ABSTRACT

Title of Dissertation: EXAMINING THE ROLE OF SPEECH RHYTHM IN NEWBORN LANGUAGE DISCRIMINATION THROUGH MACHINE LEARNING AND BIOLOGICALLY-INFORMED MODELS OF SPEECH PROCESSING

Ruolan Leslie Famularo
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Naomi Feldman
Department of Linguistics

Newborns are sensitive to the difference between the speech of some languages but not others, a phenomenon referred to as early language discrimination. While this is commonly attributed to their sensitivity to the temporal rhythm in speech, it has never been systematically tested. In this thesis, I explored the behavioral phenomenon of language discrimination through a series of simulations using machine learning models. In addition to typical models directly drawn from machine learning, I also introduced a model that is grounded in auditory neuroscience through differentiable programming.

Results from the traditional machine learning models suggest that rhythm was not necessary for any model to perform language discrimination in a humanlike manner, which implied that other mechanisms relying on global statistics alone could be possible for language discrimination and potentially used by humans during behavioral tests. Additionally, with the differentiable model with auditory neuroscience constraints, while the model uses rhythm to perform language

discrimination, the range of rhythm was much faster than what is associated with syllable rhythm. These results have implications about newborn language perception and language acquisition that follows, and may be used to drive the design of future infant studies. Additionally, the application of differentiable programming to introduce intuitions and constraints from neuroscience and cognition offers a new path of manipulating deep neural networks in the study of neural and cognitive modeling.

EXAMINING THE ROLE OF SPEECH RHYTHM IN NEWBORN LANGUAGE DISCRIMINATION THROUGH MACHINE LEARNING AND BIOLOGICALLY-INFORMED MODELS OF SPEECH PROCESSING

by

Ruolan Leslie Famularo

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
2025

Advisory Committee:
Professor Naomi Feldman, Chair/Advisor
Professor Ramani Duraiswami, Co-Advisor
Professor Shihab Shamma
Professor Jonathan Simon
Professor Rochelle Newman

Foreword

This dissertation represents collaborative work and has been supported by esteemed colleagues and institutions. Chapter 2 was conducted in collaboration with Dr. Thomas Schatz and Ali Aboelata, with portions of the work published in the Annual Cognitive Science Proceedings of 2024. Chapter 3 was the result of joint efforts with Drs. Shihab Shamma and Dmitry Zotkin.

This research was made possible through the support of the National Science Foundation (NSF), under grant BCS-2120834.

Luci perpetuae.

Acknowledgments

The completion of this thesis is thanks to many people. First and foremost, I would like to thank my advisors, Drs. Naomi Feldman and Ramani Duraiswami. They introduced me to the areas of cognitive modeling and differentiable programming and, throughout my PhD, taught me so much in terms of technical knowledge, presentation skills, and how to conduct scientific research in a responsible way.

I would also like to thank other professors that have offered me guidance during this research journey. I want to thank Dr. Thomas Schatz, who has been mentoring me from basic speech processing techniques to advanced statistics; Dr. Shihab Shamma, who has made time to answer my questions and discuss science; Drs. Jonathan Simon and Rochelle Newman, for serving on my committee and offering inspiring questions and comments on this thesis. I also want to thank Drs. Benjamin Kedem, Carol Espy-Wilson, William Idsardi, Samira Anderson, Robert Slevc, and Nai Ding, for introducing me to various disciplines related to this research through your courses, discussions, and/or collaborations. I have learned so much from each of you, without which this thesis would have been lacking.

I would also like to thank many people that inspired me to pursue a journey in research. Particularly, I want to thank Drs. Xin Xie and Florian Jaeger, who introduced me to speech research by allowing me to contribute as a junior undergraduate, and patiently catching me up on all knowledge from how to conduct a literature review to math and programming. I would

also like to thank Drs. Chigusa Kurumada and Ralf Haefner, whose research course helped me build research habits and presentation skills that prepared me well for my PhD. I would also like to thank Drs. Aaron White and Joyce McDonough, whose courses in linguistics opened up the world of computational linguistics and phonetics. Apart from the science side, I would also like to thank Drs. Honey Meconi and Bruce Frank for leading me to see the structure behind music, and encouraging me to see music as a viable part of my career.

The completion of this work is also indebted to research assistants and collaborators that accelerated the research pipeline, or rather, hammering out various details to to ensure correct implementation of various scripts that were essential to the research in this thesis. I would especially like to thank Ali Aboelata for his work on corpus curation, data collection and scripting that were indispensable to Chapter 2 of this thesis. I would also like to thank Grace Brown and Brody Montag for their contribution to this thesis. Additionally, I want to thank the students of my classes, who helped me become a better communicator.

I also owe my thanks to many colleagues who made graduate school a welcoming environment to conduct research. Particularly, I want to thank Joselyn Rodriguez and Jingyi Chen for being great office mates, as well as for scientific and career discussions over food and bubble tea. I would also like to thank members of the PIRL lab, phonology circle, and the computational cognitive science group for interesting discussions and exchange of research ideas over the years. I also want to thank others from various departments that I have encountered in my cohort, courses, and seminars for their insights and comments on my work, as well as encouragements. Last but not least, I want to thank Kim Kwok, Pam Komarek, and Claire Morse for helping me with various administrative work, and UMIACS staff for their technical support.

Lastly, I would like to thank my family for their support through happy and tough times. I

also would like to thank the Netherlands Bach Society, whose music kept me sane through many deadlines.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AM	Amplitude modulation
AN	Auditory nerve
CNN	Convolutional Neural Network
GMM	Gaussian Mixture Model
IIR	Infinite Impulse Response
MFCCs	Mel Frequency Cepstral Coefficients
RNN	Recurrent Neural Network
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
STRF	Spectrotemporal receptive field

Chapter 1: Introduction

Human newborns are able to discriminate between certain languages but not others. This phenomenon has long been attributed to sensitivity to speech rhythm, the organized temporal structure of vowels and consonants (Mehler et al., 1988; Moon et al., 1993; Nazzi et al., 1998). While this presumed link between speech rhythm and early language discrimination has substantially impacted theories of language acquisition, it was never explicitly tested.

While newborn data are constrained to stimuli and experiment designs that are exciting enough for newborns, machine learning models presents a promising complementary piece to the study of newborn speech perception. Particularly in the context of language discrimination, various models have been applied to simulate newborn language discrimination, including early day recurrent neural networks (Dominey and Ramus, 2000) and later clustering-based models that operate on real speech stimuli (Carbajal, 2018; Carbajal et al., 2016). These models serve as a proof-of-concept for modeling infant perception and behavior using computational simulations. However, some of these models leave out important aspects of perception (such as using simulated data instead of real speech) that can now be addressed by modern-day machine learning models. Furthermore, all of these models rest on the idea that syllable-level rhythm drives early language discrimination, an idea that was never systematically tested.

The current thesis seeks to explore the connection between syllable-level rhythm and early

language discrimination through a series of simulations. In addition to testing models with and without access to syllable-level rhythm, an additional model is introduced where auditory processing is explicitly encoded as the earliest steps of the model using differentiable programming. Language discrimination results show that an array of machine learning models do not require the temporal structure to perform language discrimination, and even in the differentiable auditory model, which does require slower temporal structure for language discrimination, the critical temporal modulation does not sit in the acknowledged syllable range of 4.5 Hz, but is much faster. I then discuss these results in their implications for theories of speech rhythm perception and early language discrimination, as well as the use of differentiable models to embody hypotheses in cognitive simulations.

1.1 Speech rhythm

Rhythm in speech was characterized as early as in Pike (1945), who proposed that languages can be categorically divided into rhythmic classes depending on the level carrying a basic unit of rhythm. The original theory argued that the basic unit of rhythm is isochronous (periodic). However, such isochrony cannot be found in connected speech (Dauer, 1983; Pamies Bertrán, 1999). Despite the lack of isochrony in the physical signal, humans from infants to adults are found to be sensitive to differences between language rhythm in a way predicted by rhythmic classes (Mehler et al., 1988; Nazzi et al., 1998; Ramus et al., 2003). This has led to a substantial body of research aimed at identifying acoustic correlates in the speech stream that correspond to rhythmic classes. Correlates that have been identified thus far include global metrics such as the percentage of vowels (%V), the variability of consonantal intervals (ΔC), and variation

coefficients (VarcoC and VarcoV), which characterize the variability of vowel and consonantal intervals normalized by dividing over the mean of such intervals (Dellwo et al., 2006; Ramus et al., 1999). Measures of local variability in speech such as the Pairwise Variability Index (PVIs), which characterizes the variability of neighboring vocalic intervals, capture additional perceivable differences between languages (Grabe and Low, 2002).[1]

These metrics characterize the variability of the vocalic and intervocalic intervals at different levels, but several critical problems remain. As these measures are designed to be annotated by trained phoneticians by visualizing speech on a spectrogram, they cannot be directly applied to filtered speech (Nazzi et al., 1998), synthesized speech (Ramus and Mehler, 1999a) or non-speech (Abboub et al., 2016), where humans still perceive differences based on the rhythm of the stimuli. This weakness stems from the fact that there exists no theory to rigorously characterize what the representations of speech are or how they are formed. As a result, we cannot explain by what means the summary metrics are tracked in the auditory system, but can only establish correlational relationships between these metrics and human perception. Additionally, only a subset of these metrics may characterize the rhythmic difference between certain languages, and this subset is different for different languages. While nPVIs characterize the rhythmic difference between many languages (Grabe and Low, 2002), this is the only metric among all others that fails to characterize the syllable timing of Mandarin (Lin and Wang, 2007). Similarly, while the Eastern Arabic dialects are classified as stress-timed, the %V and ΔC measurements put them closer to syllable-timed languages like French and Catalan (Hamdi-Sultan et al., 2004). Lastly, while rhythm can be defined as temporal changes in various dimensions including duration, intensity,

[1]PVIs come in two forms, a raw measure (rPVIs) and a normalized measure (nPVIs). The latter is very similar to the former except for additional normalization over syllable durations.

pitch, and spectral features (e.g., vowel quality and musical timbre), these summary metrics only explore the durational aspects of speech rhythm, while all the other aspects are up to the decision of the annotator, making it less accurate and robust. In lieu of the complexity of temporal regularities that can be present in the speech stream, the abstract method involving human judgment may be involuntarily biased by the annotator's individual judgment and language experience.

More recent work has found additional measures focusing on the slow amplitude modulation of speech, drawing from engineering techniques that improved the aforementioned problems. Specifically, they focused on the slow amplitude modulations in the speech stream. While all languages have peaks concentrated around the same frequency of 4–5 Hz compared with nonspeech stimuli such as music (Ding et al., 2017), the specific location of this peak as well as the width differ across languages of different rhythms (Varnet et al., 2017). Combining the location and width of the peaks in the amplitude modulation spectra, the two cues create well-separate clusters for three rhythmic classes. This for the first time gave an annotator-free acoustic characterization of speech rhythm (also see Tilsen and Arvaniti, 2013 and simplification by Gibbon, 2018). Amplitude modulation was further used in a model of speech perception (Goswami, 2019; Leong et al., 2014), where specific predictions about rhythmic perception can be generated hierarchically for amplitude envelopes at different frequencies of an utterance. These findings led to a representation of speech rhythm, where semi-regular oscillation at certain frequencies drives the rhythm in speech, and speech is processed by the listener through these oscillations. Closely related to these ideas, a cognitive model based on envelope tracking was proposed for infants' cognition of syllables and even words (Räsänen et al., 2018). In this model, the envelope of speech, which conceptually equals the amplitude modulation, is extracted from natural speech at different frequencies (following a gammatone filterbank to extract such frequencies). Each envelope is

then passed into an oscillator which makes the oscillations more periodic. Lastly, different frequencies are combined through coherence to output a new envelope called the sonority envelope speech, from which syllable-level segments can be extracted by slicing the speech at the valleys of the sonority envelope. This work contributes a cognitive model that is the first to connect the amplitude modulation of speech with speech perception in infants.

However, the amplitude modulation account also misses some critical aspects of the perception and function of speech rhythm. First of all, it has not examined the crosslinguistic difference in speech rhythm, from which some of the most foundational behavioral observations of speech rhythm are made. For the sonority envelope model in Räsänen et al. (2018), among several rhythmically different languages tested with this model, the performance of syllable segmentation was similar across languages, while the hyperparameters of this model (such as the preferred rate of the oscillators) are kept the same for all languages. This contradicts the abundant behavioral evidence that infants process speech with different rhythm differently (Garcia et al., 2018; Nazzi and Ramus, 2003). The hierarchical model of amplitude modulation (Goswami, 2019; Leong et al., 2014) partly answers this question, but no crosslinguistic difference has been examined using this model. In a recent study (Peter et al., 2022), initial attempts were made to compare neural tracking of speech rhythm across several different languages. Nonetheless, this exploration remains at an early stage: the stimuli used in the study were synthesized and isochronous, which took away the temporal variability of natural speech. Additionally, for the studies that reported crosslinguistic differences in speech rhythm in terms of amplitude modulation (Ding et al., 2017; Varnet et al., 2017), the languages used in these works are different from those with which infants are tested for their crosslinguistic sensitivity of rhythm. Therefore, it is difficult to compare the amplitude modulation difference to behavioral results. Lastly and very importantly, amplitude modulation is

often characterized through Fourier power spectra of the envelope, which inevitably assumes periodicity in the speech stream as Fourier transform breaks down a signal into a sum of sinusoidal, each of which is perfectly rhythmic. However, natural speech is neither rhythmic in perception nor in acoustics, unlike music or poetry. For example, the syllable rate of speech is found to peak around 4–5 Hz; this corresponds to a peak around 4–5 Hz on the amplitude modulation spectra for most languages (Ding et al., 2017; Varnet et al., 2017). However, there is no perception of beats in naturalistic speech in the same way as music (Patel, 2010). We also do not observe a sharp peak in the power spectra of speech as in music (Ding et al., 2017). Therefore, for Fourier transform or any other tool that inherently assumes periodicity, the exact meaning of such periodicity is unknown. As such, it is possible that some more important, aperiodic temporal regularity exists around the syllable level of the speech stream, but is overlooked when one only looks at the Fourier power spectrum.

1.1.1 Speech rhythm and early language discrimination

Infants as young as newborns are known to discriminate between certain languages but not others (Mehler et al., 1988; Nazzi et al., 1998; Ramus, 2002). In these experiments, infants are exposed to utterances of one language, which are played through a speaker in the lab. When the speech stream switches to a new language, infants may notice the change, their response to which is reflected through a increased sucking rate on a pacifier. A significant increase in the sucking rate would indicate successful discrimination between two languages. It has been known that infants can discriminate between certain pairs such as English and Japanese, but not other pairs such as English and Dutch (Nazzi et al., 1998). The predominant explanation to

date for the specificity in language discrimination is that discrimination occurs when the pair of languages differ in rhythm, the organized temporal structure of vowels and consonants in the speech stream (Nazzi and Ramus, 2003; Ramus et al., 1999). Extending from this observation about newborns, knowledge of the temporal regularity of one's native language is hypothesized to be acquired first and facilitate later learning such as word segmentation (Jusczyk et al., 1999) and syntactic structure, a theory called the prosodic bootstrapping hypothesis (Nazzi and Ramus, 2003). Because sensitivity to speech rhythm forms the basis of theories of language acquisition, it is important to know whether infants are indeed sensitive to speech rhythm.

Additionally, to ensure that the infants do not rely on segmental or lexical information in their discrimination, the speech stream in behavioral studies was manipulated through low-pass filtering (Nazzi et al., 1998; Ramus, 2002) or resynthesis (Ramus, 2002; Ramus and Mehler, 1999b), where segmental information was replaced with synthesized sounds from similar categories (e.g. replaced any fricatives with /s/). These manipulations remove fine spectral information about phonetic categories and words, and to some extent control for the different phonemic inventory and phonotactics across languages. The persistence of language discrimination after such heavy processing indicates that infants are sensitive to what remains in the processed speech stream, which was presumed to be rhythm (Nazzi and Ramus, 2003).

1.2 Machine learning and models of speech perception and processing

In recent years, computational simulation of infant language acquisition from speech has made significant progress. Using insights from machine learning, it has become possible to simulate models of early speech learning with large-scale naturalistic input (Matusevych et al.,

2020; Schatz et al., 2021a). The use of naturalistic input is more ecologically valid than the lab-recorded, annotated or artificially generated data that had previously been used in computational models of language acquisition, and it enables exploration of the role of previously unemphasized features in speech, such as the distribution of speakers (Li et al., 2020a) and rhythm (Carbajal et al., 2016).

In summarizing the literature on machine learning models used to simulate speech perception, I roughly divide them into two kinds, one based on unsupervised clustering and one based on deep learning. With unsupervised clustering, models such as Gaussian Mixture Models are applied to preprocessed speech, and the clustering pattern after fitting the model on speech data is used as the model's representation of speech, often in the form of a multinomial distribution. As these models have been used in speech technology for unsupervised learning over speech or music, they have also been adopted to simulate humans' (often infants') representation of speech without explicitly labeled inputs, much like speech exposure in the first few months of an infant's life. For example, Schatz et al. (2021a) used a Dirichlet Process Gaussian Mixture Model (DPGMM) to cluster over segmental acoustic representations of naturalistic speech. They trained models on either American English or Japanese in order to simulate infants growing up in one of these language environments. They then tested these models using a machine ABX task (Schatz, 2016; Schatz et al., 2013), in which the model hears three sounds from a speech corpus (A, B and X, where A and X might both be instances of [l], as in *lock*, and B might be an instance of [ɹ], as in *rock*) and needs to determine whether X is more similar to A or to B. Higher accuracies correspond to better discrimination. They showed that after training, the American English models could discriminate the [l]-[[ɹ]] contrast better than the Japanese models could (Li et al., 2020a; Schatz et al., 2021a). This replicated the behavioral trend observed in infants between 6 and 12

months of age (Kuhl et al., 2006).

In the clustering-based models, features like mel-frequency cepstral coefficients (MFCCs) were commonly used. These are features characterizing the principal components of the short-time spectrum of speech, which are computed from short slices of speech (25 ms). The MFCCs are a common engineering representation of speech that mimics basic human auditory processes, and are some of the most popular in clustering-based approaches, as they represent important features in speech such as formant locations with relatively few dimensions, and the low covariance between dimensions makes it ideal for clustering algorithms such as Gaussian Mixture Models, where covariance is often assumed to be zero to reduce computational complexity. However, since MFCCs are short-time features calculated over only 25 ms, it does not characterize temporal dependencies well. To compensate for that, differences along time are taken (referred to as deltas) and used jointly as the original MFCC features, and algorithms such as dynamic time warping were employed to align speech when necessary. These approaches make MFCC features possible to work with temporal dependencies, but can still pose differences between machine processing of temporal dependencies and how sequential information is naturally handled by the human auditory system.

With the rise of deep learning, another family of models borrows from contemporary deep learning for speech processing. In this line, deep neural networks are often used as black-box algorithms to learn any regularity about human speech in a data-driven manner. Similar to unsupervised clustering, the deep learning models are trained in a manner that does not require explicit text or categorical labels, which corresponds to a model training regime called self-supervised pretraining (Mohamed et al., 2022). In self-supervised pretraining, the model was trained with some artificial task, called a "pretext task", where the model is expected to learn about the regu-

larities and structures of speech effectively after the model learns to perform the pretext task well. Common pretext tasks involve compressing speech through a bottleneck using autoencoders, reconstructing speech from various sources of corruption such as masking part of the spectrogram (Liu et al., 2021), and predicting a missing part of the speech in the future (Chung and Glass, 2020; Oord et al., 2018) or in the middle with past and future context (Baevski et al., 2020). The hidden layer representations of the deep neural network after training models on the pretext tasks, called embeddings, have been found to contain rich information about text, speaker identity, and beyond. As a result, these embeddings can be taken to perform tasks such as speech recognition and speaker identification using a very small amount of additional data and model training, while yielding high-performing and sometimes superhuman results. In other cases, the pretrained models can also be from a supervised task, such as speech recognition (Radford et al., 2023; Tuckute et al., 2023).

Both clustering-based and deep learning models have been used to simulate various problems in speech perception and processing, including early language discrimination. While infants are sensitive to temporal regularities related to their native language right when they are born (Abboub et al., 2016), speech models often do not take rhythm into account. For example, clustering-based models do not have access to the temporal regularities at all since the clustering algorithms consider all short-time speech features independently, ignoring the temporal sequence in the model representations. To model the sensitivity to temporal regularities, early models employ primitive recurrent neural networks (Cariani, 2001; Dominey and Ramus, 2000) which are sensitive to the temporal sequence through the recurrent architecture. For clustering-based models, feature engineering is used to create delta features over many frames spanning hundreds of milliseconds, which is enough to capture full syllables of speech (Carbajal, 2018; Carbajal et al.,

2016).

1.3 Differentiable programming and speech processing

1.3.1 Differentiable programming in the context of machine learning

Audio and speech processing have nowadays moved towards deep and end-to-end architectures that learn from data, and achieved super-human performance on various tasks. However, training such models is expensive in terms of both computation and data. For example, pre-training a speech model for downstream tasks requires hundreds of hours of data and access to expensive GPU clusters (Mohamed et al., 2022). These deep models lack robustness and are vulnerable to adversarial attacks, with drastic loss of performance from imperceptible modifications to inputs (Wu et al., 2022). Also, these models are "black-box" and difficult to explain.

In comparison, the human auditory system can effortlessly perform diverse tasks, including speech/speaker recognition and separation/enhancement. Classical audio and speech processing used features for machine learning that mimicked human perceptual processing, For example, the mel-frequency cepstral coefficient (MFCC) is conceptually similar to the cochlear analysis of sound (e.g., (Meyer and Kollmeier, 2011)). Although biomimetic methods are falling out of fashion due to the performance of pure input-output deep learning, integrating them into deep models may be able to achieve the best of both worlds, allowing improved data/computational efficiency, robustness, and explainability.

Traditional models can be adapted to be an integrated part of a deep learning model through differentiability. Historically, domain scientists and mathematicians have developed highly successful *forward mathematical models* connecting inputs to outputs. Such models are theory-

driven, fast, and give actionable explanations, but usually perform poorly in the inverse settings where data has to be related to inputs. Deep learning models, however, allow learnable relationships between data and inputs via universal function approximations through optimization on appropriate cost functions. On the other hand, in the context of end-to-end deep learning, automatic differentiation ("autodiff") is the center of the spotlight, where derivatives collected from output to input allow "backpropagation", i.e., gradient calculation through autodiff, a key ingredient of stochastic gradient descent used to train deep networks. Differentiable programming frameworks such as PyTorch and JAX make this almost invisible to the programmer when using standard network models. More recently, differential Physics (Thuerey et al., 2021) extends the concept of training based on a differentiable forward map to classical mathematical physics, allowing gradient-based optimization, and training of networks that respect physics and learn the parameters. Related differentiable approaches have been applied to other fields, including Operator Learning (e.g., (Li et al., 2020b)) and Neural Radiance fields (NeRFs) (Xie et al., 2022).

1.3.2 Differentiable programming in speech perception models

In audio processing in general, differentiability has been applied toward interpretable and robust audio synthesis. In Engel et al. (2020), a differentiable vocoder was used to combine the traditional vocoder system with deep learning modules for audio generation and style transfer. In Shier et al. (2023), differentiability was specifically used for percussion sound synthesis. Across prior work, as differentiability allows physical features (e.g., pitch of the music or envelope of the percussive attack and decay) to be specified, the architecture allows for customizations that

are suitable for personalization. The current work will explore this customizability towards personalized model fitting.

On the other hand, the use of various "frontends" in audio processing has received long-lasting interest as a way to incorporate neuroscience knowledge into machine learning models. These frontends act as the gateway of input audio into the model, and perform rudimentary processing such as time-frequency decomposition, oftentimes in a neuroscientifically informed manner. In classical, non-differentiable settings, auditory models of the ear and brain (Elhilali and Shamma, 2008; Mesgarani et al., 2006; Meyer and Kollmeier, 2011) have been applied to audio processing. These models outperformed traditional features such as MFCCs, especially in the presence of noise. Recently, several audio front-ends (Ravanelli and Bengio, 2018; Zeghidour et al., 2021) have been proposed to improve audio tasks in performance and/or robustness in deep learning. Features based on cortical processing have also been combined with differentiable approaches to improve performance/robustness as a frontend (Vuong et al., 2020) and as a loss function (Vuong et al., 2021). We seek to extend upon prior work and create a multi-stage differentiable model that adheres to biological auditory processing.

1.4 Outline of Thesis

The rest of this thesis will be organized as follows. In Chapter 2, I simulate early language discrimination using various cognitive models relevant to the early speech modeling literature, including clustering-based models and deep neural networks that are trained in a self-supervised manner. The results suggest that most models that were tested discriminated languages similar to humans, but none of them required rhythm — slower temporal information beyond 25 ms — to

do so.

In Chapter 3, moving towards models that are more biologically plausible, I introduce an auditory model to deep learning models of speech processing through differentiable programming. The auditory model can be trained jointly with a deep neural network, and I demonstrate the improved performance, robustness and interpretability through two speech processing tasks: phoneme recognition and source separation.

In Chapter 4, I apply the differentiable auditory model as a biologically plausible frontend in cognitive models and examined the new model on language discrimination. Results suggest that the differentiable model, unlike other machine learning models tested in Chapter 2, relies on temporal information to perform language discrimination. This shows promising prospects of applying differentiable models to study cognitive phenomena, especially in the current case of longer, suprasegmental information.

Chapter 5 concludes this thesis.

Chapter 2: Exploring the role of speech rhythm in early language discrimination through machine learning simulations

2.1 Overview

Despite the widespread hypothesis that newborns' language discrimination relies on sensitivity to rhythm in the speech stream, existing data are also compatible with other possible interpretations. Here we focus on the possibility that newborns' discrimination may instead reflect sensitivity to global and segmental acoustic properties of the speech signal. In contrast with rhythm, global properties can be calculated by averaging short, segmental information without requiring information about any temporal order. It is already well known that there are global properties of speech that correlate with rhythmic classes, such as the mean percentage of vocalic durations (%V) and the variability of consonantal durations (ΔC) (Ramus et al., 1999). For the language pairs that infants were tested on, these are the only acoustic correlates that were directly compared with infant language discrimination. However, it is not yet known whether these global properties on their own would be sufficient to explain infants' discrimination, and most previous work that has measured these properties' correlation with rhythmic classes nevertheless assumes that what humans are sensitive to are the rhythmic properties, rather than the global properties (Dominey and Ramus, 2000; Langus et al., 2017; Ramus et al., 1999).

In this chapter, we show that early language discrimination can be simulated without using any rhythmic information. We conduct simulations of language discrimination using representations ranging from the acoustic level to higher-level representations derived through several different models of speech perception, using intact speech as well as speech that has been scrambled to remove its rhythmic structure. We show that all of these representations across speech features and models can qualitatively simulate infant discrimination, implying that infants could be relying on global properties, rather than rhythm, in discriminating between languages. This has implications for theories of early language learning that center around rhythm and its crosslinguistic differences.

We first give background on the evidence in favor of infants' rhythmic sensitivity and existing challenges to this theory. We then give an overview of the framework and models we use for simulations. Our next sections present simulation results on natural and low-pass filtered speech showing that a wide range of models and representations with or without access to rhythm can all capture newborn discrimination. Finally, we discuss the implications of these results for theories of language acquisition and their compatibility with other evidence on infants' sensitivity to rhythm.

However, even in manipulated speech, language-specific cues that are irrelevant to rhythm remain. As reviewed earlier, the two summary statistics that have been shown to correlate with human discrimination results (Ramus et al., 1999) were global statistics and can be calculated without access to the temporal sequence of segmental information. These global differences between languages likely stem from phonological properties related to timing, such as vowel duration (Ramus et al., 1999) and syllable structure (Langus et al., 2017), and it has been widely assumed in the literature that infants are sensitive to these local temporal structures in the speech

stream that generated the global statistics, rather than to the global statistics themselves. In theories of acquisition, it has been hypothesized that later acquisition of e.g. word segmentation (Jusczyk et al., 1999) and phonetic categories in bilingual infants (Sundara and Scutellaro, 2011) are based on this sensitivity to rhythm.

Furthermore, in computational models that replicated infant language discrimination through simulations, it is also often concluded without controlled testing that the computational models succeed in perceiving "rhythm". For example, in Dominey and Ramus (2000), a small connectionist model was trained to discriminate between annotated speech of various languages. The model's discrimination behavior aligns with that of newborns, which is taken as further evidence that human discrimination is also driven by local temporal regularity, even though no controls were included without access to temporal information. Additionally, in Carbajal et al. (2016), when a clustering-based model showed language discrimination results on raw speech that aligned with what was predicted for humans, the model's success was attributed to learning of temporal regularities, which was made available to the model through feature engineering. In both cases, the computational models were assumed to use rhythm towards language discrimination, but this assumption was never tested.

While the prosodic bootstrapping hypothesis assumes that sensitivity to speech rhythm is a precursor to later language development, evidence from different fields has challenged how this occurs. Firstly, the knowledge of linguistic rhythm takes place at very different developmental stages across languages (Mazuka, 2007). For example, while stress cues are available to English-learning babies as early as 7.5 months, (Jusczyk et al., 1999), the use of the mora as an overarching phonological unit does not appear in 4-year-old Japanese children (Kubozono, 2000). This makes it unclear how the cues driving early language discrimination interact with

language development. In production (Grabe et al., 1999; Payne et al., 2012; Polyanskaya and Ordin, 2015), children continue to produce languages such as English with acoustic correlates that are more similar to languages like Italian (referred to as "syllable-timed languages" in some theories) until they are 11 or 12. This raises questions about whether acquisition of language-specific rhythm really happens during the first few months of life, or is rather a lifelong process that takes until adulthood. Lastly, existing data are scarce about changes in language discrimination along development. As Gasparini et al. (2021) pointed out, among the only two studies (Johnson and Braun, 2011; Nazzi et al., 2000) that examined the effect of learning, the respective results have contradicting theoretical implications. Also, since both studies use naturalistic speech that was not manipulated, infants could discriminate between languages by relying on any combination of segmental (spectral and temporal cues) and suprasegmental (intonation and timing) information. As such, the acquisition of speech rhythm is not well-supported by these empirical studies.

2.2 Simulating Early Language Discrimination

In our simulations, we model the qualitative results from infant language discrimination. We focus on discrimination of very young (3-day-old) infants Nazzi et al. (1998), whose results inspired language discrimination studies on older infants as well as learning theories such as the prosodic bootstrapping hypothesis. In their experiment using a habituation paradigm, they found that 3-day-old French infants can discriminate between English and Japanese, but not between English and Dutch. Furthermore, in another experiment, they examined if the infants can discriminate between a mixture of languages switching to another mixture depending on whether the rhythm of the mixture is homogeneous. The authors found that discrimination was

only possible when the group of language was homogeneous in its "rhythm" characterizations.

Here, we simulate both studies using an array of computational cognitive models. In Simulation I, we simulate Experiments 1 and 2 in Nazzi et al. (1998), which tests language discrimination between a pair of languages. In Simulation II, we simulate Experiment 3 in Nazzi et al. (1998), which tests language discrimination between languages that are grouped by their supposed rhythmic similarity.

2.2.1 Models

We examine three different models which generate different representations; we can interpret these as hypotheses regarding the cognitive representations used in language discrimination and early speech perception in general. These models range from deep learning embeddings to generative clustering algorithms. While each model represents a specific hypothesis about acquisition (e.g. predictive coding, generative inference, or adaptation), they have all been used in simulating human perception and early learning (Carbajal et al., 2016; Dominey and Ramus, 2000; Feldman et al., 2013; Vallabha et al., 2007). As such, we simulate human language discrimination with all these models to highlight the generality of our findings.

Firstly, we used recurrent neural network (RNN) based on predictive coding, using a more modern architecture than Dominey and Ramus (2000). The RNN is a smaller version of the Vector-Quantized Autoregressive Predictive Coding model (Chung et al., 2020), which predicts acoustic features in the future given current and past acoustic features. Our network contains 3 recurrent layers with vector-quantizing (a bottleneck step to facilitate learning) after the last recurrent layer. For hyperparameters, the model we used contains 32 hidden units per layer and

codebook size 16, and the network predicts 9 frames in the future (i.e. 90 ms). The rest of the parameter choices are the same as Chung et al. (2020). We examine the same model trained for 10k, 100k, and 500k steps to examine the effect of learning on language discrimination. For the main results, we used embeddings from the first layer of the RNN.

Secondly, we used the i-vector model following Carbajal et al. (2016). This can be seen as an extension of GMMs with an additional step: after a model is trained, it adapts to new data in a lower-dimensional space, generating an adaptation (i.e., a shift from the original distribution) called an i-vector. The i-vector obtained this way can be used as the model's representation of this new utterance. Notably, in Carbajal et al. (2016), the i-vector model was used directly to replicate language discrimination behavior in human infants. The model operates on engineered acoustic features: mel-frequency cepstral coefficients (MFCCs) were computed from raw speech along with a pitch track, on which shifted delta coefficients (SDCs) were computed, allowing the features to contain information about speech over a longer time span (200 ms, as opposed to 10–30 ms in typical spectrogram-based features), which is enough to contain suprasegmental information such as local syllable durations. The model behaved like human infants in discriminating languages across rhythmic classes (English and French) but not within (Spanish and Catalan). The success of the i-vector model in these two studies was attributed to its access to longer temporal information over 200 ms. In our simulations, we replicate this i-vector model with the original features containing suprasegmental and pitch information ("Original" model) on novel train and test corpora. Additionally, to test the claim that suprasegmental information led to human-like language discrimination in the i-vector models, we also simulated i-vector models that cannot access slower temporal information by training and testing models with only segmental features. In one version, we only keep MFCCs in the feature without temporal features

or pitch track ("MFCC" model). In another version ("Spectrogram" model), to further remove spectral information, we extracted information using binned spectrograms similar to the other models. If suprasegmental and temporal information is necessary for language discrimination, we would expect the MFCC and Spectrogram models to not have humanlike language discrimination. We train all i-vector models using the MSR Identity Toolkit (Sadjadi et al., 2013) with 256 clusters for the UBM, and a dimensionality of 200 for the i-vectors.

Lastly, we used Gaussian Mixture Models (GMM), which cluster individual frames of the speech signals to find patterns. In cognitive science, similar models have been used to simulate early speech learning (Li et al., 2020a; Schatz et al., 2021b). This model assumes a "bag of speech frames," meaning that speech frames are seen by the model in a way that is independent of their ordering, and the temporal structure therefore is not considered by the models. In our simulations, we trained models using scikit-learn (Pedregosa et al., 2011) with 50 Gaussians initialized using k-means, trained using Expectation Maximization until convergence (threshold 1×10^{-3}). In theory, the model could be augmented to include timing information through feature engineering, but we did not perform that here.

Training and Testing We trained our models using French to simulate the language exposure of a newborn (3-day-old) infant. For the GMM and i-vector model, we trained each model using a small amount (1 hour) of French obtained from the Globalphone corpus (Schultz, 2002), which contains roughly equal amounts of four speakers (2 males and 2 females). For the RNN model, since the type of self-supervised deep learning algorithm is a lot more data-hungry, we trained each model using 100-hour subsets (sampled randomly) from the Common Voice corpus (Ardila et al., 2019). While the Common Voice corpus was crowdsourced and contained a larger amount

speakers with different recording conditions, it has enough data to train self-supervised neural networks. We trained multiple models on disjoint training data to account for individual model differences – 4 models for the GMM and i-vector model, and 5 models for the RNN.

In addition to the models above, we also performed testing using acoustic features, without any further modeling. For this, we directly take log-Mel spectrograms as the representation for test.

Test procedure We use the machine ABX task (Schatz, 2016; Schatz et al., 2013) to evaluate language discrimination in machines. In the past, this technique has been applied to segmental categories (Dunbar et al., 2019; Schatz et al., 2021b) and suprasegmental information such as language identity (Carbajal et al., 2016). In our case, we apply the ABX task to examine the models' ability to discriminate between languages. Behavioral experiments such as Nazzi et al. (1998) measured discrimination in newborns using the headturn paradigm. Our test method does not directly replicate this specific testing paradigm, but instead can be seen as a conceptual simulation of language discrimination.

The machine ABX task is conceptually similar to the human version of the ABX task, which is a 2AFC task where humans receive three tokens, A, B, and X, and are asked to decide whether X is closer to A or B. In the machine, the decision of whether the machine chooses A or B depends on a distance measure calculated over the model's representations. The algorithm works as follows. In one trial, utterances A, B, and X are randomly sampled, where A and X are from one language (e.g. Japanese) and B is from the other (e.g. English). Then, the distance $d(A, X)$ and $d(B, X)$ are calculated. If $d(A, X) < d(B, X)$, the machine is considered to be correct, since X is closer to A than B in the machine's representation. The choice of distance

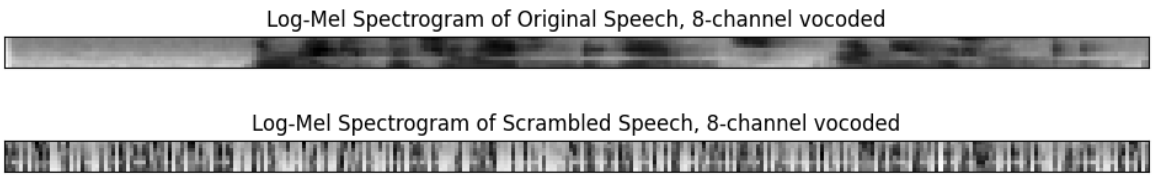

Figure 2.1: An example of the 8-channel spectrogram for the first three seconds of an English test utterance. Top: Original spectrogram; bottom: temporally scrambled spectrogram. Note how the low spectral resolution from the reduced number of frequency channels removed detailed information such as harmonics and formants, which obscures detailed spectral information such as vowel identity and pitch.

function depends on the representation of the model. For the GMMs, since the posteriogram representations are probability distributions, we used symmetrized Kullback-Leibler (KL) divergence. For all other model representations, we used cosine distance. For the i-vector model, since only one i-vector was generated for each utterance, the distance metric involves only the direct distance computation between the vectors illustrated above. For all other models, since the representations are generated for each time frame, we randomly sample one second[1] of speech material from each utterance and performed dynamic time warping. Dynamic time warping finds the optimal temporal alignment between two samples. In this case, we used the sum instead of the average of distances along the path generated by dynamic time warping, with a cost value of 1.

Since we have a large number of utterances, we randomly sampled A, B, and X in each trial and made 3000 independent draws for each model and condition, with A and B being the correct answer half of the time (i.e. in 1500 draws). The ABX error rate was calculated over all the draws to represent the model's ability to discriminate between languages. A lower ABX error rate indicates better discrimination, and chance performance is at 50%.

[1] We also ran tests with 0.5 and 2 seconds of speech material, with qualitatively similar results.

2.2.2 Feature Preprocessing

We used 8 MFCC features in the i-vector model following Carbajal et al. (2016), and spectrogram representations for all other models. For simplicity and to remove most of the spectral contents, we used only 8 frequency channels on a logarithmic scale.[2] An example of the resulting spectrogram is shown on the top of Figure 2.1. Additionally, we create the two following conditions through manipulating the speech features.

Temporal Scrambling We temporally scramble (or shuffle) the segmental frames in our test stimuli to examine whether language discrimination is driven by rhythm. Since the speech signal is transmitted in the time domain, any information in speech is characterized through temporal regularities. Among all temporal regularities, the faster ones correspond to segmental information such as formant and harmonics, and the slower ones correspond to syllable, word, and phrase-level rhythm. In our study, we focus on the effect of slower regularities, which we reflect through the ordering of segmental information, such as the order of individual frames of a spectrogram. Since the spectrogram frames are 10 ms apart from each other, any temporal information that is slower than 100 Hz is included in this range, including syllable rate (4–8 Hz) and slower regularities such as stress and prosodic phrase. To remove the slower rhythmic information, we randomly shuffled the acoustic features to disrupt any temporal structure that is slower than 10 ms in the speech signal. An example of a scrambled utterance is shown in Figure 2.1. We introduce this condition to remove temporal information either directly used by the model (i.e. the RNN and the Full version of the i-vector model) or used during the test procedure through dynamic time warping (i.e. in the RNN and GMM).

[2] Replication using 80 frequency channels showed qualitatively similar results.

Low-pass filtering In behavioral studies, the speech stimuli were often low-pass filtered to remove information such as formants. In all our simulations, we run tests on test stimuli that are unfiltered ("raw") as well as low-pass filtered. For low-pass filtering, we used a 4th-order Butterworth filter with cutoff frequency at 400 Hz.

2.2.3 Simulation I: discrimination of language pairs

In the first set of simulations, we simulate language discrimination between two languages. Following the first two experiments in Nazzi et al. (1998), we tested English-Japanese and English-Dutch. In the behavioral study, 3-day-old infants dishabituated at a switch between English and Japanese, but not between English and Dutch. In this simulation, we replicate this effect in models that can or cannot access rhythmic information.

To test the models, we used utterances from the Common Voice corpus (v. 13.0) (Ardila et al., 2019). Among all sentences, we discarded sentences with downvotes (i.e., rated by online participants to have noise, dis-fluency, or bad otherwise; around 20%). For each set, we selected utterances that are between 4 and 10 seconds long. The utterances were shuffled to avoid selecting multiple clips from the same speaker. Then, we manually listened to the audio files and selected the first 100 utterances without significant noise, disfluencies, or obvious non-native accents. Through the manual selection process, the criteria filtered out roughly 50% of the utterances. The IDs of the selected utterances are included in the code release. All utterances are root-mean-square normalized before any further processing. The results are shown in Figure 2.2.

All models showed significantly better discrimination on English-Japanese than English-Dutch, indicating that the models' performance replicates that of human infants in e.g. Nazzi

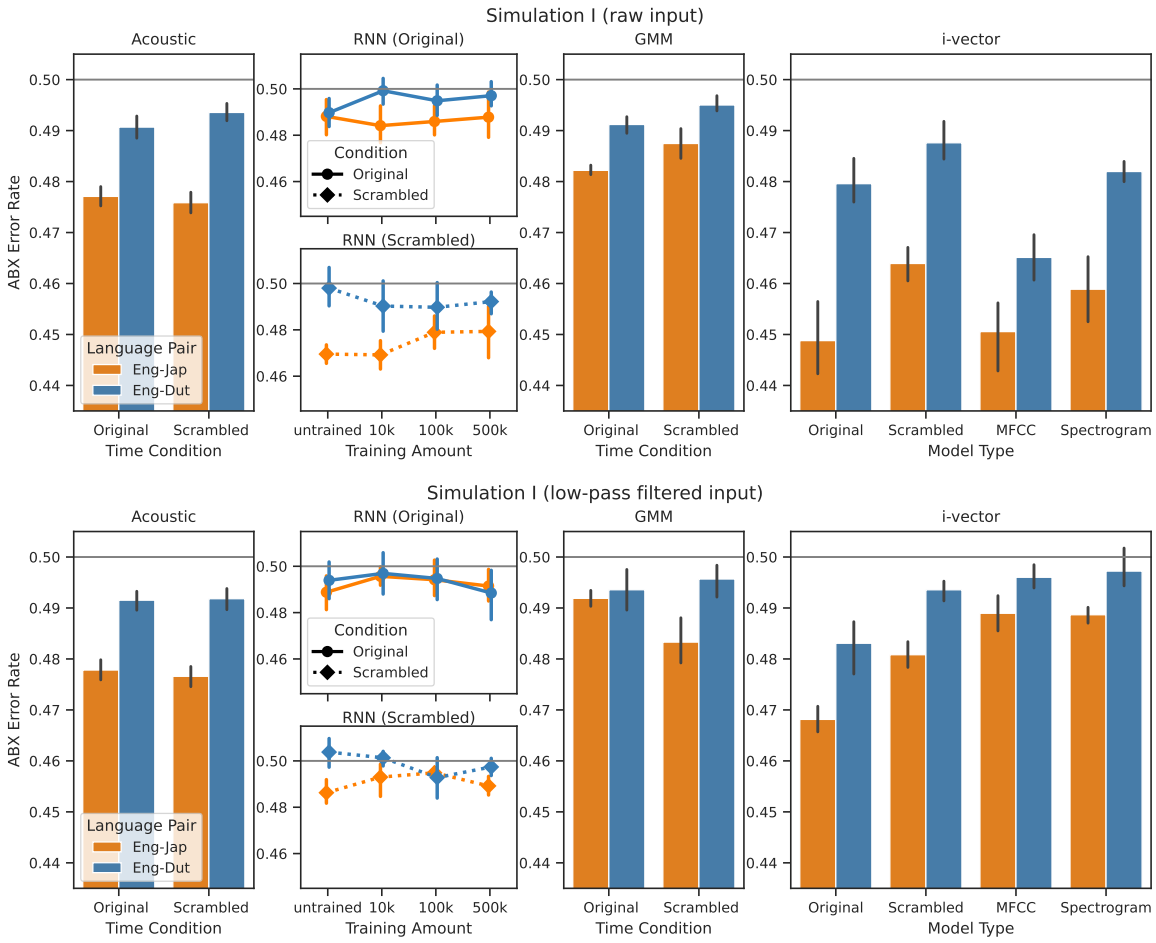

Figure 2.2: Language discrimination results on single languages. The top panel shows results with unfiltered test stimuli, and the bottom panel shows results with low-pass filtered test stimuli. A lower ABX error rate indicates better discrimination, where chance level (50%) was marked by grey horizontal lines. For the RNN model, we also displayed tests run over model at different stages of training, from untrained (i.e., random initialization) to models trained for 500k steps. Errorbars represent 95% CI calculated over models trained on disjoint training data. For the acoustic representations, since no models were trained, we performed five statistically independent tests over which the error bars were calculated.

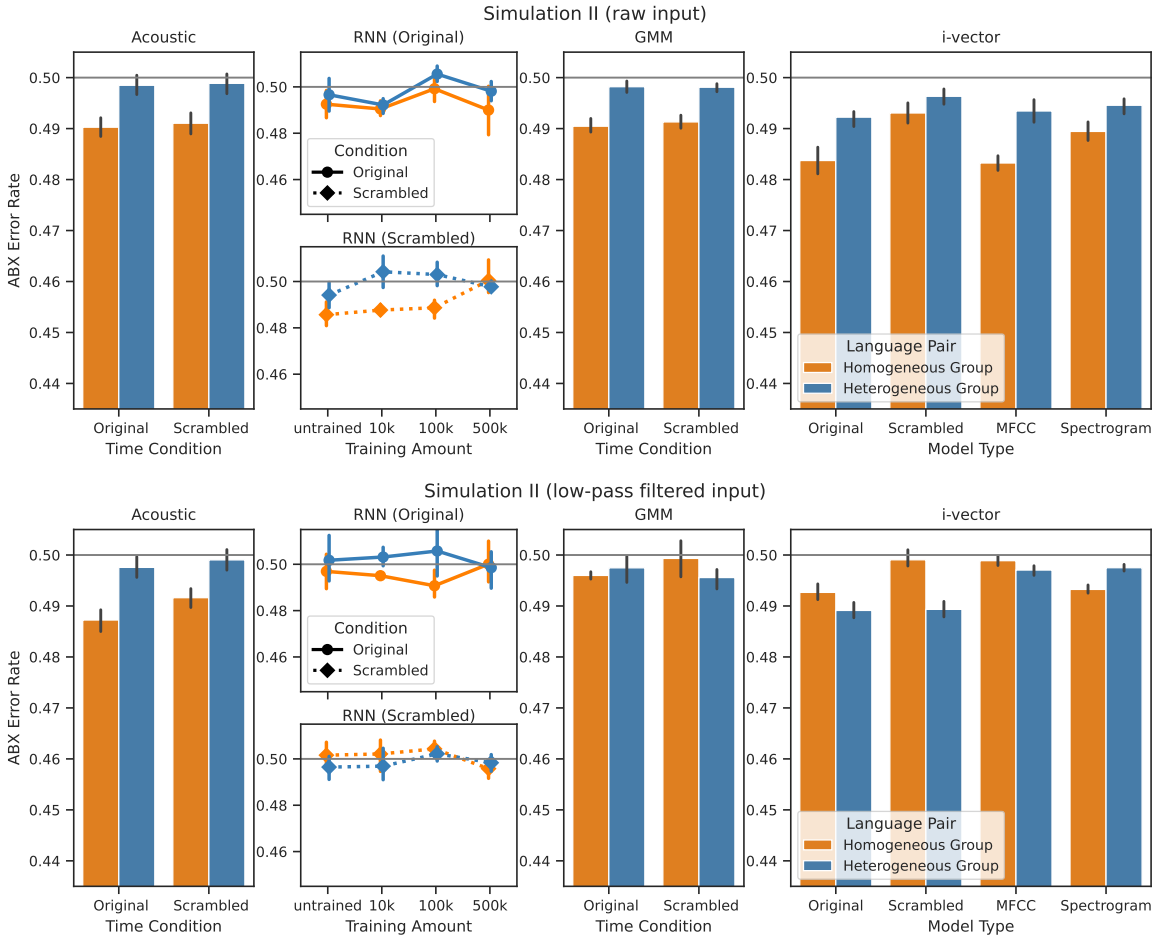

Figure 2.3: Language discrimination results on grouped languages. The top panel shows results with unfiltered test stimuli, and the bottom panel shows results with low-pass filtered test stimuli. A lower ABX error rate indicates better discrimination, where chance level (50%) was marked by grey horizontal lines. For the RNN model, we also displayed tests run over model at different stages of training, from untrained (i.e., random initialization) to models trained for 500k steps. Errorbars represent 95% CI calculated over models trained on disjoint training data. For the acoustic representations, since no models were trained, we performed five statistically independent tests over which the error bars were calculated.

et al. (1998). This is true for both intact speech and temporally scrambled speech, indicating that rhythm is not important to the effect of language discrimination. For the i-vector models, the crosslinguistic effect was observed not only in the Full model which contains slower temporal features, but also in the Scrambled model where temporal information was shuffled, as well as the MFCC and Spectrogram models where slower temporal information was unavailable. This indicates that language discrimination can be achieved using segmental information alone, without using slower temporal regularities that are crucial for rhythm.

In certain cases (e.g. in the case of some RNN models and the GMM when stimuli were low-pass filtered), scrambling in fact led to stronger humanlike results. This result seems counterintuitive for two reasons. First, for both models, dynamic time warping at test time must align the test utterances, which could be harder if the stimuli are scrambled and lack syllabic and phrasal structures for alignment. Secondly, for the RNN model, since the model is sensitive to the temporal sequence due to its recurrent structure, the temporally scrambled stimuli are out-of-distribution compared with the training data. However, given our hypothesis that language discrimination can be achieved by relying on global information (e.g. %V), temporally scrambling the frames would lead to a more even distribution of global information. This would make the global statistics more consistent when computed over different stretches of speech, leading the A and X stimuli in the ABX task to be more reliably similar. Therefore, the increase in effect size after scrambling the stimuli serves as evidence for our hypothesis that language discrimination in these models may rely on global properties rather than temporal structure.

Additionally, humanlike language discrimination can be achieved with both scrambled and unscrambled acoustic representations. Since the acoustic features are 8-channel spectrograms, the spectral information is highly degraded, with cues requiring fine frequency details (e.g. for-

mants and frication centroids) removed. However, global information such as the distribution of vowels and consonants should remain intact. This again suggests that discrimination is possible using global information alone, even based on acoustics without further processing.

Discussion In discriminating language pairs, the models tested were generally successful in replicating human-like results. Specifically, among all the model conditions that displayed humanlike effect in temporally ordered speech, they all displayed an effect in scrambled speech. This suggests that temporal regularities are not necessary for language discrimination in these models.

However, when discriminating low-pass filtered speech, some models showed mixed (in the GMM) or null (in the RNN model) effects. Considering that low-pass filtered speech (in test) was greatly different from unfiltered speech (in training), especially in that the higher-frequency channels were set to zero after low-pass filtering, the distributional shift between training and test data might lead to greater difficulty to obtain any meaningful effect from these models. For example, in the RNNs, the embeddings of the test utterances could generally be distant from each other regardless of language, thus leading to discrimination scores that are not crosslinguistically different. Human infants, on the other hand, are not subject to such numerical constraints (Nazzi et al., 1998). This may be due to other normalizing methods in human perception and cognition, or prenatal exposure to speech which resembles low-pass filtered speech more. Nonetheless, even under numerical difficulty, temporal scrambling seemed to help the model behave like human infants, which cannot be explained if temporal regularities are necessary for humanlike language discrimination.

2.2.4 Simulation II: discrimination of mixed languages

In the first set of simulations, we tested the model on discriminating between language pairs, and found the results to replicate language discrimination in human infants. However, discrimination between two single languages can be confounded with factors specific to the languages chosen (sound inventory, language-specific prosody, etc.) and the data collection process of each corpus (recording device, background noise, etc.). In the second set of simulations, we model Experiment 3 in Nazzi et al. (1998) to test the discrimination between two groups of languages. The original study found infants to dishabituate more when languages switch across a homogeneous group (i.e., from a group containing English and Dutch to a group containing Spanish and Italian, and vice versa) than across a heterogeneous group (i.e., English and Spanish vs. Dutch and Italian). We simulate language discrimination using computational models and test whether the models replicate human behavior. We test two pairs of groups: in the homogeneous group, we grouped English with Dutch, which switches into a group containing Spanish and Italian (and vice versa); in the heterogeneous group, we grouped English with Spanish, which switches into a group containing Dutch and Italian. Since all the utterances used in the two conditions are the same, only switching between different sets, this can eliminate some confounds specific to the speech material. We used the same corpus and selection criteria for test stimuli as Simulation I. The results are shown in Figure 2.3.

When the test stimuli were unfiltered, language discrimination for the models mostly aligned with that of humans, which is similar to Simulation I. For the Acoustic model and the GMM, the significance remained regardless of scrambling. In the RNN model, with unscrambled stimuli, language discrimination was in the right direction yet insignificant. With scrambling, the results

were significant in the RNN across all three training amounts. This is similar to the RNN's behavior in Simulation I: when the test stimuli were scrambled to become temporally uniform, the effect size increased, likely because global information helped with discrimination more than temporal regularity. In the i-vector model, we also observed humanlike discrimination in the Full and MFCC models, suggesting that temporal regularity was not necessary for humanlike behavior. The Full Scrambled model, however, showed a marginal yet insignificant outcome. Instead, in the MFCC and Spectrogram models, when temporal information was removed altogether directly, discrimination aligned with human behavior.

When the stimuli were low-pass filtered, the results were less consistent across models. In the Acoustic model and the Spectrogram version of the i-vector model, however, the results were human-like. For the other two models (RNN and GMM), however, the crosslinguistic effect was inconsistent regardless of whether there is temporal information.

From the second set of simulations, we conclude that different models were generally able to discriminate the rhythmic group better than the nonrhythmic group, even when temporal structure is inaccessible to the model. However, when the test stimuli were low-pass filtered, some of the models failed to replicate human behavior completely, while the successful models were more likely to replicate human behavior with the absence of temporal regularities. Overall, in no case do we see a model consistently succeed with temporal information available, and consistently fail without it.

2.3 General Discussion

We used computational models to simulate language discrimination using naturalistic speech as well as manipulated conditions to remove temporal regularities. We find various models to be able to replicate the behavioral discrimination pattern, even when temporal regularities are not accessible to the models. This suggests that language discrimination can be achieved through information other than temporal regularities, a result that calls for further investigation into the nature of language discrimination in human infants.

Our results have implications for theories of language development that center on the role of speech rhythm in early acquisition. First of all, it directly challenges the prosodic bootstrapping hypothesis (Nazzi and Ramus, 2003), which proposes that an early sensitivity of rhythm drives acquisition of metric units in one's native language. The prosodic bootstrapping hypothesis specifically assumes that sensitivity of speech rhythm narrows during the first year of life, which our results challenge directly by calling into question whether speech rhythm plays a role at all during newborn language discrimination. Our results also challenge the theory in bilingual acquisition that rhythmic similarity between the languages being learned affects the timing of vowel discrimination and word learning (Sundara and Scutellaro, 2011). While languages that are considered "rhythmically similar" are more similar in acoustic correlates of speech rhythm, they share other global and segmental properties compared with "rhythmically dissimilar" languages, which the current study supports as likely causes of language discrimination. Lastly, our findings can potentially influence the narrative of language development, which currently emphasizes speech rhythm as the earliest available cue to infants that drives further language development (e.g., Werker and Hensch, 2015; Gervain et al., 2021).

In our simulations, one major difference between our test paradigm and the behavioral experiment in Nazzi et al. (1998) was the difference in the distribution of test speakers. In behavioral studies like Nazzi et al. (1998), the test stimuli were obtained from two female speakers with minimal voice quality difference between each speaker. On one hand, factors such as voice quality are subjective and difficult to replicate. On the other hand, additional variability such as language background (Li and Post, 2014; Lin and Wang, 2008) and age (Pellegrino et al., 2021), are not typically reported in behavioral studies but are known to affect speech rhythm. To address these sources of variability, in our simulations, we used utterances from a wide range of speakers selected from crowd-sourced corpora. While this introduces additional noise in the test stimuli, thereby making the discrimination task harder, we still observe significant crosslinguistic languages, suggesting that we successfully replicated the language discrimination phenomena while ruling out the confound variables noted above.

One difference between our results and the behavioral studies is the absolute existence of an effect. In Nazzi et al. (1998), infants showed an absolute inability to discriminate between English and Dutch, while being able to discriminate between heterogeneous language groups. In our simulations, many models did not follow this exact pattern, with some models showing chance-level discrimination for heterogeneous language groups, or significantly above chance for English-Dutch. We attribute this to two fundamental differences between the human study and our simulations. First, while the machine ABX error rate can be different with any detected differences between the two languages tested, human perception may require certain differences to reach certain thresholds before showing a behavioral response. Therefore, the machine ABX results could be oversensitive and its absolute values less useful than relative comparisons, such as English-Japanese against English-Dutch. Second, while the machine ABX

task uses three utterances per trial, infants are able to hear many more utterances through the habituation-dishabituation paradigm such that they may be able to bootstrap in-group similarity better than the machine, allowing them to detect a difference between heterogeneous rhythm groups while the computational models failed to do so. Both explanations suggest that rather than the absolute value of scores, a better comparison between behavioral data and simulation results is to compare relative scores in a controlled setting, which is how we analyzed the results.

2.3.1 Language discrimination with altered segmental properties

In past studies on the behavioral and neural response in young infants to speech rhythm, various manipulations on the speech stimuli have been used. The fact that language discrimination remains despite the removal of some segmental properties (such as low-pass filtering and resynthesizing) but not when speech is temporally manipulated (such as playing speech backward) has been traditionally taken to establish infants' sensitivity to rhythmic cues in speech. Here, we argue against the claim that these baselines provide evidence of rhythm sensitivity per se.

In behavioral studies about newborn language discrimination, stimuli were often low-pass filtered to remove information at higher frequencies. This practice stems from earlier theories that prenatal experiences play a role in early language discrimination, and the evidence that prenatal exposure to language has only lower frequencies available due to the attenuation of higher frequencies *in utero* (Mehler et al., 1988). Additionally, since lowpass filtering (e.g., at 400 Hz) also removes spectral information such as most formants and consonants like sibilants, it is also assumed that lowpass filtering removes phonetic and phonotactic information but retains prosodic

(and as a subset of prosody, rhythm) information. These assumptions motivate behavioral studies such as Nazzi et al. (1998) to low-pass filter the stimuli as a way to retain prosodic information while greatly removing spectral information. The fact that infants were able to discriminate between rhythmically different languages after lowpass filtering has been taken as evidence that they used suprasegmental, and likely rhythmic, cues for this discrimination task.

However, lowpass filtering speech is not well-justified to extract rhythm, and the exact purpose of using lowpass filtered stimuli to examine rhythm perception remains ambiguous. While lowpass filtering removes segmental information, it retains information like pitch contour, which is known to be a factor that contributes to human language discrimination (Chong et al., 2018). Additionally, since the cutoff frequency of 400 Hz may be higher than some vowels and different numbers of harmonics depending on the speakers' pitch, the exact content being removed and retained from the speech stream is unclear. As a result, low-pass filtering may not be the best way to test for rhythm sensitivity.

Another manipulation on speech was to resynthesize the speech to remove confounding segmental or prosodic information. In Ramus and Mehler (1999b), French adults were able to discriminate between English and Japanese when speech is resynthesized to "SASASA" with pitch contour removed, but not when speech is resynthesized to "AAAA" with original pitch contour. The only two differences between these two conditions are the removal of sibilants, which are high-frequency, and the addition of pitch contours, which are available in lowpass filtered speech. However, for the "AAAA" condition which is more similar to lowpass filtered speech, discrimination between English and Japanese was not achieved. This suggests that low-pass filtered stimuli, which is conceptually similar to the "AAAA" manipulation, may not be the best manipulation for the observed language discrimination effect. On the other hand, the

8-channel spectrogram in our study greatly reduced spectral information such as pitch and formants, but retains the difference of e.g. vowels and sibilants, and is more conceptually similar to the "SASASA" manipulation. Additionally, it is shown that pitch contour may play a role in infant language discrimination as well. In Chong et al. (2018), 7-month-old infants' discrimination between English and German was attributed to sensitivity to different intonation (i.e. pitch contour) between the two languages. These results indicate that the perception of rhythm or temporal regularities in speech may be separate from intonation or pitch contour, where lowpass filtering is not a technique that can tease apart the two. In addition to behavioral studies, Tilsen and Arvaniti (2013) used a data-driven approach to analyze the envelope of various languages and found correlates of syllable-level and stress-level rhythm, but critically, speech envelopes in this study were obtained by high-pass filtering at 400 Hz to remove the influence of energy from the fundamental frequency on the envelope. This further questions whether frequencies below 400 Hz serve as a useful cue to speech rhythm, or rather contain separate information such as pitch change.

In our simulations, we tested the model with both unfiltered and low-pass filtered speech to address the practice in behavioral studies. Compared with tests performed on unfiltered stimuli, models tested on low-pass filtered stimuli generally showed weaker cross-linguistic differences in language discrimination. In addition to the arguments above that cast doubt on the validity of using lowpass filtered speech to examine language discrimination and rhythm, we also note some more constraints specific to the nature of the computational models used. In generative models (which GMMs and i-vector models belong to), stimuli are assumed to be generated from hidden distributions, whose parameters are learned through training. When the models were trained on original speech but tested on lowpass filtered speech, the test stimuli differed greatly from the

distribution of the training data. As a result, the models would assign low likelihood for all values in the test data, making it hard to compare any two utterances for language distance. On the other hand, since newborn infants are exposed to speech *in-utero* in a manner that is similar to hearing lowpass filtered speech, the lowpass filtered stimuli may not be out of distribution for them. Also, additional normalization in perception may be present to help infants extract information different from lowpass filtered speech, which is an extra step that is missing in the current computational models. Due to the difference between machines and humans, we expect the models to perform in a more noisy way with lowpass filtered stimuli.

Some behavioral studies also use backward speech, to which infants do not show crosslinguistic differences in language discrimination, to show that the temporal structure of real speech is important for discrimination (Mehler et al., 1988). However, backward speech would be identical to forward speech in terms of any known acoustic correlates (e.g., %V, ΔC, amplitude modulation). If we expect infants to be sensitive to language differences because of these metrics, then infants should be able to discriminate when speech is played backward. Therefore, the most likely explanation is that backward speech is out of distribution for infants, similar to how lowpass filtered speech is out of distribution for the computational models in question. Since backward speech is not an ecologically valid listening experience, infants may not be able to extract the segmental or suprasegmental cues in the same way as when they listen to naturalistic speech. Similarly, if the temporally scrambled stimuli were able to be reconstructed in the time domain and played to humans, it would likely sound jarring and difficult to extract any meaningful information for discrimination judgment.

Lastly, conceptually similar results were observed in visual language discrimination. In visual language discrimination (Weikum et al., 2007), infants are tested on their ability to dis-

criminate between different languages, but from silent video recordings of the speaker instead of the auditory speech stream. In follow-up work that has not yet been published, the visual recording was scrambled at 200 ms, with any slower information removed. While 4-month-old infants successfully discriminated between scrambled English and French, 8-month-olds failed to do so (Weikum et al., unpublished data). This offers behavioral evidence to support the hypothesis that younger infants are able to discriminate between languages with the removal of rhythm to some degree. Particularly, older infants' discrimination condition on the intact slower rhythm, which suggests a change in the importance of rhythm along development. This aligns with our findings and further challenges the theory that rhythm sensitivity precedes other cues in language development.

To summarize, although various manipulations have been tested in behavioral studies, in comparing human and machine behavior, it is necessary to consider the ecological validity of the stimuli for both humans and machines. While our models were less consistent in simulating infants' language discrimination when tested on low-pass filtered speech than on unfiltered speech, we have argued in this section that when considered together, experimental controls such as low-pass filtering that have removed various segmental properties from speech still do not provide strong evidence that infants are attending to rhythm in tests of language discrimination.

2.3.2 Relation to other literature

While newborn infants may be using cues other than rhythm in language discrimination, we also examine the relationship between our stimuli and slow amplitude modulations, which have been described to reflect rhythmic differences. Specifically, all spoken languages have been

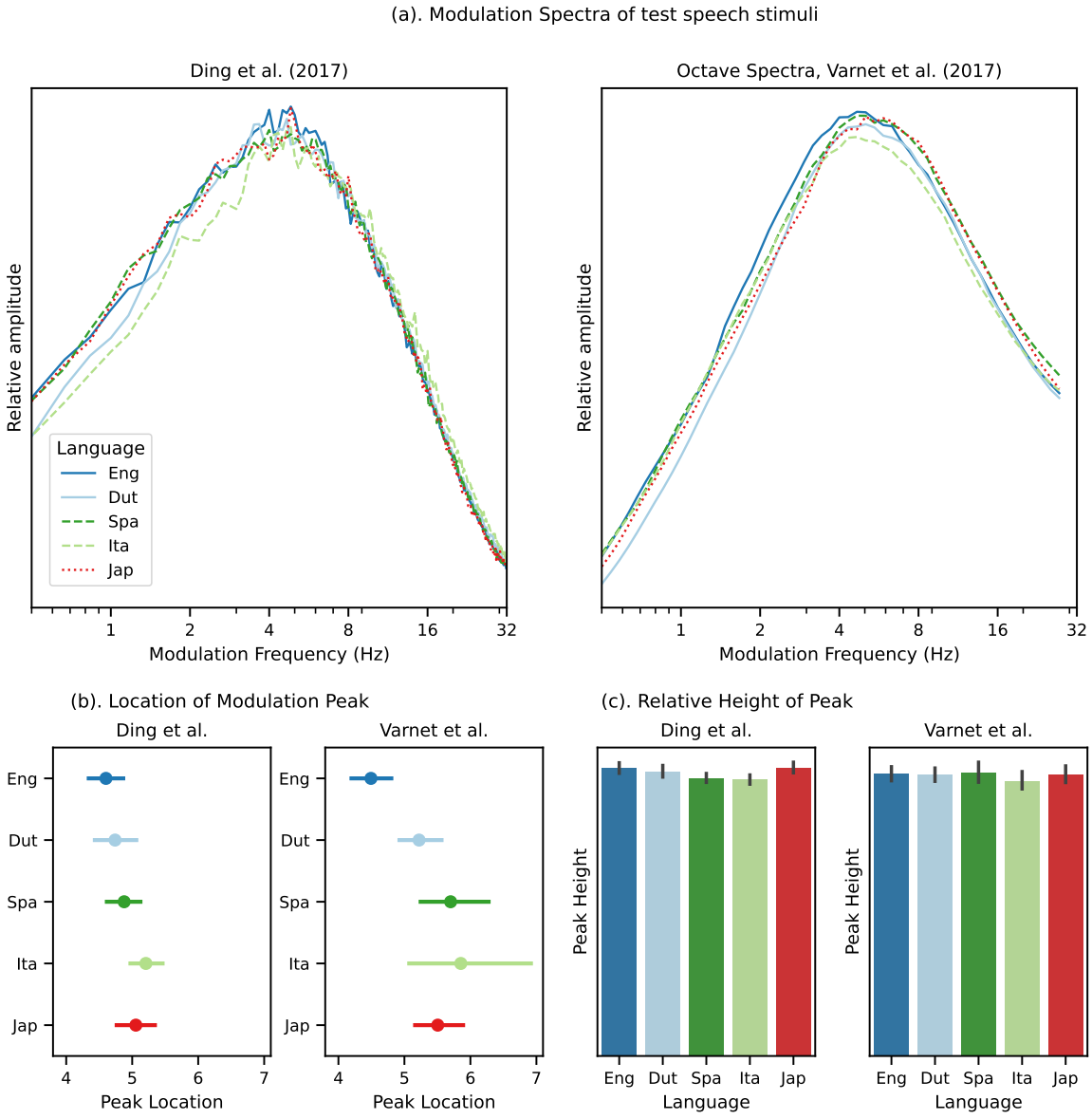

Figure 2.4: **(a).** Amplitude modulation spectra of the test stimuli used in the current study, using the methods from Ding et al. (2017) (left) and the octave spectra method in Varnet et al. (2017) (right) **(b).** Peak locations of the modulation spectra for each test language. Dots represent average over all utterances, and the error bars denote 95% CI. **(c).** Relative height of the peaks. Bars represent average over all utterances, and the error bars denote 95% CI.

found to have amplitude modulation peaks around 4.6 Hz (Ding et al., 2017), where the specific location and height of the modulation spectrum peak was correlated with rhythmic similarity (Varnet et al., 2017). In theory, the location and height of the peak would respectively reflect the average syllable rate and the variability around the average, which would be greater for languages such as English and Dutch. However, evidence exists against the robustness of using such measures (Zhang et al., 2023). Namely, the correlation between language rhythm and measurements of the modulation spectrum is not robust to varied speaking style and less than several minutes of speech. In our simulations, we also analyzed the amplitude modulations using the approaches from both Ding et al. (2017) and Varnet et al. (2017) (see Figure 2.4). We discovered little difference between all five languages in terms of the location and height of the peak, and the variability in the location and height of the peak is not organized by rhythmic typology. Therefore, for the current set of stimuli, while our models are able to replicate human perception, the amplitude modulation spectra do not correlate with rhythmic similarity. We therefore present a case where languages did not differ in temporal modulation at the syllable rate, but are nonetheless discriminable by models tested in this study, many of which do not have access to any temporal regularities.

In the cognitive neuroscience literature, there are a few related studies that explored crosslinguistic differences in early language perception through neural recordings. Our simulation results are compatible with their observations. For example, during passive listening, infants' EEG activity was shown to be different in gamma band when listening to different languages (Peña et al., 2010), which corresponds to the segmental (phoneme and subphoneme) level of information. Meanwhile, no difference was found in slower modulations that correspond to syllable-level rhythm, such as the delta and theta bands (Giraud and Poeppel, 2012). In other passive listen-

ing studies to multiple languages, while both evoked and phase-based analyses were performed on the EEG data, only evoked responses were found to be different across languages, while the phase-based analyses stayed the same (Barajas et al., 2021; Nacar Garcia et al., 2018). Based on the presumed connection between rhythm and amplitude modulation (Varnet et al., 2017), differential markers should be expected in frequency-domain analyses instead of evoked responses, since the presumed amplitude modulation difference across languages were found in the frequency domain, and a significant evoked relationship can indicate anything in the speech stream, including anything prosodic or segmental. Nonetheless, neural recording studies that are more closely related to language discrimination are needed to arrive at stronger conclusions.

While our experiments call into question whether rhythm is the driving factor for language discrimination, they do not challenge whether infants are sensitive to rhythm at all. In fact, infants are sensitive to even pure-tone rhythm patterns subject to the iambic-trochaic law even at birth (Abboub et al., 2016). Unlike naturalistic speech which is complex and contains many confounding factors, the controlled pure-tone stimuli lead to the same representation after scrambling. Therefore, infants' behavioral effect must arise from actual sensitivity to temporal regularity. However, the specific connection between the observed effect on pure tones and older infants' ability to segment words from the speech stream (Jusczyk et al., 1999; Polka and Sundara, 2012) should be considered more carefully.

2.4 Conclusion

This chapter has argued that newborns' ability to discriminate rhythmically different languages does not provide strong evidence of their sensitivity to rhythm. Our simulations show

that a wide variety of models that can predict the relative ease with which newborns discriminate languages can still do so when rhythmic information was eliminated from the stimuli. We then argued that across various manipulations with intact and low-pass filtered speech, global properties such as the percentage of vowels are equally accessible and can be driving the discrimination behavior. As the only direct acoustic correlate that is directly compared with human performance contain these summary statistics that are only related to rhythm in theory (Ramus et al., 1999), existing evidence about rhythm being important in early language discrimination is not strong enough. Lastly, we discuss our results with respect to amplitude modulation, a measure of temporal structure in speech that has been found to correlate with rhythmic classes of languages, is unlikely to be available as a cue in the types of tasks that are used with infants. Together, these results call for a reconsideration of the nature of language discrimination in human infants.

As theories of language acquisition often center around rhythm, our results push the field towards a reconsideration of the role of rhythm in early language acquisition. Specifically, speech rhythm has been considered as one of the earliest accessible cue to infants, and has been presumed to be learned by the infants prior to learning segmental and lexical information. As our results suggest that speech rhythm may not be relevant to language discrimination, or even accessible to infants, the role of rhythm in language acquisition may be different than what has previously been supposed.

Chapter 3: A neurally grounded front-end for deep audio processing

3.1 Overview

In Chapter 2, I simulated newborn language discrimination using models directly drawn from machine learning and automatic speech processing. From the representations of these machine learning models, the results suggest that humanlike language discrimination can be achieved without rhythm at all. While I discussed the implication this have for speech perception and representation in humans, it is also important to note that the representations and processing of these models differ from what we know about human auditory perception. One important distinction is that most models only ever have access to spectrogram-like features. While spectrograms, when binned along the mel scale and with logarithms taken, have a close resemblance to the representation humans have for further processing of speech (Rahman et al., 2020), it may lack some temporal fine structure that humans have access to due to the removal of phase information in general. Additionally, putting spectrogram features through clustering algorithms or neural networks allows any generic statistics from spectrogram features to be learned, while computations performed by the human auditory system may be more specific and constrained. Additionally, while some clustering models have access to MFCCs, which involves additional processing in the cepstral domain, the cepstral features were often seen as temporally independent frames by the clustering algorithms. This assumption greatly differ from human perception,

where spectral information is perceived in the temporal order.

To address these inherent differences between the current models and human perception, I introduce a differentiable and lightweight audio front-end informed by auditory neuroscience from the ear to the brain. It allows joint learning of cochlear and cortical parameters, as well as parameters of a backend neural network. I implemented the model and showcased example applications, highlighting the models' robustness and interpretability.

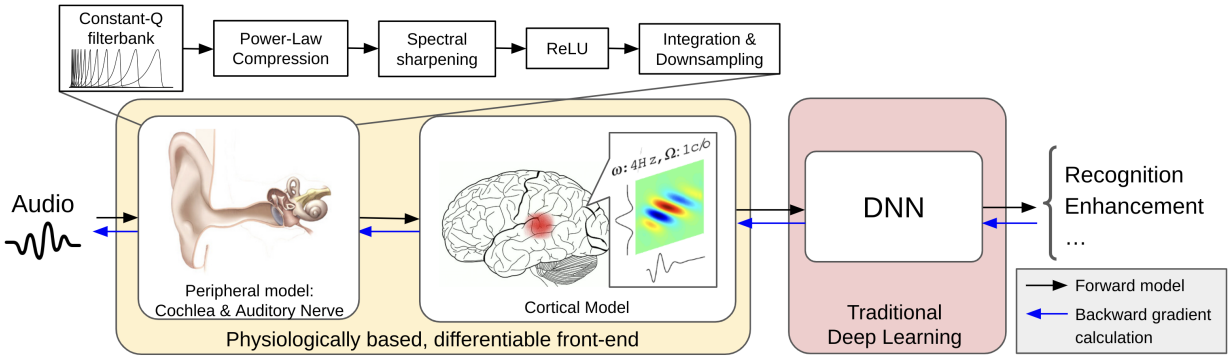

Figure 3.1: Auditory processing from cochlea to cortical representations is shown from left to right. Black arrows indicate the forward model. Blue arrows indicate the direction of gradient calculation using the chain rule.

3.2 The auditory processing model

The differentiable model is based on the forward model in (Chi et al., 2005; Elhilali, 2004; Shamma et al., 1986), which is one of a host of models for peripheral auditory processing. These models vary considerably in the details of the various stages of filtering and transduction (Farhadi et al., 2023), with emphasis on cochlear mechanics (de Boer and Nuttall, 2000; Skrodzka, 2005), hearing-aid design (Moore, 1996; Zhang and Gan, 2011), or simplifications towards efficient engineering applications (Ravanelli and Bengio, 2018; Zeghidour et al., 2021). The version used here provides a mathematically tractable model of early auditory processing that supplies all the

essential details but remains mathematically tractable. To foster replicability and new applications, the code and parameters are available at https://github.com/pirl-lab/diffAudNeuro.

3.2.1 Step 1. A Model of the Cochlear and Peripheral Hearing

The first stage performs computations of the auditory periphery, and converts input audio into a time-frequency representation, in four steps: a filterbank that performs frequency decomposition, nonlinear compression, lateral inhibition network (filter along frequency dimension), and lowpass pooling. The filterbank step consists of a bank of constant-Q filters, corresponding to frequency decomposition in the cochlea. The filters were shaped to have a frequency domain response that approximates a rounded exponential (roex) function (Chi et al., 2005; Lyon, 2017):

$$\text{if } x \in [0, x_h],\ |H(x)| = (x_h - x)^{\alpha} e^{-\beta(x_h - x)} \text{ else, } H(x) = 0 \tag{3.1}$$

where x_h represents the highest frequency in the passband for one filter on a normalized frequency scale, and $\alpha = 0.3$ and $\beta = 8$ are constants chosen to match the lower and upper skirts from neurobiological measurements (Allen, 1985), and resemble the asymmetric sharp-cutoffs around the center frequency seen in cochlear filters in psychoacoustics and physiology (Bowman et al., 1998; Glasberg and Moore, 1990).

This is followed by power-law compression, a lateral inhibition step highlighting change across frequencies, and short-term integration. These mimic processes in the cochlea and auditory nerve. In compression, I used $y = x^{\alpha}$ where α is a learnable parameter initialized at 1.0, and x and y are the input and output. Lateral inhibition was done via a linear filter (initialized at $[1 - 1]$) along the frequency followed by ReLU, which detects positive changes. For short-term

integration, a leaky integrator with output $y(t) = e^{-t/\tau}x(t)$, with time constant τ was used. The output is downsampled to 200 Hz.

The resulting auditory spectrogram differs from the conventional STFT spectrogram as it approximates a wavelet transform of the signal, better matches human auditory processing (Chi et al., 2005), and preserves spectrotemporal details for cortical analyses (Shamma and Lorenzi, 2013). Further, parameter choices of this model have been shown to improve performance in different listening environments (Shamma and Lorenzi, 2013), making the model suitable as a differentiable frontend, where task-specific parameters are learned.

3.2.2 Step 2. Cortical Features

In the cortical step, joint spectrotemporal modulation features are extracted via filters from the auditory spectrogram. Each filter is characterized by spectral (Ω) and temporal modulations (ω). The spectrogram pattern that maximally excites the cortical spectrotemporal filter (or the STRF) roughly takes the shape of a Gabor kernel parameterized by Ω and ω. The activation $r_f(t)$ for a single cortical neuron, tuned around a frequency f, at time instant t can be computed as a function of frequency, scale, and rate by convolving the input auditory spectrogram $s(f, t)$ with the STRF kernel:

$$r_f(t) = \text{STRF}(f, t; \Omega, \omega) *_{f,t} s(f, t) \tag{3.2}$$

where $\text{STRF}(.)$ stands for the 2D modulation band-pass filter.

This model is based on neuroscience experiments that explored transformations from the auditory spectrogram to cortical representations (David et al., 2007), where modulation filters

were designed to capture the important features in the cortical neural response. In this way, these filters decompose modulations of the auditory spectrogram into different pass-bands over a range of spectrotemporal modulations. This closely resembles the scattering transform (Bruna and Mallat, 2013; Mallat, 2012).

3.2.3 Making the Model Differentiable

I implemented the pipeline from waveform to cortical representations in a fully differentiable way using JAX (Bradbury et al., 2018). Our implementation is conceptually based on Chi et al. (2005), adding vectorization to utilize GPU speedup. Additionally, all IIR filters were implemented in the Fourier domain, which is more tractable for gradient calculation than recursive filtering in the time domain. This allows parameters to be updated through back-propagation, using training schemes typical for neural networks such as mini-batching and optimizers.

The parameters were initialized with values usual in auditory models, and updated jointly with backend neural network parameters through back-propagation. The values for the poles and zeros of the constant-Q filterbank are taken from Chi et al. (2005). All values were initialized to 1.0 for power-law compression, $[1 - 1]$ for the lateral inhibition filter, and $\tau = 8$ for short-term integration. In the cortical stage, 40 spectrotemporal filters were initialized with both classical values ("log-spaced", spaced logarithmically with more values at low spectral and temporal modulations) as well as uniform random initialization ("random" within the range of $(0, 9]$ in spectral and temporal modulations). Through training, all the parameters initialized above were allowed to be freely updated except for the parameters in the constant-Q filterbank, since allowing these updates would generate too many degrees of freedom.

The choice of the auditory model on which the differentiable model was based balances the adherence to biological principles and simplicity. As prior work (Meng et al., 2023) showed that cochlear filterbank parameters differed little from the Mel initialization when they are made learnable, we opt for a more complex and biologically plausible filterbank structure and do not update its values. In this way, it adheres closely to neural processes and offers more stable training and interpretability. For compression, the power-law compression only costs one parameter per frequency channel, and is a simplified version of PCEN (Wang et al., 2017). For lateral inhibition and short-term integration, the parameters were shared across channels, covering all channels with 3 parameters. We chose to focus the learnable parameters at the compression step based on previous work showing its relative importance in audio frontends (Meng et al., 2023). The frontend in the current model has just 212 learnable parameters, and can be tweaked to increase or reduce parameters as needed.

3.3 Applications of the Differentiable Frontend

In this chapter, I show results on two types of tasks: classification (phoneme recognition) and enhancement (speech enhancement). I selected these tasks for the following reasons. First, both tasks are of interest to both technologists and neuroscientists, which means the parameters can be interpreted using theories and human data (see §3.4). Second, instead of limiting our front-end to classification like (Zeghidour et al., 2021), I also applied it to speech enhancement (an end-to-end problem). While some approaches (e.g., Conv-TasNet, Luo and Mesgarani, 2019) used learnable frontends, we show the benefits of constraining them to biologically plausible operations.

Cortical features that extract different spectrotemporal modulations from the acoustic signal could improve recognition of phonetic categories (Mesgarani et al., 2010) and separate speech from various types of noise (Mesgarani and Shamma, 2007), based on the hypothesis that different phonemes and speech sources generate distinct spectrotemporal modulation profiles, and thus are separable in the cortical representation. However, such feature-engineering is now less favored as they are outperformed by end-to-end methods. Here, we show that our differentiable model achieves comparable and often better performance compared with larger data-driven models. Additionally, the differentiable model shows superb robustness compared with their end-to-end counterparts.

3.3.1 Phoneme Recognition

We test the differentiable frontend on phoneme recognition, where sound categories are predicted from the speech signal. While accuracy in phoneme recognition often positively correlates with ASR performance (Oh et al., 2021), phoneme recognition allows testing on small models without needing a language model.

Phoneme recognition with our differentiable frontend takes in waveform input and outputs the spectrotemporal modulation values, which are fed into a 3-layer CNN, followed by a linear layer projecting to the number of phonemes (14.6k total parameters). The CNN layers have 3x3 filters with 10, 20, and 40 channels, each followed by GeLU. All parameters are learned jointly using the Adam optimizer, initial learning rate of 0.001, a batch size of 4, and for 200k steps. These parameters were obtained through a small grid search, although the results are qualitatively similar to other hyperparameters.

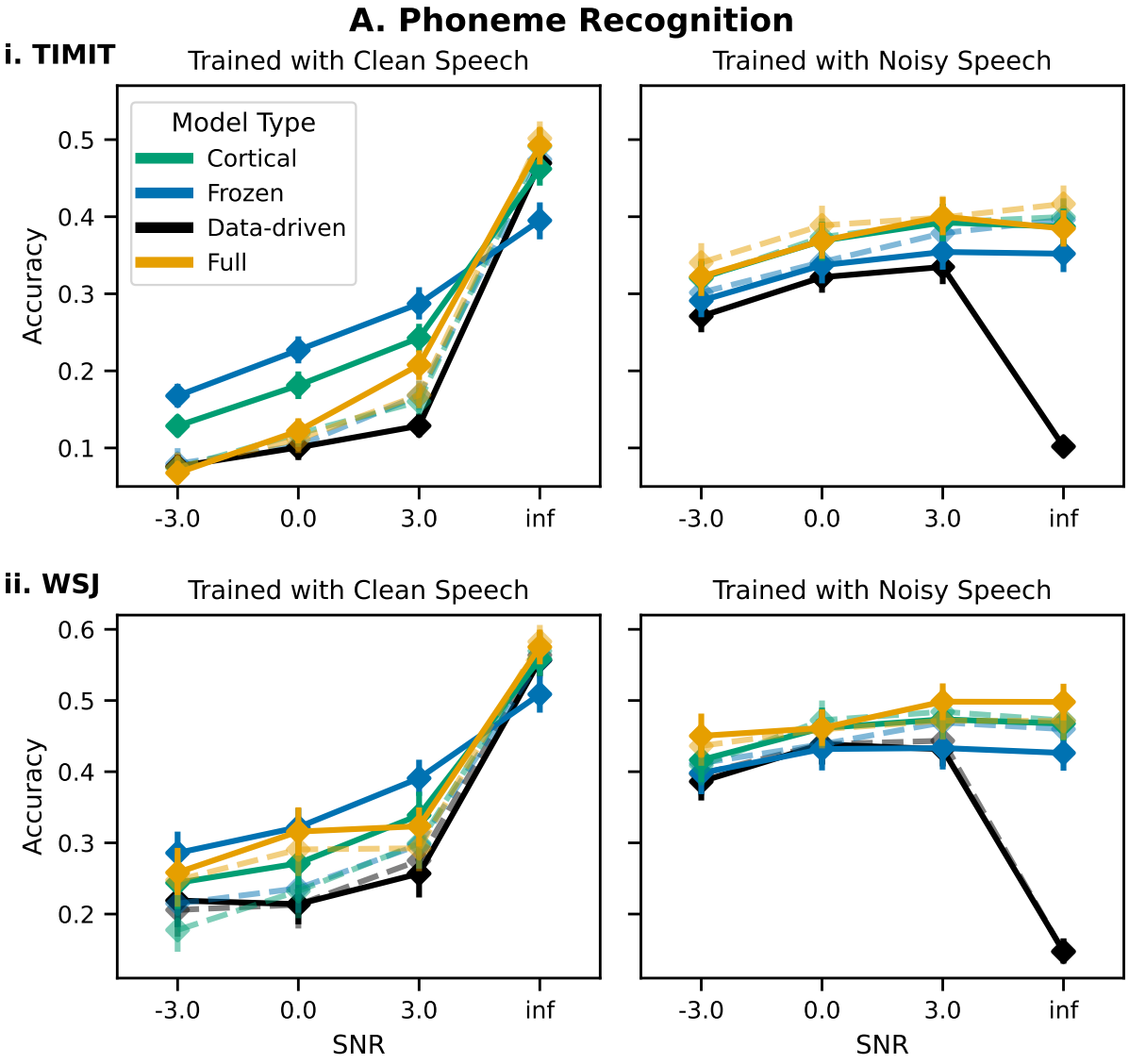

Figure 3.2: Results for Phoneme Recognition. The x-axis shows test conditions where test speech was mixed with pink noise at -3, 0, and 3 dB SNR or clean. The y-axis shows classification accuracy. Solid lines denote models initialized with random cortical parameters, and dashed lines denote models initialized with log-spaced cortical parameters. Error bars denote 95% CI.

Ablation-test: In the *CNN* model, I replaced the cortical step with an extra 3×3 convolution layer with 40 channels, increasing the number of parameters since each cortical filter contains two parameters, while each CNN filter contains 10. This architecture performs feature extraction in a more data-driven way since the filters are not bound to bandpass the modulations.

In the *Frozen* model, both cochlear and cortical parameters are frozen during training and not updated, resembling the classical feature-engineering approach. In the *Cortical* model, only cortical parameters were updated.

The model was trained and tested on two corpora, TIMIT (Garofolo et al., 1993) and WSJ (Paul and Baker, 1992) (with phone labels generated from forced alignment (Povey et al., 2011)). The training data for TIMIT consists of 4620 utterances, and WSJ was randomly subset to 4940 utterances to keep the duration comparable. We trained the model on a modest GPU (a single RTX2080ti with 5 GB RAM), needing up to 12 hours of training time. In each training sample, one second of speech is randomly sampled from the training data. All models are tested on 100+ utterances of hold-out data (test split in TIMIT and randomly sampled for WSJ).

The test accuracy is shown in Fig. 3.2. Here, I highlight two observations. First, differentiable models (Full and Cortical) outperformed their data-driven counterparts (CNN) in most cases. This shows the advantage of differentiability in performance and robustness, and aligns with observations in classical (Mesgarani et al., 2010) and differentiable (Vuong et al., 2020) modulation-based models. Second, choices in parameter initialization (log-spaced vs random) influenced the models' performance. When trained with clean speech, the randomly initialized models generalized better to noisy conditions. This is supported by the distribution of the learned cortical filters, which I will elaborate on in Section 3.4.

3.3.2 Speech Enhancement

The model was tested on speech enhancement in a low-resource setting. I selected this task to show that differentiable frontends need not only be used for classification, as they were in the

literature Vuong et al. (2020); Zeghidour et al. (2021), but also in end-to-end tasks. Also, while separation and enhancement algorithms are often trained with big models and data (Scheibler et al., 2023), real-world hearing-aid/headphone processing scenarios require small models and data due to deployment constraints for personalized denoising. Here, we consider a task where speech needs to be separated from music. This setting is particularly relevant to our architecture since modulations are informative about auditory objects (Ding et al., 2017; Elhilali and Shamma, 2008; Gu and Stern, 2008).

Table 3.1: SI-SDR for speech enhancement, higher is better. Best models in each row are bolded. 95% CI are shown in brackets.

	Full		Cortical		Frozen		CNN
Initialization	Log	Random	Log	Random	Log	Random	-
Original	8.28(.24)	**8.31(.23)**	8.05(.24)	7.83(.2)	7.80(.24)	7.60(.2)	7.91(.25)
NewTarg.	**5.85(.19)**	5.09(.2)	5.30(.19)	5.09(.2)	5.22(.19)	4.57(.2)	5.39(.2)
NewNoise	**16.2(.19)**	15.5(.2)	16.0(.2)	15.6(.2)	15.2(.19)	15.0(.2)	16.1(.21)

The model uses the differentiable front-end followed by a CNN. Instead of classification labels, this CNN outputs a mask to be multiplied element-wise with the complex input STFT spectrogram (with window length of 256 and hop length 80), which is transformed back to waveform domain. The CNN has four layers with 20, 40, 10, 1 channels, with GeLU after each layer. The CNN outputs to a fully connected layer that projects the 129 frequency channels to the 256 STFT frequency channels, followed by sigmoid activation, to generate masks. As the loss function, I used the sum of L1 waveform loss and L1 multi-scale complex STFT spectrogram loss at window lengths [256, 512, 1024] and hop lengths at $1/4$ of the window length as loss function. For training, two hours of speech was used (from one female speaker, WSJ S002, 1000 utterances) and music ((Rafii et al., 2019), 50 songs). In each training sample, 1 second of speech

and music were randomly selected and mixed at 0 dB SNR. Other hyperparameters are identical to phoneme recognition. At test time, the model was evaluated on holdout data, as well as new distributions involving a new target speaker (WSJ S001, male) or a new type of noise (car noise from DEMAND (Thiemann et al., 2013)). The model in each condition was tested with 500 samples mixed at 0dB SNR.

The results (Table 3.1) show that the fully differentiable frontend was superior to all other ablations and data-driven counterparts. The log-spaced initialization significantly improved model performance. This is the opposite of the phoneme recognition results, which I will discuss further below.

3.4 Explainability

In addition to efficiency and robustness, since the differentiable front-end is based on neuroscientific processes, the parameters are directly interpretable. Here, we focus on the cortical parameters, while the full parameters including the learned cochlea parameters are published on our GitHub site. As the cortical stage band-passes in modulation domain, it serves as a bottleneck that discards modulations irrelevant to the task. Furthermore, the values of the parameter directly match different features in the audio: high spectral modulation (> 5 cycles/octave) corresponds to narrow bandwidths related to spectral harmonics and pitch, while low spectral modulation (< 5 cycles/octave) is related to spectral envelope information such as formants and timbre (Elliott and Theunissen, 2009). Thus, the learned parameters (shown in Fig. 3.3) are indicative of how the model performs the task.

The performance difference between two different initialization methods in the two tasks

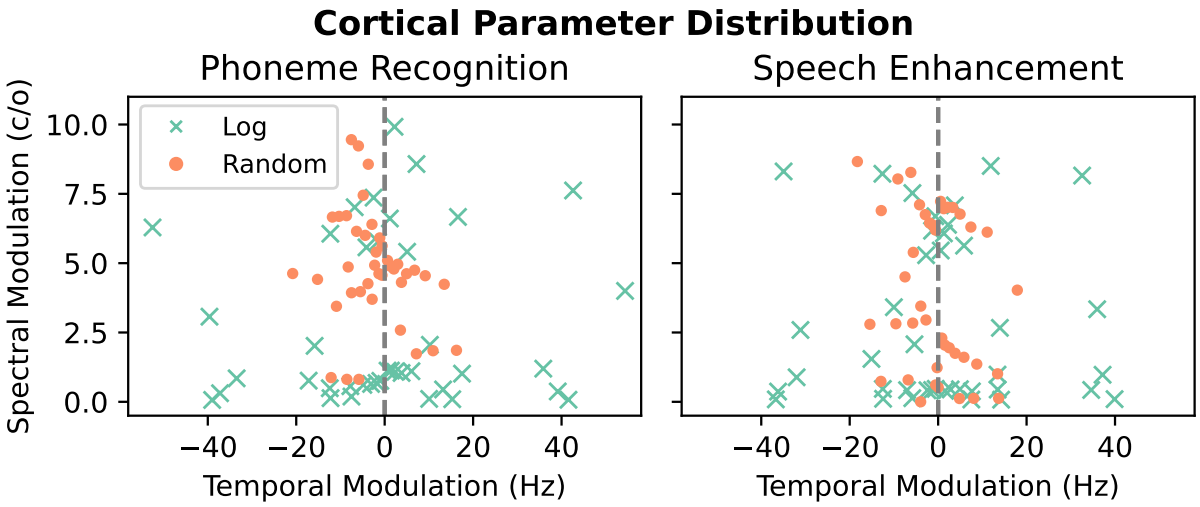

Figure 3.3: Distribution of trained spectrotemporal filter parameters. Left: phoneme recognition in quiet; right: speech enhancement. Each point represents one of 40 filters. Two models are shown in each panel, respectively initialized from log-spaced values (cross) and randomized values (dot). The sign of temporal modulation encodes the direction of the spectrotemporal modulation – positive indicates upward-tilting Gabors and vice versa.

can be explained by the difference in the learned filter distributions. In Fig. 3.3, the log-initialized models converged to filters spanning across a wider range of temporal modulations, up to > 50 Hz, which are unrealistic for human cortical analysis or for modulation energy distribution in speech (Elliott and Theunissen, 2009). Performance-wise, the randomly initialized model, which converged towards realistic distributions, yielded better results. This observation aligns with prior differentiable modulation models in Vuong (2023), that successful models converge to lower spectrotemporal modulations. However, in speech enhancement, the log-initialized models outperformed randomly initialized models. In the parameter distribution, the log-initialized model converged to temporal modulations within the biologically plausible range (≈ 30 Hz), and the higher temporal modulations likely helped separate speech from noise effectively through finer temporal detail. In conclusion, the learned output of such differentiable models is sensitive to initialization and worthy of further study. Our results can also inform specific initialization choices.

Furthermore, the learned distribution is closely related to the nature of the task. In the speech enhancement model, spectral modulations were contained within a lower range (< 8 cycles/octave), which we attribute to the model learning the higher pitch of the target female speaker with lower spectral modulations.

3.5 Discussion and Conclusions

In this chapter, I present a differentiable front-end model of auditory processing combining signal processing, neuroscience, and deep learning, and apply it to speech classification and enhancement tasks. Compared with traditional deep speech processing, the advantages of the differentiable model approach include the following. First, the differentiable frontend is lightweight (212 parameters total) and can be adapted to tasks with only a few hours of training data. Secondly, compared with non-differentiable and data-driven counterparts, the differentiable frontend achieved better and generalizable performance. Lastly, the model parameters are interpretable. As such, this frontend can be used for low-resource settings such as personalized denoising, where only minutes or a few hours of user data can be collected. Additionally, since the model only employs linear filters, it can be adapted to be fully convolutional for arbitrary speech durations on modest hardware.

Audio and speech processing has seen revolutions in performance since deep learning has become an inherent part of the model architecture. However, as deep learning scales up with model size and data, training becomes slow and expensive, and the trained models are "black-box" and difficult to explain. In comparison, the human auditory system can effortlessly perform diverse tasks after having listened to much fewer hours of audio or speech compared with a

modern-day pretrained model.

The current model presents a middle ground between non-differentiable, spectrogram-based models and fully data-driven, end-to-end frontends (e.g., Luo and Mesgarani, 2019; Baevski et al., 2020). Among similar cochlear (Lyon et al., 2024; Zeghidour et al., 2021) and cortical (Vuong, 2023) models, our approach is rooted in auditory neuroscience, employing a biologically plausible filterbank and joint learning of cochlear and cortical stages, making it a better candidate for interpretability and hearing aid fitting. Also, our model does not employ any recursion, making back-propagation fast and stable. Lastly, the parameters made differentiable were informed by prior work (Meng et al., 2023), pruning hundreds of parameters that deviate little after initialization.

The model can be further optimized in the implementation of forward model and gradient calculations. Although the current implementation focuses on differentiability, the model is fully compatible with CPU-based processing in deployment by simply changing the hand-engineered parameters with those learned through training. For this purpose, I have published Python implementation based on NumPy and SciPy on GitHub. Also, gradient calculations could benefit from custom gradients to improve time and memory efficiency, where auto-differentiation increases the memory footprint in an uncontrolled fashion for novel or complex functions.

Since this differentiable model is built based on theories of auditory neuroscience and human perception, this model can also be applied to hearing loss measurements and audio personalization. The cochlear parameters (e.g., compression) are directly related to hearing loss in different frequencies, allowing hearing loss to be characterized from user-collected data. Additionally, our forward model that predicts cortical responses can be adapted to be fitted with brain recordings, some of which are indicative of hearing loss through only a few electrodes

(Shinn-Cunningham et al., 2017). This allows us to fit parameters from brain recordings instead of cochlear measurements (Drakopoulos and Verhulst, 2023), allowing for more degrees of freedom. This would lead to new types of supervised listening device fitting.

Chapter 4: Applying the differentiable auditory front-end to cognitive modeling: A case on language discrimination

4.1 Overview

In Chapter 2, I simulated language discrimination with various cognitive models, but all models become more humanlike with temporal information removed. This contradicts the theory that rhythm — the ordered temporal structure in speech — drives the human behavior of language discrimination. Such contradiction can be attributed to two potential reasons: language discrimination in humans is also not driven by rhythm, or at least not rhythm alone. Alternatively, all the computational models tested in Chapter 2, which were chosen based on its relevance in this literature and spanned across a wide representation of algorithms, process speech differently from humans. In the discussion section of Chapter 2, I explored the former possibility and related it to other supporting evidence in the literature. In this chapter, I consider the alternative, that the existing models we used process speech differently from humans, and therefore produced results that did not generalize to human behavior.

In the current chapter, I simulate language discrimination using the differentiable model developed in Chapter 3, which is a neuro-biologically grounded model. Particularly, the cortical part of this model processes speech by breaking the time-frequency representation down into

different spectrotemporal modulations. As reviewed in Chapter 3, this was first introduced as a model of the receptive fields of a single neuron in the auditory cortical processing of animal models. Since then, STRFs have been used as a linear and simplistic model of subcortical and cortical processing in cognitive neural recording as well (Hullett et al., 2016; Jenison et al., 2015). Additionally, introducing STRF-related processing has improved a series of machine speech processing tasks (Mesgarani and Shamma, 2007; Mesgarani et al., 2010). Here, I adapt the differentiable cortical model, which performs STRF-based feature extraction with fully learnable joint spectral and temporal modulations, as a part of a cognitive model to study early language discrimination. Contrasting previous cognitive models of speech perception that commonly use input features based on MFCCs (Carbajal et al., 2016; Schatz et al., 2021b) and modern-day deep learning models that commonly use log-mel spectrograms (Chung et al., 2020; Radford et al., 2023) the addition of a differentiable cortical model as the frontend not only provides a constraint to the model by limiting audio processing to be more similar to that of humans, but also allows the cortical parameters to be learnable, hence adding flexibility and interpretability to the model.

In the context of early language discrimination, the choice to add the differentiable cortical model was also motivated by its unique relevance to speech rhythm. By definition, the cortical STRF model analyzes audio stimuli by breaking them down into a set of joint spectrotemporal modulations, of which the temporal modulations are closely related to the notion of speech rhythm. As reviewed in Chapter 1, the prevailing theory of language discrimination states that the behavioral effect is driven by infants' sensitivity to syllable-level rhythm (Mehler et al., 1988; Nazzi and Ramus, 2003). It has also been shown across studies that the speech has an amplitude (i.e., temporal) modulation peak at around 4.5 Hz crosslinguistically (Ding et al., 2017; Varnet et al., 2017), which coincides with the range of syllable level rhythm. Additionally, from stud-

ies studying the spectrotemporal modulation composition of naturalistic speech, it was found that spectrotemporal modulations in speech are most concentrated in ranges that are low in both spectral and temporal modulation, and mostly below 4.5 Hz in temporal modulation (Elliott and Theunissen, 2009). In a study where a differentiable linear autoencoder was trained to decompose speech (in the form of log-mel spectrogram) into a limited number of STRF representations and reconstructed, the learned STRFs also converged to low temporal modulations (Vuong et al., 2021), which corresponds to the observed spectrotemporal modulation distribution in speech. Lastly, newborns are found to be sensitive to the stress pattern (trochaic vs. iambic) between pure tones alternating on pitch, duration, or amplitude at roughly the syllable-level rate (Abboub et al., 2016). While the summary statistics proposed in Chapter 2 would be the same between trochaic and iambic pure tone stimuli, the STRFs would be distinct in either the combination of temporal modulations (on duration and amplitude variations) or joint spectrotemporal modulations (for pitch variations). This suggests that STRF-based cortical features could well align with existing knowledge of newborn auditory perception as well. These prior results justify the application of STRFs to study slow, syllable-level temporal regularities and particularly, their role in early language discrimination.

4.2 Model Overview

I explored two types of architectures to incorporate the differentiable auditory front-end to simulate the early exposure to speech in infants and their learning from such exposure. Below, I introduce the architecture and training regime of each model.

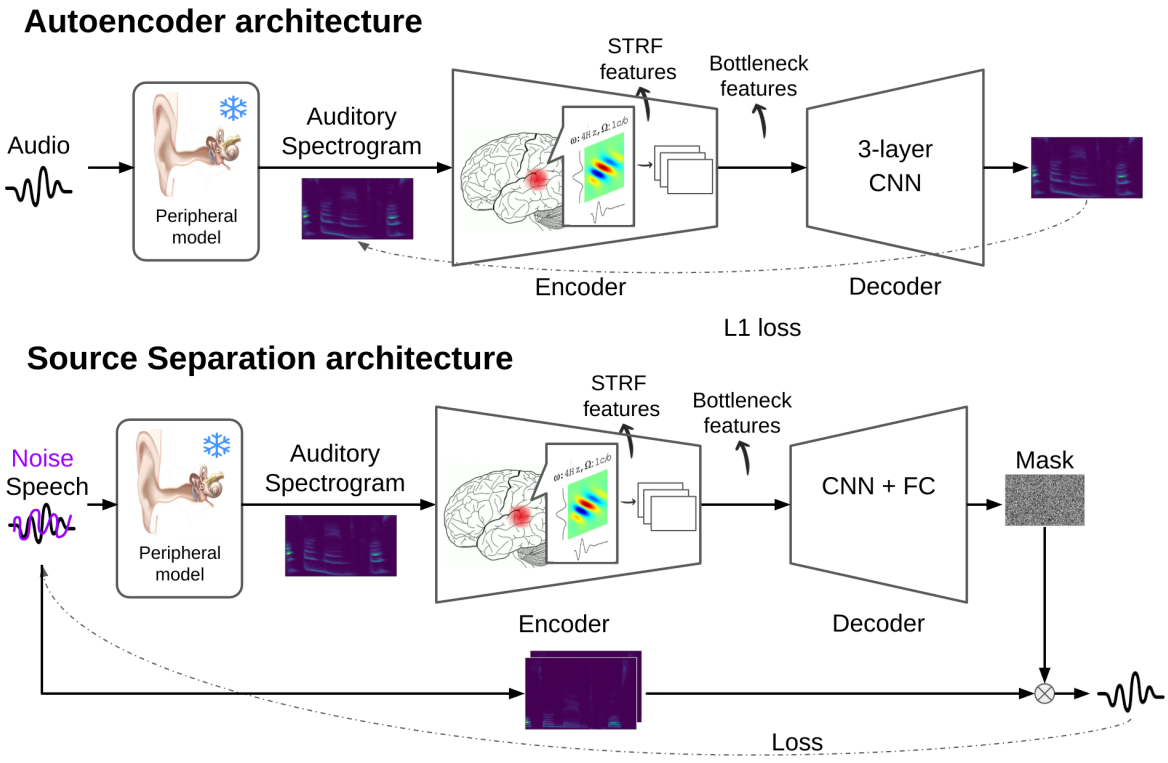

Figure 4.1: Architecture of the autoencoder model (top) and source separation model (bottom).

4.2.1 Autoencoder Model

The autoencoder model aims to reconstruct the original speech material when there is a bottleneck (see Figure 4.1) in the middle of the model architecture, where the part of the model before and after the bottleneck are respectively referred to as the encoder and decoder. In this case, I constructed the encoder by using two sequential steps to create an information bottleneck. First, even though the time-frequency representation (i.e., the auditory spectrogram) contains all the original spectrotemporal modulations in clean speech, I constrict the cortical part of the differentiable front-end to only extract 5 STRF output features.[1] This implies that the model needs to learn the most important spectrotemporal modulations in order to reconstruct the original speech. Secondly, right after the STRF layer, I use a convolutional layer to compress the 5-channel outputs from the STRF layer to 3 channels. This further reduces the dimensionality of the input and forces the model to learn a mapping that generates the optimal representation at the bottleneck to be reconstructed in the decoder. The decoder is composed of a 3-layer convolutional neural network with 20, 10, and 1 channels in each layer. 2D convolution was performed in each layer with a `3x3` kernel and stride size of `1`. While some larger models with more channels in the intermediate layers were explored during the architectural search, the results were qualitatively similar. Therefore, to explore the cognitive simulations that a model with as few parameters as the above can achieve, I elected to use the smallest architecture attempted.

The autoencoder model was trained with only the cortical part of the auditory frontend made differentiable. This is because there is little expectation in the literature about how peripheral hearing may change during language acquisition and development. As a result, I initialized

[1] I also explored models trained with 4, 6, 7, or 8 STRFs at the bottleneck. The results were qualitatively similar. In this thesis, I will only report results from models with 5 STRFs.

the peripheral parameters in the model to values typically used in the literature and did not update them during training. The cortical parameters were initialized uniformly randomly between $[0, 9]$ cycles/octave in spectral modulation and $[0, 9]$ Hz in temporal modulation. Therefore, the peripheral part of the frontend can be seen as a feature extractor that preprocesses speech into auditory spectrograms, which serve as the input to the encoder. The decoder's output was trained to match the same auditory spectrogram that feeds into the encoder using L1 loss.[2] The model was trained on the entire set of Globalphone French (Schultz, 2002; around 20 hours) for 300k steps, with a minibatch size of 4. The initial learning rate for all updated parameters were set to be 0.001 and an ADAM optimizer was used during training. In the results, five models trained with different random initializations (controlled by random seeds) were shown.

4.2.2 Source Separation Model

The source separation model was modified from the same model in Chapter 3, reducing the size of the model to match the autoencoder model above. As shown in the bottom half of Figure 4.1, the model takes speech waveform in and passes it through the frontend, with the (frozen) peripheral model acting as a feature extractor and a cortical model that performs STRF feature extraction. Then, the representation is passed through one convolution later to compress the number of channels to 5. The decoder is a 2-layer CNN followed by one fully connected layer along the frequency dimension. The fully connected layer was designed to match any differences in frequency binning between the model's input (auditory spectrogram with logarithmically distributed constant-Q filters) and the mask to be multiplied with the linear-frequency STFT spectrogram.

[2] I also tried L2 loss, with qualitatively similar distribution on the learned STRF distributions; I elected to use L1 loss as the model training was slightly more stable across different STRF initializations.

GeLU activation function was used after each convolution layer, and the fully connected layer was followed by sigmoid nonlinearity to map the model to a mask between $[0, 1]$. The mask is then multiplied by the magnitude of the STFT spectrogram, which was lastly transformed into time domain using iSTFT with the original phase. The STFT and iSTFT used an FFT sample size of 256 and a hop length of 80.

The source separation model was trained using the same hyperparameters (learning rate, optimizer, batch size, and number of steps) as the autoencoder model. To perform source separation, I mixed the target speech (Globalphone French dataset, same as the autoencoder model) with a distractor (noise) that is composed of popular music (Rafii et al., 2019; using the mixed tracks). The input was 1-second samples of French speech and music that were mixed at 0 dB SNR. The model output in time domain is related to the clean speech segment using the same loss function as the source separation model in Chapter 3, with L1 waveform loss and L1 multiscale complex STFT spectrogram loss.

4.2.3 Testing

The trained models were tested in their ability to discriminate between languages. The same stimuli as Chapter 2 were used, where each model discriminated between two pairs of languages (English-Japanese and English-Dutch) and also between groups of languages that were judged to be rhythmically similar within the group (homogeneous) or rhythmically dissimilar within the group (heterogeneous). To achieve humanlike performance, a model is expected to discriminate English-Japanese better than English-Dutch, and the homogeneous groups better than heterogeneous. Similar to Chapter 2, the model's discrimination was tested using the ma-

chine ABX task, where 2000 trials are performed with 1-second-long speech materials that are randomly sampled from the same dataset as Chapter 2.

Different from Chapter 2, I explored four testing methods: cosine distance with normalized dynamic time warping, cosine distance with unnormalized dynamic time warping, cosine distance with mean pooling (where the bottleneck representation was averaged across time), and cosine distance with max pooling (where the max value was taken for each feature dimension along time). Pooling was introduced to test the STRF-related representations because the STRF representations are much larger in dimensionality than the features tested in Chapter 2. For example, 5 STRFs with 128 frequency channels per STRF yielded 640 dimensions, compared with 32 for RNN embeddings and 8 for vocoded acoustic features in Chapter 2. Although the ABX task has been used on high-dimensional data (Li et al., 2020a; Schatz et al., 2021b), the test was performed over tri-phone segments that are usually much shorter than 1 second. The current test setup, which is high in both feature dimensionality and long in duration, likely poses extra difficulty for the test paradigm to obtain crosslinguistic effects beyond chance level. I hope to use pooling to alleviate this problem by removing the time component, and therefore dynamic time warping, from the test.

While in Chapter 2, only cosine distance with unnormalized dynamic time warping showed a robust discrimination effect better than chance, in the models tested in this chapter, the discrimination effect was robust and qualitatively similar across all metrics except mean pooling, where the results were noisy and inconsistent across different models. Therefore, the mean pooling distance metric was excluded from the analysis, and the remaining three metrics are aggregated in the following analyses to simplify the analysis. This was done to decrease the number of comparisons needed in interpreting the results. While it remains an interesting question to explore the

effect of specific distance metric on the crosslinguistic results on different types of embeddings, I leave this for future work with a larger sample size (in this context, the number of models trained on disjoint training data) for robust statistical interpretation.

From each model, I extracted two types of representations to be tested with ABX. As the encoder contains two parts, one being STRFs and the other being a convolution layer, I extracted representations after each part of the encoder. The STRF features represent crude spectrotemporal encoding without further processing, and the features after the convolution layer ("bottleneck features") reduce the dimensionality slightly further and represent some form of further processing based on the STRF features.

Lastly, similar to Chapter 2, the models were tested with the following two manipulations, scrambling and low-pass filtering. In the scrambling condition, the input auditory spectrogram was shuffled along time, creating a representation with slower temporal information removed. In the low-pass filtering, the stimuli were low-pass filtered with a cutoff frequency of 400 Hz.

4.3 Model performance on language discrimination

The language discrimination results for the autoencoder model are shown in Figure 4.2. Firstly, for single language discrimination, when full-band speech was used, the model showed a significant difference in its ability to discriminate between English-Japanese better than English-Dutch, and the effect, unlike all models in Chapter 2, diminished or disappeared when the stimuli were scrambled. This suggests that temporal structure was necessary for the model's humanlike performance. On low-pass filtered speech, the autoencoder model could not seem to discriminate reliably when speech was lowpass filtered, showing only a marginal crosslinguistic effect with

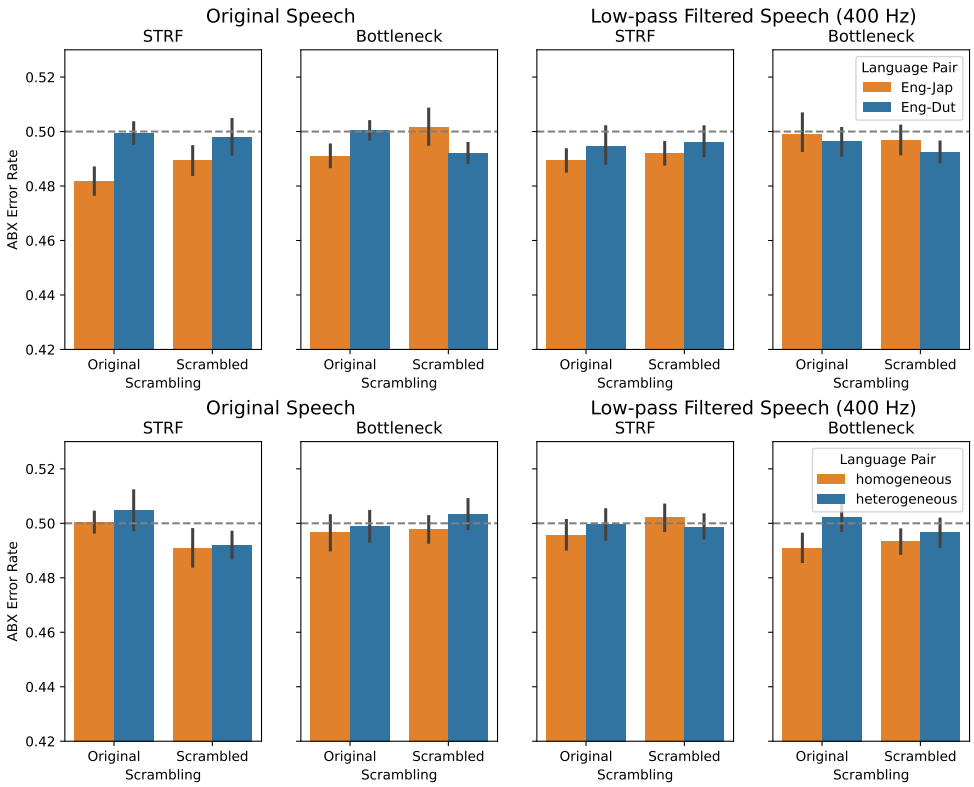

Figure 4.2: **Top:** Performance of the autoencoder model on single-language discrimination. **Bottom:** Performance of the autoencoder model on grouped language discrimination. Error bars denote 95% CI. The dashed horizontal line denotes chance-level performance (50% correct).

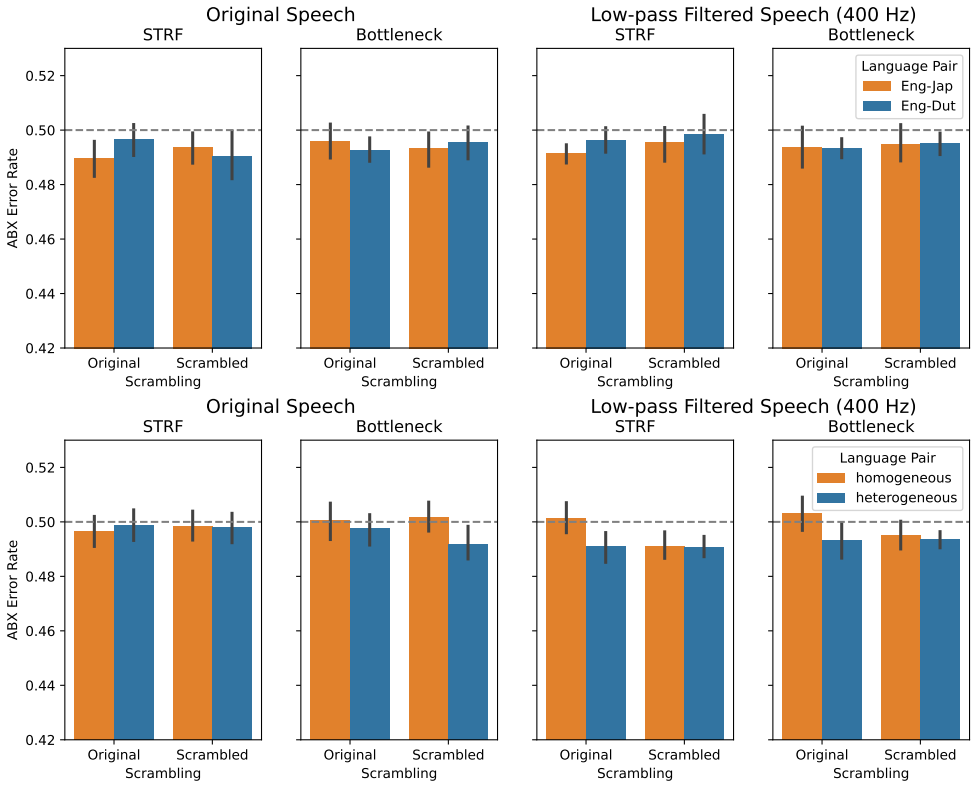

Figure 4.3: **Top:** Performance of the source separation model on single-language discrimination. **Bottom:** Performance of the autoencoder model on grouped language discrimination. Error bars denote 95% CI. The dashed horizontal line denotes chance-level performance (50% correct).

the STRF features and no effect with the bottleneck features. With the grouped languages, on the other hand, a crosslinguistic effect was observed for low-pass filtered speech but not for full-band speech. Overall, while not completely consistent with humanlike predictions, the autoencoder model was able to discriminate between languages in a humanlike direction.

The language discrimination results for the source separation model were shown in Figure 4.3. On single languages, only the STRF features seem to marginally show a humanlike crosslinguistic effect. With grouped languages, none of the features seem to be able to capture the humanlike crosslinguistic pattern. In fact, for grouped languages, a crosslinguistic effect in the opposite direction was observed, suggesting that the source separation model shows very dif-

ferent behavior compared with the autoencoder, despite having the same front-end architecture.

4.4 Interpreting model performance using spectrotemporal ablations

One benefit of using the differentiable auditory frontend lies in its interpretability. Figure 4.4 shows the distribution of learned STRFs when trained with different tasks. First of all, for both pretext tasks, the learned distributions were generally condensed in low temporal modulations and low-medium spectral modulations, which qualitatively matches the modulation pattern reported in literature for speech in general (Elliott and Theunissen, 2009) and in past linear models (Vuong et al., 2021). Secondly, it can be observed that the autoencoder and source separation models behaved differently in the trade-off of high spectral and temporal modulations. While the autoencoder STRFs are more densely distributed within high temporal modulations, the source separation models are more densely distributed within high spectral modulations. This can be explained by the specificity in the pretext tasks – as the autoencoder has the goal of reconstructing the *exact* auditory spectrogram as the input, having information that is high in temporal modulation can track temporal fine structures that is otherwise not captured in low temporal modulation ranges. On the other hand, the source separation model has the task of separating speech from music, and pitch differences, which are mostly reflected in the range of high spectral modulations, may be one of the most informative cues in separating speech from music. Overall, the STRF distributions from the two pretext tasks are reasonable given the difference in their pretext tasks.

In addition to interpreting the STRF distributions alone, since the auditory frontend presents as a gateway to later processing stages of the model, by manipulating the STRF representations

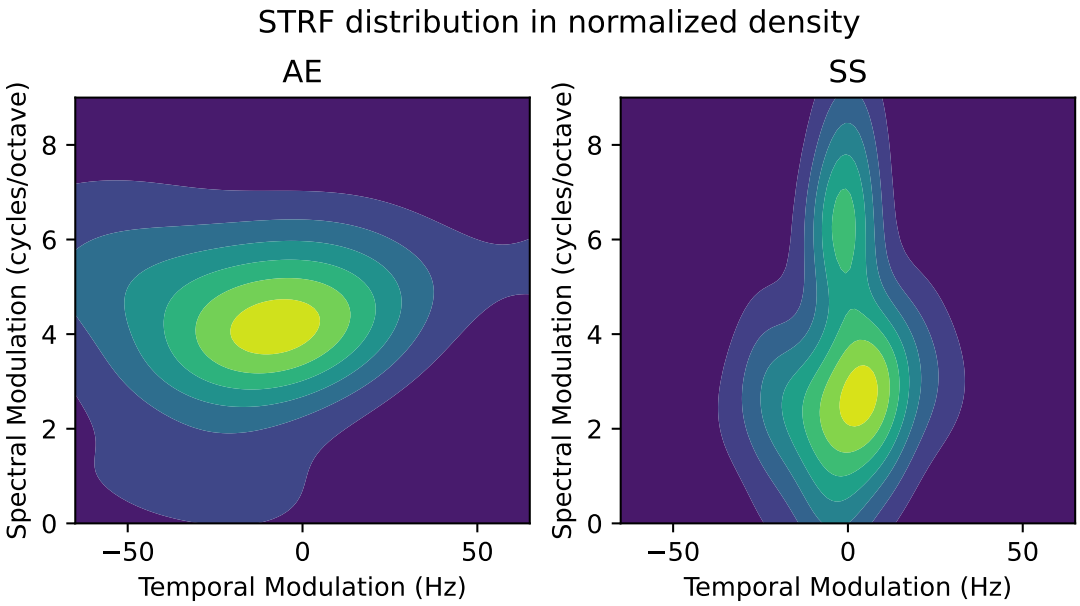

Figure 4.4: Distribution of learned STRFs for the autoencoder model (left) and the source separation model (right). The density was calculated across STRFs aggregated over different initializations (6 models for the autoencoder model, and 5 models for the source separation model).

that the rest of the model has access to, it is possible to expose causal relationships of how features across spectral and temporal modulations affect the models' performance on language discrimination. In the case of language discrimination, as the models were shown to require temporal information to perform the discrimination tasks in a humanlike way, it would be informative to further explore what kind of information in time drives this crosslinguistic effect.

To achieve this, I performed the following ablations to the model at the STRF layer. With the output of the cortical model, I set the features outside a range of temporal modulations to be zeros, allowing only information in certain temporal modulation ranges to be passed through. I then performed the ABX test using STRF features that were ablated in this way, as well as bottleneck features obtained from ablated STRF representations. I chose three ranges of temporal modulations, allowing only less than 5 Hz of temporal modulations, between 5 and 25 Hz, and

above 25 Hz. This division was made to roughly correspond to syllable-level (around 4.5 Hz) and slower, stress and phrasal level rhythm that corresponds to the theta and delta band in EEG terminology, fast, transient information corresponding to the (low) gamma range in EEG, and anything intermediate including the beta band in EEG, potentially correlating with temporal regularities related to onset and rime (Giraud and Poeppel, 2012; Goswami, 2022). Since the source separation model did not show a robust and humanlike effect of language discrimination, only the autoencoder model results are further reported here with ablations.

The ablation results for single languages are shown in Figure 4.5. Since the main results showed a significant and robust effect only with full-band speech, I will focus on this part in the ablated results, where the rest of the results are shown for completion. From the literature reviewed earlier about speech rhythm and their correspondence with slow temporal modulations (4.5 Hz for "syllable" and slower for stress and phrasal prosody), it would be expected that ablations allowing spectrotemporal modulation that are less than 5 Hz in temporal modulation would best correspond to rhythm at the syllable level, and therefore, ablations outside of this range should minimally affect the crosslinguistic effect observed in the full model. Relatively, in other ablations where the temporal modulations related to syllable rhythm are removed, the crosslinguistic effect should disappear. However, this does not seem to be the case. When temporal modulations less than 5 Hz were retained, discrimination for English-Japanese was actually not reliably different compared with English-Dutch. On the other hand, a consistent crosslinguistic pattern was observed when temporal modulations greater than 25 Hz were retained.

In the grouped language results (shown in Figure 4.6), I will focus on the low-pass filtered results as those were found in the full model to show a significant effect. Again, when the modulations at the syllable level in time were allowed through, the discrimination effects were

not humanlike. While the effects were marginal, it was observed in both STRF and bottleneck features when modulations greater than 25 Hz were allowed through, corresponding to the single language results.

The results from ablations by temporal modulations contradict prior assumptions about what range of temporal modulations drive language discrimination, and suggest that syllable level rhythm may not be the driving factor for language discrimination after all. Additionally, these results may also explain why the source separation models did not behave humanlike in their language discrimination results. As shown in Figure 4.4, the STRF distribution of the source separation model was very scarce in the range greater than 25 Hz in temporal modulation. Combined with the ablation results on the autoencoder, which found temporal modulations in this range drive humanlike language discrimination effects, it further explains why the source separation model performed poorly in terms of humanlike performance (as shown in Figure 4.3).

4.5 Discussion

In this chapter, I explored language discrimination using the differentiable auditory model as a biologically informed frontend. This was done by incorporating the differentiable auditory frontends into two deep learning models of speech processing, respectively trained using pretext tasks of autoencoder (reconstructing information through a bottleneck) and source separation (keeping clean speech and removing noise). The results suggest that of the models that were able to discriminate languages in a humanlike manner, rhythm was necessary for the humanlike effect, unlike other models tested in Chapter 2. When the models are examined in terms of what range of modulations drive the humanlike crosslinguistic effect, the ablated models were not able

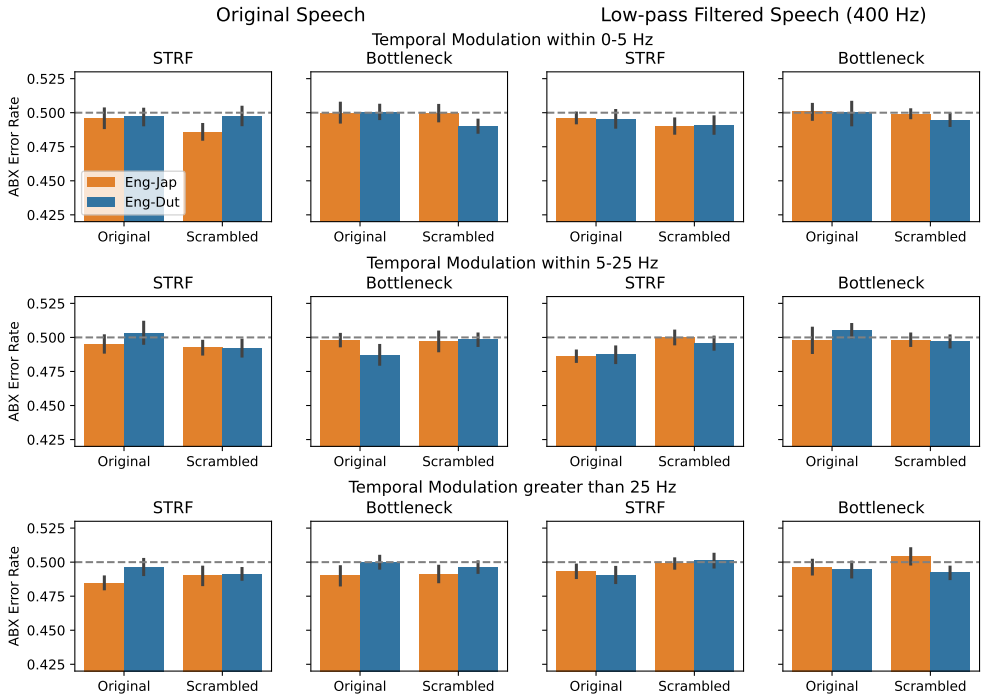

Figure 4.5: Performance of the autoencoder model on single-language discrimination, with ablations in terms of spectrotemporal modulation. Each row shows the models' performance where only the specified range in temporal modulation was allowed through the STRF features.

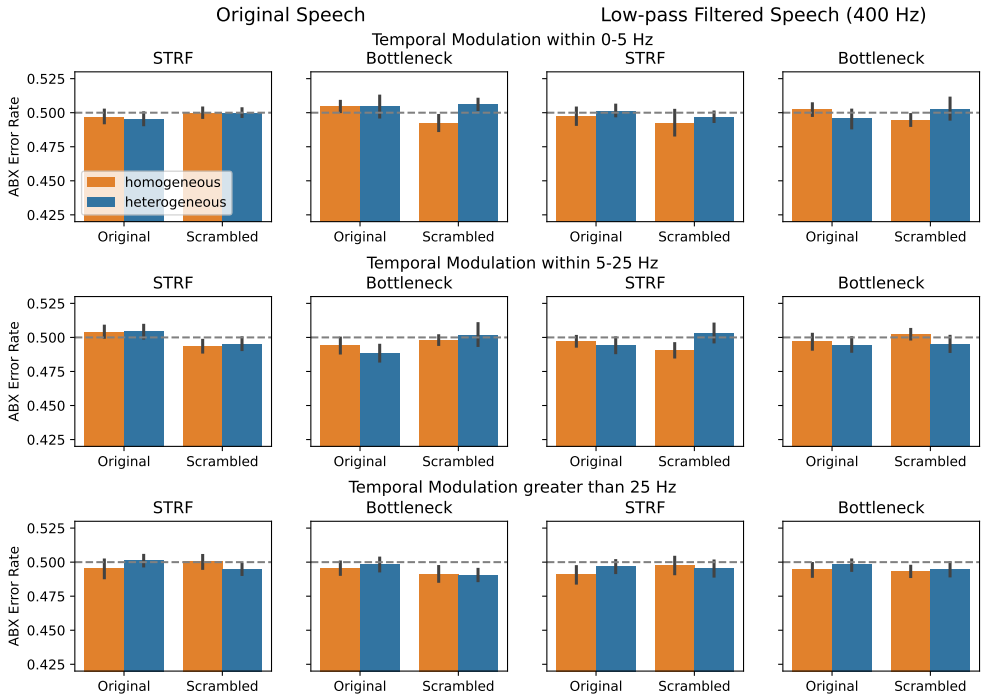

Figure 4.6: Performance of the autoencoder model on grouped language discrimination, with ablations in terms of spectrotemporal modulation. Each row shows the models' performance where only the specified range in temporal modulation was allowed through the STRF features.

to retain the humanlike effect using syllable-level rhythm alone. Instead, multiple results suggest that temporal modulation beyond 25 Hz best corresponds to the models' humanlike discrimination effect. Additionally, the two models trained with different pretext tasks showed different learning outcomes as reflected by the learned STRF distribution, which also aligns with the previous conclusion that temporal modulation beyond 25 Hz is important for humanlike language discrimination.

When considered together with the results reported in Chapter 2, the current results suggest that models with explicit neurobiological constraints may use different processing mechanisms on the same task, even on simulations of behavioral, computational-level questions such as newborn language discrimination. Among all the models that were tested in Chapter 2, rhythm was not necessary for humanlike discrimination. However, in models in this chapter where deep neural networks must operate on learnable spectrotemporal modulation features, rhythm becomes necessary for humanlike discrimination. As studies across species (including birds, ferrets and humans) and neural recording methods (single neuron recordings on animals to non-invasive population recordings in humans) support STRFs as a simplistic method to model auditory processing in the subcortical and cortical regions, the current results imply that a model that is explicitly set up to use auditory neuroscience processes may show different results from models directly borrowed from machine learning and engineering. This suggests that cognitive modeling may benefit from more incorporation of neurobiology intuitions.

As the results in Chapter 2 across different types of models strongly suggest that rhythm was not used in any of the models, it is not conclusive what is the case for human perception. One possible explanation is that humans use various cues to perform language discrimination, including potentially segmental cues, global statistics, and sequential, temporal information. This can

be supported in the literature, where newborns were able to discriminate between Japanese and Dutch when the speech was resynthesized with flat intonation (i.e., pitch track), but the effect size was much smaller than the same pair of languages resynthesized, but with the original intonation. This suggests that the distribution or temporal sequence of pitch may be one cue to newborns to discriminate between Japanese and Dutch, but not the only cue. Similarly, information that is purely spectral and segmental may be used together with temporal and sequential information in human language discrimination, where different models represent potentially a subset of processing mechanisms used by humans. In order to tease apart these relationships to further understand what cues are important to newborns, further behavioral studies with more controlled language pairs would be beneficial, as was also pointed out by (Gasparini et al., 2021). Unfortunately for behavioral studies, it is impossible to directly test infants with scrambled spectrograms, since magnitude spectrograms are not audible, and combining the scrambled phase information and transforming the scrambled complex spectrogram into time domain leads to noisy sounds that likely will not appeal to newborns. Nonetheless, other methods could be explored to remove certain temporal regularities from speech in a controlled manner. While it is beyond the scope of this thesis, some preliminary experiments and stimuli are underway to test language discrimination without certain temporal regularities behaviorally.

While the ablation results are unexpected given the predictions based on syllable-level rhythm, they align with some results from the cognitive neuroscience literature. As reviewed in Chapter 2, when infants passively listened to continuous speech of different languages, EEG recordings showed a difference in only gamma band activity, but not in theta or delta bands which correspond to the rates of syllable and stress. Granted, it is not always well-justified to compare frequency-domain EEG data to that of speech, as activity across different frequency bands in

speech acoustics and cortical EEG measurements do not necessarily operate in the same fashion, or arise from the same mechanism. However, at least within syllable-level rates (i.e., the theta band), entrainment (i.e., synchronization in phase) has been known to occur in EEG towards the speech stimuli, and correlates with attention and intelligibility (Ding and Simon, 2014; Peelle et al., 2013). The fact that a crosslinguistic effect was not observed in this frequency band for infants aligns with the observation that the differentiable auditory models did not use temporal modulation at this range for language discrimination.

The current results also reveal the importance of pretext tasks in cognitive simulations. While the autoencoder models (Matusevych et al., 2023) and denoising (Tuckute et al., 2023) have both been used to model speech perception and cognition in the past, the current results suggest that the different pretext tasks lead to different learning strategies as reflected by the STRF distributions, as well as different behavior on language discrimination compared with humans. While different pretext tasks may all reflect some aspect of processing performed by the human brain, the current results suggest that the specific implementation and architecture may lead to differences in the model's strategy in solving the task and hence model performance on tasks of interest. Therefore, it would be of interest in future cognitive modeling studies to explore a variety of pretext tasks to facilitate the field's understanding of what pretext tasks may be correlated to what kind of behavior. With more information about how pretext tasks are linked to specific learning outcomes, the choice of pretext alone can serve as a type of manipulation to test different cognitive theories. On the other hand, while the autoencoder models aligned with human language discrimination better than the source separation models, neither perfectly replicated the data of human newborns, suggesting that potentially more tasks or variations of these tasks could be explored.

Chapter 5: Conclusion

5.1 Summary and Main Contributions

In this thesis, I built a differentiable model of auditory processing, and applied it to simulations of early language discrimination which I compared with other machine learning models in the literature. Specifically, I focus on the widely-assumed notion that rhythm, or the temporal structure at the syllable level, drives behavioral language discrimination. The results suggest that the traditional machine learning models, spanning from clustering-based models to self-supervised deep neural networks, do not require temporal structure to perform language discrimination in a humanlike manner. On the other hand, a model built with the differentiable auditory model needs the temporal structure to perform language discrimination in a humanlike manner.

Two main contributions were made in this thesis. First, I examined the role of rhythm in early language discrimination through a series of computational simulations. While the role of rhythm in early language discrimination has often been assumed but never systematically examined, this thesis explored several aspects of rhythm, including whether rhythm is the sole or primary factor that drives language discrimination, and what scale of temporal regularity may be the most important to drive the human behavioral pattern of discriminating certain languages but others. The results revealed unexpected patterns — many models do not use temporal order at all to produce humanlike discrimination patterns, even when speech was low-pass filtered to remove

spectral and segmental details. This suggests an alternative that certain global information such as the overall distribution of vocalic and consonantal segments may suffice to drive language discrimination in humans. Then, in some models that are explicitly coded to process speech using temporal regularities, temporal modulations that are much faster than the syllable rate were found to drive the models' humanlike discrimination pattern, contrary to the expectation informed by theories that syllable- and stress-level temporal regularities are associated with human language discrimination.

Secondly, I introduced a new paradigm for cognitive simulations where hypotheses and known constraints about human neurobiology and cognition can be embodied through differentiable programming in deep learning. While deep learning has seen revolutionary improvements in the last several years, achieving human-level and superhuman performances in many tasks of audio, speech, and language processing, the application of such models to simulate perception and cognition is limited by its nature as a universal function approximator, which implies that a neural network that is large and deep enough, when trained on ample data, can approximate any input-output mapping in theory. This limits the use of deep learning in reflecting specific constraints known to human perception and cognition. Furthermore, it implies that deep models lack interpretability by nature, since it is difficult to understand what specific computations are performed in a given layer, head, or channel of a neural network. Differentiable programming, on the other hand, allows linear and nonlinear functions to be added to the computational graph of a neural network, such that the basic building blocks can be custom-defined outside of popular architectures such as CNN, RNN, and transformers. These custom-defined functions can embody theories and models from cochlear processing to cortical models, and even specific cognitive hypotheses. The parameters in these functions can then be optimized in a task-driven manner, using

modern-day deep learning optimizing algorithms and hardware. I have shown in this thesis that a differentiable auditory model with such biological intuitions and constraints can improve the performance, robustness, and interpretability of deep speech processing. Additionally, I have also demonstrated its potential in cognitive simulations by explicitly defining spectrotemporal modulation-based processing, and use the transparency of this architecture to perform ablation tests to examine the contribution of various temporal modulation ranges in language discrimination.

On a greater scale, this thesis also establishes some frameworks in training and testing large-scale models that operate on real speech inputs to simulate suprasegmental features. So far, the machine ABX test has been used to test a series of clustering-based (Li et al., 2020a; Schatz et al., 2021b) and deep learning-based (Matusevych et al., 2023) models. However, since suprasegmental features are much longer in time with varying segmental content across tokens, using machine ABX tests to probe the models' representation on such suprasegmental features may be unrealistic. In Carbajal et al. (2016), the problem of feature length was cleverly avoided by using i-vector models, which generate low-dimensional i-vectors that are low in feature dimensionality and do not contain a time dimension. In a series of simulations in Chapter 2, the i-vector models were indeed found to have lower ABX error rates, and greater effect size overall. However, the other models tested also revealed better than chance discrimination performance on language pairs that humans discriminate, suggesting that machine ABX test with dynamic time warping is successful in evaluating a model's sensitivity to suprasegmental information such as rhythm. Additionally, in Chapter 4, I introduced pooling as another method to perform ABX. This resembles the i-vectors in removing the time dimension from the computation, and was shown to be another method to explore deep and STRF embeddings for suprasegmental representations.

Additionally, this thesis contributes a method for compiling test utterances from crowd-sourced data, and a dataset compiled from such public domain data that can be further used in future research for replicability. Unlike studies that require only one high-resource language such as English, the study of language discrimination is often constrained by a lack of test stimuli that contain all languages as tested in humans. Although languages can in theory be collected from a variety of corpora, the difference in semantic context, speaker demographics and recording conditions that vary more across corpora than within often poses additional confounds. This constrains past computational studies to use a combination of languages that are not directly comparable to human data. For example, in past i-vector models (Carbajal, 2018; Carbajal et al., 2016), the test language pairs included English-French and English-Italian. While French was the same as the training data, and therefore may pose additional familiarity to the model, English-Italian was not a pair that was directly tested in infants. In this thesis, to directly compare the models' performance to human data, stimuli were directly selected from a crowd-sourced speech database (Ardila et al., 2019), and hand-selected to control for utterance duration, recording noise, and disfluencies. This showcases a new way to simulate crosslinguistic phenomena, where obtaining native speakers from an array of languages for stimulus recording is no longer a prerequisite. Combined with contemporary speech transcription (Radford et al., 2023) and alignment (McAuliffe et al., 2017) models and software, other phenomena in crosslinguistic phonetic effects and language processing can also be studied using this data compilation method in the future.

5.2 Future Directions

The conclusions of this thesis suggest potential alternatives to what humans may be sensitive to in the speech stream, which inform future perceptual studies. Specifically, the computational models support representations other than syllable rhythm to replicate humanlike language discrimination, such as global statistics and faster temporal fluctuations. From these alternative hypotheses, experiments could be designed to test adults or young infants on how various components that differ across speech of different languages may separately drive language discrimination.

Another future direction is to extent the current computational paradigm to a broader range of cognitive phenomena, including time-dependent phonetic categories and other suprasegmental features. As the current thesis demonstrates a case of using a differentiable auditory model to simulate human perception of suprasegmental features in the speech stream, the model may also be adapted to explore other perceptual phenomena that are time-dependent. For example, while clustering-based models are known to capture well-known perceptual attunement phenomena on the English /l/-/ɹ/ contrast, which is mostly based on spectral information, the clustering models have been found to capture human discrimination poorly on duration-based contrasts such as geminates and vowel duration in Japanese (unpublished data). As no behavioral correlates were reported to correspond to this discovery in the models, it is possible that the poor discrimination performance on durational contrasts arises from the mathematical assumption in clustering-based models, which operate on MFCCs and do not naturally characterize the duration of sound. As such, these durational contrasts could be interesting to be tested with the differentiable auditory model, which processes sound through spectrotemporal modulations by design. Another time-

dependent phenomenon is the perception of illusory vowels (Dupoux et al., 1999), where native speakers of Japanese can perceive an extra syllable out of nonwords like "ebzo" as in "ebuzo". While various theories exist to characterize this phenomenon from pure transitional probability to syllable-based structures (Dupoux et al., 2011), the differentiable auditory model may be able to represent some of these theories in its use of specific temporal modulations.

Lastly, the introduction of auditory neuroscience in deep learning models for cognitive simulations invites other methods that embody neuroscientific or cognitive theories in deep learning models. For example, the current differentiable auditory model can be directly modified to contain more detailed models that more closely adhere to audition on an algorithmic or implementation level. The cochlear filterbank model can be replaced by more time-dependent models such as the CARFAC (Lyon, 2011; Lyon et al., 2024) model or the transmission line model (Altoè et al., 2014). Between peripheral and cortical processing, subcortical processes could also be added including brainstem processes that give rise to measurable elements such as the auditory brain response (ABR). In addition to different models of auditory neuroscience, another area to be explored involves cognitive hypotheses. The current thesis explores specific pretext tasks such as autoencoder reconstruction and denoising (source separation). In the deep learning literature, existing architectures such as autoregressive predictive coding (Chung and Glass, 2020) and masked reconstruction and autoencoders Mohamed et al. (2022), among other pretext tasks in self-supervised pretraining, provide interesting candidates to explore as hypotheses of how perception, cognition and learning occurs in humans. Introducing and comparing more pretext tasks through differentiability can offer insights into whether these theories are realistic models of human perception and learning.

Bibliography

Abboub, N., Nazzi, T., and Gervain, J. (2016). Prosodic grouping at birth. *Brain and language*, 162:46–59.

Allen, J. B. (1985). Cochlear modeling. *IEEE AssP MAGAziNE*, 2(1):3–29.

Altoè, A., Pulkki, V., and Verhulst, S. (2014). Transmission line cochlear models: improved accuracy and efficiency. *The Journal of the Acoustical Society of America*, 136(4):EL302–EL308.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Barajas, M. C. O., Guevara, R., and Gervain, J. (2021). The origins and development of speech envelope tracking during the first months of life. *Developmental cognitive neuroscience*, 48:100915.

Bowman, D., Eggermont, J., Brown, D., and Kimberley, B. (1998). Estimating cochlear filter response properties from distortion product otoacoustic emission (DPOAE) phase delay measurements in normal hearing human adults. *Hearing research*, 119(1-2):14–26.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood: A multidisciplinary approach*. PhD thesis, Université de recherche Paris Sciences et Lettres.

Carbajal, M. J., Fér, R., and Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *CogSci*.

Cariani, P. (2001). Temporal codes, timing nets, and music perception. *Journal of New Music Research*, 30(2):107–135.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.*, 118(2):887–906.

Chong, A. J., Vicenik, C., and Sundara, M. (2018). Intonation plays a role in language discrimination by infants. *Infancy*, 23(6):795–819.

Chung, Y.-A. and Glass, J. (2020). Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3497–3501. IEEE.

Chung, Y.-A., Tang, H., and Glass, J. (2020). Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*, 11(1):51–62.

David, S. V., Mesgarani, N., and Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network*, 18(3):191–212.

de Boer, E. and Nuttall, A. L. (2000). The mechanical waveform of the basilar membrane. II. From data to models–and back. *J. Acoust. Soc. Am.*, 107(3):1487–1496.

Dellwo, V., Karnowski, P., and Szigeti, I. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In *Language and language processing: Proceedings of the 38th linguistic colloquium*, pages 231–241. Peter Lang.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81:181–187.

Ding, N. and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, 8:311.

Dominey, P. F. and Ramus, F. (2000). Neural network processing of natural language: I. sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1):87–127.

Drakopoulos, F. and Verhulst, S. (2023). A neural-network framework for the design of individualised hearing-loss compensation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2395–2409.

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., et al. (2019). The zero resource speech challenge 2019: Tts without t. *arXiv preprint arXiv:1904.11469*.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance*, 25(6):1568.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., and Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of memory and language*, 64(3):199–210.

Elhilali, M. (2004). *Neural basis and computational strategies for auditory processing*. PhD thesis, University of Maryland.

Elhilali, M. and Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *J. Acoust. Soc. Am.*, 124(6):3751–3771.

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

Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (2020). Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

Farhadi, A., Jennings, S. G., Strickland, E. A., and Carney, L. H. (2023). Subcortical auditory model including efferent dynamic gain control with inputs from cochlear nucleus and inferior colliculus. *J. Acoust. Soc. Am.*, 154(6):3644–3659.

Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, 120(4):751.

Garcia, L. N., Guerrero-Mosquera, C., Colomer, M., and Sebastian-Galles, N. (2018). Evoked and oscillatory EEG activity differentiates language discrimination in young monolingual and bilingual infants. *Scientific reports*, 8(1):1–9.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Gasparini, L., Langus, A., Tsuji, S., and Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, 213(August 2020):104757.

Gervain, J., Christophe, A., and Mazuka, R. (2021). Prosodic bootstrapping. In Gussenhoven, C. and Chen, A., editors, *The Oxford Handbook of Language Prosody*, pages 563–573. Oxford University Press, Oxford.

Gibbon, D. (2018). The future of prosody: It's about time. *CoRR*, abs/1804.09543.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517.

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

Goswami, U. (2019). Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences*, 1453(1):67–78.

Goswami, U. (2022). Language acquisition and speech rhythm patterns: an auditory neuroscience perspective. *Royal Society Open Science*, 9(7):211855.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In *Laboratory Phonology*, volume 7, pages 515–546.

Grabe, E., Post, B., and Watson, I. (1999). The acquisition of rhythmic patterns in English and French. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1201–1204. University of California Berkeley, CA.

Gu, L. and Stern, R. M. (2008). Single-channel speech separation based on modulation frequency. In *ICASSP*, pages 25–28. IEEE.

Hamdi-Sultan, R., Barkat-Defradas, M., Ferragne, E., and Pellegrino, F. (2004). Speech timing and rhythmic structure in arabic dialects: A comparison of two approaches. In *International Speech and Communication Association*, pages 1613–1616.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026.

Jenison, R. L., Reale, R. A., Armstrong, A. L., Oya, H., Kawasaki, H., and Howard III, M. A. (2015). Sparse spectro-temporal receptive fields based on multi-unit and high-gamma responses in human auditory cortex. *PLoS One*, 10(9):e0137915.

Johnson, E. and Braun, B. (2011). The role of intonation in language discrimination by 4.5-month-olds [poster]. *Society for Research in Child Development, Montreal, Canada*.

Jusczyk, P. W., Houston, D. M., and Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3-4):159–207.

Kubozono, H. (2000). Children's shiritori word games and the acquisition of mora: A case study. *Bulletin of the Faculty of Letters, Kobe University*, (27):587–602.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2):F13–F21.

Langus, A., Mehler, J., and Nespor, M. (2017). Rhythm in language acquisition. *Neuroscience & Biobehavioral Reviews*, 81:158–166.

Leong, V., Stone, M. A., Turner, R. E., and Goswami, U. (2014). A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America*, 136(1):366–381.

Li, A. and Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm: Evidence from l1 Mandarin and German learners of English. *Studies in Second Language Acquisition*, 36(2):223–255.

Li, R., Schatz, T., Matusevych, Y., Goldwater, S., and Feldman, N. H. (2020a). Input matters in the modeling of early phonetic learning. In Denison, S., Mack, M., Xu, Y., and Armstrong, B. C., editors, *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 578–584.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2020b). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Lin, H. and Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of chinese language and computing*, 17(3):127–140.

Lin, H. and Wang, Q. (2008). Interlanguage rhythm in the English production of Mandarin speakers. In *Proceedings of the 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers*, pages 18–20.

Liu, A. T., Li, S.-W., and Lee, H.-y. (2021). Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM TASLP*, 27(8):1256–1266.

Lyon, R. F. (2011). Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function. *The Journal of the Acoustical Society of America*, 130(6):3893–3904.

Lyon, R. F. (2017). *Human and machine hearing: extracting meaning from sound*. Cambridge University Press.

Lyon, R. F., Schonberger, R., Slaney, M., Velimirović, M., and Yu, H. (2024). The carfac v2 cochlear model in matlab, numpy, and jax. *arXiv preprint arXiv:2404.17490*.

Mallat, S. (2012). Group invariant scattering. *Commun. Pure Appl. Math.*, 65(10):1331–1398.

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., and Goldwater, S. (2020). Evaluating computational models of infant phonetic learning across languages. In Denison, S., Mack, M., Xu, Y., and Armstrong, B. C., editors, *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 571–577.

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., and Goldwater, S. (2023). Infant phonetic learning as perceptual space learning: A crosslinguistic evaluation of computational models. *Cognitive Science*, 47(7):e13314.

Mazuka, R. (2007). The rhythm-based prosodic bootstrapping hypothesis of early language acquisition: Does it work for learning for all languages? *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 132:1–13.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Meng, H., Sethu, V., and Ambikairajah, E. (2023). What is Learnt by the LEArnable Front-end (LEAF)? Adapting Per-Channel Energy Normalisation (PCEN) to Noisy Conditions. In *INTERSPEECH 2023*. ISCA.

Mesgarani, N. and Shamma, S. (2007). Denoising in the domain of spectrotemporal modulations. *EURASIP J. ASMP*, pages 1–8.

Mesgarani, N., Slaney, M., and Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE TASLP*, 14(3):920–930.

Mesgarani, N., Thomas, S., and Hermansky, H. (2010). A multistream multiresolution framework for phoneme recognition. In *Eleventh Annual Conference of the International Speech Communication Association*.

Meyer, B. T. and Kollmeier, B. (2011). Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication*, 53(5):753–767.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., et al. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Moon, C., Cooper, R. P., and Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development*, 16(4):495–500.

Moore, B. C. (1996). Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear Hear.*, 17(2):133–161.

Nacar Garcia, L., Guerrero-Mosquera, C., Colomer, M., and Sebastian-Galles, N. (2018). Evoked and oscillatory eeg activity differentiates language discrimination in young monolingual and bilingual infants. *Scientific reports*, 8(1):2770.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756.

Nazzi, T., Jusczyk, P. W., and Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43:1–19.

Nazzi, T. and Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech communication*, 41(1):233–243.

Oh, D., Park, J.-S., Kim, J.-H., and Jang, G.-J. (2021). Hierarchical phoneme classification for improved speech recognition. *Applied Sciences*, 11(1):428.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pamies Bertrán, A. (1999). Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. *Language design: Journal of theoretical and experimental linguistics*, 2:103–130.

Patel, A. D. (2010). *Music, language, and the brain*. Oxford university press.

Paul, D. B. and Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York*.

Payne, E., Post, B., Astruc, L., Prieto, P., and Vanrell, M. d. M. (2012). Measuring child rhythm. *Language and Speech*, 55(2):203–229.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peelle, J. E., Gross, J., and Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, 23(6):1378–1387.

Pellegrino, E., He, L., and Dellwo, V. (2021). Age-related rhythmic variations: The role of syllable intensity variability. *TRANEL-Travaux neuchâtelois de linguistique*, 74:167–185.

Peña, M., Pittaluga, E., and Mehler, J. (2010). Language acquisition in premature and full-term infants. *Proceedings of the National Academy of Sciences*, 107(8):3823–3828.

Peter, V., van Ommen, S., Kalashnikova, M., Mazuka, R., Nazzi, T., and Burnham, D. (2022). Language specificity in cortical tracking of speech rhythm at the mora, syllable, and foot levels. *Scientific Reports*, 12(1):1–12.

Pike, K. L. (1945). *The Intonation of American English*. ERIC.

Polka, L. and Sundara, M. (2012). Word segmentation in monolingual infants acquiring canadian english and canadian french: Native language, cross-dialect, and cross-language comparisons. *Infancy*, 17(2):198–232.

Polyanskaya, L. and Ordin, M. (2015). Acquisition of speech rhythm in first language. *The Journal of the Acoustical Society of America*, 138(3):EL199–EL204.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE ASRU*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. (2019). MUSDB18-HQ - An uncompressed version of MUSDB18.

Rahman, M., Willmore, B. D., King, A. J., and Harper, N. S. (2020). Simple transformations capture auditory input to cortex. *Proceedings of the National Academy of Sciences*, 117(45):28442–28451.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, 2(1):85–115.

Ramus, F., Dupoux, E., and Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proc. 15th International Congress of Phonetic Sciences*, pages 337–342. Universitat Autònoma de Barcelona.

Ramus, F. and Mehler, J. (1999a). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1):512–521.

Ramus, F. and Mehler, J. (1999b). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1):512–521.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292.

Räsänen, O., Doyle, G., and Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171:130–150.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE.

Sadjadi, S. O., Slaney, M., and Heck, L. (2013). Msr identity toolbox. *Seattle, WA, USA: Microsoft*.

Schatz, T. (2016). *ABX-discriminability measures and applications*. PhD thesis, Université Paris 6, Paris, France.

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., and Dupoux, E. (2021a). Early phonetic learning without phonetic categories. *Proceedings of the National Academy of Sciences*, 118(7):e2001844118.

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., and Dupoux, E. (2021b). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7).

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Scheibler, R., Ji, Y., Chung, S.-W., Byun, J., Choe, S., and Choi, M.-S. (2023). Diffusion-based generative speech source separation. In *ICASSP*, pages 1–5. IEEE.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*.

Shamma, S. and Lorenzi, C. (2013). On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *The Journal of the Acoustical Society of America*, 133(5):2818–2833.

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., and Rinzel, J. (1986). A biophysical model of cochlear processing: intensity dependence of pure tone responses. *J. Acoust. Soc. Am.*, 80(1):133–145.

Shier, J., Caspe, F., Robertson, A., Sandler, M., Saitis, C., and McPherson, A. (2023). Differentiable modelling of percussive audio with transient and spectral synthesis. *arXiv preprint arXiv:2309.06649*.

Shinn-Cunningham, B., Varghese, L., Wang, L., and Bharadwaj, H. (2017). Individual differences in temporal perception and their implications for everyday listening. *The frequency-following response: A window into human communication*, pages 159–192.

Skrodzka, E. B. (2005). Mechanical passive and active models of the human basilar membrane. *Appl. Acoust.*, 66(12):1321–1338.

Sundara, M. and Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, 39(4):505–513.

Thiemann, J., Ito, N., and Vincent, E. (2013). The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *POMA*, volume 19.

Thuerey, N., Holl, P., Mueller, M., Schnell, P., Trost, F., and Um, K. (2021). Physics-based deep learning. *arXiv preprint arXiv:2109.05237*.

Tilsen, S. and Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1):628–639.

Tuckute, G., Feather, J., Boebinger, D., and McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., and Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, 142(4):1976–1989.

Vuong, T., Xia, Y., and Stern, R. (2020). Learnable spectro-temporal receptive fields for robust voice type discrimination. *arXiv preprint arXiv:2010.09151*.

Vuong, T., Xia, Y., and Stern, R. M. (2021). A modulation-domain loss for neural-network-based real-time speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6643–6647. IEEE.

Vuong, T. M. T. (2023). *Incorporating Modulation Information into Deep Neural Networks for Robust Speech Processing*. PhD thesis, Carnegie Mellon University.

Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A. (2017). Trainable frontend for robust and far-field keyword spotting. In *ICASSP*, pages 5670–5674. IEEE.

Weikum, W. M., Vatikiotis-Bateson, E., and Werker, J. F. (n.d.). Speech properties infants use to discriminate languages visually. Unpublished data.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., and Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, 316(5828):1159–1159.

Werker, J. F. and Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, 66:173–196.

Wu, H., Zheng, B., Li, X., Wu, X., Lee, H.-Y., and Meng, H. (2022). Characterizing the adversarial vulnerability of speech self-supervised learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3164–3168. IEEE.

Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., and Sridhar, S. (2022). Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library.

Zeghidour, N., Teboul, O., Quitry, F. D. C., and Tagliasacchi, M. (2021). LEAF: A learnable frontend for audio classification. *arXiv preprint arXiv:2101.08596*.

Zhang, X. and Gan, R. Z. (2011). A comprehensive model of human ear for analysis of implantable hearing devices. *IEEE Trans. Biomed. Eng.*, 58(10):3024–3027.

Zhang, Y., Zou, J., and Ding, N. (2023). Acoustic correlates of the syllabic rhythm of speech: Modulation spectrum or local features of the temporal envelope. *Neuroscience & Biobehavioral Reviews*, page 105111.