evidence That Cue Weighting Requires More Than Long Term Statistics

Prior research often assumed that humans either entirely or partially rely on cues in speech perception because of how reliable these cues are when they are observed in their long term statistical distributions. Much research has also shown that listeners typically rely more on one cue than another when categorizing speech sounds, as described in chapters 2 and 3. Here I ask whether observations of the long term statistics really are indicative of how reliable the cues are like presupposed by a prior computational model by Toscano and McMurray, 2010.

In this chapter, I present a rate distortion theory model of feature weighting and use it to ask whether human listeners select feature weights simply by mirroring the feature reliabilities that are present in their input. I experiment with an additional component (that for simplicity will be named cue weighting parameter) that listeners appear to use that is not re ected by the input statistics. This is important because prior computational modeling (Toscano & McMurray, 2010) explicitly modeled cue weighting as based on reliability coming from distribution of the data.

I rst present the corpora that will simulate possible long term statistical distributions. Then, I present various simulations with the rate distortion theory model, both on extracted speech cues, as well as on actual speech signal.

Second, I show that cue weights can change to be more human like when we add a pa-

rameter to the model that preferentially guides model learning. This suggests that an internal mechanism is at play in governing listeners' weighting of different aspects of the speech signal, in addition to tracking statistics.

These explorations will rst serve as a step to a more exible model – we will see that the kind of model presented in this chapter cannot exibly change cue weighting, and as such, a novel solution to that is presented in chapter 6. And second, they will lead us to investigate cue weights important to humans in a bigger state-of-the-art speech to text model (that is not part of rate distortion theory models), with the goal of seeing whether a bigger model exhibits any exibility (in next chapter 7).

5.1 Introduction

In speech perception, listeners rely on different aspects of the signal to map it to categories. In chapter 2 we have seen that listeners like to rely more on some features than on others. This is often thought (for example by Toscano and McMurray, 2010) to be an effect of observing the world: if the listener tracks statistics of the acoustics they have heard in their lives, some features may be more reliable than others simply by being easier to track.

The relevance, according to Toscano and McMurray, 2010, is in the fact that cues enable separation between categories. If cue values of two categories cluster close together (for example, because the means of their distributions are close), they should be less reliable. If cue values are very dispersed (i.e., the variance of the distribution is high), this will also reduce the reliability of that cue. For more clarity on reliability based on separability of distributions, see gure 5.1. For an assumed distribution of two cues by Toscano and McMurray, 2010 where one cue is more

reliable than another, see gure 5.2.

Figure 5.1: Cue Reliability Based On Statistics (Illustrative) - If means are further apart or the variances are smaller, there is more separability on that dimension (as in A and B). Inversely, means closer together or more variance introduce more confusion on which distribution the black dots belongs to (C and D).

However, speech used in Toscano and McMurray, 2010 is carefully articulated speech of four talkers. In addition, studies that test cue weighting in humans and establish that a cue is primary also presuppose nice and clean distributions for a single voice (Idemaru & Holt, 2014; Lehet & Holt, 2020; Liu & Holt, 2015; Wu & Holt, 2022; X. Zhang et al., 2021; Zhou et al., 2024).

This seems like a natural extrapolation on the observed behavior and on some carefully selected data. However, real speech is never so nicely distributed (Hillenbrand et al., 1995).

This motivated the rst part of the following simulations, where I ask how much can data statistics predict cue reliability, when data statistics is taken out of a real speech corpus. Speech corpora typically contain multiple speakers that vary in their pronunciations, and hence also in

Figure 5.2:Assumed Cue Distributions (Illustrative) - Here we see assumed cue distributions that cause cue 2 to be more reliable than cue 1. As we can see, cue 2 shows good separability as indicated with the green dotted line (that is, if all of the decision would be based on the dimension of cue 2. Meanwhile, cue 1 shows overlap between categories, and as such becomes more uncertain as a perceptual boundary. We can see this with the illustrative blue dotted line: parts of category 1 and 2 would be categorized as the opposite category if the decision were based entirely on cue 1.

their cue distributions. This is relevant because the cue distributions on a group of speakers may not be as easily separable. Furthermore, speech corpora also contain read or conversational speech that is a lot more natural than laboratory produced syllables or short words.

To address this question, a rate distortion theory model was developed and trained on three different corpora. This serves as an optimal cue weighting model based only on statistics of the data it was trained on. It allows us to see how the cues are optimally weighted for each corpus the model was trained on. As we will see, the cue weights thus obtained do not match humans entirely.

5.2 The Simulated Phenomenon

In the following simulations we focus on the question of whether cue weights respond uniquely to observed long term statistics. To evaluate whether they do, we compare our models

to human cue weights as observed in quiet from Wu and Holt, 2022. This study has shown that humans rely more on VOT to identify [p] and [b] than they do on F0, and they rely more on spectral information than on duration to identify [ɛ] and [æ].

Both cases for vowels and consonants share the study design where two cues are manipulated between two categories. In addition, human listeners rely more on the primary cue than on the secondary cue. Although both cases are comparable on the experimental design, in this chapter I only show simulations with vowels, and this is because pitch is very variable across utterances and it was very challenging to normalize it properly. To be speci c, extracted pitch did not show a bimodal distribution where each distribution depended on one category, but rather bimodal distribution already within the cue. As such, it was hard to interpret, given prior assumptions about bimodal pitch distribution (one distribution per category). Henceforth we will be looking at comparisons between models and humans in identi cation of [ɛ] and [æ].

An illustrative human cue weighting is reproduced in gure 5.3. For more details on the design of the study, see chapter 3, section 3.2.

Figure 5.3: Human Cue Weights (Illustrative) - Here we see reliance of cue illustratively reproduced from (Wu & Holt, 2022). Listners rely more on spectral information than on duration when identifying [ɛ] and [æ].

First, we look at a model that only depends on data statistics and we consider it an optimal cue weighting model, given speci c data. We show that this does not mimic human behavior well overall. To mimic human behavior better, we propose a model adjustment with a cue weighting parameter called α after Bates and Jacobs, 2021 is added to guide learning. In this way, the model's learning is biased to be more “attentive” towards a speci c feature. This makes the model learn differently than just from the data, and the resulting cue weights resemble human behavior better. We will see that with such a simulation we obtain biased cue weighting, geared towards a primary cue. This indicates that an internal mechanism can drive cue weighting that are observed in quiet speech, not only cue statistics. This will also serve as a necessary step to later introduce a exible cue weighting model implementing selective attention in chapter 6.

5.3 Model

To mimic cue weighting, we use a rate distortion theory framework. This framework represents a channel that is trained with a trade off: on the one hand, the channel receives the input and reconstructs it. As such, it tries to minimize the distortion or errors in reconstruction. On the other hand, the channel must reduce the dimension of the data. To do so ef ciently, it tries to develop parsimonious latent representations. One can train such a model to be better at one of these two extremes by adjusting the parameters.

As has been explained in chapter 4, rate distortion theory offers a possible explanation of perceptual encodings of sensory signal, where the nature of encodings can depend on the chosen goal. For example, the goal could be to reduce redundancy like in ef cient coding (Barlow, 1961), or it could be to encode signals as faithfully as possible while being constrained in information

processing.

Our model is implemented as a beta variational auto-encoder (β-VAE), which is a neural network model. Its loss can be understood in terms of rate and distortion. Specifically, β-VAE is a probabilistic generative model that consists of an encoder and a decoder. These two components are trained jointly. The encoder receives the input in its original dimensions and then reduces its dimensions. From there, a latent representation is extracted (to be precise, it is in fact sampled from the encoder's distribution). The latent representation serves as input to the decoder, which then reconstructs the input.

Specifically, the base β-VAE we use has the following loss (repeated from equation 5.2 from chapter 4):

$$L(\ ;\ ;x) = \ R + D = \ D_{KL}\ \ e(z_i|x_i)||p(z_i)\ \ + (x_i\ \ y_i)^2 \qquad (5.1)$$

The architecture as such will allow us to receive the input, extract latent representations from it and reconstruct it. However, since adult listeners extract features so that they can identify what they are hearing, another model is added.

This model, that we call category model, is designed to model listener's mapping of signal via speech features to a phonological category. It takes as input the latent representation coming from the encoder and categorizes the input based on that. Since the adult listener already have developed their perceptual space, this model is learning in a supervised way: it is trained with a pair of latent representation from the β-VAE and a ground truth label. The objective function used for learning is binary cross entropy. In addition, this model is trained jointly with the β-VAE.

The total loss for our model is then:

$$L(\ ;\ ;x) \quad D_{KL}\ \ e(z_i|x_i)||p(z_i) \quad + (x_i \quad y_i)^2 + BCE(cat(z_i); l_i) \tag{5.2}$$

where cat denotes the category model, l_i is the i-th label associated with the input x_i and BCE is binary cross entropy loss, which can be calculated as:

$$L_{BCE} = \frac{1}{N} \sum_{i=1}^{N} l_i \log(p_i) + (1 \quad l_i)\log(1 \quad p_i) \tag{5.3}$$

Here, N is the number of observations in the batched input and p_i is the probability of the i-th observation being in class 1.

With this final objective, we can find out what cue weights are when the model is simply trained on data, without any interventions on our side. Therefore, our hypothesis is that if the cue weights depend on long term statistics, this kind of a model should show cue weighting that resembles human behavior.

Each model used in our simulations had convolutional layers with rectified linear activation. The number of layers and their dimensions depended on the size of the input, but all convolutional layers had convolution window of size 3 and step of size 2.

5.4 Long Term Distributions: Corpora

To investigate how are cues weighted with realistic input, we simulate that with data coming from speech corpora. These corpora serve as an approximation of how the observed long term statistics can look like because they contain lots of speakers talking in a more ecological scenarios (read or conversational speech, not necessarily carefully slow articulated speech).

In the present simulations we focused on three corpora that contain American English speech: TIMIT (Garofolo et al., 1993), Buckeye (Pitt et al., 2005) and Wall Street Journal (WSJ) corpora (Paul & Baker, 1992).

In the ascending order of the size of the corpora:

1. TIMIT corpus (Garofolo et al., 1993) is a carefully constructed corpus, initially for automatic speech recognition, built by MIT, SRI International and Texas Instruments. It containts about ve hours of speech from two genders (70 % male and 30 % female) and about 630 speakers of 8 major dialects of American English. Each of the speakers read approximately 10 sentences. Its transcriptions are time aligned and have been done so manually, so this is the corpus with the most precise alignments in our study. The transcriptions are on word and phonetic level.

2. Buckeye corpus (Pitt et al., 2005) is a corpus of completely conversational speech, recorded by researchers at Ohio State University. Approximately 40 speakers (half male, half female) from or near Columbus, Ohio freely conversate with the interviewer (either male or female). Transcriptions were done both on word and phonetic level. Phonetic alignments were rst done automatically and then hand corrected. Buckeye contains about 40 hours of speech total, with male and female speakers. Speakers are largely from middle and upper class.

3. WSJ corpus (Paul & Baker, 1992) is a corpus built by DARPA and contains approximately 141 hours of speech, where approximately half of speakers are male, and half female. While there was some spontaneous speech included, mostly this corpus consists of read speech. Speakers were reading from texts extracted from Wall Street Journal news. This

corpus has machine aligned ortographic and phonetic information.

5.4.1 Input Data Pre-Processing

To mimic humans we look at speci c cues, namely duration and the rst formant (F1). When constructing the stimuli interpolations, Wu and Holt, 2022 used the rst ve formants and used that as a single cue which they called spectral quality cue. However, for simplicity we only use F1, which shows the most difference between the two categories. This does not intend to say that the other formants show no category separability, but according to our investigations the differences between formants beyond F1 were marginal.

To prepare the data for modeling, all of the data was rst downsampled to 16000Hz. After this, we either:

1. extract features (rst formant and duration) from the speech signal for vowels,
2. transform speech into lterbanks,
3. transform speech into MFCCs, or
4. transform speech into articulatory features.

The rst step was done not only to visualize distributions of formant and duration values, but also as input to the model (see section 5.6.3). This kind of input allowed us to change the objective of the model and to manipulate more human like cue weights.

F1 and duration were extracted with the help of phonetic alignments that each of these corpora had. F1 was extracted as a median of all F1 values found between the beginning and the end of the vowel with Praat in Python. Some computational errors are expected since Praat

occasionally returns a wrong value for the rst formant, depending on how clear the signal is. Duration was obtained by simply calculating the difference between when the vowel started and ended. Therefore, if there are any misalignments, our errors re ected that. Furthermore, for machine aligned corpora (Buckeye and WSJ) time stamps were always given only up to fourth decimal, therefore our durations are also cropped to the fourth decimal. Finally, we excluded any vowel with an F1 greater than 1200Hz or a duration less than 30ms to account for any misalignment errors.

Furthermore, each value was then normalized per speaker with z-scoring (similar to Toscano and McMurray, 2010). Z-scoring is done by looking at all values of for examples F1 and nding their mean and standard deviation. From every value, the mean is subtracted and everything is then divided by the standard deviation. Z-scoring assumes a normal distribution, which is likely an incorrect assumption, but it scales all the numbers to uniform levels and removes of some of the speaker information.

Figures 5.4, 5.5 and 5.6 show how the extracted values of F1 and duration look for TIMIT, Buckeye and WSJ respectively.

It is easy to notice that none of these distributions look quite as nicely separated as assumed. In fact, TIMIT may have the most separable categories with less overlap, which may be due to careful construction of the corpus – the other two corpora are recordings of more naturalistic speech and hence have more overlapping categories.

Figure 5.4: TIMIT: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker).

5.4.2 Evaluation Stimuli

At test time, we do not use the experimental stimuli from Wu and Holt, 2022 to evaluate the models. This is because both their duration and F1 are exaggerated and fall far from the distributions of duration and F1 of each of the corpora, even when normalized. We tested models' categorization along the entire span of possible combinations of duration and F1 for a speci c corpus, with a step of 0.1 when stimuli were simply extracted as F1 and duration. For stimuli transformed with one of the signal processing techniques, we use 10% of the data that covers different productions of both vowels. For each stimulus we extract its F1 and duration which we normalize by z-scoring per speaker. This gives us different steps of duration and F1 on which we can test the cue weights.

Figure 5.5: Buckeye: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker).

5.5 Evaluation: Cue Weights

To obtain perceptual feature weights from our models, we follow the procedure in Wu and Holt, 2022: we use linear regression[1] with features as predictors and category as the dependent variable. When we use simple input (F1 and duration only), we tested models' categorization along the entire span of possible combinations of duration and F1 for a speci c corpus, with a step of 0.1. When we use full sized input, we test with withheld stimuli not used in training.

To obtain comparable human feature weights, we used human data and experimental stimuli from Wu and Holt, 2022. For each corpus, we rst z-scored the experimental stimuli using the

[1] Human behavior in identifying phonemes resembles logistic regression curve, not a linear one. However, since qualitatively we obtain similar results, we follow the procedure as used with humans in the original behavioral experiments.

Figure 5.6: Wall Street Journal: F1 and duration distributions – vowels [ɛ] in pink and [æ] in green on z-scored values of duration and F1 (normalized per speaker).

average mean and average variance of all female speakers of the corpus. We then used the step sizes in these z-scored duration and F1 values to recompute the human feature weights based on raw data from Wu and Holt, 2022.

5.6 Simulations

5.6.1 Acoustic Input

First, let us look at the base-VAE with acoustic features, namely with MFCCs and filterbanks. These representations have been used for decades in various computational treatments of speech. As described in section 5.4, for each corpus, vowels were extracted according to their time stamps from alingment documentation and then transformed separately into MFCCs and

Figure 5.7: Cue Weights (MFCCs)- Cue weights of [ɛ] and [æ] extracted from each corpus and then transformed into MFCCs. Cue weighting with these featural representations does not seem do mimic humans, but rather shows reverse cue weights for each of the three corpora.

into filterbanks.

Filterbanks were obtained computationally with 25 milisecond window, 10 milisecond step size and 26 filters for filterbanks. Mel Filter Cepstral Coefficients (MFCCs) were obtained with 30 milisecond window, 15 milisecond step size and 13 cepstral values for MFCCs. Since vowels have different durations, each vowel was padded with zeros to achieve maximum length of the longest vowel in the corpus. This was done because the model needed equal sized input.

The input to each model was a single corpus in one of the formats (filterbanks or MFCCs). Each vowel was used in conjunction to its label to train the category model in addition to the β-VAE. After a search it was established that β is best small to achieve learning, and specifically was $\beta = 0.001$ for all three corpora.

When we have trained the models, we simply test them with 10% of withheld data. To

Figure 5.8:Cue Weights (Filterbanks)- Cue weights of [ɛ] and [æ] extracted from each corpus and then transformed into lterbanks. Cue weighting with these featural representations does not seem do mimic humans, but rather shows reverse cue weights for each of the three corpora.

obtain cue weights we train 10 models per each corpus and follow cue weighting extraction as per Wu and Holt, 2022 (see section 5.5). Each model serves as a simulated subject so that we can compare models to humans.

In gure 5.7 we see results for speech extracted with MFCCs and in gure 5.8 we see results for speech extracted with lterbanks. None of the cue weighting thus obtained resembles human behavior: in fact, duration is uniformly weighted much higher than F1.

5.6.2 Articulatory Input

To check whether articulatory features will allow us to extract more human like cue weights, we extract articulatory features with model from Siriwardena et al., 2022 on TIMIT corpus. We then train 10 models per corpus. We x= 0:0025

Figure 5.9:Cue Weights (Filterbanks Vs. Articulatory) - Cue weights of [ɛ] and [æ] extracted from TIMIT and then transformed into lterbanks (called Acoustic on the gure) and articulatory features extracted from (Siriwardena et al., 2022).

Articulatory features were obtained from a model from Siriwardena et al., 2022. This is a multi-task learning based bidirectional Gated Recurrent Neural Network model that performs acoustic-to articulatory speech inversion. It learns from MFCCs and predicts six articulatory features: lip aperture, lip protrusion, tongue body constriction location, tongue body constriction degree, tongue tip constriction location and tongue tip constriction degree. These features were taken to experiment whether they would make cue weighting more human a model trained on classic acoustic features.

Results show that articulatory features do not exhibit human like cue weighting, but they at least do not show as extreme opposite cue weighting as acoustic representations (for comparison, see gure 5.9).

5.6.3 Low Dimension Input

Figure 5.10: Cue Weights (extracted feature values) Cue weights of [ɛ] and [æ], when first formant and duration are extracted from each corpus and used as model input. Cue weighting based solely on input data matches humans only for a model traines with WSJ corpus.

None of the models so far showed human trend in cue weighting for vowels presented in quiet. Specifically, duration is always upweighted in comparison to F1. Because VAE needs equal sized inputs and we pad the input with zeros to match the longest vowel, VAE may simply resort to compute the difference between where there is information in comparison to where it is not. This is only a hypothesis, because both duration and F1 separate the two vowels and we have no evidence about why the VAE separates categories as it does.

We next test how VAE weights cues when we use a very simple input: only F1 and duration. This input is too simple to be reduced much more since it is two dimensional, but it can tell us something about separability of F1 and duration distributions. This is also more similar

to simulations performed by Toscano and McMurray, 2010 who also used extracted cues from speech.

The extraction of cues is the same as described in section 5.4: we computationally obtain F1 and duration for each vowel, then z-score it per speaker to obtain normalized values.

We train each corpus with a separate model. The results can be seen in gure 5.10: only a model trained on WSJ performs with a human like trend of cue weighting (F1 weighted above duration). However, models trained either on Buckeye or TIMIT do not exhibit this trend.

5.6.4 Multiple Vowel Classi cation

Figure 5.11: Multiple Vowels: extracted feature values- Cue weights of [ɛ] and [æ], when we use all vowels of TIMIT during training. The input is simply the extracted and normalized rst formant and duration for each vowel.

To see whether having more vowels in our channel would help with human like feature weighting, we now expand the number of vowels in the training, as well as we expand our cat-

Figure 5.12: Multiple Vowels: MFCCs) - Cue weights of [E] and [æ], when we use all vowels of TIMIT during training. The input is all vowels of TIMIT transformed into MFCCs.

egory model. The rationale is that perhaps F1 dimension would be more separable for English vowels than duration.

For these simulations, we only train on TIMIT, because it has most carefully aligned data. In addition, it is smallest which speeds up simulations. We use all 19 vowels in TIMIT: [A], [I], [i:], [Q], [æ], [i], [1], [E], [I], [u], [U], [0], [@], [Ä], [Ç], [a1], [aU], [e1], [oU], [O1].

We train two different models with two types of input. First, we use extracted z-scored single z-scored values of F1 and duration for each corpus as in section 5.6.3. Second, we separately use speech transformed into MFCCs as in section 5.6.1.

The main difference (besides the input) is the expanded category model. We now expand the number of hidden layers (three instead of one) and the number of possible decision labels: we can now categorize the input into 19 different vowels. The total objective function is still the same as in equation 5.2, but instead of binary cross entropy we use categorical cross entropy,

which is a generalized case of binary cross entropy.

To see cue weighting results with multiple vowels, see gure 5.11 for simple input and gure 5.12 for MFCC input. For each of these two types of stimuli we see that duration is upweighted: slightly more for MFCC inputs than for extracted F1 and duration, but nonetheless none of them resembles human cue weighting.

Our optimal cue weighting model that depends only on data then does not seem to resemble humans, not even when we expand the number of categories. This leads us to introduce a new adjustment to the model, described in the following section.

5.7 Primary Cue Bias With A Cue Weighting Parameter

None of the previous simulations gave us cue weighting as we see in humans in Wu and Holt, 2022, no matter what the input format was. This suggests that with our RDT model, we cannot nd that data statistics alone can explain human cue weighting. Therefore, we experimented with input that only uses extracted F1 and duration.

To mimic human behavior better, a cue weighting parameter called after Bates and Jacobs, 2021 is added to guide learning. In this way, the model's learning is biased to be more “attentive” towards a speci c feature. This makes the model learn differently than just from the data, and the resulting cue weights resemble human behavior better.

This small sized input also allowed us to use a “forced” cue weighting parameter, because we can decide which part of the input should be more “important” to the model. In other words, we penalize the model more for poorly reconstructing one part of the input than another. With this, we can bias the model to “focus” more on one part of the stimulus. To do that, we scale the

Figure 5.13: Cue Weights (extracted feature values) Cue weights of [ɛ] and [æ], when first formant and duration are extracted from each corpus and used as model input. Cue weighting based solely on input data does not show the same trend as humans who largely upweight the first formant for TIMIT nor Buckeye. When we add the cue weighting parameter, we see that cue weighting mimics humans.

distortion part of the loss with a free parameter

$$D = \sum_i (x - y)^2 \tag{5.4}$$

The total new forced cue weighting objective then is:

$$L(\ ;\ ;x) \quad D_{KL}\ \ e(z_i|x_i)||p(z_i) \ + !_i(x_i \quad y_i)^2 + BCE(cat(z_i); l_i) \qquad (5.5)$$

Which is equal to equation 5.2, except that the distortion part of the loss is now scaled with ! .

With this we compare two types of models: one, that behaves optimally given the data, and the other, that is forced to learn one part of the input better, both trained on extracted F1 and duration data. To compare the learning of the two versions of models, we train each with a separate corpus so that in total we obtain six trained models (two per corpus). The input to the model are extracted F1 and durations that are z-scored as described in section 5.4. We predict that ! will force the models to learn the F1 better when we adjust it appropriately. To do so, we choose an extreme value, = [0:999; 0:001], where the rst value (0.999) is for scaling reconstruction error of F1, while the second value is for scaling reconstruction error of duration. The value for all these models was set to 0.0025.

Once we train our six models, we can see that now we see human like feature weigthing with ! for all three corpora (see gure 5.13. When we do not use , we see human like feature weighting only in WSJ corpus. This is interesting because it seems that if the listener found themselves in three different situations (conversational (i.e. Buckeye) vs. read speech (i.e. WSJ), for example), they would learn to weight cues differntly according to these simulations.

5.8 Conclusions

In this chapter, we have seen computational simulations of human cue weighting in the framework of rate distortion theory. To evaluate our models, we simulate cue weighting in humans as observed in Wu and Holt, 2022 and focus only on vowels [ɛ] and [æ].

To so so, we used real data from three different speech corpora of standard American English to train our models: TIMIT, Buckeye and WSJ. Every model was trained with a separate corpus. Our input was either simple and contained only extracted F1 and duration from each vowel, or was full vowel transformed into either lterbanks, MFCCs or articulatory features.

Our simulations show that models trained with fully dimensional input (lterbanks, MFCCs, articulatory features) do not exhibit human like feature weighting. With these simulations, cue weighting is actually inverse of human cue weighting.

When we use simple input (F1 and duration), we only see human like feature weighting for models trained on WSJ. However, when we add a cue weighting parameter to direct learning in our model, we observe more human like behavior.

This cue weighting parameter can be understood as forced attention to a particular cue for whichever reason, for example because that cue comes in rst temporally (i.e. F1 can be estimated before the entire vowel is produced). Since our model receives the entire input at once, ! can serve to mimic that.

Therefore, likely cue weights are not simply learned from statistics, not even in quiet, at least looking at our own simulations. To approximate human behavior across all corpora we need the cue weighting parameter. However, the model with ! has a fault: while it approximates human cue weighting for speech in quiet, it cannot simulate the changes in cue weighting seen

in accented or distorted or noisy speech adaptation. For this, we need to propose a model that incorporates selective attention or a mechanism that is capable of quick cue reweighting, which we do in chapter 6.

Chapter 6: Switch: Simulating Flexible Feature Weighting

In this chapter, we present simulations done within a rate distortion theory with some additions, namely an adjusted beta variational auto-encoder that can model listener's quick and effortless adaptation to new speakers and listening situations, all with a single model.

This model is built on work presented in chapter 5, where we introduced a cue weighting parameter λ. This parameter enabled us to capture human behavior of upweighting one cue more than another (primary cue) for speech in quiet. However, although that kind of model shows some internal mechanism in addition to tracking statistics, it cannot change its cue weighting, unless retrained.

To this end, we introduce further adjustments to the VAE model with β. The model's two important characteristics are multiple encoders that act as feature detectors and a weighting mechanism that implements selective attention mechanism. As such, the listener can quickly adapt on a moment to moment basis.

6.1 Introduction

Speech perception is an active process of extracting information from a highly complex and variable signal. When mapping the speech signal to phonetic categories, such as [p] or [b], listeners can make use of information that is spread across several aspects of the speech signal

(features) (Lisker, 1986). For example, one commonly used feature that distinguishes [p] and [b] in English is voice onset time (VOT), which denotes a difference between the burst and the start of voicing. Another feature signaling the [p] and [b] distinction is pitch (F0). In English, [b] has shorter VOT and lower F0, whereas English [p] has longer VOT and higher F0. In typical listening situations, listeners rely more on VOT than on F0 (Wu & Holt, 2022). Previous models have aimed to capture this unequal reliance on different speech features as optimal inference under uncertainty, and have hypothesized that long-term input statistics determine feature weighting (Kleinschmidt & Jaeger, 2015; Toscano & McMurray, 2010).

However, recent data suggest that listeners are extremely flexible and fast at reweighting features when encountering new speakers and new listening conditions (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Davis et al., 2005; Guediche et al., 2016; Idemaru & Holt, 2014, 2020; Lehet & Holt, 2020; Liu & Holt, 2015; Wu & Holt, 2022; H. Zhang et al., 2022). Specifically, Wu and Holt, 2022 showed that listeners primarily rely on VOT in a clear listening condition, but they rely primarily on F0 when VOT is obscured (noisy listening condition). In addition, after hearing miscorrelated features (i.e. VOT typical for [b] and F0 typical for [p]), listeners rely even more on their preferred feature (VOT in the clear condition, F0 in the noisy condition) and ignore the information given by the other feature.

Previous models cannot easily capture this pattern because they do not take into account that speech can be momentarily perturbed. Any flexibility in feature weighting in those models requires training a separate model for each listening condition. This clearly cannot account for cases where listeners quickly adapt to miscorrelated features (a set of input statistics that they have not previously experienced).

In this chapter, we introduce a neural architecture and show that it can facilitate rapid

changes in perceptual feature weighting. Our model instantiates the idea of selective attention (Holt et al., 2018), a flexible cognitive process prioritizing one feature over another that enables listeners to adapt to the speaker and situation without conscious control. Mathematically, our model is based on rate distortion theory (RDT) (Sims, 2016), implemented as a multiple encoder beta variational autoencoder (ME-β-VAE). We show how speech features can be weighted flexibly on a moment to moment basis within this neural architecture without needing to retrain the network for each individual situation. This flexible weighting allows the model to switch between different conditions and provides insight into how listeners can achieve the fast, flexible reweighting that has been observed empirically.

6.2 A model based on rate distortion theory

We propose a new neural architecture that can allow listeners to flexibly shift their attention. Our model is based on an idea from information theory known as rate distortion theory (RDT).

RDT is a probabilistic model of a system (often also called a channel) trying to maximize its performance with capacity constrained information processing (Barlow, 1961; Sims, 2016, 2018). Its objective is to minimize perceptual errors. Therefore, the encoding channel extracts as much information as possible that is relevant to the task: in this case, reconstructing the acoustics of the speech signal and mapping the speech signal to a category, like [p] or [b]. RDT assumes that the constrained capacity is the source of what is often called internal or sensory noise. The information that passes through the channel is chosen so that task performance is maximized, subject to the constraints on channel capacity.

Our neural architecture is based on the β-VAE (Kingma & Welling, 2013), which has

been shown mathematically to implement the RDT framework of ef cient information processing (Alemi et al., 2018). The -VAE is a probabilistic deep neural network model trained to optimize the loss containing the rate (forcing the encoder to learn meaningful latent representations) and the reconstruction terms (forcing the model to reconstruct the input as best possible):

$$L(\ ;\ ; x_i; y_i) = \quad D_{KL}(e(z_i|x_i)||p(z_i)) + MSE(x_i; y_i) \tag{6.1}$$

where bolded symbols are vectors and acronyms denote speci c losses. In particular, is the input, is the output, is the latent information found at the end of the encoder. Losses: denotes KL divergence, MSE denotes mean square error, are parameters of the probability distributions; $e()$; $p()$ denote the probability distributions, where $e()$ is an approximation of $p()$. is a parameter scaling the KL divergence proportionally to the MSE loss.

Our -VAE incorporates several advances relative to the original architecture from Kingma and Welling, 2013. First, properties of the optimal channel in RDT depend on the loss function, and the channel would attend more to a speci c part of the input if deviations in reconstructing that part of the input were penalized more in its loss function. This idea has been implemented in research on visual attentional allocation (Bates & Jacobs, 2021) and we adopt it here to simulate asymmetric feature weighting (i.e., reliance on one primary feature). Speci cally, we obtain this by scaling each dimension of the reconstruction loss (MSE) with a prede ned asymmetric weight which we call , so that the second term in the loss function becomes $MSE(x_{i_j}; y_{i_j}) \cdot$. Setting a high value of for the reconstruction of a particular feature, such as VOT, leads the network to prioritize that feature in its latent encoding.

Second, following Bates et al., 2019, we add a supervised categorization model, enabling

category mapping based on the features. This allows us to extract the information in such a way that it best serves the mapping of the input to one of the categories. Any information not contributing to the categorization is as such less important to send through the channel. The categorization model is trained jointly with the VAE on binary cross-entropy (BCE) loss (see full loss in Equation 6.2): the loss compares the output of the categorization model given latent information z_i with the ground truth binary label l_i (either [p] or [b]) fed to the model during training: $BCE(cat(z_{i_j}); l_i)$.

Third, our model has multiple encoders as seen in Ternes et al., 2022. Each encoder extracts different feature information. This is achieved with feature weighting as in Bates and Jacobs, 2021: we force each encoder-decoder combination to have higher reconstruction accuracy on one of the dimensions — VOT for encoder 1, and F0 for encoder 2 — and inversely, lower reconstruction accuracy on the other. This unequal accuracy reconstruction makes the model learn different information with each encoder. This models internal manipulations that we hypothesize listeners to be performing on the input when optimizing their perceptual system based on experience.

In other words, we show how listeners can build a perceptual system that allows them to allocate their focus more to one feature rather than another, given their needs in any given listening situation. The joint training of encoders is achieved by getting KL and MSE losses for each encoder and then averaging over them (see full loss in Equation 6.2). The model is jointly optimized to extract features, weight them by how reliable they are in the moment and map them to a category:

Figure 6.1: Multiple encoder -VAE training - Schematic architecture of training of multiple encoder -VAE with an added categorization model. The number of encoders can be extended. Each encoder encodes the same input. A sample z_j from each encoder E_i is decoded separately in training. The category model receives a random proportion of data from each encoder.

$$
\begin{aligned}
L(\ ;\ ; x_i; y_i) = \frac{1}{n} \sum_{j=1}^{n} & D_{KL}(q_j(z_j \mid x_i) \,||\, p(z_j)) \\
& + MSE(x_i; y_i) \cdot \tilde{\ }_j \\
& + BCE(cat(\sum_{j=1}^{n} \ _j z_j); l_i)
\end{aligned}
\tag{6.2}
$$

Each training step consists of each encoder encoding the same input x. The last layer of each encoder is a sampling step where the latent variable z is obtained. Each z_i is pushed through the decoder to obtain the reconstructed input, which is in turn used to calculate the loss function. $_j$ is a randomly sampled weight from $[0; 1]$ interval, such that $\sum_{j=1}^{n} \ _j = 1$. $_j$ de nes the proportion of the information taken from each encoder, such that their combination is pushed through the category model $cat(\sum_{j=1}^{n} \ _j z_j); l_i))$.

To motivate our architecture choice we also show results on a single channel -VAE, that