The Representation of Speech Variability and Variation in Deep Neural Networks

Badr Mohammed Badr Abdullah

A dissertation submitted towards the degree
PhD in Computational Linguistics
at the Faculty of Philosophy
Saarland University

Saarbrücken, June 25, 2024

DAY OF COLLOQUIUM:
March 1, 2024

DEAN OF THE FACULTY:
Prof. Dr. Stefanie Haberzettl

EXAMINATION BOARD:
Chair – Prof. Dr. Ingo Reich, Saarland University
Advisor, Reviewer – Prof. Dr. Dietrich Klakow, Saarland University
Reviewer – Dr. Odette Scharenborg, Delft University of Technology
Professorial member of the committee – Prof. Dr. Bernd Möbius, Saarland University
Postdoc member of the committee – Dr. Wei Xue, Saarland University

LOCATION:
Saarbrücken

To the Noor of my eyes.
To my loving parents.

Abstract

The central aim of this thesis is to bridge between the study of human speech variability and representation learning, focusing on how modern deep neural networks (DNNs) process and encode speech variability and variation in their latent representations. Diverging from prior machine learning research which has primarily focused on improving model performance in the face of variability, this thesis seeks to provide better insights into how different dimensions of speech variability shape neural network representations. The first part of this thesis, concerned with neural models of spoken language identification, introduces two studies investigating the model's adaptability to domain variability and the extent to which the model representations capture cross-linguistic variation. The second part of this thesis focuses on neural models of spoken-word representations, presenting three studies that explore various dimensions of variability including: the encoding of word-form variability in the model representational geometry, the variability of linguistic experience and its role in shaping non-native spoken-word representations, and the integration of high-level lexical knowledge into the model to abstract from variability in word acoustic realization. The third and final part of this thesis analyzes the latent discrete representations in transformer-based speech models trained with self-supervision and codebook learning, and demonstrates that information-theoretic metrics reflect acoustic-phonetic variability in segment realization. In summary, this thesis makes tangible contributions by uncovering how neural models encode domain, acoustic-phonetic, and cross-linguistic variation, exploring the role of L1/L2 similarity on non-native spoken-word processing, and characterizing the relationship between discrete speech representations and abstract phonetic categories such as phonemes. Throughout six diverse studies, this thesis takes an interdisciplinary perspective and demonstrates the utility of machine learning models as a potent scientific tool to answer novel and linguistically-informed research questions that are grounded in the fields of sociolinguistics, speech perception, and cognitive modeling research.

Zusammenfassung

Das zentrale Ziel dieser Dissertation ist es, die Forschungslücke zwischen der Untersuchung von Variabilität und Variation in der menschlichen Sprache und der maschinellen Verarbeitung von Sprache auf der Grundlage von Repräsentationslernen zu schließen, um neue Erkenntnisse darüber zu gewinnen, wie moderne tiefe neuronale Netze (DNNs) verschiedene Dimensionen der Sprachvariabilität in ihren Repräsentationen verarbeiten und kodieren. Obwohl einige Aspekte der Variabilität in früheren Forschungsarbeiten zur computergestützten Sprachverarbeitung behandelt wurden, lag der Hauptschwerpunkt bei vorherigen Ansätzen des maschinellen Lernens stets auf der Entwicklung von Modellen, die robust gegenüber Variationen in den Aufnahme- und Akustikbedingungen sind, sowie auf der Generalisierungsfähigkeit gegenüber Unstimmigkeiten zwischen Trainings- und Testdaten aufgrund von Domänen-, Sprecher- und linguistischen Variationen. Daher konzentrierten sich die Forschungsbemühungen in der bisherigen Sprachrepräsentationsforschung in erster Linie auf die Verbesserung der Leistungsmetriken für eine bestimmte Aufgabe bei Vorhandensein einer Variabilitätsquelle. Anstelle dieses leistungsorientierten Ansatzes nimmt diese Dissertation eine andere Perspektive ein und zielt darauf ab, zu analysieren und zu verstehen, wie das Repräsentationsprofil von neuronalen Sprachnetzwerken durch verschiedene Dimensionen der Sprachvariabilität geformt wird, wie z.B. Domänenvariabilität, sprachübergreifende Variation, Variabilität innerhalb der Kategorie, Variabilität in der sprachlichen Erfahrung und akustische Variabilität abstrakter phonetischer Kategorien

In dieser Dissertation werden sechs Studien vorgestellt, die in drei verschiedene Teile gegliedert sind, wobei jeder Teil einer Sprachverarbeitungsaufgabe gewidmet ist. Im ersten Teil der Dissertation stelle ich zwei Studien vor, die sich mit neuronalen Modellen zur Identifikation gesprochener Sprache (SLID) befassen, um ihre Anpassungsfähigkeit an Domänenvariabilität zu untersuchen (Studie I) und zu analysieren, inwieweit sie sprachübergreifende Variationen darstellen (Studie II). In Studie I zeige ich, dass DNNs - wie erwartet - nicht robust gegen Domänenvariabilität sind, jedoch können bestimmte Trainingsstrategien (z.B adversarial learning) effektiv sein, um zu verhindern, dass das Modell Abkürzungen in den Daten lernt, um seine domänenübergreifende Generalisierung zu verbessern. In Studie II zeige ich, dass die Repräsentationen neuronaler Netze sprachübergreifende

Ähnlichkeit erfassen und in einer Weise geclustert sind, die Sprachverwandtschaft widerspiegelt.

Im zweiten Teil der Dissertation stelle ich drei Studien vor, die sich mit neuronalen Modellen des Keyword-Spotting und der akustischen Worteinbettung befassen, um die Variabilität von gesprochenen Wortrealisierungen zu untersuchen. Zunächst gehe ich näher auf die Geometrie des Repräsentationsraums für gesprochene Wörter ein, um zu untersuchen, wie er die Variabilität von Beispielen innerhalb einer Kategorie kodiert und wie sich die Variabilität in den Anfangsbedingungen des Modells auf die Repräsentationen auswirkt, sobald sie konvergiert sind (Studie IV). Anschließend wird eine Studie vorgestellt, die darauf abzielt, die Variabilität der sprachlichen Erfahrung und ihre Rolle bei der Verarbeitung nicht-muttersprachlicher Sprache zu modellieren (Studie V). Konkret wird in dieser Studie die sprachliche Erfahrung als die Muttersprache (L1) des Modells während des Trainings charakterisiert und die Verarbeitung nicht-muttersprachlicher gesprochener Wörter simuliert, indem das Ausmaß gemessen wird, in dem nicht-muttersprachliche Modelle muttersprachliche Repräsentationen von gesprochenen Wörtern erzeugen. Schließlich stelle ich ein Berechnungsmodell für die Repräsentation gesprochener Wörter vor, das von der menschlichen Sprachverarbeitung inspiriert ist und eine Zuordnung zwischen der akustischen Form und einer semantischen Repräsentation auf abstrakter Ebene erlernt, die lexikalisches Wissen kodiert (Studie V). Ich zeige, dass die Integration von lexikalischem Wissen in das Training gesprochener Wortrepräsentationen die Fähigkeit des Modells verbessert, zwischen lexikalischen Kategorien zu unterscheiden, und das Modell ermutigt, von der Variabilität des Sprechers und des lexikalischen Kontexts zu abstrahieren.

Im dritten Teil konzentriere ich mich auf die diskreten Repräsentationen von Sprache, die sich beim Training von Transformer-Modellen durch Selbstüberwachtes- und Codebuchlernen entstehen. In diesem Teil wird ein Ansatz zur Charakterisierung der Beziehung zwischen diskreten Sprachrepräsentationen und abstrakten phonetischen Kategorien wie Phonemen vorgestellt. Konkret schlägt das Kapitel zunächst einen informationstheoretischen Rahmen vor, in dem jede phonetische Kategorie als eine Verteilung über diskrete Einheiten dargestellt wird. Die Studie zeigt, dass die Entropie phonetischer Verteilungen die akustisch-phonetische Variabilität der zugrunde liegenden Sprachlaute widerspiegelt, wobei Sonoranten im Durchschnitt entropischer sind als Obstruenten. Darüber hinaus zeigt sich, dass phonetisch ähnliche Laute auf niedriger Ebene ähnliche Verteilungen aufweisen, während eine Clusteranalyse zeigt, dass die höchste Ebene der Aufteilung Obstruenten und Sonoranten trennt.

Insgesamt bietet diese Dissertation wertvolle Einblicke in die Art und Weise, wie DNNs Sprachvariabilität über mehrere Dimensionen hinweg verarbeiten und kodieren. Dies verbessert unser Verständnis von Sprachverarbeitung und trägt zur Entwicklung robusterer und linguistisch informierter Sprachtechnologieanwendungen bei.

Most of the problems of the world stem from linguistic mistakes and simple misunderstandings. Don't ever take words at face value. When you step into the zone of love, language as we know it becomes obsolete. That which cannot be put into words can only be grasped through silence.

— ***Elif Shafak***

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dietrich Klakow, for his invaluable support and guidance throughout my PhD journey. His consistent encouragement to follow my curiosity and his availability during difficult times have been instrumental in my progress.

Special thanks go to the principal investigators of the SFB C4 project, Prof. Bernd Möbius and Prof. Tania Avgustinova, for their motivating and insightful feedback on the work presented in this thesis.

I would also like to acknowledge my first academic mentor, Prof. Sameen Fatima, for introducing me to the field of natural language processing and computational linguistics, and for her encouragement to pursue a an academic degree in Europe.

Furthermore, I am grateful to all the wonderful people I had the pleasure of meeting at the LST department, particularly my corridor mates during my PhD research journey in the LSV group: Marius, Anupama, Zena, Dawei, Alex, Aravind, Miaoran, Dave, Dana, Nico, JJ, Micael, Vagrant, Julius, David, Paloma, Claudia, and Florian.

Special thanks to my friends in the Coli community: Miriam, Iza, Koel, Hali, Mario, Polina, Samantha, Nikos, Insa, Stalin, Saad, Anna, Katie-Ann, Andrew, Fraser, Guadi, Kathryn, Katia, Stefan, and Natascha.

I extend my gratitude to the lovely people I met during the LCT journey: Ruixue, Svetlana, Masha, Ludmila, Joan, Guido, Adam, Aria, Andrea, Gosse, and Nina.

I would also like to express my appreciation to the European Union for granting me the opportunity to study as an Erasmus Mundus scholarship holder with the LCT program. Being a member of the LCT community has been a fantastic and life-changing experience. Special thanks to Bobbye Pernice for her kindness and support from the first day I arrived in Europe until the submission of this thesis.

I would additionally like to extend a heartfelt thank you to all the professors, colleagues, and fellow students whom I encountered during my course of study.

Finally, I would like to thank my parents and my sisters for their unconditional love and support, without which this thesis would not have been possible.

Contents

Part I

FOUNDATIONS

1

Introduction

Prior to the advent of deep learning, the field of modeling **speech signal processing** relied on hidden Markov models (HMMs) and Gaussian mixture models (GMMs). While these traditional models were widely used, they had inherent limitations that hindered their adoption in various speech technology applications. One significant drawback was the need for task-specific feature engineering, which required experts to manually design and engineer task-relevant features. This process was time-consuming, labor-intensive, and often involved domain expertise. Moreover, the complexity of the data pipelines involved in these approaches made it challenging to scale, adapt, and maintain the systems, particularly when faced with different languages, dialects, or acoustic conditions. The reliance on handcrafted features limited the models' ability to capture intricate patterns and complex temporal dependencies in speech signals, leading to suboptimal performance in challenging conditions. As a result, the success of **speech technology applications** that relied on these traditional approaches was rather limited, mainly due to their complexity and inefficiency (Hinton et al., 2012).

Deep neural networks (DNNs) have emerged in the last decade as the leading paradigm in modeling speech processing, thanks to advances in representation learning and the availability of high-performing computing processors. DNNs have overcome the limitations of earlier traditional methods, leading to their widespread adoption for different speech processing tasks, ranging from language identification and keyword spotting to automatic speech recognition (ASR) and spoken dialogue systems. By automatically extracting task-relevant features from acoustic data, **speech representation learning** eliminates the need for manual feature engineering in acoustic models. The flexibility and adaptability of neural networks facilitate the transfer of their weights and representations across different tasks and languages, while earlier approaches had little transferability. End-to-end neural architectures offer scalable and elegant solutions that simplify the complex data pipelines involved in speech processing. For example, and unlike traditional

ASR systems that required separate acoustic and lexical models, along with language models to guide the search process, end-to-end architectures seamlessly integrate bottom-up pattern recognition and top-down linguistic decoding (e.g., Bérard et al., 2016; Chan et al., 2016; Graves, Fernández, et al., 2006; Toshniwal et al., 2018; Ying Zhang et al., 2016). By consolidating different components into a single architecture, the complexity and inter-dependencies of the data pipelines have been substantially reduced. Throughout multiple layers of non-linear transformations, DNNs learn to hierarchically integrate low-level acoustic cues to build high-level, distributed representations of linguistic units that are (perceptually) discrete (e.g., phonemes, syllables, words).

The success of deep neural networks (DNNs) in modeling human speech is particularly impressive if we consider the inherent **complexity** and **variability** of human speech. Speech variability and variation have been extensively studied in the field of human language processing and speech perception research (Bent and Holt, 2017; Clopper and Pisoni, 2021; Klatt, 1989; Luce and McLennan, 2005; Pisoni, 1993, *inter alia*). A key question in this area is how humans can effortlessly comprehend speech despite the **lack of acoustic-phonetic invariance** (Klatt, 1979). In other words, despite the absence of simple mapping between the various acoustic realizations of a phonetic segment and its underlying perceptual unit—the phoneme, humans exhibit a remarkable ability in decoding the speaker's communicative intent encoded in a spoken utterance, even in adherent conditions where the speech signal is noisy or incomplete. The variability in speech goes beyond phonetic segment variability and encompasses various factors, including **speaker-related** factors such as vocal tract shape, gender, age, and **context-related** factors such as coarticulation and predictability of linguistic units within a given context. Additionally, sociolinguistic factors, such as dialect and linguistic experience of the speaker, further contribute to the variability in speech production. As a result, it is highly unlikely for two acoustic realizations of the same linguistic unit to be identical, even if produced by the same speaker.

The problem of speech variability and the intriguing fact of its negligible impact on human listeners during speech comprehension have received considerable attention in human language processing research. However, in machine learning and speech technology research, the study of speech variability has been limited to only a few factors. The primary focus in machine learning research has been on developing models that are robust against variations in recording conditions and acoustic environments, as well as generalization capabilities against training-testing data mismatch scenarios (Benzeghiba et al., 2007; Sriram et al., 2018; Tatman and Kasten, 2017; Tripathi et al., 2018; Z. Wang et al., 2003). Another area of

interest has been the adaptation of acoustic models to account for cross-speaker variability (Hansen and Hasan, 2015; Hazen and James R Glass, 1997; Liao, 2013; Meng, J. Li, et al., 2018; Saon et al., 2013). These research directions primarily focused on improving performance metrics for the given task in the presence of variability. However, the analysis of neural representations in response to speech variability has not received much attention in prior work. In this thesis we argue that analyzing neural network models through the lens of speech variability enables us to understand what research questions need to be asked for a better-informed exploration of speech representations. Towards this end, this thesis presents several studies that address dimensions of variability that have been overlooked, including adaptability to domain variability, cross-linguistic variation, within-category and cross-category spoken-word variability, and acoustic variability of abstract phonetic categories. By grounding its research questions on speech variability, the thesis establishes a connection between different areas of research that have remained so far unconnected—the machine learning of speech processing on one side, and human speech perception, sociolinguistics, cognitive modeling on the other.

1.1 Thesis Statement

Various efforts within the machine learning community have aimed to develop robust models for speech and natural language processing. However, these efforts have traditionally viewed variability as a source of "undesirable noise", and focused primarily on improving model performance in its presence. While these endeavors have indeed improved the robustness and generalization abilities of speech processing models, these models remain "black boxes" that are opaque and difficult to interpret. As a result, we have recently witnessed significant progress in analytic methods for neural network interpretability (See, for example, Alishahi et al., 2017a; Belinkov and J. Glass, 2017; Chung, Belinkov, et al., 2021; Scharenborg, Gouw, et al., 2019; Shah et al., 2021). Despite this progress, the community's primary interest has shifted towards large-scale speech and language models, with an emphasis on their emergent capabilities and their encoding of linguistic structure (e.g., Pasad et al., 2021; G. Shen et al., 2023). Unarguably, understanding the dynamics of these large deep neural network models is an intriguing scientific pursuit. However, since these models are not trained in settings that control for domain and linguistic factors, their analysis falls short in understanding how speech variability shapes their representations. This thesis sets itself apart from ongoing efforts in the research community in two key aspects: (1) it places speech variability at the core, considering deep neural networks as scientific tools to analyze the influence of

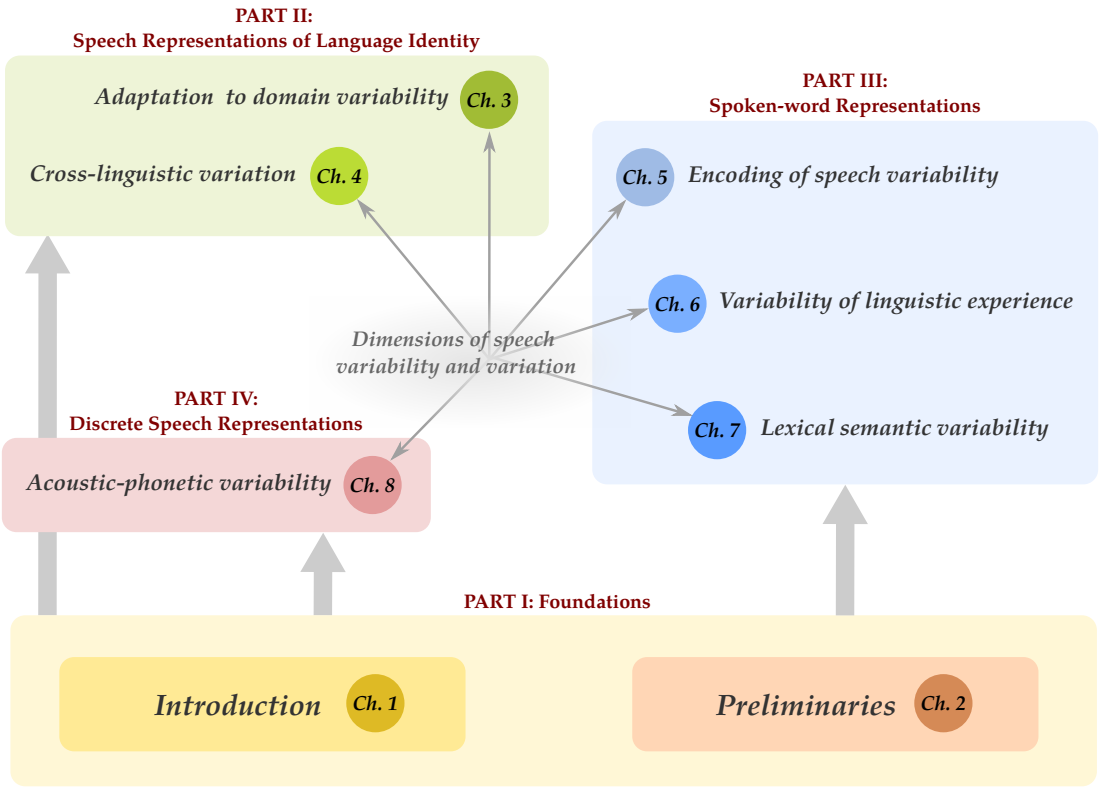


Figure 1.1: A visual illustration of the thesis organization. Chapters are organized into three parts where each part is dedicated a speech processing task. Each chapter presents a study that addresses one dimension of speech variability and variation.

multiple dimensions of variability, and (2) it presents well-controlled experiments that isolate specific sources of variability while controlling for others. Throughout several studies, the thesis aims to bridge the gap between the study of human speech variability and representation learning, with a specific focus on *understanding how modern deep neural networks process and encode the different dimensions of speech variability in their latent representations*. The thesis comprises six studies investigating the role of speech variability across various dimensions, shedding light on how it shapes neural network representations and addressing research questions that are grounded in the fields of sociolinguistics, speech perception, and cognitive modeling research. Approaching the topic through an interdisciplinary lens, this thesis highlights the utility of machine learning models as powerful scientific tools for exploring better-informed research questions within these areas.

1.2 Thesis Overview

This thesis consists of six different studies unified by the overarching objective of analyzing and understanding how deep neural networks process, represent, and encode speech variability and variation. Despite their diversity, all studies in this thesis are concerned with speech processing models that take an untranscribed, continuous acoustic signal as input and produce latent speech representations. In addition to PART I, Foundations, the rest of the thesis is structured into three parts, each dedicated to a specific speech processing task as follows

- In PART I, which consists of chapter 1 and chapter 2, we present an introduction to the thesis as well as the essential knowledge required to understand the experimental studies in the following parts of thesis.
- In PART II, which consists of chapter 3 and chapter 4, we study neural models of spoken language identification and develop a novel approach to mitigate the effect of domain variability and analyze the encoding of cross-linguistic variation in the intermediate model representations.
- In PART III, which consists of chapter 5 through chapter 7, we focus on neural models of spoken-word processing. Through three different studies, we explore how speech variation is encoded in the their representations, investigate the impact of linguistic experience variability (as characterized by the language of the training data) on their representational profile, and examine how semantic supervision facilitate the abstraction from variability in the acoustic realization of spoken words.
- In PART IV, which consists of chapter 8, we develop a framework based on information-theory to analyze how self-supervised discrete representations of speech reflect acoustic-phonetic variability.

Each chapter in this thesis is summarized as follows

Part I: Foundations

Chapter 1 presents an introduction to the thesis, its structure, contributions, and the publications it was based on.

Chapter 2 presents some preliminaries that are used in the rest of this thesis. These preliminaries include essential theoretical knowledge regarding human speech as well as some technical foundations on the representation of speech in machines.

Part II: Speech Representations of Language Identity

Chapter 3 is concerned with the problem of domain variability and its impact on neural network representations for spoken language identification (SLID). The impact of domain variability is quantified by the extent to which a SLID model trained on one dataset (i.e., source domain) can generalize to speech samples from an unseen dataset (i.e., target domain). The study in this chapter first shows that the representations of convolutional models for SLID do not transfer well across different domains that vary in their recording conditions (i.e., read speech and broadcast speech). The chapter then presents a novel approach based on unsupervised adversarial adaptation to encourage the model to build domain-invariant speech representations. Further analysis in this study shows that adversarial training prevents neural networks from exploiting dataset-specific artifacts as predictive features for the language, thus leading to better cross-domain generalization.

Methods: unsupervised domain adaptation, domain adversarial learning, gradient reversal, convolutional neural networks.
Sources of variability: domain variability, discrepancy in recording conditions.
Relevant publications:

- Badr M. Abdullah, Tania Avgustinova, Bernd Mobius, and Dietrich Klakow. **Cross-Domain Adaptation of Spoken Language Identification for Related Languages: The Curious Case of Slavic Languages**. *In the proceedings of Interspeech 2020.*

Own contributions:

- The implementation of the data preprocessing, training, and evaluation pipeline.
- The code development for the neural models and adversarial training.
- The analysis and visualization of the results.
- Writing the paper with assistance from co-authors.

Chapter 4 is concerned with the encoding of cross-linguistic variation in DNN-based representations of spoken language identity. While DNN-based models have been shown to perform very well on the task of discriminating related languages from acoustic speech signals, it remains unknown whether they capture cross-linguistic variation in their intermediate representations. This chapter presents a

case study on the related Slavic languages that investigates the degree to which the model's representational similarity among languages reflect objective measures of language similarity. Even though the model does not have access to any signal regarding how the languages relate to each other, this study demonstrates that the model representations exhibit a cluster structure that corresponds to the phylogenetic groups within the Slavic language family, even for languages that are not observed during training. The findings of this study provide further evidence that neural networks learn to faithfully encode the hierarchical grouping of the data in a way that largely corresponds to our linguistic intuitions.

Methods: exploratory visualization analysis, geographic correlation analysis, clustering analysis.
Sources of variability: cross-linguistic variation.
Relevant publications:

- Badr M. Abdullah, Jacek Kudera, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. **Rediscovering the Slavic Continuum in Representations Emerging from Neural Models of Spoken Language Identification**. *In Proceedings of VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, COLING 2020.*

Own contributions:

- The conceptualization of the research study.
- The code development for representation extraction, data visualization, and analytic methods.
- Writing the paper with assistance from co-authors.

Part II: Spoken-Word Representations

Chapter 5 is the first of three chapters that study neural models of spoken-word processing and auditory-lexical representations. In this part, each word is modeled as an abstract category consisting of several acoustic exemplars that vary across speakers and within-speaker given different lexical contexts. Spoken-word processing models encode each acoustic exemplar in a representational space such that different exemplars of the same word category are nearby. The chapter begins this part by presenting an analytic study from a neural network interpretability perspective. Concretely, the study in this chapter takes a closer look into the geometry of spoken-word representation space to investigate how it encodes cross-category variability. The first analysis in this study shows that

models of spoken-word representations tend to be highly anisotropic, which means that the variation within the speech samples is encoded in a small fraction of all possible dimensions. The second analysis shows word discriminability positively correlates with word acoustic distinctiveness—operationalized as phonological surprisal and word length—but does not correlate with word frequency. The third analysis demonstrates that (trivial) variability in the initial conditions (i.e., random initializations) yield neural networks for spoken-word representations to exhibit substantial individual differences in their geometry, specially for models trained with contrastive objectives. This chapter concludes with a few recommendations on using DNN-based spoken-word representations as cognitive models for spoken-word processing.

Methods: convolutional and recurrent neural networks, geometric isotropy analysis, representational similarity analysis.
Sources of variability: cross-category and within-category variability of spoken-words.
Relevant publications:

- Badr M. Abdullah and Dietrich Klakow. **Analyzing the Representational Geometry of Acoustic Word Embeddings**. *In Proceedings of BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, EMLNP 2022.*
- Badr M. Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, and Dietrich Klakow. **Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study**. *In Proceedings of Interspeech 2021.*

Own contributions:

- The implementation of the data preprocessing, training, and evaluation pipeline.
- The code development for neural models of spoken-word representations.
- Supervising the research assistant, Iuliia Zaitova, who was involved in preprocessing the data for the languages in the study.
- The visualization of the concepts and analysis of the results.
- Writing the papers with assistance from co-authors.

Chapter 6 presents a comprehensive study that aims to model the role of variability in the linguistic experience and its impact on non-native speech processing. Concretely, this study characterizes the linguistic experience as the "native"

language of the model during training and it simulates human spoken-word processing by monolingually training neural networks models on different languages. The main contribution of this chapter is introducing a framework based on representational similarity analysis that quantifies the extent to which non-native models produce native-like representations. Furthermore, the chapter presents a case study on the closely related West Slavic languages (namely Czech and Polish) and demonstrates that representational similarity correctly predicts higher mutual intelligibility among West Slavic languages in comparison to other Slavic languages (e.g., Russian, an East Slavic language) as well as other languages spoken in Europe (e.g., German, a non-Slavic language). Since the models trained for this study do not have access to high-level linguistic information such as sentence context or lexical semantics, these findings provide evidence that cross-linguistic intelligibility can be partly attributed to form-based processing.

Methods: recurrent neural networks, representational similarity analysis, centered kernel alignment.
Sources of variability: variability in the linguistic experience, cross-linguistic variation.
Relevant publications:

- Badr M. Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. **How Familiar Does That Sound? Cross-Lingual Representational Similarity Analysis of Acoustic Word Embeddings**. *In Proceedings of BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, EMLNP 2021.*

- Badr M. Abdullah, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. **Representational Similarity Predicts Cross-linguistic Intelligility: Insights from Computational Modeling**. *A journal article under-review.*

Own contributions:

- The conceptualization of the research study.
- The development of the experimental pipeline and the underlying codebase.
- The visualization of the concepts and analysis of the results.
- Conducting and maintaining the experiments on the local university servers.
- Writing the papers with assistance from co-authors.

Chapter 7 presents a study that takes inspiration from human speech processing to encourage computational models of spoken-word processing to abstract away from speaker and word enviroment variability. In chapter 5 and 6, models are trained with supervision signals that only capture low-level, form-based information about the word. That is, spoken-word representations are learned in a bottom-up approach whereby acoustic-phonetic cues are combined in the model to encode its acoustic and phonological structure. Nevertheless, a host of psycholinguistic studies on human listeners have shown that top-down, high-level lexical properties—such as word semantics—not only interact with the word recognition process but also facilitate discrimination between word competitors. Inspired by prior psycholinguistic findings, this chapter presents a model that integrates form-based and meaning-based supervision via multi-task learning. The proposed model learns a mapping from the acoustic input onto a lexical representation in addition to bottom-up form-based supervision. This chapter presents experiments on three languages and empirically demonstrates that integrating high-level lexical knowledge into training spoken-word representations improves the ability of the model to discriminate between word categories.
Methods: recurrent neural networks, phonological decoding, form-to-meaning regression, dimensionality reduction.
Sources of variability: cross-category and within-category variability of spoken-words, variability in word semantic content.
Relevant publications:

- Badr M. Abdullah, Bernd Möbius, and Dietrich Klakow. **Integrating Form and Meaning: A Multi-Task Learning Model for Acoustic Word Embeddings**. *In Proceedings of Interspeech 2022.*

Own contributions:

- The Implementation of the data preprocessing, training, and evaluation pipeline.
- The code development for the semantically-enriched model of spoken-word representations.
- The visualization and analysis of the results.
- Writing the paper with assistance from co-authors.

Part III: Discrete Speech Representations

Chapter 8 focuses on latent, discrete representations of speech that emerge while training transformer-based models via self-supervision and codebook learning. This chapter aims to characterize the relationship between discrete speech representations and abstract phonetic categories such as phonemes. Concretely, the chapter first proposes an information-theoretic framework whereby each phonetic category is represented as a distribution over discrete units. The chapter then presents a case study on how two different self-supervised models (namely, wav2vec 2.0 and XLSR) encode American English speech as discrete units. The study shows that the entropy of phonetic distributions reflects the acoustic-phonetic variability of the underlying speech sounds, with sonorants being more entropic on average than obstruents. In addition, phonetically similar sounds are found to exhibit similar distributions at the low level while a clustering analysis shows the highest level of division separates obstruents and sonorants. The findings of this study suggests the characterization of discrete units as sub-phonemic events, rather than high-level categories such as phonemes.

Methods: self-supervised speech models, information theory, information entropy and surprisal, Jensen-Shanon divergence.
Sources of variability: acoustic-phonetic variability, segment realization variability.
Relevant publications:

- Badr M. Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, and Dietrich Klakow. **An Information-Theoretic Analysis of Self-supervised Discrete Representations of Speech**. *In Proceedings of Interspeech 2023*.

Own contributions:

- The conceptualization of the research study.
- The visualization and analysis of the results.
- Supervising the research assistant, Mohammed Maqsood Shaik, who extracted the discrete representations and aligned them to phonetic categories in TIMIT dataset.
- Writing the papers with assistance from co-authors.

1.3 Additional Publications

In addition to the studies discussed in next chapters, the author of this PhD thesis has been involved in a few collaborations that have contributed the following publications:

- Marius Mosbach, Stefania Degaetano-Ortlieb, Marie-Pauline Krielke, Badr M. Abdullah and Dietrich Klakow. **A Closer Look at Linguistic Knowledge in Masked Language Models: The Case of Relative Clauses in American English**. *In Proceedings of International Conference on Computational Linguistics, COLING 2020.*

- Alexandra Mayn, Badr M. Abdullah and Dietrich Klakow. **Familiar words but strange voices: Modelling the influence of speech variability on word recognition**. *In Proceedings of the student research workshop, EACL 2021.*

- Nicole Macher, Badr M. Abdullah, Harm Brouwer and Dietrich Klakow. **Do we read what we hear? Modeling orthographic influences on spoken word recognition**. *In Proceedings of the student research workshop, EACL 2021.*

- Elizabeth Salesky, Badr M. Abdullah, Sabrina J. Mielke, Elena Vitalievna Klyachko, Oleg Serikov, E. Ponti, Ritesh Kumar, Ryan Cotterell and Ekaterina Vylomova. **SIGTYP 2021 Shared Task: Robust Spoken Language Identification.** *In Proceedings of the SIGTYP Workshop for Research in Computational Linguistic Typology and Multilingual, EACL 2021.*

- Iuliia Zaitova, Badr M. Abdullah and Dietrich Klakow. **Mapping Phonology to Semantics: A Computational Model of Cross-Lingual Spoken-Word Recognition**. *In Proceedings of VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, COLING 2022.*

- Badr M. Abdullah, Mohammed Maqsood Shaik, and Dietrich Klakow. **On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models**. *In Proceedings of the SIGTYP Workshop for Research in Computational Linguistic Typology and Multilingual, EACL 2023.*

- Julius Steuer, Badr M. Abdullah, Johann-Mattis List, and Dietrich Klakow. **Information-Theoretic Characterization of Vowel Harmony: A**

Cross-Linguistic Study on Word Lists. *In Proceedings of the SIGTYP Workshop for Research in Computational Linguistic Typology and Multilingual, EACL 2023.*

2

Preliminaries

While each part of this thesis is self-contained, it is important to have a foundational understanding of the motivation and methods behind the studies presented in the next chapters. Therefore, this chapter serves as an overview of the theoretical and technical aspects that form the basis of this thesis.

2.1 The Dual Nature of Human Speech

One of the most fascinating characteristics of human speech is its dual nature (James Robert Glass, 1988; Ladd, 2011; L. J. Lee, 2004). On the one hand, speech is **discrete**. That is, listeners perceive speech as a sequence of linguistic units that are discrete and categorical, usually in the form of phonemes or syllables. The phoneme as a concept is well-established in speech science as "the smallest unit of speech that can distinguish one word from another in a particular language". For instance, the words *bill* and *pill* differ by one phoneme in word-initial position. Experimental studies have shown that human listeners who did not practice an alphabetic writing system can discriminate between phonemic categories, and this effect has been observed in both preliterate children and illiterate adults. Furthermore, each language is characterized by a finite inventory of phonemes that are combined to form syllables, and then words. The word formation process is governed by language-specific phonotactic rules that make some sound combinations more probable than others. To form complete utterances, words are combined according to morphosyntactic rules of the language. These linguistic units form a hierarchy that represents the linguistic description of speech (see Figure 2.1).

On the other hand, speech is **continuous**. That is, the physical realization of human speech is a continuous stream of energy with no explicit boundaries

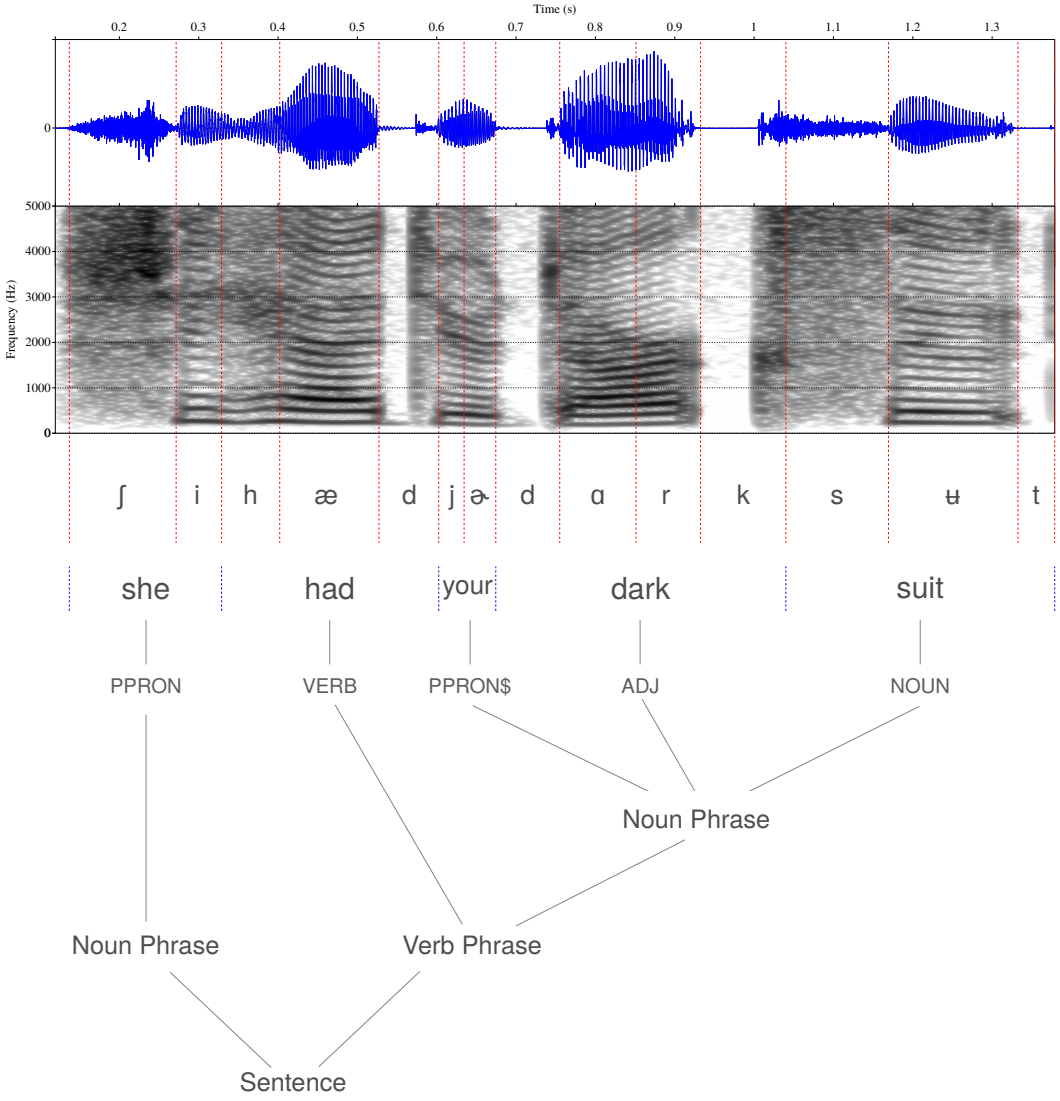


Figure 2.1: From the continuous to the discrete: the linguistic description of speech.

between adjacent linguistic units. Speech processing can thus be viewed as a process of decomposing the complex continuous speech signal into form-based representations. These representations then make contact with higher-level linguistic representations at the levels of syntax, semantics, and pragmatics, with the ultimate goal of decoding the communicative intent of the speaker. Given this definition of speech processing, the acoustic speech signal can be thought of as a stream of energy that encodes three sources of information (Clopper and Pisoni, 2005): (1) the linguistic message, (2) individual speaker attributes, and (3) the communication channel. Each source of information in the signal is inherently subject to variability. For example, the encoding of a linguistic message may substantially vary due to in-context predictability, leading to the reduction of predictable expressions either lexically or phonetically, or at both levels. On the other

hand, speakers vary in their vocal tract shapes, which affects speech production, as well as in their dialect or language variety. Lastly, the communication medium itself can induce changes in the acoustic manifestation of speech due to variability in the environment or electronic transmission. Considering all these sources of variability in speech, the ability of humans to effectively communicate in everyday life is indeed remarkable.

2.2 Speech Variability and Variation

The traditional approach to the study of speech perception and language processing has often relied on abstract phonemic descriptions of speech that largely ignore its inherent variability across different utterances, speakers, and contexts. In this paradigm, the main assumption is that the human auditory system normalizes the acoustic signal into an abstract representation that is largely invariant to the previously discussed sources of variability. Proponents of this approach view the sources of variability in speech as natural consequences of language variation and argue that variability is not something to be discarded but rather should be incorporated into conceptual and computational models of speech perception and processing (e.g., Clopper and Pisoni, 2021; Pisoni and Levi, 2012). According to this alternative approach, human listeners encode these "**indexical**" properties of speech in their memory, leveraging the variability in speech signals to help them better understand and interpret linguistic messages. For example, exemplar-based models of speech perception posit that every experience of a spoken word leaves a memory trace (or exemplar) that is rich in acoustic-phonetic detail, including speaker-specific and context-dependent information (e.g., Hawkins, 2003; Pierrehumbert, 2002; Port, 2007). Then, the recognition of linguistic units involves the activation of and comparison among these exemplars in memory.

Furthermore, the problem of language variation has been a long standing topic of interest to sociolinguists. One can identify two lines of research that have emerged in the sociolinguistics literature which delve into different dimensions of variations. The first line of research is concerned with the descriptive study of language variation as it occurs in social strata, geographical regions, and ethnic groups. Researchers along this line of research are faced with questions about the social implications of such variation (Clopper and Pisoni, 2021). The second line of research is concerned with how variation in language is processed by listeners in order to decode the intended message by the speaker. This line of research focuses on language processing in the face of dialect variation and develops objective measures to quantify the mutual intelligibility of related language varieties (Gooskens,

2017; Van Heuven, 2008). The latter line of sociolinguistic research bears greater relevance to the work presented in this thesis with respect to the representation of cross-linguistic variation in neural networks and the role of linguistic experience non-native spoken word processing.

While neural networks have been adopted as cognitive models for speech processing in recent research (Magnuson et al., 2020; Matusevych, H. Kamper, et al., 2021, *inter alia*), the study of speech variability under the framework of representation learning has been somehow limited. This situation is largely due to the predominant emphasis within the machine learning community on improving performance metrics for predefined benchmarks, which are typically used as indicators of progress. As a result, a majority of the research efforts have been channeled towards the development of models that exhibit robustness against variability in recording and acoustic conditions. Moreover, considerable focus has been dedicated to improving the models' ability to generalize, thereby enabling them to effectively handle discrepancies between training and evaluation datasets. However, an area that remains relatively underexplored is the investigation of how neural network representations are shaped by speech variability. While a few factors such as recording conditions and speaker-specific variations have sufficiently been studied, the subtler aspects of speech variability—for example, acoustic-phonetic and cross-linguistic variations—are not adequately explored and analyzed. This reveals an essential gap in our understanding that requires further comprehensive research that takes a closer look into the nuances of speech variability under the representation learning framework.

2.3 Spectral Representations of Speech

The (digital) speech signal can be described as a sequence of feature vectors where the dimensions of each vector correspond to the energy of different frequency bands. This low-level feature representation is produced via a signal processing pipeline transforming the temporal audio waveform into a frequency domain representation consisting of spectral vectors, each representing the information in a small temporal window of the signal. Consider a digitized audio signal $s[n]$ consisting of N samples, where n is an integer that is an index over time (i.e., the n^{th} sample). The corresponding spectral representation can be described as a sequence $\boldsymbol{A} = \boldsymbol{a}_{1:T} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)$ consisting of T acoustic vectors, where each vector $\boldsymbol{a}_t \in \mathbb{R}^k$ and k represents the number of frequency bands. The most common spectral representations in ASR research are Mel-frequency filter banks, also known as Mel-Frequency Spectral Coefficients (MFSCs), and Mel-Frequency Cepstral

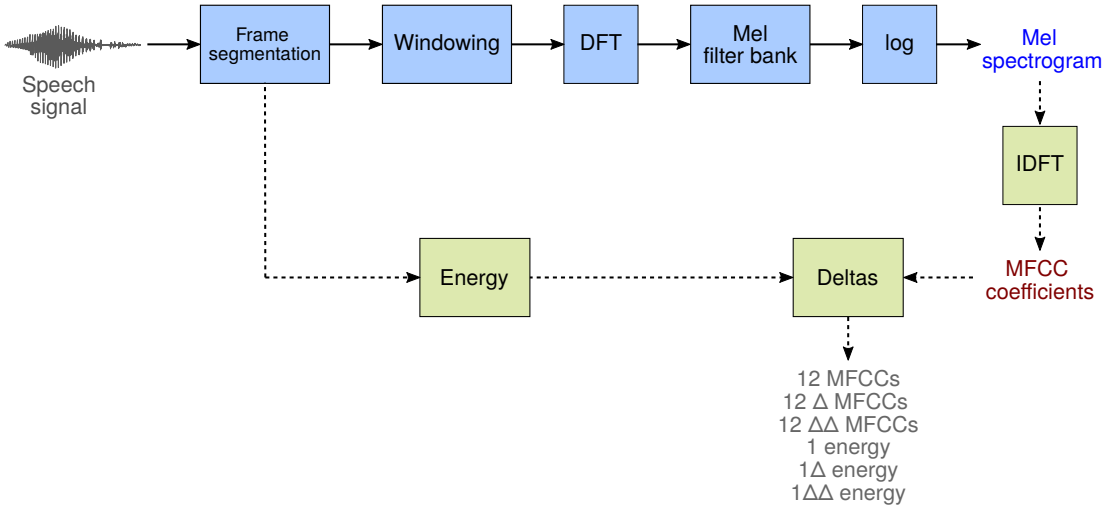


Figure 2.2: A schematic that illustrates the different processing steps in extracting spectral representations from a speech waveform. Dashed lines indicate processing steps that are required only for MFCCs. The figure is based on the extraction pipeline from Jurafsky and Martin (2000).

Coefficients (MFCCs). The preprocessing pipeline for these feature representations typically involves the following steps (illustrated in Figure 2.2):

- **Frame segmentation:** The temporal speech signal $s[n]$ is first divided into T number frames each consisting of J samples, typically with a stride of K samples between the start of adjacent frames ($J < K$). Even though speech is a non-stationary, time-varying signal, this segmentation assumes that each frame is a quasi-stationary signal with negligible variation in energy characteristics in the corresponding time window. Therefore, the frame size J and stride K are chosen such that the properties of the audio signal are fairly time-invariant within each frame.

- **Windowing:** To minimize discontinuities at the edges of the frames, each frame is processed by a window function such as a Hamming window, which numerically reduces the values of the signal toward zero at the window boundaries.

- **Discrete Fourier Transform (DFT):** The DFT step is applied to extract spectral information of each windowed frame and convert it from the time domain to the frequency domain. The goal of this step is to compute the amount of energy in each frequency band.

- **Mel Filter Bank and Log:** The human auditory system is known to exhibit various sensitivities to different frequency bands. That is, the human ear is

more sensitive to frequency bands below 1000 Hz. Modeling this property in the feature extraction pipeline has been shown to improve speech recognition performance. To this end, the powers of the spectrum are computed and then mapped using a set of triangular filters that are spaced according to the Mel scale. The Mel scale is a perceptual scale of pitches where equal distances correspond to equal perceived differences in pitch. Then, and since the human auditory system seems to be logarithmic, the log operation is applied to each of the Mel values. These operations together result in a Mel filter bank representation, or MFSCs.

- **Discrete Cosine Transform (DCT):** The amount of energy at different frequency bands can exhibit statistical correlations that are undesirable for earlier models of speech processing such as Gaussian Mixture models. To decorrelate the spectral features, a Discrete Cosine Transform (DCT) is applied to the log Mel filter bank energies to decorrelate the energies. Typically only the first 12-13 DCT coefficients are kept in order to decompress the data.

- **Delta and Delta-Delta features:** It is a common practice to append the delta (first derivative) and delta-delta (second derivative) of the MFCCs to the feature vector to incorporate some information that captures the dynamics of the speech signal. This step is optional and can be omitted for models that are designed to capture the temporal dynamics of speech such as deep neural networks.

This signal processing pipeline has been the subject of extensive engineering in the last few decades. Although various modifications to this pipeline have been proposed in the literature, the improvements due to the feature extraction pipeline seem to be trivial compared to modeling innovations in the architectural system design and training data, specially when modeling speech using deep neural networks.

2.4 Neural Network Representations of Speech

In this section, we present three neural network architectures that are widely employed as representation models for speech processing.

2.4.1 Convolutional Neural Networks

Although convolutional neural networks (CNNs) have been initially developed in the computer vision research, they have been adapted for language problems in NLP (Collobert et al., 2011; Kalchbrenner et al., 2014; Kim, 2014; Ye Zhang and Wallace, 2015) as well as speech processing (Abdel-Hamid et al., 2014; Merkx and Scharenborg, 2018; Palaz et al., 2015; Sainath et al., 2015). One-dimensional convolutional networks are often employed as a front-end processor to identify local patterns and extract relevant phonetic features from a raw speech input or a spectral representation of the audio signal. In this section, we consider the case of a spectral representation in the form of a sequence of acoustic feature vectors such as MFCCs. Let us denote the spectral representation of the speech input as a matrix $\mathbf{A} \in \mathbb{R}^{k \times T}$, where k is the number of frequency bands, or channels, and T is the number of frames in the signal. This matrix can be formally described as

$$\mathbf{A} = \boldsymbol{a}_{1:T} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_T \\ | & | & | & | \end{bmatrix} \tag{2.1}$$

To extract high-level features that are predictive for the speech processing task, a convolutional filter $\mathbf{W} \in \mathbb{R}^{k \times M}$ is applied to a window of M acoustic vectors to obtain a new latent feature

$$\mathbf{c}_t = \alpha(\mathbf{W} * \boldsymbol{a}_{t:t+M-1} + \mathbf{b}) \tag{2.2}$$

where $\mathbf{c}_t$ is a scalar, $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters that are shared across time (i.e., same set of parameters applied to all possible convolutional windows $\boldsymbol{a}_{t:t+M-1}$), the $*$ operation is an element-wise matrix multiplication followed by summation (in order to obtain a scalar), and α is a non-linear activation function. The parameters of the filter $\mathbf{W}$ are learned when training the neural network and back-propagating the error in a (self)-supervised learning task. The convolutional filter is then applied to all successive convolutional windows of width M by sliding

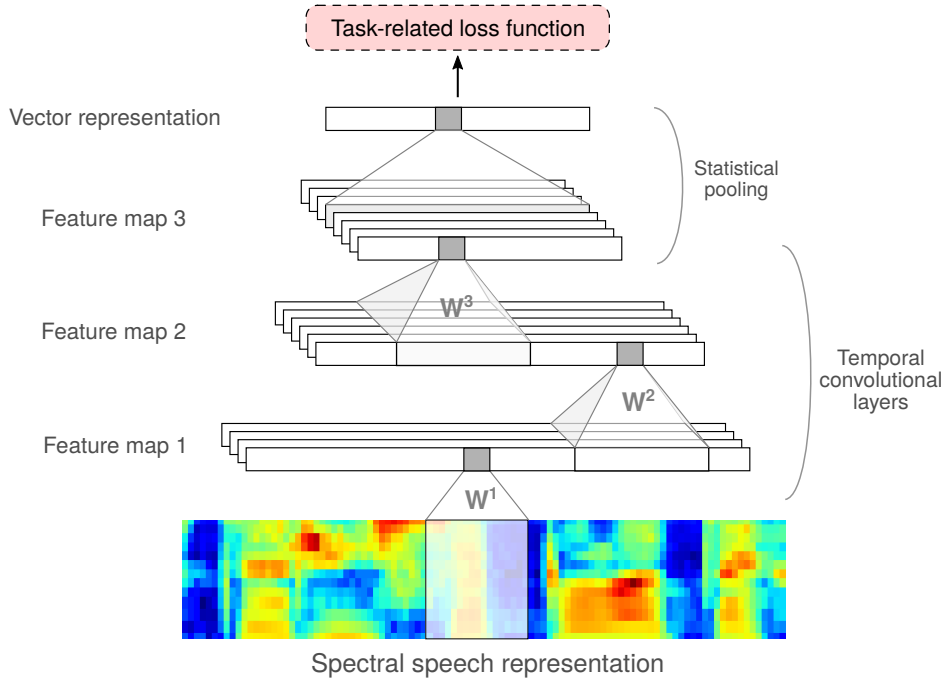


Figure 2.3: A visual illustration of a convolutional neural network for learning high-level representations of speech. The convolutional block in this example network consists of three convolutional layers followed by statistical pooling operation.

the filter over the $k \times T$ matrix that represents the input speech to induce a feature map $\mathbf{C} \in \mathbb{R}^{(T-M+1)}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_{T-M+1} \end{bmatrix} = \begin{bmatrix} | \\ \mathbf{c}^1 \\ | \end{bmatrix} \tag{2.3}$$

Note that multiple convolutional filters are simultaneously applied on the input so that different filters will specialize in detecting various phonetic and lexical features. That is, instead of applying a single convolutional filter $\mathbf{W}$, a set of filters $\{\mathbf{W}_i\}_{i=1}^{F}$ are applied and jointly learned during training. The resulting representation from applying multiple filters is the feature map $\mathbf{C} \in \mathbb{R}^{(T-M+1)\times F}$, which can be described as follows

$$\mathbf{C} = \begin{bmatrix} | & | & | & | \\ \boldsymbol{c}^1 & \boldsymbol{c}^2 & \dots & \boldsymbol{c}^F \\ | & | & | & | \end{bmatrix} \tag{2.4}$$

The presented computation so far only describes the processing of speech in a single convolutional layer. Oftentimes, the input goes through several layers of convolutions (e.g., l convolution operations) that learn higher levels of abstraction of the input signal as a set of feature maps $\{\mathbf{C}^1, \cdots, \mathbf{C}^l\}$. At the end of the convolutional block, the last feature map $\mathbf{C}^l$ is down-sampled using a statistical pooling operation over time to obtain a statistical summary of the feature that is captured by the filter

$$\hat{\mathbf{c}} = \texttt{stat-pool}(\mathbf{C}^l) \tag{2.5}$$

Here, $\hat{\mathbf{c}} \in \mathbb{R}^{F^l}$ is a high-level feature representation, where F^l is the number of filters in the last convolutional layer. The pooling operation makes it possible to obtain fixed-size representations for variable length sequences. For a speech classification task, the final vector representation is then passed to a few fully-connected layers followed by a softmax function to obtain a probability distribution over the set of output labels. For tasks that require learning a vector representation of the speech input as a task of its own, the resulting high-level feature representation is passed into an objective function that computes a loss, for example via contrastive learning. However, modern CNN architectures may potentially involve many other operations (e.g., batch normalization, skip connections, unit dropout, etc.), which can be quite complex compared to the one described in this section.

2.4.2 Recurrent Neural Networks

In representation learning, the conventional architecture for processing variable-length sequences is a recurrent neural network (RNN). Given a sequence of spectral feature vectors $\mathbf{A} = \boldsymbol{a}_{1:T} \in \mathbb{R}^{k \times T}$, an RNN is recursively defined as follows

$$\mathbf{h}_t = \begin{cases} \mathcal{F}_{RNN}\Big(\mathbf{h}_{t-1}, \boldsymbol{a}_t; \theta\Big) & \text{if } t \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

where $\mathbf{h}_t$ is the hidden state of network at time step t and θ are the parameters of the recurrent function. This recursive definition enables the RNN to process sequences of arbitrary length, and that is the case with speech data. In theory, the hidden state ht is a summary of the information that has been observed in the sequence up to timestep t. In practice, vanilla RNNs fails to encode all information in the hidden state when sequences are long. It was shown that this limitation of RNNs is caused by the vanishing gradient problem (Y. Bengio et al., 1994; Jozefowicz et al., 2015). To address this problem, gated recurrent structures with memory elements have been proposed. The most notable variants of gated recurrent structures are the long-short term memory (LSTM, Hochreiter and Schmidhuber, 1997) and the gated recurrent unit (GRU, Cho et al., 2014). Since GRUs are used in this thesis, we introduce the typical computational dataflow for a GRU in this section. A GRU is characterized by two gates: reset gate $\mathbf{r}_t$ and update gate $\mathbf{z}_t$. The motivation to use these two gates is to adaptively control how much information from the hidden state should be carried to the next state and how much information from the current input should be taken in the next hidden state. The vectorized computation of GRU can be expressed as follows

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \boldsymbol{a}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \tag{2.7}$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \boldsymbol{a}_t + \mathbf{U}_z \mathbf{h}_{t-1}) \tag{2.8}$$

$$\widetilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{v}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \tag{2.9}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \widetilde{\mathbf{h}}_t \tag{2.10}$$

The recurrent computation presented so far considers processing the input sequence in one direction. The last hidden state in the RNN can be considered as a representation that compresses, or summarizes, the entire input sequence in a vector. However, it was shown that processing the sequence in two directions (i.e., forward and backward) can improve the performance of many sequence modeling tasks such as machine translation (Bahdanau et al., 2014) and speech recognition (Graves, A.-r. Mohamed, et al., 2013). In the bidirectional recurrent structure, the last hidden states in the forward direction and backward direction are merged (either by concatenation or element-wise addition) to form a single vector that represents the sequence. For speech classification tasks, this vector is transformed via a fully connected layer into a softmax vector that represents a probability distribution over the output space. In this thesis, we focus on unidirectional recurrent networks. A schematic view of a unidirectional RNN for speech processing is depicted in Figure 2.4.

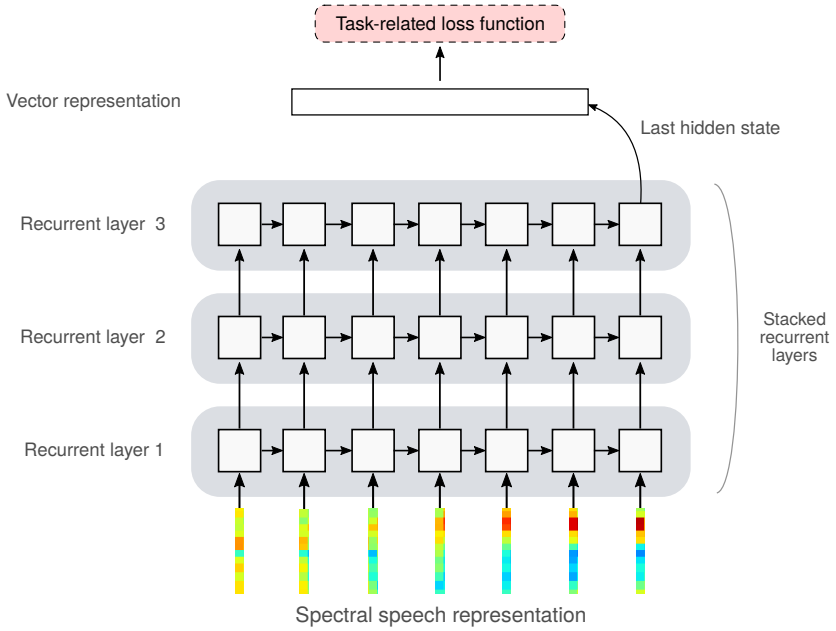


Figure 2.4: A visual illustration of a unidirectional recurrent neural network for learning high-level representations of speech. The recurrent block in this example network consists of three (stacked) recurrent layers.

2.4.3 Transformer Neural Networks

The transformer architecture has been a revolutionary innovation in the field of NLP as well as speech processing (Vaswani et al., 2017). Transformer-based models are particularly effective for sequence-to-sequence tasks where the input and output sequences can be of different lengths, such as in speech recognition, text-to-speech synthesis, or speaker diarization. They are also effective for speech classification tasks, including spoken language identification and emotion recognition. Moreover, large speech transformer-based models for speech can be pre-trained via self-supervision without explicit transcriptions or labels. Since the last study of this thesis presents an analysis of the transformer-based speech representations from wav2vec 2.0 (Baevski et al., 2020), this section describes the computations within this model.

Contrary to earlier neural models, wav2vec 2.0 does not operate on spectral speech representations, but directly on the raw audio. To reduce the complexity of the acoustic signal, the input is first passed through a series of convolutional layers that downsample it. Each acoustic frame consists of a 25 msec of speech interval (with a 10 msec stride), which is then transformed into a single representation that is suitable to be further processed by the transformer layers. This step produces a feature vector sequence at a rate of 50 frames per second. It has been shown that

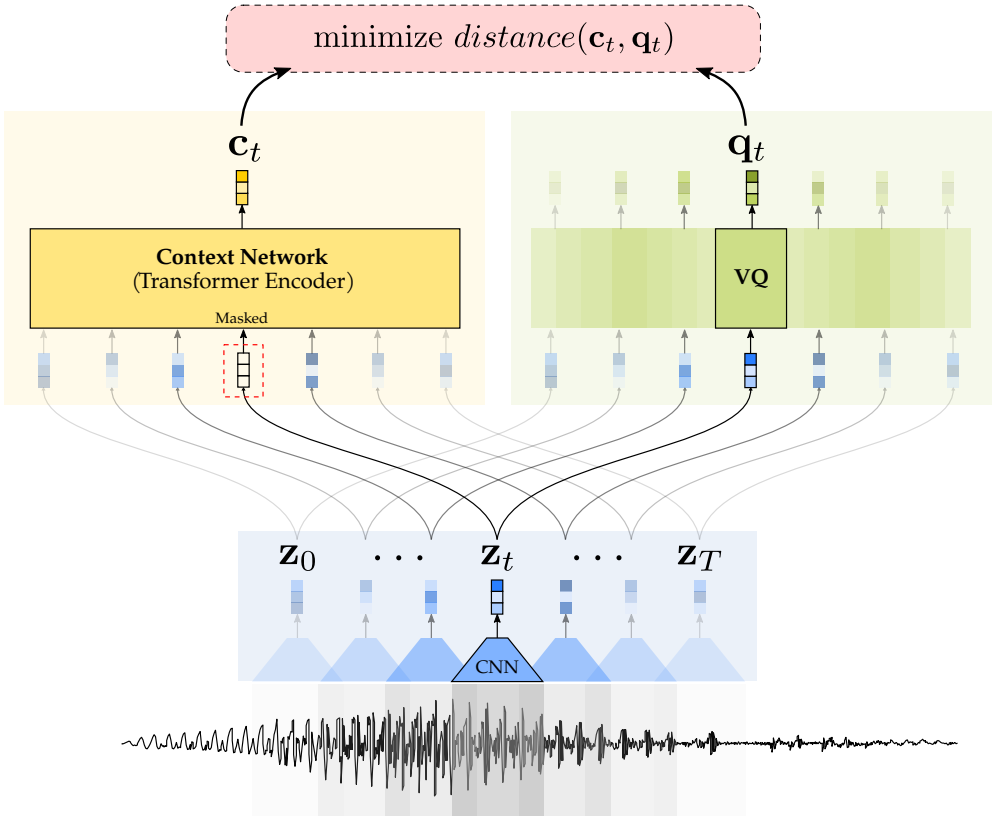


Figure 2.5: A visual illustration of the wav2vec 2.0 model. The vector quantization module (VQ) is only used during pre-training. Adapting the model to a downstream task requires fine-tuning the Transformer network using labeled speech data.

the output of the convolutional layers is highly correlated with spectral speech representations (Pasad et al., 2021).

Formally, consider a continuous acoustic signal represented as a sequence of T acoustic frames denoted as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. As previously mentioned, here $\mathbf{x}_t$ is an interval of the raw waveform that corresponds to 25 msec of speech. Within the wav2vec model, the signal $\mathbf{x}$ is first transformed via a local, temporal convolutional encoder $f : \mathcal{X} \mapsto \mathcal{Z}$ into a sequence of latent speech representations in a continuous space as $f(\mathbf{x}) = \mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$, where $\mathbf{z}_t \in \mathbb{R}^d$. During pre-training, the feature sequence is fed into two different blocks:

- A Transformer network $g : \mathcal{Z} \mapsto \mathcal{C}$ that contextualizes the feature sequence and build representations $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_T)$ via the self-attention mechanism to (ideally) capture information at the global level (i.e., entire sequence).

- A quantization module $q : \mathcal{Z} \mapsto \mathcal{Q}$ that discretizes the feature sequence and to induce quantized representations $\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_T)$ via product quantiza-

tion. These quantized representations are the targets of the self-supervised learning objective during pre-training.

Note that the quantization module is only used during pre-training and completely discarded when fine-tuning wav2vec 2.0 on a (labeled) downstream task such as speech recognition or language identification. Nevertheless, the nature of the discrete unit that emerge in the model during pre-training are interesting from an interpretability point of view. Therefore, we dedicate the last chapter of this thesis to the phonetic analysis of these discrete units. The details of the computations within the quantization module are explained in-depth in Chapter 8. In this section, we focus on the computations of the contextualization network based on the Transformer architecture.

The Transformer block consist of L number of layers (i.e., usually 12 or 24 layers in speech models). The output of the convolutional feature extractor $\mathbf{z}_t$ is passed through the L layers of the transformer network. Each layer consists of a multi-head self-attention mechanism and a position-wise feed-forward network. For example, the first Transformer's layer (i.e., $l = 1$) takes the output of the convolutional encoder as input and produces a contextualized representation as output. At the high-level, this operation can be formalized as follows

$$\mathbf{c}^{(1)} = \texttt{FNN}(\texttt{self-attention}(\mathbf{z})) \tag{2.11}$$

Here, `self-attention`() is s the scaled dot-product attention operation introduced in Vaswani et al. (2017) and `FNN`() represents a feedforward neural network. The `FNN`() function applies a linear transformation followed by a non-linear activation to each position separately, and then another linear transformation. Generalizing from the first layer to the other layers in the Transformer network, the contextualized representation at any layer l can be described as follows

$$\mathbf{c}^{(l)} = \texttt{FNN}(\texttt{self-attention}(\mathbf{c}^{(l-1)})) \tag{2.12}$$

A detailed mathematical description of `self-attention`() operation is beyond the scope of this section, and we therefore refer an interested reader to the work of Vaswani et al. (2017) to get a deeper view of the self-attention mechanism. However, at the high-level, the `self-attention`() operation enables a representation $\mathbf{c}_t^{(l)}$ in a particular temporal position t to be interact with the representations of the lower layer $\mathbf{c}^{(l-1)} = (\mathbf{c}_1^{(l-1)}, \ldots, \mathbf{c}_T^{(l-1)})$ across all temporal positions. The self-attention

mechanism eliminates the need of a recurrent function to model the temporal dynamics of the speech signal. The temporal range of the attention mechanism in Transformer network is unbounded in theory, which enables the Transformer layers to capture longer-term dependencies such lexical and discourse-level context of a spoken sentence.

2.5 Information Theory

Information theory, at its core, is a mathematical framework that aims to study and quantify the storage and communication of information. As a framework, information theory was first pioneered by Claude Shannon and was fundamentally concerned with quantitatively defining information transmission and compression, as well as understanding the limits and capabilities of information systems. Moreover, information theory has been shown to be applicable to the study of linguistic structure, explaining why human languages are the way they are and how communication pressures shape human languages(Futrell and Hahn, 2022; Mahowald et al., 2013; Piantadosi et al., 2011, *inter alia*). Therefore, information theory can be viewed as a quantitative framework for measuring the amount of information conveyed by linguistic units, such as phonemes or words. Information theory has been adopted as a framework to study various aspects of linguistic structure, including phonology, morphology, and syntax. Due to its relevance to the work presented in this thesis, this section introduces some fundamental concepts of information theory. This section is based on the definitions and notations introduced in the textbook of MacKay (2003).

Consider a random variable X associated with a triple $(x, \Omega_x, \mathcal{P}_x)$, where x represents the outcome of the random variable. The outcome can take a value from a set of possible values defined by $\Omega_x = \{\omega_1, \ldots, \omega_I\}$, where each value is associated with probability from the set $\mathcal{P}_x = \{p_1, p_2, \ldots p_I\}$. We define a probability mass function such that

$$P(x = \omega_i) = p_i \qquad \text{and} \qquad \sum_{\omega_i \in \Omega_x} P(x = \omega_i) = 1 \tag{2.13}$$

Note that the set Ω_x represents the outcome space of an experiment. For example, when randomly sampling a letter from an English book, the set Ω_x is the English alphabet. In general, the set Ω_x can represent any set of discrete linguistic events such as phonemes or words. For convenience and brevity of notation, we write $P(x = \omega_i)$ as $P(x)$. Moreover, we can quantify the uncertainty associated with a

random variable using information theoretic notation of **surprisal**. For a specific outcome, we measure its Shannon information content or surprisal is computed as

$$\eta(x) = -\log_2 P(x) \tag{2.14}$$

This quantity is measured in bits, and it quantifies the unexpectedness of an outcome being observed as a value of the random variable X. When an outcome is certain to occur, it is surprisal is minimal (i.e., $\eta(x) = 0$). The uncertainty or "randomness" of the associated probability distribution can be further quantified as the average surprisal, or **entropy** as follows

$$H(X) = \sum_{x \in \Omega_x} P(x)\, \eta(x) \tag{2.15}$$

where $0 \leq H(X) \leq \log_2 |\Omega_x|$ and it is also measured in bits. If one outcome is certain to occur, then entropy associated with the random variable is minimal such that $H(X) = 0$. On the other hand, a distribution associated with a random variable is maximally entropic (i.e., $H(X) = \log_2 |\Omega_x|$) when all outcomes are equally likely to occur. For convenience, the entropy $H(X)$ may also be written as $H(\boldsymbol{p})$, where $\boldsymbol{p}$ can be viewed as a "vectorized" representation of $\mathcal{P}_x$.

2.6 Representational Similarity Analysis

Representational Similarity Analysis (RSA) is a data-analytical framework developed in the neuroscience community to enable comparison of neural activity patterns across brain regions and computational models of information processing (Kriegeskorte et al., 2008). The RSA framework abstracts away from the activity patterns themselves and operates on the geometry of the representation or feature space. This makes it applicable for interpretability and analysis of neural networks when the correspondence between neurons across different layers or architectures is unknown. In NLP research, RSA has previously been employed to study the correlation between neural network and symbolic representations of language (Chrupała and Alishahi, 2019) and analyze word representations in language models (Abdou et al., 2019; Abnar et al., 2019; Beinborn and Choenni, 2020; Lepori and McCoy, 2020; J. Wu et al., 2020). In computational speech processing research, RSA has been further employed to analyze the representations of speech recognition models (Chrupała, Higy, et al., 2020; Chung, Belinkov, et al., 2021) and quantify their similarity to human brain activations while listening to speech (Magnuson et al., 2020). In this thesis, the RSA framework is used in two different

studies to: (1) analyze the effect of initial model conditions on the consistency of its word representations, and (2) quantify the impact of linguistic experience, characterized by the language of exposure, on non-native spoken-word processing. Therefore, we introduce some fundamental information on the representational similarity analysis framework in this section.

Consider a set of K acoustic stimuli $\{\boldsymbol{A}^1, \ldots, \boldsymbol{A}^K\}$ in addition to two neural representation models $\mathcal{F}^y$ and $\mathcal{F}^z$. Here, the two models $\mathcal{F}^y$ and $\mathcal{F}^z$ have the same architecture but differ across one dimension of variability. For instance, the two models are initialized with the same initial weights but trained on two different languages, or trained on the same data but initialized differently. Now, $\mathcal{F}^y$ and $\mathcal{F}^z$ can be used to induce representations of the speech stimuli, resulting in two views of the data: $\mathbf{Y} \in \mathbb{R}^{K \times D}$ and $\mathbf{Z} \in \mathbb{R}^{K \times D}$, respectively. Since the two models are trained independently, their representations do not correspond at the neuron-level. Therefore, comparing the two the views at the individual representation level using a metric such as the cosine distance is not feasible. To solve this problem, the RSA framework was developed the so that the representation similarity can be quantified using pairwise similarities instead of direct sample-to-sample comparison.

Although several variants of the RSA framework have been introduced within the neuroscience and machine learning community, we use Centered Kernel Alignment (CKA) as a representational similarity measure between the two views of the same input samples in this thesis.(Kornblith et al., 2019). CKA abstracts away from the representations themselves and operates on pairwise distances between the sample representations. Moreover, CKA has been shown to be invariant to orthogonal transformation and isotropic scaling which makes it suitable for our analysis when analyzing the effect of one of dimensions of speech variability. Given two different views of the input stimuli as the matrices $\mathbf{Y}$ and $\mathbf{Z}$, each view matrix is multiplied by a centering matrix $\mathbf{H} = \mathbf{I}_K - \mathbf{1}_K/K$ to make each column's mean equal to zero and obtain centered second moment matrices as

$$\begin{aligned} \mathbf{G_Y} &= \mathbf{H}\mathbf{Y}\mathbf{Y}^\top\mathbf{H}^\top/D, \\ \mathbf{G_Z} &= \mathbf{H}\mathbf{Z}\mathbf{Z}^\top\mathbf{H}^\top/D \end{aligned} \tag{2.16}$$

Then, the representational similarity of the two views is computed using CKA as

$$\text{CKA}(\mathbf{Y}, \mathbf{Z}) = \frac{\langle \text{vec}(\mathbf{G_Y}), \text{vec}(\mathbf{G_Y}) \rangle}{||\mathbf{G_Y}||_F ||\mathbf{G_Z}||_F} \tag{2.17}$$

where vec(.) represents the vector-reshaped matrix, $\langle ., . \rangle$ is the inner product, and $||.||_F$ is the Frobenius norm. This ensures that CKA $\in [0, 1]$, where values close to

1 indicate high similarity between the two views of the data, while values close to 0 indicate low similarity.

In summary, RSA aims to characterize the representational geometries by abstracting away from the representations themselves and analyzing the correlation between pairwise dissimilarity matrices derived from different views of a given set of stimuli. This way, RSA allows for both qualitative and quantitative comparisons between different representation models, enabling us to gain insights about how different dimensions of speech variability shape the representational profile of neural networks.

Part II

SPEECH REPRESENTATIONS OF LANGUAGE IDENTITY

3

Domain-Invariant Speech Representations for Language Identification

This chapter begins the second part of the thesis, which is concerned with representations of spoken language identity. The study presented in this chapter focuses on the problem of domain variability and how it affects the transferability of neural network representations across different datasets. Concretely, we first show that the representations of convolutional models for spoken language identification do not transfer well across different datasets that vary in their recording conditions. To solve this problem, we propose an approach based on unsupervised adversarial adaptation to encourage the model to build domain-invariant speech representations. We demonstrate that adversarial training prevents neural networks from exploiting dataset-specific artifacts as predictive features for the language, thus leading to better performance across domains.

3.1 Introduction

Spoken language identification, henceforth **SLID**, is the problem of determining the identity of the language in a spoken utterance (H. Li, Ma, and K. A. Lee, 2013). In today's globalized world, SLID systems can facilitate a wide range of cross-lingual speech and communication technologies such as spoken language translation (Bangalore et al., 2012; Fügen et al., 2007; Waibel et al., 2000) and multilingual spoken document retrieval (Chelba et al., 2008). Furthermore, robust SLID systems can be crucial in assisting the field linguistics community and their ongoing efforts in the preservation, documentation, and categorization of the world's endangered languages (Levow, Ahn, et al., 2021; Levow, Bender, et al., 2017).

Earlier work has addressed the SLID task using the so-called phonotactic approach. In this paradigm, the acoustic signal is first transduced into a sequence of discrete symbols (e.g., phones), then probabilistic models are utilized to obtain language likelihoods (Lamel and Gauvain, 1994; H. Li and Ma, 2005). This approach has been outperformed by approaches that operate directly on the acoustic signal without discretizing it. Acoustic approaches used to be based on Gaussian Mixture Models (GMMs) and the ***i***-vector framework, which has been applied to both speaker and language identification (Garcia-Romero and Espy-Wilson, 2011; Kenny, 2010; Martinez et al., 2011; Su and Wegmann, 2016). Currently, end-to-end deep neural networks (DNNs) are predominant for SLID and outperform GMMs, especially for short utterances (e.g., Gonzalez-Dominguez et al., 2014; Lopez-Moreno et al., 2014; Mateju et al., 2018; P. Shen et al., 2018; Shon et al., 2018).

Discriminating between **closely related languages** is a difficult task for human listeners due to their phonetic and phonological similarity (Skirgård et al., 2017). On the other hand, SLID models based on neural networks have shown striking performance discriminating between spoken varieties of Arabic (Bulut et al., 2017; Grégory Gelly et al., 2016; Shon et al., 2018), Slavic languages (Mateju et al., 2018), and languages in accented speech samples from multilingual speakers (Titus et al., 2020). For instance, the best neural SLID model in the work of (Mateju et al., 2018) has reported an error rate as low as 1.2% when discriminating between 11 Slavic languages. Generally speaking, the impressive performance of DNN-based SLID reported in the literature gives the impression that SLID is almost a solved problem.

3.1.1 The Problem of Domain Variability

Despite their success in solving many challenging problems in the field of artificial intelligence, DNNs still have some limitations. One of the well-known limitations of DNN-based models is that their performance usually degrades when tested on data samples with different characteristics than the training data. This problem is referred to in the machine learning community as a **data distribution shift** (Ben-David et al., 2010; Ganin, Ustinova, et al., 2016). In other words, DNN-based models usually do not perform well when they encounter test samples that are drawn from a different distribution even if it is similar to the training data distribution. In the speech modality, this problem is encountered when the recording conditions vary between training and test samples. For example, a model trained on read speech recorded in a quiet environment will not perform

well if tested on speech recordings with a microphone or background noise. This discrepancy in the recording environments exemplifies a **domain mismatch** problem whereby the training distribution can be regarded as the source domain and the test distribution be regarded as the target domain. In this chapter, we view the domain mismatch problem as one instance of **intersession variability** in the acoustic realization of human speech (H. Li, Ma, and K. A. Lee, 2013).

In addition to the domain mismatch problem, neural networks tend to exploit dataset biases as **shortcuts** when learning to perform a task from data (Geirhos, Jacobsen, et al., 2020). That is, instead of recognizing the underlying structure of the data and how it relates to the prediction task, neural networks have a strong tendency to learn decision rules based on spurious correlations between the input features and the labels (e.g., Beery et al., 2018; Geirhos, Rubisch, et al., 2018; Glockner et al., 2018; Y. Goyal et al., 2017; Kavumba et al., 2019). As a result, a strong performance on a benchmark task where the training and test samples are random splits of the same dataset does not usually transfer to other datasets with more challenging conditions in real-world scenarios. To measure the true ability of a model to solve a certain task, we argue that the evaluation procedure should always consider **out-of-domain** samples in addition to the in-domain samples. Since spurious correlations are usually dataset-specific (or domain-specific) and very unlikely to exist in other datasets (or domains), out-of-domain evaluation can deliver insights about the limitations of the model and whether or not it has learned the task by recognizing the true patterns that are predictive of the labels.

3.1.2 Research Questions

Most of the previous work in SLID has addressed the task without considering the **out-of-domain generalization** problem. That is, SLID models in previous work were trained and evaluated using disjoint splits of the same dataset where the training and evaluation samples have similar, if not identical, recording conditions (i.e., same domain). Therefore, the impact of **dataset bias** on SLID robustness has not yet been investigated with a systematic evaluation across datasets. For the case where the languages are closely related and perceptually similar, picking up superficial correlations in the data could be an easier task for neural networks than learning robust discriminative patterns between the languages. Therefore, it remains unclear whether the success of SLID models in an in-domain setting reflects the true ability of the models in identifying features that are truly predictive of the languages under study. This chapter presents a study whereby we bridge this gap and focus on the challenging case of SLID for short utterances of related

Slavic languages in a cross-domain setting. We investigate the following research questions:

- **RQ1.** To what degree do neural SLID models for related languages generalize to another domain with different recording conditions?
- **RQ2.** Are different low-level speech features equally robust under domain mismatch?
- **RQ3.** Can we adapt SLID models to a new domain without using labelled data in the new domain? If yes, what are the factors that affect the adaptability of the model?

To address these research questions, we present a series of SLID experiments with datasets from two domains: (1) Read speech recordings from the Slavic subset of the GlobalPhone speech database (Schultz et al., 2013), and (2) Slavic broadcast recordings collected and distributed by Mateju et al., 2018 for SLID (**RQ1**). We also compare the performance of spectral (MFSCs) and cepstral (MFCCs) speech features within the same domain and across different domains (**RQ2**). Finally, we propose a novel SLID model based on unsupervised domain adaptation with adversarial learning (Ganin and Lempitsky, 2015) to improve the model robustness and cross-domain transferability, analyze predictions from the adapted model, and visualize its representations compared to the baseline (**RQ3**).

3.2 SLID with Deep Neural Networks

3.2.1 Problem Definition

We define the SLID task as a discriminative sequence classification problem. First, a variable-length utterance is transformed by an acoustic front-end into a sequence of acoustic observations $\mathbf{x}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathbb{R}^k$ is a frame-level feature vector at timestep t. To simplify the notation, we use a bold symbol without subscripts (i.e., $\mathbf{x}$) to denote the entire input sequence. Given a sequence $\mathbf{x}$, the goal is to predict the spoken language $\hat{y}$. Using a deep neural network as a classification model, the SLID problem can be defined as

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \,|\, \mathbf{x};\, \boldsymbol{\theta}) \tag{3.1}$$

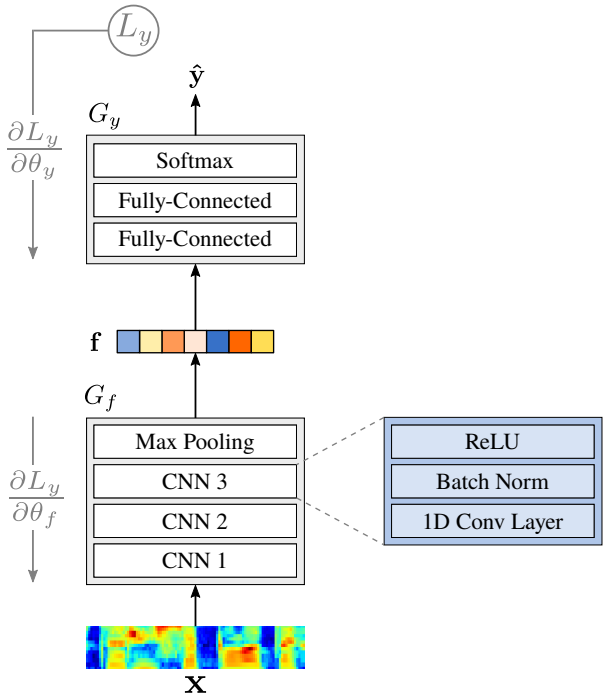


Figure 3.1: A schematic view of our baseline SLID model. The model can viewed as two components trained jointly: (1) a high-level feature extractor $\mathcal{G}_f$ and (2) a language classifier $\mathcal{G}_y$

.

where $\mathcal{Y}$ is a finite set of languages, $\boldsymbol{\theta}$ is the model's parameters learned in a supervised approach, and $P(y|\mathbf{x};\boldsymbol{\theta})$ represents a posterior probability of the language label y.

3.2.2 Baseline SLID

Our SLID model consists of a 1D 3-layer convolutional network followed by 3-layer fully-connected feed-forward network as schematized in Figure 3.1(a). We describe the mapping between the input and the output in the neural network as follows: given an input sequence $\mathbf{x}$ sampled from a spectro-temporal space $\mathcal{X}$, the input is first transformed by a high-level feature extractor $G_f : \mathcal{X} \rightarrow \mathbb{R}^D$ into a D-dimensional feature vector as

$$\mathbf{f} = G_f(\mathbf{x};\boldsymbol{\theta}_f) \in \mathbb{R}^D \tag{3.2}$$

Then, the feature vector $\mathbf{f}$ is transformed by a language classifier $G_y : \mathbb{R}^D \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ into a logit vector via a series of non-linear transformations as

$$\hat{\mathbf{y}} = G_y(\mathbf{f};\boldsymbol{\theta}_y) \in \mathbb{R}^{|\mathcal{Y}|} \tag{3.3}$$

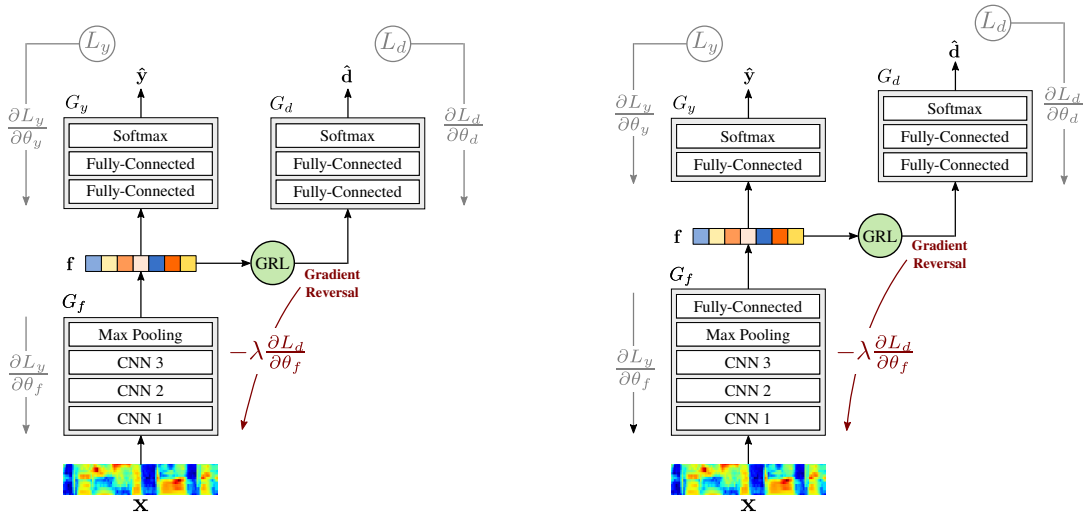


Figure 3.2: A schematic view of our domain-adversarial neural networks for SLID: (a) the architecture of DA-SLID I, and (b) the architecture of DA-SLID II

By applying a softmax function on the logit vector $\hat{\mathbf{y}}$, we obtain an empirical probability distribution over the language space. The parameters of the network $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_y$ are learned jointly in an end-to-end approach given a dataset $\mathcal{D}_{\mathcal{S}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\mathcal{S}}}$ of $N_{\mathcal{S}}$ labelled samples from a single domain. The objective function is to minimize

$$J(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\mathcal{S}}} L_y\Big(G_y\big(G_f(\mathbf{x}_i; \boldsymbol{\theta}_f); \boldsymbol{\theta}_y\big), y_i\Big) \tag{3.4}$$

where L_y is the loss of the language classifier. For the sake of simplicity, we consider the convolutional block as the high-level feature extractor G_f while the feed-forward block as the language classifier G_y in our baseline SLID model.

3.2.3 Domain-Adversarial Neural Network for SLID

In this study, we employ domain-adversarial neural network (DANNs) which have been successfully applied to many vision and speech recognition problems (Ganin and Lempitsky, 2015; Meng, Zhuo Chen, et al., 2017; Shinohara, 2016). DANNs aim to minimize the discrepancy between two domains given a dataset $\mathcal{D}_{\mathcal{T}} = \{\mathbf{x}_i\}_{i=1}^{N_{\mathcal{T}}}$ of $N_{\mathcal{T}}$ unlabelled samples in the target domain, in addition to the source labelled samples $\mathcal{D}_{\mathcal{S}}$. This adaptation technique is unsupervised because it only requires unlabeled samples in the target domain, while the supervision is transferred from the labeled samples of the source domain.

To improve the SLID model's out-of-domain generalization, the feature representations emerging from the model should be both language-discriminative and domain-invariant. This objective can be achieved if the model is encouraged during training to build up representations that are good predictors of the spoken language but do not encode domain-related information. To this end, a fully-connected feed-forward block $G_d : \mathbb{R}^D \rightarrow [0, 1]$ is added to the network to predict the domain given $\mathbf{f}$ (see Figure 3.2 (a) and (b)). We view G_d as a domain classifier with a separate set of parameters $\boldsymbol{\theta}_d$ which are learned by exploiting the domain labels of source and target samples. That is, each training sample in the source domain $(\mathbf{x}_i, y_i)$ is augmented with a domain label $d_i = 0$, while each training sample in the target domain $\mathbf{x}_j$ is augmented with a domain label $d_j = 1$. We seek the parameters $\boldsymbol{\theta}_d$ that minimize the loss of the domain classifier. On the other hand, the feature extractor G_f is trained such that $\mathbf{f}$ is uninformative for the domain classifier. Thus, we seek the parameters $\boldsymbol{\theta}_f$ that maximize the domain classifier loss, encouraging the feature representation $\mathbf{f}$ to be domain-invariant. This procedure is an instance of adversarial learning where different blocks in the network are trained with competing objectives. The overall objective function is to minimize

$$J(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y, \boldsymbol{\theta}_d) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_\mathcal{S}} L_y\Big(G_y\big(G_f(\mathbf{x}_i; \boldsymbol{\theta}_f); \boldsymbol{\theta}_y\big), y_i\Big) - \lambda \sum_{(\mathbf{x}_i, d_i) \in (\mathcal{D}_\mathcal{S} \cup \mathcal{D}_\mathcal{T})} L_d\Big(G_d\big(G_f(\mathbf{x}_i; \boldsymbol{\theta}_f); \boldsymbol{\theta}_d\big), d_i\Big) \quad (3.5)$$

where L_y is the loss of the language classifier, L_d is the loss of the domain classifier, and λ is a trade-off hyperparameter that controls the contribution of the domain classifier's loss to the overall loss which is computed as

$$\lambda = \frac{2}{1 + \exp(-10 \cdot \alpha)} - 1 \quad (3.6)$$

where α is a progress parameter that is proportional to the iteration index and changes from 0 to 1 in a linear scale during the training procedure. Therefore, the adaptation hyperparameter λ is initiated at 0 to suppress the noisy signal from the language classifier during the early training iterations. In practice, this adversarial learning strategy is realized with a special layer that behaves as an identity function in the forward pass but during backpropagation it reverses the direction of the gradient signal coming from the domain classifier's loss into the

feature extractor. Following previous literature on adversarial domain adaptation (Ganin and Lempitsky, 2015), we refer this layer as as a gradient reversal layer.

Prior work on adversarial domain adaptation has made a design decision to consider the convolutional block as the feature extractor $G_f(.)$ and the feed-forward block as the label predictor $G_y(.)$, and we also adopt the same design for our baseline SLID model. In this study, we argue that this design decision is rather arbitrary since it is solely based on the network architecture of the internal blocks and not supported by empirical evidence. Moreover, this design decision could have non-trivial consequences on the DANN performance since the gradient reversal layer is plugged in between the feature extractor $G_f(.)$ and the domain classifier $G_d(.)$. Therefore, we experiment with two variants of the domain adversarial SLID model: (1) DA-SLID I: an identical configuration to Ganin and Lempitsky (2015), where the convolutional block of the model is considered as the feature extractor as illustrated in Figure 3.2 (a), and (2) DA-SLID II: we consider the feature extractor as the convolutional block as well as the first layer of the fully-connected block Figure 3.2 (b). Therefore, in DA-SLID II the reversed gradient signal from the domain classifier is back-propagated into all layers of the network, except the two final layers before the softmax of the language classifier.

3.3 Experimental Data and Setup

3.3.1 Datasets for Slavic SLID

GlobalPhone Read Speech (GPS) We use the Slavic portion of the multilingual GlobalPhone speech database (Schultz et al., 2013) which includes read speech recordings from native speakers of six Slavic languages: Bulgarian (BUL), Croatian (HRV), Czech (CZE), Polish (POL), Russian (RUS), and Ukrainian (UKR). The utterances vary in length and quality across languages. We set the minimum utterance length to 3 seconds and segment longer utterances into non-overlapping 3-second speech segments. Our final training subset consists of 8,000 utterances per language. We use the same splits as in (Tachbelie et al., 2020).

Radio Broadcast Speech (RBS) A large collection of Slavic recordings were collected by harvesting online radio broadcasts in (Mateju et al., 2018; Nouza et al., 2016). The original dataset contains recordings for 11 Slavic languages. We use the same subset of six languages as in the GPS dataset. The extracted utterances are either segments of professional news reports or of spontaneous speech during discussions. Occasionally, the utterances include background music

and different sorts of acoustic noise. We sample 8,000 and 500 utterances per language from the training split as our training and validation sets, respectively. This dataset does not include any speaker IDs. Thus, we cannot confirm whether training and evaluation speakers are disjoint, which may have an impact on the model's performance.

3.3.2 Low-level Feature Extraction

In our experiments, we use the first 13 coefficients of MFSCs and MFCCs, with the zeroth coefficient being the average frame energy, as low-level speech features. While previous works usually refer to MFSCs as mel-filterbanks (Shon et al., 2018), we use the term MFSCs to refer to mel-frequency spectral features that are correlated (A. Mohamed, 2014). Since both datasets in our study are sampled at 16 kHz, we extract frames of 400 samples with 160 samples stride, which corresponds to 25 ms and 10 ms, respectively. We apply standardization on the features to have zero mean and unit variance.

3.3.3 Model Architecture and Hyperparameters

Baseline Architecture. We use 1D 3-layer convolution over the temporal dimension with 128, 256, and 512 filters and widths of 5, 10, and 10 for each layer and keep stride step at 1. We apply batch normalization and ReLU non-linearity following each convolutional operation. We apply max pooling to downsample the representation only at the end of the convolution block. For the feed-forward block, we use 3 fully-connected layers ($512 \rightarrow 512 \rightarrow 512 \rightarrow 6$) before the softmax for both the non-adapted and the adapted SLID models.

Domain-Adversarial SLID. For our SLID models with domain adaptation, we use an identical architecture for the convolutional and feed-forward blocks as the baseline model. The only difference is an additional feed-forward block with the gradient reversal layer where we employ a 3-layer feed-forward network ($512 \rightarrow 1024 \rightarrow 1024 \rightarrow 2$) as the domain classifier.

Training Details. In this study, we focus on three experimental variables: (1) model type (Baseline SLID vs. DA-SLID I vs. DA-SLID II), (2) adaptation direction (GPS $\rightarrow$ RBS vs. RBS $\rightarrow$ GPS), and (3) low-level acoustic features (MFSCs vs. MFCCs). For each possible combination, we train 25 neural network instances that differ in their random initialization which we control using different random seed for each run. Therefore, we have 300 neural network instances in

Table 3.1: Cross-domain evaluation of SLID models in accuracy (%). Δ indicates the relative difference.

Source Dataset		Evaluation		
		In-domain	Out-of-domain	Δ
GlobalPhone Speech (GPS)	MFSCs	96.59	**40.87**	-57.69
	MFCCs	96.39	39.53	-58.99
Radio Broadcast Speech (RBS)	MFSCs	96.07	**56.49**	-41.20
	MFCCs	95.60	53.12	-44.44

total. For the loss function, we use cross-entropy loss for both L_y and L_d. The ADAM optimizer is used with learning rate of 0.001. We train our models with a batch size of 256 for 50 epochs and observe the in-domain validation performance during training. The epoch that yields the top performing model on the in-domain validation is used for both in-domain as well as out-of-domain evaluation on held-out test sets.

3.4 Experimental Results

We now present and discuss the results of our experiments. To make the results comparable across datasets and prevent undesirable effects due to potential biases caused by utterance length mismatch, we train and evaluate each of our SLID models on 3-second utterances. Since the GPS evaluation data is imbalanced, we use balanced accuracy (Brodersen et al., 2010) as our evaluation metric to obtain a better estimate of the model performance. We observe that balanced accuracy scores highly correlate with equal error rate (EER) and average cost (C_{avg}), which we do not report for the sake of conciseness.

3.4.1 Cross-Domain Evaluation

Table 3.1 presents the results of the cross-domain evaluation on both datasets without adaptation. For this evaluation, we report the maximum accuracy score we obtain across all 25 runs for each model type. Even though our SLID models are not heavily regularized, the in-domain performance is always above 95%, while MFSC and MFCC features yield a comparable performance. On the other hand, out-of-domain (OOD) evaluation shows a considerable drop in accuracy in each cross-domain setting. It is interesting to observe that the drop in accuracy is more

Table 3.2: OOD performance of adapted models in accuracy (%).

Transfer task		Adaptation Model		
		Baseline SLID	DA-SLID I (Δ)	DA-SLID II (Δ)
GPS → RBS	MFSCs	40.87	44.67 (+09.30)	**47.43** (+16.05)
	MFCCs	39.53	41.93 (+06.07)	43.23 (+09.36)
RBS → GPS	MFSCs	56.49	70.47 (+24.75)	88.78 (+57.17)
	MFCCs	53.12	73.85 (+39.02)	**91.86** (+72.93)

pronounced for MFCC features, while the correlated MFSCs seem to be more robust under domain shift. To verify this observation, we conduct a paired student t-test across the results of all 25 neural networks instances of the baseline models trained on MFSCs and MFCCs. The result of the test shows that the difference is statistically significant for the model trained on the GPS and evaluated on RBS ($t(24) = 6.28, p < 0.0001$) and for the model trained on RBS and evaluated on GPS ($t(24) = 8.67, p < 0.0001$). Moreover, the impact of domain shift is more pronounced in when training GPS and evaluating on RBS.

3.4.2 Domain Adaptation Results

In our adaptation experiments, we investigate two transfer tasks; GPS → RBS and RBS → GPS. The results are shown in Table 3.2 where we report the performance of the model with the maximum accuracy score across the 25 different runs for each model type. The adapted models consistently improve the out-of-domain accuracy compared to the source-only non-adapted baseline with both features and in both directions. Our DA-SLID II model yields the best results, which suggests that the domain discrepancy is present not only in the convolutional layers, but also in the fully-connected layers that are more distant from the input. We present and discuss the results for both directions.

RBS → GPS Both domain adaptation models yield significant improvements over the baseline models. While the MFSC-based DA-SLID II model improves OOD accuracy from 56.49% to 88.78% with a relative accuracy gain of 57.17%, its MFCC-based counterpart improves OOD accuracy from 53.12% to 91.86% with a relative accuracy gain of 72.93%.

GPS → RBS Even though adapted models improve over the baseline, the improvements in this direction are less impressive than what is observed in the RBS → GPS direction. The MFSC-based DA-SLID II model improves OOD

accuracy from 40.87% to 47.43% with a relative accuracy gain of only 16.05%. On the other hand, the MFCC-based DA-SLID II model improves OOD accuracy from 39.53% to 9.36% with a relative accuracy gain of only 9.36%.

3.4.3 Result Discussion

The performance gap between the two directions in our experiments seemed surprising at the beginning. In retrospective, this should not be surprising as the two directions are not equally challenging. The RBS dataset is more diverse in terms of the number of unique speakers and background noise. An SLID model trained on the RBS dataset has to learn to extract language ID features from noisy speech signals, thus it is expected to be more generic and perform well on clean speech signals even under domain shift. This finding is consistent with what has been reported in the domain adaptation literature on how source domain diversity affects adaptability of the model to new domains (Ganin and Lempitsky, 2015). On the other hand, if the model has not been exposed to noisy speech signals during training, it is unlikely to perform well on noisy signals even if the representation discrepancy has been minimized, which is the case in the GPS $\rightarrow$ RBS direction. This suggests that alternative adversarial training procedures that add noise to the input representation could be explored to encourage the model to transform the noisy input signals into noise-robust representations. Moreover, our experiments show that MFCCs are more sensitive to input variations due to domain shift, thus MFCC-based models in both directions tend to benefit more from adaptation in terms of relative accuracy gain compared to their MFSC-based counterparts, with only one exception case.

3.5 Stability Analysis

In the previous section, the performance of the neural network instance that performs best on the validation set was reported. That is, we report only the performance of a single run out of 25 different instances that were trained for each model type. However, adversarial learning, which involves training two competing networks, is known to be unstable and sensitive to different random initializations. As a result, the performance of the models may vary across different runs. Therefore, we analyze the results across all neural network instances in this section.

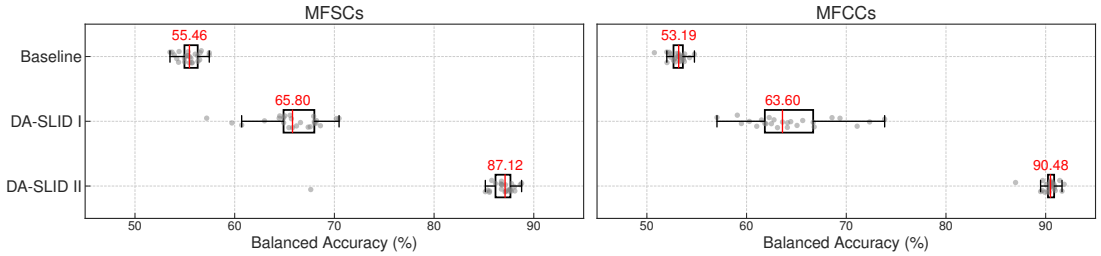


Figure 3.3: Stability analysis of the model in the RBS → GPS transfer task: (left) MFSC features and (right) MFCC features. Each data point corresponds to out-of-domain evaluation accuracy of a single run. The value of the median is annotated on top of each box plot.

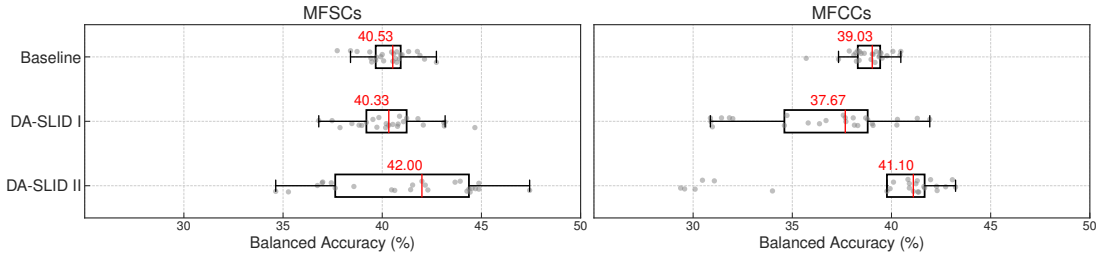


Figure 3.4: Stability analysis of the model in the GPS → RBS transfer task: (left) MFSC features and (right) MFCC features. Each data point corresponds to out-of-domain evaluation accuracy of a single run. The value of the median is annotated on top of each box plot.

RBS → GPS The stability analysis of this transfer task is shown in Figure 3.3. By visualizing the out-of-domain results of the 25 different runs for each model as samples from a distribution, one can visually observe that the three distributions seem to be different. To verify this visual observation by means of statistical test, we conduct paired sample t-tests. For the models based on MFSCs, the t-test reveals that both differences between the DA-SLID II vs. the baseline ($t(24) = 36.31, p < 0.0001$) as well as the DA-SLID I vs. the baseline are statistically significant ($t(24) = 15.42, p < 0.0001$). Likewise, for the models based on MFCCs, statistically significant differences are observed between the DA-SLID II vs. the baseline ($t(24) = 148.17, p < 0.0001$) as well as the DA-SLID I vs. the baseline ($t(24) = 12.96, p < 0.0001$). These results demonstrate that the domain adaptation performance of this transfer task is consistent and stable across different runs, albeit some outliers.

GPS → RBS The stability analysis of this transfer task is shown in Figure 3.4. For both feature types MFSCs and MFCCs, the three distributions appear to be overlapping and the domain adaptation performance does not seem to be consistent across different runs. Statistical tests based on a paired sample t-test did not reveal any statistically significant differences between the distributions with

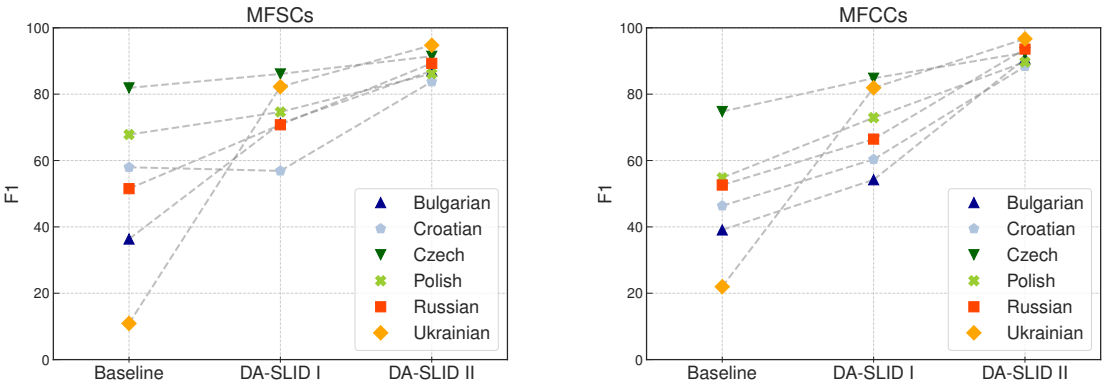


Figure 3.5: Out-of-domain F_1 score (%) per language of our MFSC-based model (left) MFCC-based model (right) in the RBS → GPS direction.

p-value < 0.01, indicating that the improvements in the GPS → RBS direction are not consistent across different runs. Therefore, it seems that adversarial domain adaptation improves only the maximum across 25 runs for this transfer task, but the improvement is not consistent and there are many failed runs in the adapted models where the performance is even below than the baseline. This analysis demonstrates that the diversity of the speech samples in the source domain plays a crucial role in the success of domain adaptation using adversarial learning. A more diverse source domain provides a broader range of acoustic variations and linguistic characteristics, enabling the model to learn more robust and generalizable language ID features.

3.6 Why Does Adversarial Domain Adaptation work?

In this section, we seek to understand why unsupervised adaptation with adversarial training improves out-of-domain performance. To this end, we analyze the results and the representations of the RBS → GPS transfer task to get insights into the factors that lead to the significant improvements in this direction.

3.6.1 Fine-grained Performance Analysis

Figure 3.5 shows the performance per language measured by F_1 score. In the non-adapted case, we observe a much higher variance between languages compared to the adapted models. For example, while the non-adapted MFCC-based model achieves up to 74.8% F_1 on Czech, F_1 is only 21.9% for Ukrainian, which is barely above than chance-level F_1 (16.7%). We inspected the performance on Ukrainian in the other direction and found that the F_1 is even worse than chance-level.

We hypothesized that the acoustic conditions of the Czech recordings in the two domains are similar, while the discrepancy is maximal in the case of Ukrainian. To validate this hypothesis, we manually inspected several Ukrainian utterances from the GPS dataset. We found that most utterances are characterized by unnatural pauses and hesitations that distort the speech signal and are uniformly distributed across Ukrainian training and evaluation speakers in the GPS dataset. This effect adds to the discrepancy due to domain shift since RBS utterances are more naturally flowing speech than the read speech from the GPS dataset, despite the occasional background noise. In particular, this effect creates abnormal patterns that hinder non-adapted SLID performance in two ways: (1) if these patterns are not uniformly distributed across languages and observed during training, the network exploits them as shortcuts to discriminate between languages, and (2) if these patterns are encountered during out-of-domain inference, the distorted signal causes a failure because the model has not been exposed to such patterns during training. Both cases lead to poor out-of-domain generalization when training on a single-domain dataset without adaptation. However, since these patterns are only present in one dataset, they are good predictors of the domain. Therefore, adversarial training with domain confusion prevents the models from exploiting such dataset-specific artifacts, which consistently yields a better out-of-domain generalization. The advantage of adversarial training is demonstrated in Figure 3.5. For instance, our MFCC-based adapted model boosts the F_1 score on Ukrainian from 21.9% to 96.7%, which is surprisingly the highest in this transfer direction.

3.6.2 Visualizing the Representations

In Figure 3.6, we visualize the representations using the t-SNE algorithm (Maaten and Hinton, 2008). We sample a set of 1800 data points from each domain and obtain the representations from the penultimate hidden layer of the MFCC-based SLID models: (a) source-only non-adapted SLID, (b) DA-SLID I, and (c) DA-SLID II. Figure 3.6 demonstrates how adversarial domain adaptation encourages the neural network to build language-discriminative representations that are also domain-invariant. Therefore, we can attribute the success of cross-domain transfer learning to the domain-invariant nature of the emergent representations in the neural network.

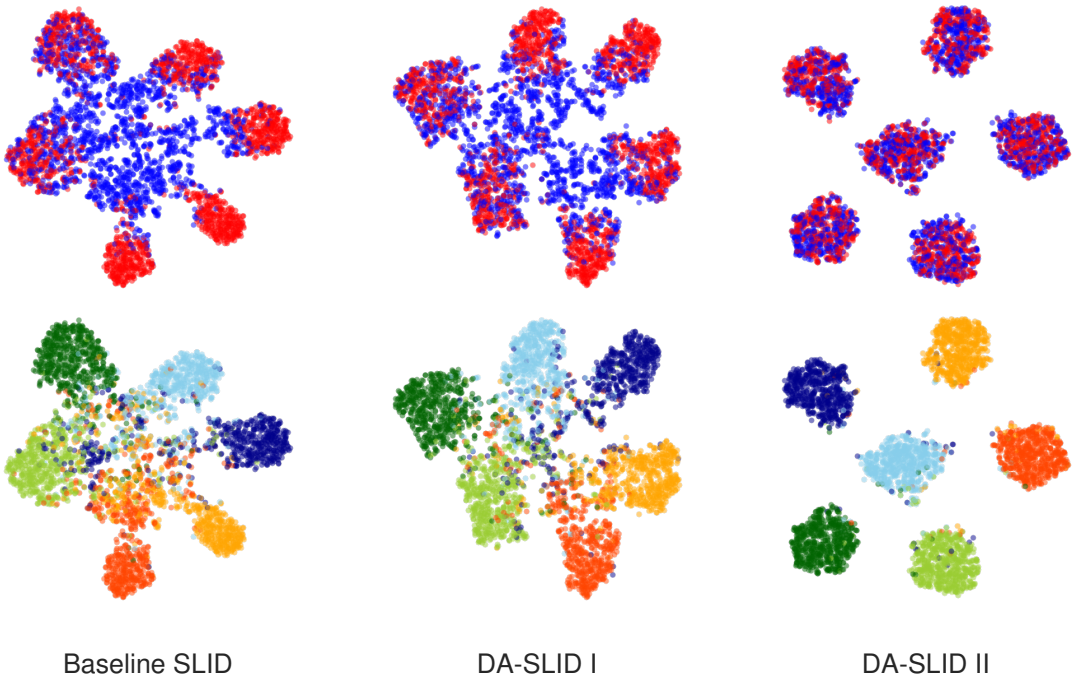


Figure 3.6: t-SNE visualization: (Top) data points are colored by domain, red points correspond to source domain samples while blue points corresponds to target domain samples, and (Bottom) data points are colored by language.

3.7 Summary

In this study, we have investigated the problem of spoken language identification for closely-related languages in a cross-domain setting, using deep convolutional neural networks as discriminative models. While our experiments have confirmed that they perform very well within-domain, our cross-domain evaluation has revealed that neural models poorly generalize to a novel dataset with acoustic conditions that differ from those that have been observed during training. To improve the robustness of our models against domain mismatch, we have applied unsupervised domain adaptation with gradient reversal and shown that our adaptive models generalize better across domains. Our analysis has shown that adversarial training prevents the model from exploiting dataset-specific artifacts, thus leading to better out-of-domain generalization. We have identified the diversity of the speech samples in the source domain as the major factor that affects the adaptability of the model to a new domain. Given a diverse source dataset, our adaptive models achieved relative accuracy improvements of up to 72.9%.

4

Language Representations and Cross-Linguistic Variation

This chapter is concerned with the encoding of cross-linguistic variation in neural network representations of spoken language identity. While neural models have been shown to perform very well on the task of discriminating related languages from acoustic speech signals, it remains unknown whether they capture cross-linguistic variation in their intermediate representations. This chapter presents a case study on the related Slavic languages that investigates the degree to which the model's representational similarity among languages reflects objective measures of language similarity. Even though the model does not have access to any signal regarding how the languages relate to each other, this study demonstrates that the model representations exhibit a cluster structure that corresponds to the phylogenetic groups within the Slavic language family, even for languages that are not observed during training.

4.1 Introduction

The relationship between a group of human languages can be characterized across several dimensions of variation (Skirgård et al., 2017), including (1) the temporal dimension, wherein languages have diverged from a common historical ancestor as in the case of Romance languages; (2) the spatial dimension, wherein the speaker communities are geographically adjacent as in the case of the Indo-Aryan and Dravidian languages of India; and (3) the socio-political dimension, wherein languages have evolved under shared political and/or religious forces as in the case of Malay and Swahili. Languages, or language varieties, can be related across all these dimensions, which often results in a dialect continuum. In some cases, speakers of languages that constitute a **dialect continuum**—

for example, North Germanic languages of mainland Scandinavia— can usually communicate with each other efficiently using their own mother tongue. The degree of **intercomprehensibility** between speakers of different language varieties within a continuum is mainly determined by linguistic similarities (Gooskens, 2007). In this study, we focus on the representation of **cross-linguistic variation** in neural models of spoken language identification for the related Slavic languages, which are known to be mutually intelligible to various degrees depending on how "distant" the languages are.

One of the goals of linguistics is to study and categorize languages based on objective measures of linguistic distance. The degrees of similarity at different levels of the linguistic structural organization can be seen as preconditions for, as well as predictors of, successful oral intercomprehension. For closely-related languages, similarities at the pre-lexical, that is the acoustic-phonetic and phonological, level have been found to be better predictors of cross-lingual speech intelligibility than lexical similarities (Gooskens, W. Heeringa, and Beijering, 2008; W. Heeringa et al., 2009). In a different, yet relevant research direction, Skirgård et al. (2017) have investigated non-linguists' perception of language variation using data from the popular spoken language guessing game, the Great Language Game (GLG). By analyzing the confusion patterns of the GLG's human participants, the authors have shown that factors predicting players' confusion in the game correspond to objective measures of similarity established by linguists. For example, both phylogenetic relatedness and overlap in phoneme inventories have been identified as factors of perceptual confusability (and by implication, similarity) of languages in GLG.

4.1.1 Research Question

The development of automatic systems that determine the identity of the language in a speech segment has received attention in the automatic speech processing community (see H. Li, Ma, and K. A. Lee (2013) for an overview). State-of-the-art approaches for automatic spoken language identification (SLID) are based on multilayer deep neural networks (DNNs). DNN-based LID systems are parametric models that learn a mapping from spectro-temporal acoustic features of (untranscribed) speech to high-level feature representations in geometric space where languages are linearly separable. These models have shown tremendous success not only in discriminating between distant languages but also closely-related language varieties (Grégory Gelly et al., 2016; Mateju et al., 2018; Shon et al., 2018). Nevertheless, none of the studies in previous work has analyzed neural SLID models

to investigate whether or not and to what extent they capture cross-linguistic variation in their emerging representations. Thus, it is still unknown whether the distances in these representation spaces correspond to objective measurements of linguistic similarity and variation. In this study, we aim to fill this gap and consider the family of Slavic languages as a case study in our analysis.

- **RQ** Do the neural SLID models capture cross-linguistic similarity in their emergent representations? If yes, Which (linguistic) factors can best explain the similarity in the emergent language representation space?

4.2 Background

4.2.1 Slavic Languages

The Slavic language family is a branch of Indo-European languages that is conventionally divided into three subgroups: West-, East-, and South-Slavic (see 4.1). Apart from being related across the temporal dimension by sharing a common ancestor, Slavic languages form a spatial continuum of variation in a relatively connected geographic area across Europe and Northern Asia, except for the region where the Romance and Finno-Ugric wedge separates the South-Slavic from the West- and East-Slavic subgroups. Beside this traditional division (see Ethnologue, 23ed.), alternative classifications can be found in the Slavistics literature (cf. Bednarczuk, 2018; Dalewska-Greń, 2020; Lehr-Spławiński et al., 1954; Mańczak, 2004; Nalepa, 1968; Pianka and Tokarz, 2000). Nevertheless, and despite differences in taxonomies among various proposals, the development of contemporary Slavic languages from a common historical ancestor is uncontroversial. The supporting arguments are based on historical phonology and comparative studies of the phoneme inventories (Sawicka, 1991), as well as on studies of loanwords and Slavic toponyms. The high number of cognates as well as cross-linguistically shared features, such as lexical aspect, phonemic jotation and complex consonant clusters, provide strong evidence for common roots. In terms of diachronic phonology, the Common-Slavic era ends with the vocalization and reduction of the *yers* – the so-called "half-vowels" or "reduced vowels", [ъ] and [ь]. The outcomes of these alternations consistently define the most common division of Slavic. Similarly, the reflexes of *yat* [ě] provide a clear distinction between East-, West- and South-Slavic. The results of common phonological processes, such as liquid metathesis, palatalization and sibilarization, also support the tripartite division. Moreover, these regularities of sound changes allow us to precisely trace the phonological

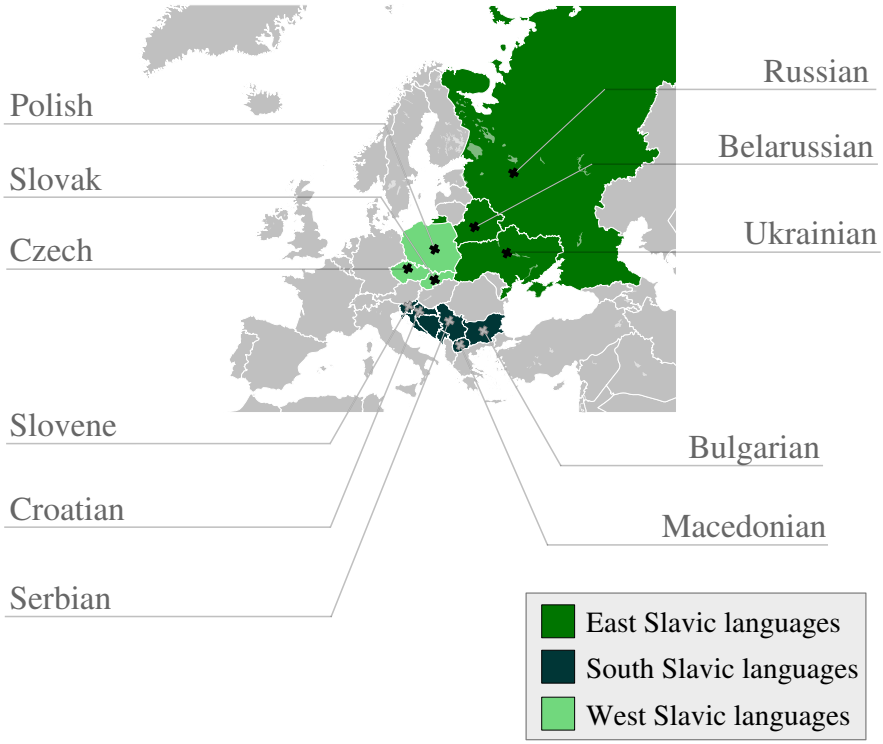


Figure 4.1: A political map that illustrates the set of Slavic languages in this study. The languages are classified based on widely accepted tripartite division of the Slavic languages.

development within the language family not only in the core standardized varieties but also in vernaculars and dialects. One of the objectives of our study is to assess the extent to which neural models of spoken language learn to detect such regularities from acoustic realizations of contemporary Slavic speech.

4.2.2 Language Identification in Speech Signals

Research in automatic identification of the language in a speech signal (i.e., SLID) is mainly concerned with the development of computational models that take an acoustic realization of a short utterance (usually a few seconds of speech) and predict the spoken language as output (H. Li, Ma, and K. A. Lee, 2013). Currently, end-to-end deep neural networks are the predominant paradigm for SLID and have shown tremendous success in prior studies (Gregory Gelly and Gauvain, 2017; Gonzalez-Dominguez et al., 2014; Lopez-Moreno et al., 2014). In this paradigm, the SLID problem has been modelled as a temporal sequence classification problem in which a spectro-temporal representation of a spoken utterance (e.g., a sequence of spectral feature vectors) is transformed via a multi-layer neural network into a

high-level vector representation that captures language identity. Other studies in the literature have addressed SLID for closely-related spoken language varieties including Arabic dialects (Grégory Gelly et al., 2016; Shon et al., 2018), Iberian languages (Grégory Gelly et al., 2016), and Slavic languages (Mateju et al., 2018).

At the intersection of speech recognition and linguistic typology, Gutkin et al. (2018) have trained a neural network on a large-scale multilingual speech database to predict typological features of the World Atlas of Language Structure (WALS) (Dryer and Haspelmath, 2013) for a language given a speech segment. The authors have shown that the speech modality contains enough signals to predict typological features of a held-out set of languages without explicit linguistic annotations. Their findings indicate that neural networks trained on multilingual speech could capture linguistic regularities and generalize beyond the languages observed in the training data. However, we are not aware of any prior work that has analyzed the emerging representations from SLID models or investigated whether or not the distance in these representation spaces reflect the linguistic distance.

4.2.3 Language Representations in Continuous Vector Spaces

Inspired by the advances in representation learning for NLP, multilingual neural models have been explored in the literature to induce real-valued language vectors, also known as *language representations* or *language embeddings*, where a single vector ($\mathbf{v} \in \mathbb{R}^d$) is associated with each language. Even though it has been motivated from different points of view, the main idea of this stream of research is to train a single NLP model on many languages whereby the language representation space is learned by exploiting the multilingual signal. For example, M. Johnson et al. (2017) introduced a multilingual neural machine translation (NMT) model in which the required target language of the translation was specified by a language embedding. Other studies have either scaled this approach to a massively multilingual setting (Malaviya et al., 2017; Östling and Tiedemann, 2017) or explored other NLP tasks such as linguistic structure prediction (Bjerva et al., 2019) and grapheme-to-phoneme conversion (Peters et al., 2017). Furthermore, Rabinovich et al. (2017) and Bjerva et al. (2019) have analyzed the learned language representations and shown that the distance in the representation space reflects the phylogenetic distance between Indo-European languages. However, Bjerva et al. (2019) have argued that structural syntactic similarities between languages are a better predictor of the language representation similarities than phylogenetic relatedness.

The most relevant analysis to ours is the recent work by Cathcart and Wandl (2020), in which the authors have trained a neural sequence-to-sequence model on a Slavic etymological dictionary. Their model was trained to take as input a reconstructed Proto-Slavic word form and a language embedding, then produce a word-form in the modern language specified by the language embedding. The authors have applied a clustering analysis on the learned language embeddings and were able to approximate the phylogenetic Slavic family tree. Our study complements this line of research with one fundamental difference: we perform our analysis on contemporary realizations of Slavic speech instead of the historically reconstructed phonological data without explicitly training our model to capture systematic sound changes.

4.3 Analytical Methodology

In this study, we analyze one of the models that have been developed from the previous chapter. Concretely, we analyze the emergent language representations in the domain-adaptive spoken language identification model (DA-SLID II) which takes mel-frequency spectral coefficients (MFSCs) as input. This SLID model was trained on the Slavic radio broadcast speech (RBS) as the source domain (i.e., labelled data) and the GlobalPhone speech (GPS) as the target domain (i.e., unlabelled data) for adaptation. The RBS dataset is a large collection of Slavic speech recordings that were collected by crawling online radio stations in prior work (Nouza et al., 2016; Mateju et al., 2018). The RBS dataset includes speech segments in 11 Slavic languages from the three subgroups:

1. South Slavic languages: Bulgarian (BUL), Croatian (HRV), Serbian (SRP), Slovene (SLV), and Macedonian (MAC).
2. West Slavic languages: Czech (CZE), Polish (POL), and Slovak (SLO).
3. East Slavic languages: Russian (RUS), Ukrainian (UKR), and Belarusian (BEL)

Even though the model was trained to identify only six of these 11 languages, we use all 11 languages in this analytical study to investigate whether or not the model can correctly identify the language subgroup for an unseen languages. To obtain vector-space language representations, we feed the utterances of the evaluation set to the model and extract the representations of the last non-linear layer of the language classifier (see Figure 4.2). Therefore, we use the language classification

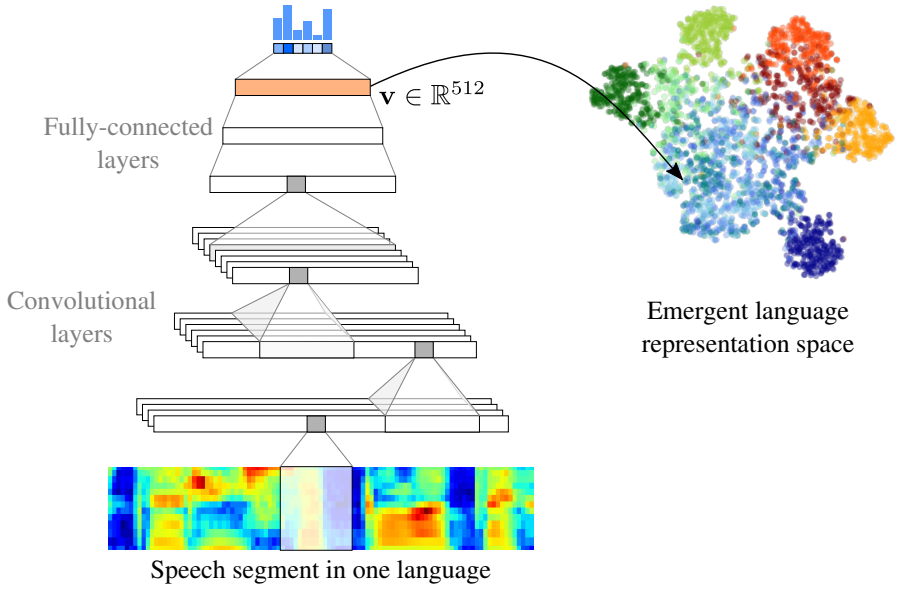


Figure 4.2: A spoken language identification (SLID) model used an encoder to represent a speech segment in a vector representation space.

model as an encoder $\mathcal{F}_{\text{SLID}} : \mathcal{X} \rightarrow \mathbb{R}^D$ to represent each speech segment $\mathbf{x}$ as a vector representation $\mathbf{v}_l$, which can be formally described as follows

$$\mathbf{v}_l = \mathcal{F}_{\text{SLID}}(\mathbf{x}; \boldsymbol{\theta}) \in \mathbb{R}^D$$

where $\boldsymbol{\theta}$ are the parameters of the encoder and D is the dimensionality of the representation of the last non-linear layer (here $D = 512$). The evaluation set consists of 500 utterances for each language where each utterance is a 5-second speech segment. Therefore, we obtain 500 representations for each language which we use in the exploratory visualization analysis in section 4.4. However, the analyses presented in sections 4.5 and 4.6 require a single vector representation for each language. To this end, we obtain a single prototypical vector representation for each Slavic language in our study by taking the average over the representations of the evaluation speech segments of language L as

$$\mathbf{v}_L = \frac{1}{|\mathcal{E}_L|} \sum_{\mathbf{x} \in \mathcal{E}_L} \mathcal{F}_{\text{SLID}}(\mathbf{x}; \boldsymbol{\theta}) \in \mathbb{R}^D$$

where $\mathcal{E}_L$ is the evaluation speech segments for language L. The distance in the representation space between two languages is then computed using two metrics: (1) Euclidean distance, and (2) cosine distances. Given two language representations $\mathbf{v}$ and $\mathbf{u}$, the Euclidean distance is defined as

$$d_E(\mathbf{v}, \mathbf{u}) = \|\mathbf{v} - \mathbf{u}\| = \sqrt{(\mathbf{v} - \mathbf{u})^\top (\mathbf{v} - \mathbf{u})} \in \mathbb{R}^+ \tag{4.1}$$

On the other hand, the cosine distance is defined as

$$d_c(\mathbf{v}, \mathbf{u}) = 1 - \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\|\mathbf{v}\|\|\mathbf{u}\|} = 1 - \frac{\mathbf{v}^\top \mathbf{u}}{\|\mathbf{v}\|\|\mathbf{u}\|} \in [0, 2] \tag{4.2}$$

where the $\|\mathbf{v}\| := \sqrt{\mathbf{v}^\top \mathbf{v}}$ is the L_2 norm of the vector $\mathbf{v}$ and $\langle \mathbf{v}, \mathbf{u} \rangle$ is the inner product between the vectors $\mathbf{v}$ and $\mathbf{u}$.

4.4 Analysis 1: Exploratory Visualization

In our first analysis, we visually inspect the emergent language representation space by means of data visualization. To this end, we use dimensionality reduction techniques to obtain two-dimensional projections of the evaluation samples and visualize the resulting data points. We use two dimensionality reduction techniques; t-SNE (Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018). The motivation for using two different techniques in this analysis is that t-SNE and UMAP have different optimization objectives that complement each other. That is, the t-SNE algorithm preserves the local structure of the space; thus, it mainly reveals the cluster structure within the representation space. On the other hand, the UMAP algorithm preserves the global structure of the space. The resulting graphs are illustrated in Figure 4.3. We observe that both t-SNE and UMAP plots in Figure 4.3 show very similar trends since the emerging language space shown in the figure correspond to the conventional sub-grouping of Slavic languages into East-Slavic, West-Slavic, and South-Slavic.

4.5 Analysis 2: Correlation with Geographic Distance

In the field of sociolinguistics, and specifically the sub-field of dialectology, geographic proximity is hypothesized to play a role in similarity of language varieties as well as a predictors of mutual intelligibility. That is, the closer the speaker communities of two language varieties A and B in the spatial sense, the more similar A and B are due to historical factors and the effect of borrowing linguistic features. This hypothesis was tested on the intelligibility of Dutch dialects and it was observed that geographic distance is not only a good predictor of dialect intelligibility but also highly correlates with measures of linguistic similarity such as lexical similarity (Gooskens and W. Heeringa, 2004; W. J. Heeringa, 2004).

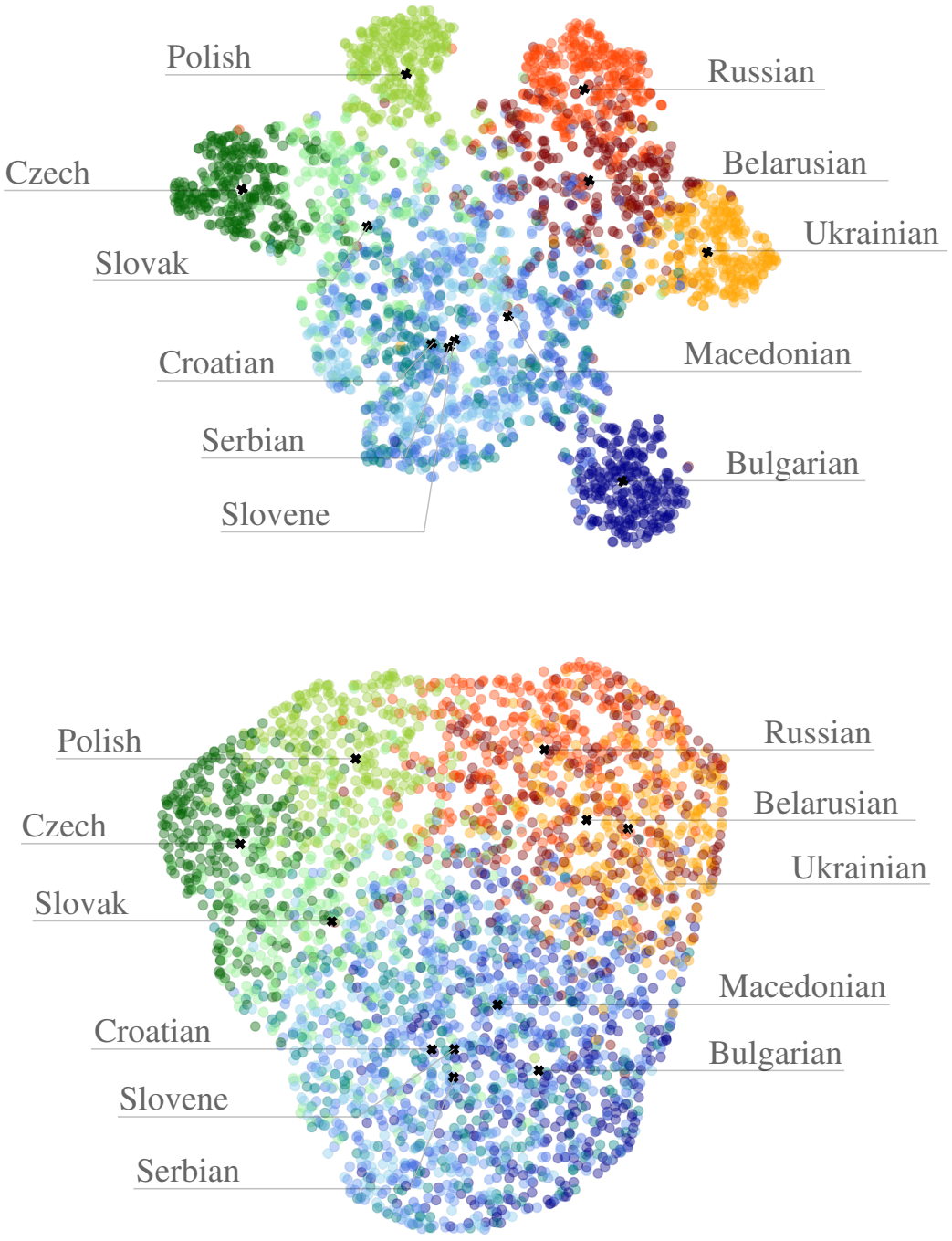


Figure 4.3: Two-dimensional visualization of representations of evaluation speech segments: (top) t-SNE projections, and (bottom) UMAP projections (best viewed in color).

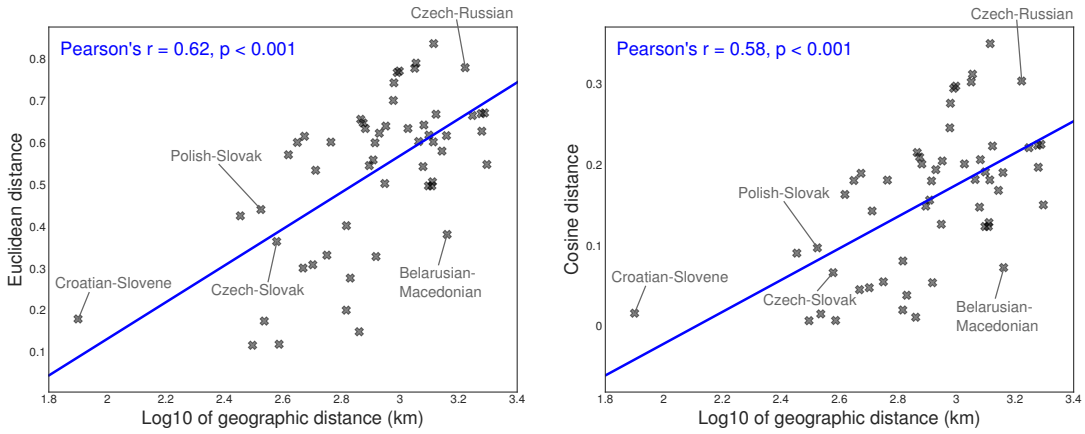


Figure 4.4: Correlation between geographic distance and distances between prototype language representations measured by Euclidean distance (left) and cosine distance (right).

However, in the same study the correlation in the case of Norwegian dialects was lower. A follow up study has proposed to use travelling time as predictor of linguistic similarity has shown as higher correlation with the intelligibility of Norwegian dialects (Gooskens, 2005). In the field of NLP and computational linguistics, Bjerva et al. (2019) have investigated the degree to which the emergent language representations in neural language models reflect the geographic proximity of their respective speaker communities.

To compute geographic distance, we follow a similar approach as in Skirgård et al. (2017) and Bjerva et al. (2019). First, each language is characterized by a pair of geographical coordinates in terms of latitude and longitude on the map, which we obtain from the ASJP linguistic database and they are intended to reflect the cultural and/or historical center of each language. We then compute the pairwise distances between the coordinates on the map (in kilometers) using the Haversine formula and convert them into $\log_{10}$ scale. Figure 4.4(a) shows a scatter plot between the data points in which the x-axis represents the geographic distance and the y-axis represents the cosine distance in the representation space. We observe a strong positive correlation between the geographic distance and the representational distance when measured in Euclidean distance (Pearson's $r = 0.62, p << 0.0001$) as well as cosine distance (Pearson's $r = 0.58, p << 0.0001$). This finding indicates that the distance in the representation space does indeed reflect the geographic distance.

Admittedly, representing a language as a pair of geographical coordinates is a cruel simplification in the context of modeling language variation and similarity. One can therefore rise the question whether the observed correlation between the language representational distance and geographic distance is a robust cor-

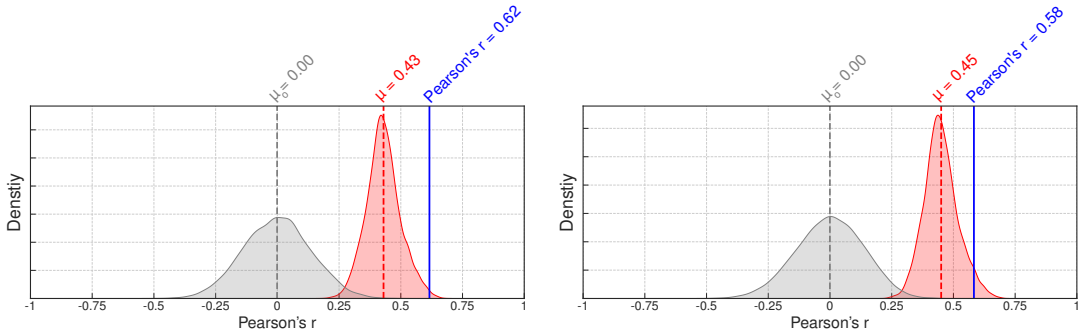
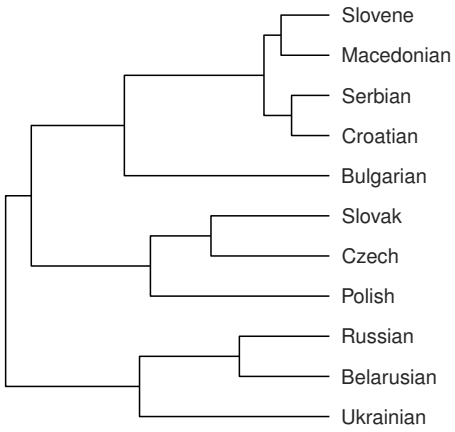


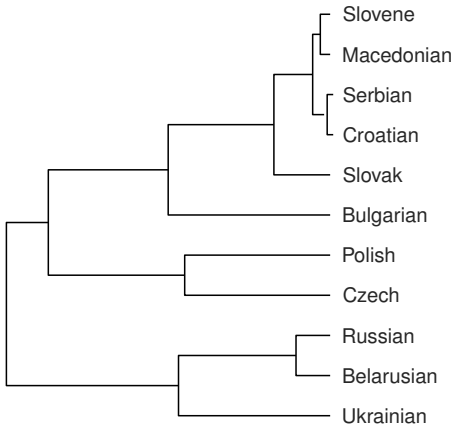
Figure 4.5: Testing the robustness of the correlation between geographic distance and representational distance measured by Euclidean distance (left) and cosine distance (right).

relation, and not just due to chance. To test the robustness of this correlation, we generate correlation coefficients using a sampling method to obtain the geographical coordinates that represent a language under two different assumptions: (1) the geographical coordinates are randomly sampled from the geographic area within the borders of the country where that language is spoken, and (2) the geographical coordinates are randomly sampled from the Slavic speaking area, without considering country borders. For each assumption, we sample a set of geographical coordinates for all languages and repeat this procedure 100,000 times to obtain two distributions. In assumption (1), the geographical coordinates are always restricted to be within the geographic area associated with the language (controlled by country borders). This restriction does not exist in assumption (2) and therefore we can obtain a null distribution where we are certain that the representational distance and geographic distance are unrelated. The results of the robustness test are shown in Figure 4.5 where the red and grey distribution curves represent the correlation values under assumption (1) and assumption (2), respectively. We observe that the two distributions are visually distinct where the mean of the null distribution (grey, assumption (1)) is centered at 0 while the mean of the "geographically restricted" distribution (red, assumption (2)) is centered approximately at 0.44. Moreover, none of the sampled sets of geographical coordinates under assumption (1) yields a negative correlation value. This clearly shows that the positive correlation we observe is a robust indication of the relatedness between the representational language similarities and geographic proximity and cannot be attributed to chance.

(a) Slavic tree generated by performing hierarchical clustering over the distance matrix computed by ***Euclidean distance***



(b) Slavic tree generated by performing hierarchical clustering over the distance matrix computed by ***cosine distance***



(c) A reference Slavic tree proposed by ***Novotná and Blažek (2007)*** using a classical glottochronological approach and cross-lingual lexical databese

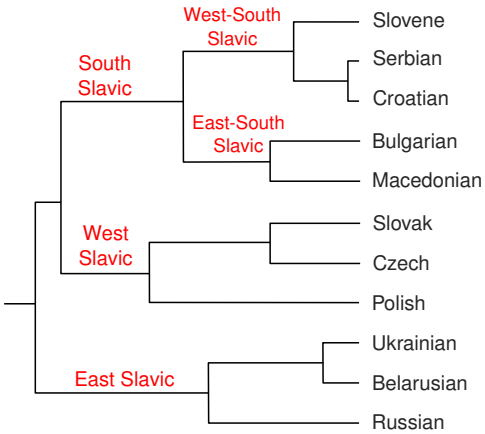


Figure 4.6: (a) a genetic tree generated from pairwise distances of the language representations measured by Euclidean distance. (b) (a) a genetic tree generated from pairwise distances of the language representations measured by cosine distance. (c) a ground-truth genetic tree.

4.6 Analysis 3: Probing the Genetic Signal

Similar to the analysis in Bjerva et al. (2019) and Cathcart and Wandl (2020), we investigate the genetic signal in the representation space. To this end, the pairwise representational distances computed in the previous section are first converted into a distance matrix for both Euclidean and cosine distances. Then, we generate a tree for each distance measure by performing hierarchical clustering on the distance matrix using the Ward algorithm (Ward, 1963). The generated trees are depicted in Figure 4.6(a) and Figure 4.6(b). We provide a reference phylogenetic tree from the work of Novotná and Blažek (2007) in Figure 4.6(c). Novotná and Blažek (2007) have used a classical glottochronological approach to propose a genetic a tree for the Balto-Slavic languages where the authors constructed a phylogenetic tree based on pairwise distances between languages using a string-based distance algorithm over a cross-linguistic word lists. We observe that the generated tree using Euclidean distance as a measure of dissimilarity (see Figure 4.6(a)) better approximates the reference tree proposed by Novotná and Blažek (2007) (see Figure 4.6(c)) for two main reasons: (1) the generated tree groups South- and West-Slavic languages into one major cluster, and (2) all South-, West-, and East-Slavic languages represent pure clusters as all languages are grouped into their respective sub-group. Nevertheless, there are minor, yet notable minor differences in the clustering within sub-groups. The most notable difference in our perspective is the the grouping of Sloven and Macedonian together before joining the Serbo-Croatian group, while Bulgarian seems a bit distant than the other South Slavic languages. Moreover, Belarusian and Ukrainian are grouped together first within the East Slavic languages in the reference tree, while in our reconstructed tree Belarusian is first grouped with Russian. On the other hand, the reconstructed tree using the cosine distance (see Figure 4.6(b)) is less similar to the reference tree since Slovak was incorrectly grouped with the Slavic languages. Other than that, the resulting (sub-)clusters are identical to the reconstructed tree using the Euclidean distance in Figure 4.6(a). Therefore, we conclude that it is necessary to test different measures of representational distance when performing phylogenetic inference over a continuous-space (computational) representation.

4.7 Discussion

A number of recent studies have shown that deep neural networks are adequate models of human perception. For example, R. Zhang et al. (2018) and J. C.

Peterson et al. (2018) have shown that emerging representations from neural models trained on visual recognition tasks are predictive of human similarity judgments. For auditory recognition, neural speech recognition models have been shown to capture human-like behavior in cross-lingual phonetic perception (Schatz and Feldman, 2018). Inspired by this line of research, our objective in this study presented in this chapter is to investigate the extent to which neural models of spoken language identification capture language similarity and relatedness. Nevertheless, and because of the complex space in which language variation can be realized, the similarity between languages is a multidimensional phenomenon that cannot be expressed in a single number (Van Heuven, 2008). We therefore do not consider a single reference as a "ground truth" in our analysis, but consider several reference criteria of linguistic distance and relatedness, namely geographic proximity and genetic relatedness.

Our analysis of the emerging language representation space has shown that the SLID model in our study captures language similarity to a great extent. The representation visualization illustrated in Figure 4.3 demonstrates the generalization ability of our model to project speech segments of held-out languages into subspaces of their respective subgroups. That is, even though the model has not observed any speech sample from Belarusian or Macedonian, Belarusian was projected onto the East-Slavic cluster while Macedonian was projected onto the South-Slavic cluster. Given that the data in our study constitute contemporary realizations of Slavic speech that do not explicitly encode diachronic sound changes, we first hypothesized that the geographic distances between the speaker communities would be a good predictor of the distances in the representation space of the SLID model. This turns out to indeed be the case as we observe a strong positive correlation at Pearson's $r = 0.62, p << 0.0001$ between geographic distances and representational distances among the prototype language representations.

On the other hand, we were less optimistic about our SLID model capturing the genetic signal between Slavic languages (that is, whether the language representation space can accurately predict the historical relationships between languages). Prior studies in the literature that have investigated computational approaches to generating genetic language trees have either employed historical etymological data capturing phonological sound changes (Cathcart and Wandl, 2020), sequences reflecting syntactic patterns in different languages (Bjerva et al., 2019; Rabinovich et al., 2017), or word lists reflecting lexical similarity (Serva and Petroni, 2008). Arguably, these sources of language data are more likely to preserve the relationship between languages across the temporal dimension compared to the contemporary Slavic speech we use in this study. Therefore, our

initial intuition was that the resulting tree would reflect variation across the spatial dimension more than the temporal dimension. Nevertheless, the tree generated by our analysis is an adequate approximation of the Slavic genetic tree given the contemporary nature of the data sources.

4.8 Summary

This chapter explored the encoding of cross-linguistic variation in neural network representations of spoken language identity. While neural models have demonstrated impressive performance in discriminating related languages based on acoustic speech signals, it remained unknown whether they captured and represented cross-linguistic variation in their intermediate representations. To this end, we presented a linguistically-informed exploration of neural representations of spoken languages. We focused on the Slavic language family and investigated the degree to which the model's representational similarity among languages aligned with objective measures of language similarity. Importantly, the model did not have any access to explicit information regarding the phylogenetic relationships between the languages under consideration. Through this study, it was demonstrated that the model's representations exhibited a cluster structure that corresponded that largely reflects the phylogenetic groups within the Slavic language family. Remarkably, this held true even for languages that were not included in the model's training data. These findings highlights of ability of neural models to capture and reflect the underlying linguistic relatedness of languages, despite the absence of direction supervision signals or knowledge regarding their phylogenetic relations.

Part III

SPOKEN-WORD REPRESENTATIONS

5

On the Geometry of Spoken-Word Representations

This chapter begins the third part of the thesis, which is concerned with neural models of spoken-word representations. In this part, each word is modeled as an abstract category that consists of several acoustic exemplars that vary across speakers and within-speaker, given different lexical contexts. Spoken-word representation models encode each acoustic exemplar in a representational space such that different exemplars of the same word category are nearby. This chapter presents an analytical study that takes a closer look at the representational geometry of acoustic word embeddings and how they encode spoken-word variability. Specifically, we analyze the uniformity of the representation space, propose a metric to quantify word category discriminability, and employ the concept of representational consistency to investigate whether acoustic word embeddings models exhibit individual differences.

5.1 Introduction

Distributed representations such as semantic word embeddings are nowadays a central component in natural language processing (NLP). Inducing word embeddings from text yields representations such that words occurring in similar contexts are nearby in the vector space (Mikolov, Sutskever, et al., 2013; Pennington et al., 2014). Therefore, the representational geometry of text-based word embeddings captures lexical similarity and semantic relatedness at multiple levels of granularity. Word embeddings, and their underlying distributional semantic models, have also been adopted as models of human semantic memory in cognitive science research (Grand et al., 2022; Nematzadeh et al., 2017; Pereira et al., 2016).

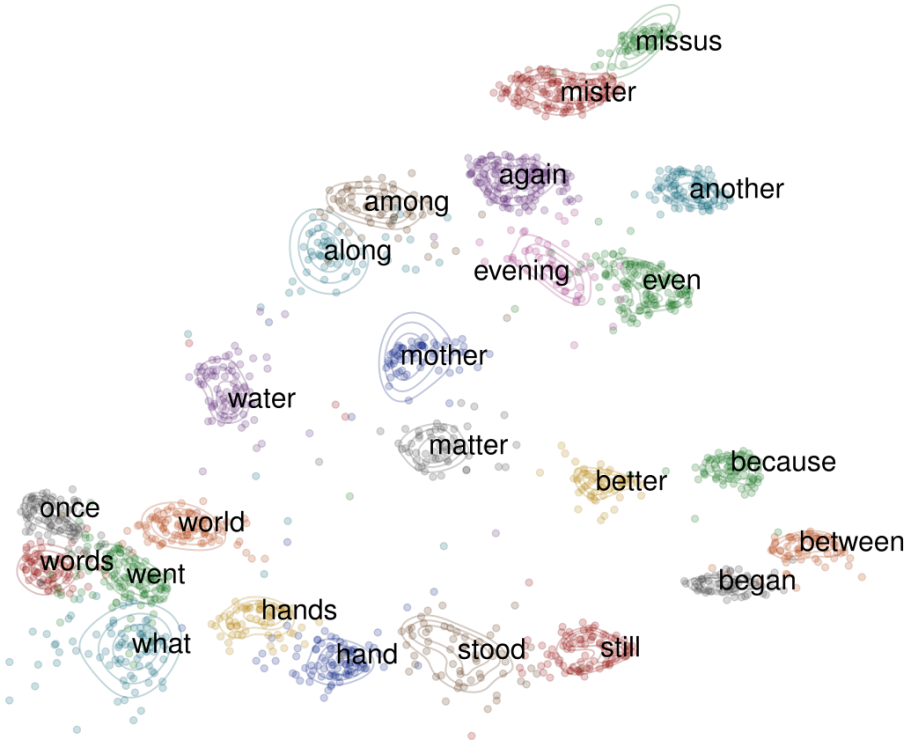


Figure 5.1: UMAP two-dimensional projection of a sample of acoustic word embeddings (AWEs) produced by a correspondence autoencoder (CAE) model trained on English speech. AWE models project different exemplars of the same word type closer in the embedding space while abstracting away from speaker and context variability.

In the speech processing domain, researchers have independently developed representations of acoustic segments that correspond to linguistic units (S. Bengio and Heigold, 2014; Herman Kamper, W. Wang, et al., 2016; Levin et al., 2013; Settle and Livescu, 2016b, *inter alia*). A notable example of such representations are acoustic word embeddings (AWEs)—vector representations that encode the sound structure of spoken words, not their semantic content or syntactic structure—see Figure 5.1. AWEs support voice-based speech technology applications such as query-by-example spoken term discovery (Jansen and Durme, 2012; Metze et al., 2013; Yaodong Zhang and James R Glass, 2009) and keyword spotting (Myers et al., 1980; Rohlicek, 1995). In addition, AWEs can facilitate access to speech recordings of endangered spoken languages that might lack standardized writing systems (Bird, 2021; San et al., 2021).

However, there are fundamental differences between text-based and speech-based word representations that have to do with the degree of **variability** between the two modalities. Contrary to written words which have **context-invariant**

orthographic realizations,[1] spoken words are **notoriously variable**. The acoustic realizations of words vary across speakers due to differences in vocal tract shapes, genders, and dialects, among many other factors. In addition, two acoustic instances, or exemplars, of the same word vary in different phonological and semantic contexts even if they are produced by the same speaker (Jurafsky, 2003). Therefore, spoken-word representations such as acoustic word embeddings are not static vectors in a lookup table, but rather they are computed "on the fly" given a spoken-word segment as input. To project different acoustic exemplars of the same word onto the same point of the representation space, spoken-word representation models have to abstract away from speaker and context variability.

5.1.1 Research Questions

State-of-the-art spoken-word representation models for acoustic word embeddings are based on deep neural networks (DNNs). In addition to their speech technology applications, AWEs have been adopted as models of human speech processing and analyzed from a cognitively motivated angle in several recent studies. For example, it has been shown that AWEs exhibit a human-like word onset bias where distinct words are more likely to be perceived as similar if they begin with the same sound (Matusevych, Herman Kamper, et al., 2020b). AWE models have also been shown to predict non-native perceptual difficulties in phonetic categorization (Matusevych, Schatz, et al., 2020) as well as cross-linguistic effects in auditory-lexical processing (Matusevych, H. Kamper, et al., 2021). Furthermore, models of AWEs have been reported to capture lexical production patterns among Japanese-speaking learners of English as a second language (Ando et al., 2021). Nevertheless, AWEs as well as their underlying neural architectures and learning objectives have not yet been extensively studied from a neural network interpretability point of view. Therefore, it remains unclear how DNN-based models of AWEs encode speech variability and whether or not their representational geometry is affected by variability in initial conditions of the model. In this study, we aim to answer the following research questions:

- **RQ1.** How do AWE models encode spoken-word variability in within their representations? That is, to which degree is the variance uniformly distributed across all dimensions of the representation space?
- **RQ2.** By considering a word as a category comprised of several acoustic exemplars, how can we quantify the compactness and separability of each

1 although some orthographic variation exists in informal, user-generated text such as tweets.

category? What lexical properties could potentially predict word discriminability within the representation space?

- **RQ3.** Does the variability in the initial conditions of the models influence their representational profile? In other words, do different initializations result in individual differences among AWE models?

To address these research questions, we employ several analytic methods from machine learning and neuroscience to take a closer a look at the representational geometry of acoustic word embeddings. Concretely, we analyze the **uniformity** of the representation space (**RQ1**, §5.5), propose a metric to quantify **word category discriminability** (**RQ2**, §5.6), and employ the concept of **representational consistency** to investigate whether AWE models exhibit individual differences (**RQ3**, §5.7).

5.2 Acoustic Word Embedding Models

Given an acoustic signal that corresponds to a spoken word represented as a temporal sequence of T acoustic feature vectors, i.e., $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T)$, the goal of an AWE model is to transform $\boldsymbol{a}$ into a fixed-dimensionality vector representation $\mathbf{x}$. Due to the variability in speech production (i.e., speech rate, emotional state, etc), the length of the acoustic segment T varies between different exemplars, or instances, of the same word type. Therefore, this task is modeled as a mapping $\mathcal{F} : \mathcal{A} \rightarrow \mathbb{R}^D$, where $\mathcal{A}$ is the (continuous) space of acoustic sequences and D is the dimensionality of the embedding. Formally, transforming a variable-length acoustic input into a D-dimensional AWE is described as

$$\mathbf{x} = \mathcal{F}(\boldsymbol{A}; \boldsymbol{\theta}_{\mathcal{F}}) \in \mathbb{R}^D \tag{5.1}$$

where $\boldsymbol{\theta}_{\mathcal{F}}$ are the parameters of the encoder function $\mathcal{F}$. In a supervised setting of training AWE models, one assumes a dataset $\mathcal{D} = \{(\boldsymbol{a}^1, w^1), (\boldsymbol{a}^2, w^2), \ldots, (\boldsymbol{a}^N, w^N)\}$ of N spoken word instances where w^i is the word type, or word category, of the i^{th} acoustic sample. In this study, we experiment with two architectures—recurrent and convolutional—and employ four different learning objectives for training AWE models that were proposed in the literature. Next, we formally describe each of the objectives.

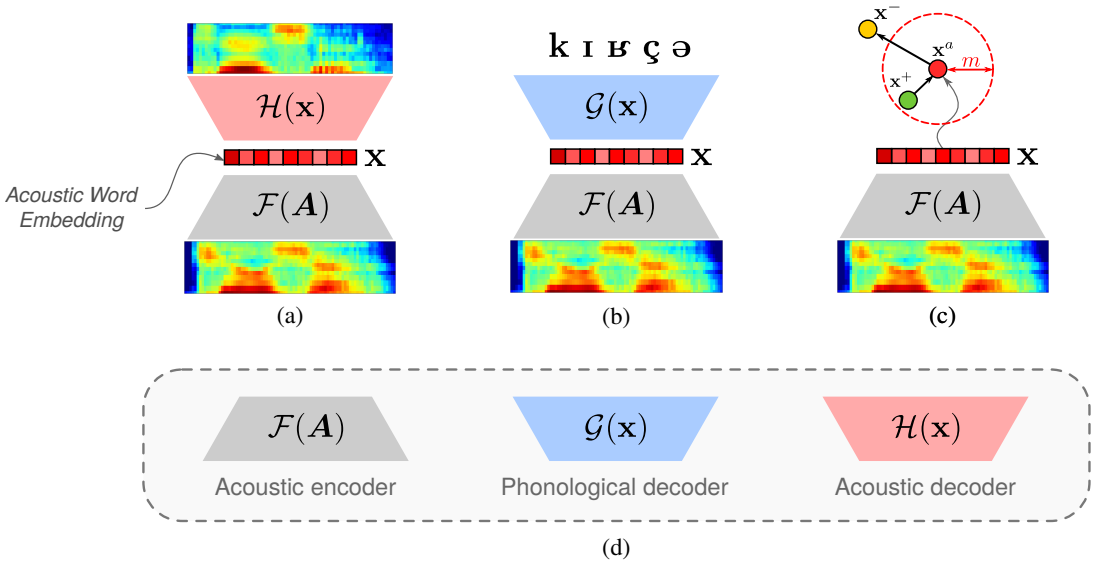


Figure 5.2: A visual illustration of the different learning objectives for training AWE encoders: (a) correspondence Autoencoder (CAE): a sequence-to-sequence network with an acoustic decoder, (b) phonologically guided encoder (PGE): a sequence-to-sequence network with a phonological decoder, and (c) contrastive siamese encoder (CSE): a contrastive network trained via triplet margin loss. After training the model, only the encoder component of the model $\mathcal{F}$ is used to produce AWEs. (d) Individual components of the models.

5.2.1 Correspondence Autoencoder

In the correspondence autoencoder (CAE; Herman Kamper, 2019), each training acoustic word sample $\boldsymbol{A}$ is paired with another sample that corresponds to the same word type $\boldsymbol{A}^+ = (\boldsymbol{a}_1^+, \boldsymbol{a}_2^+, ..., \boldsymbol{a}_S^+)$. The acoustic encoder $\mathcal{F}$ takes $\boldsymbol{A}$ as input and produces an embedding $\mathbf{x}$, which is then fed to an acoustic decoder $\mathcal{H}$ that aims to sequentially reconstruct the corresponding acoustic sequence $\boldsymbol{A}^+$—see Figure 5.2(a). The objective is to minimize the L_2 distance at each timestep in the decoder, which is equivalent to

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{H}}) &= \sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \texttt{dist}\Big(\boldsymbol{A}^{i+}, \mathcal{H}\big(\mathcal{F}(\boldsymbol{A}^i; \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{H}}\big)\Big) \\ &= \sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \sum_{t=1}^{S} \|\boldsymbol{a}_t^+ - \mathcal{H}(\mathbf{x})_t\|_2\end{aligned} \tag{5.2}$$

where $\texttt{dist}(.,.)$ can be viewed as a distance function between two sequences, $\boldsymbol{a}_t^+$ is the ground-truth acoustic feature vector at timestep t, $\mathcal{H}(\mathbf{x})_t$ is the reconstructed acoustic vector at the same timestep as a function of the embedding $\mathbf{x}$, and $\boldsymbol{\theta}_{\mathcal{G}}$

are the parameters of the decoder. Matusevych, Herman Kamper, et al. (2020a) have hypothesized that learning the correspondence between different acoustic realizations of the same word category seems to encourage the encoder to build up speaker-invariant word representations while preserving linguistically-relevant phonetic information . When the target acoustic sequence to generate is the same as the input signal $\boldsymbol{A}$, this corresponds to a conventional autoencoder (AE), which we consider as one of our baseline learning objectives in this study.

5.2.2 Phonologically Guided Encoder

The phonologically guided encoder (PGE) is trained as component in a sequence-to-sequence model to map a sequence of continuous acoustic vectors onto a sequence of discrete phonological units. First, the acoustic sample $\boldsymbol{A}$ is transformed by the encoder into an embedding $\mathbf{x}$. Given the output of the encoder $\mathbf{x}$, a phonological decoder $\mathcal{G}(.;\boldsymbol{\theta}_{\mathcal{G}})$ is trained to decode the corresponding phonological sequence $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_\tau)$ of the word-form—see Figure 5.2(b). The objective is to minimize the surprisal of the corresponding phonological sequence. This learning objective is realized by a categorical cross-entropy loss at each decoder timestep, which is equivalent to minimizing the term

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{G}}) &= -\sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \log \mathbf{P}\big(\boldsymbol{\varphi} \,|\, \mathcal{F}(\boldsymbol{A}^i; \boldsymbol{\theta}_{\mathcal{F}});\, \boldsymbol{\theta}_{\mathcal{G}}\big) \\ &= -\sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \sum_{t=1}^{\tau} \log \mathbf{P}\big(\varphi_t \,|\, t, \mathbf{x}^i;\, \boldsymbol{\theta}_{\mathcal{G}}\big)\end{aligned} \tag{5.3}$$

where $\mathbf{P}\big(\varphi_t \,|\, t, \mathbf{x}^i;\, \boldsymbol{\theta}_{\mathcal{G}}\big)$ is the probability of the phoneme φ_t at the tth timestep, conditioned on the previous phoneme sequence $\boldsymbol{\varphi}_{<t}$ and the AWE $\mathbf{x}$, and $\boldsymbol{\theta}_{\mathcal{G}}$ are the parameters of the decoder. The underlying intuition of this learning objective is as follows: despite the variation in acoustic realizations due to speaker and context variability, different exemplars of the same word category should have identical phonological sequences. Consequently, we expect the encoder to project exemplars of the same word category close together in the embedding space. Ideally, the degree of representational similarity should correlate with phonological similarity.

5.2.3 Contrastive Siamese Encoder

The contrastive siamese encoder (CSE) has been explored in the context of AWEs with both recurrent and convolutional architectures in several studies (Jacobs,

Matusevych, et al., 2021; H. Kamper et al., 2016; Settle and Livescu, 2016a). Contrary to the previously described objectives, the CSE explicitly minimizes the distance between exemplar embeddings of the same word type—see Figure 5.2(c). First, each acoustic word instance is paired with another instance of the same word type $(\boldsymbol{A}, \boldsymbol{A}^+)$. Given their embeddings $(\mathbf{x}^a, \mathbf{x}^+)$, the objective is then to minimize a triplet margin loss

$$\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}) = \max\big[0, m + d(\mathbf{x}^a, \mathbf{x}^+) - d(\mathbf{x}^a, \mathbf{x}^-)\big] \tag{5.4}$$

Here, $d(.,.)$ is the cosine distance and $\mathbf{x}^-$ is an AWE that corresponds to a different word category sampled from the mini-batch such that the term $d(\mathbf{x}^a, \mathbf{x}^-)$ is minimized. This learning objective aims to bring acoustic instances of the same word type closer in the embedding space while pushing away instances of different word types, with the extent of separation controlled by the margin hyperparameter m.

5.3 Experimental Setup

5.3.1 Data

The data in our study is drawn from the the LibriSpeech dataset which contains read speech recordings of American-English (Panayotov et al., 2015). LibriSpeech is a public dataset under the CC BY 4.0 license. We sample 384 speakers from for training and 128 speakers for evaluation—disjoint sets—and obtain word-aligned speech samples using the Montreal Forced Aligner (McAuliffe et al., 2017). To make our models comparable with prior work, which has focused on AWEs for low-resource languages, we sample $\sim$ 39.4k samples for training and $\sim$ 9.7k for evaluation. The phonetic transcription for each word is produced using the online *WebMaus* G2P tool (Strunk et al., 2014). Then, each acoustic sample is parametrized as a sequence of 39-dimensional Mel-frequency spectral coefficients (or Mel filter-banks representations) of 25ms frames a stride of 10ms—the conventional feature representation of speech in automatic speech recognition (ASR). It is worth pointing out that in this study we consider each morphological variant of a lexeme as a separate word category. For example, different inflections of the lexeme MAKE such as {MADE, MAKING, MAKER, etc.} represent different word categories, each consisting of several exemplars.

5.3.2 Architectures, Hyperparameters, and Training Details

CNN Acoustic Encoder. We employ a 3-layer temporal convolutional network (1D-CNN) with 256, 384, and 512 filters and widths of 4, 8, and 16 for each layer and keep stride step at 1. Following each convolutional operation, we apply batch normalization, `ReLU` non-linearity, and dropout. We apply average pooling to downsample the representation at the end of the convolution block, then apply one non-linear layer with `Tanh` on the CNN output, which yields a 512-dimensional AWE.

RNN Acoustic Encoder. We employ a 3-layer unidirectional Gated Recurrent Unit (GRU) with a hidden state dimension of 512, then apply one non-linear layer with `Tanh` on the GRU last hidden state, which yields a 512-dimensional AWE. We apply layer-wise dropout with a probability of 0.1.

Phonological Decoder $\mathcal{G}(.;\boldsymbol{\theta}_{\mathcal{G}})$. We employ a 1-layer GRU of 512 units hidden state that takes the 512-dimensional AWE as the initial hidden state and decodes the corresponding phonological sequence without teacher forcing.

Acoustic Decoder $\mathcal{H}(.;\boldsymbol{\theta}_{\mathcal{H}})$. We employ a 1-layer GRU of 512 units hidden state that takes the 512-dimensional AWE as the initial hidden state and decodes the corresponding acoustic sequence with a teacher forcing ratio of 0.2.

Contrastive Loss. For the CSE, we experiment with different values of the margin hyperparameter $m = \{0.2, 0.3, 0.4, 0.5\}$, out of which 0.4 yields the best performance on the validation set.

Training Details. All models in this study are randomly initialized with each parameter drawn uniformly from $[-0.05, 0.05]$. Then, each model is trained for 100 epochs with a batch size of 256 using the ADAM optimizer (Kingma and Ba, 2015) and an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.5 if the mAP on the validation set does not improve for 10 epochs.

Implementation. We build our models using PyTorch (Paszke et al., 2019) and use FAISS (J. Johnson et al., 2017) for efficient similarity search.

5.4 Evaluation: Acoustic Word Discrimination Task

We conduct an intrinsic evaluation for the AWEs to assess the performance of our models using the same-different acoustic word discrimination task measured by the mean average precision (mAP) metric (Algayres et al., 2020; Carlin et al., 2011; Herman Kamper, Elsner, et al., 2015; Settle, Audhkhasi, et al., 2019). The

word discrimination task mainly evaluates the ability of a model to determine whether or not two given spoken-word instances correspond to the same word category. Following the definition of Müller (2015), we conceptualize this task as an exemplar retrieval problem: given a query spoken-word instance $\overline{\boldsymbol{q}}$ and a candidate set of k spoken-word instances $\mathcal{S} = \{\overline{\boldsymbol{s}}_1, ..., \overline{\boldsymbol{s}}_k\}$, the goal is to rank spoken-word candidates in $\mathcal{S}$ in such a way that those which belong to the same word type as the query $\overline{\boldsymbol{q}}$ are highly ranked among other candidates. To this end, a vector-based search index is built by mapping each word candidate in $\mathcal{S}$ into an embedding. Then, the cosine similarity between the query embedding $\overline{\boldsymbol{q}}$ and each embedding in the search index is computed which yields a ranked list, or an ordering, of spoken-word instances based on the cosine similarity score. The average precision metric is used to evaluate the quality of the ordering for a single query as

$$\text{AP} = \frac{1}{|\mathcal{S}_{\boldsymbol{q}}|} \sum_{r=1}^{k} P_{\boldsymbol{q}}(r) \times \mathcal{I}_{\boldsymbol{q}}(r) \tag{5.5}$$

where $\mathcal{S}_{\boldsymbol{q}}$ are the spoken-word candidates in $\mathcal{S}$ that are true exemplars of the same word category as the query instance $\overline{\boldsymbol{q}}$, $P_{\boldsymbol{q}}(r)$ is the precision at rank r, and $\mathcal{I}_{\boldsymbol{q}}(r)$ is a relevance function such that $\mathcal{I}_{\boldsymbol{q}}(r) = 1$ if the candidate at rank r corresponds to the same word category as the query, or $\mathcal{I}_{\boldsymbol{q}}(r) = 0$ otherwise. The arithmetic average over all AP values in the test set yields the mean average precision (mAP) metric.

The results of the evaluation is shown in Figure 5.3. We observe that each recurrent encoder outperforms its convolutional counterpart within each objective. Moreover, the performance largely depends on the strength of the supervision signal where the contrastive encoders outperform other objectives that lack explicit loss to group exemplars of the same category closer in the embedding space. Note that the CAE model is pre-trained as autoencoder for 10 epochs, following prior work (Herman Kamper, 2019).

5.5 Analysis 1: Uniformity of Representation Space

In our first analysis, we take a closer look at the uniformity of representational spaces of AWE models by analyzing the distribution of cosine similarity for each model type and quantifying the degree to which the embeddings are isotropic.

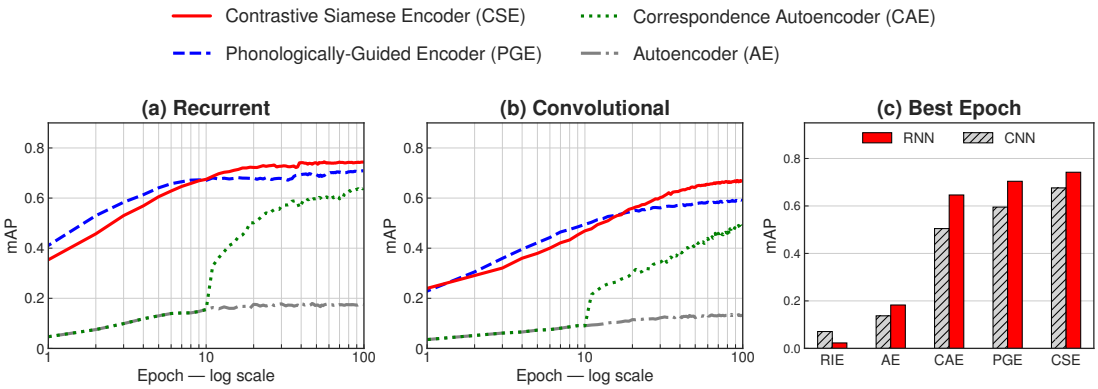


Figure 5.3: Evaluation on the same-different acoustic word discrimination task quantified by the word discrimination task and the mAP metric: Learning curves of 100 training epochs for (a) the recurrent encoder and (b) convolutional encoders. (c) mAP of the best epoch.

5.5.1 Distribution of Cosine Similarity

Analyzing the geometry of representation spaces in the acoustic domain can be achieved by examining the **cosine similarity distributions** among instances of the same word category versus those across different categories. From a practical point of view, high cosine similarity among exemplars of the same category is desirable (i.e., minimizing within-category variability), while instances from different categories should ideally exhibit low cosine similarity (i.e., maximizing cross-category separability). These concepts and how they are reflected in the representation space are visually illustrated in the Figure 5.4

For this analysis, we compute the within-category and cross-category cosine similarity scores from a large sample of spoken-word pairs derived from the training data. The results of this analysis are presented in Figure 5.5. We observe that the difference between the means of the within-category and cross-category distributions largely depends on the strength of the supervision signal, with the randomly initialized encoders (RIE) having the smallest mean differences for both architectures. The contrastive encoders have the largest mean difference with mean cross-category scores centered at the zero. This behavior of the contrastive encoders is not surprising considering the explicit supervision signal they receive in grouping exemplars of the same category closer in the embedding space. On the other hand, it is surprising that the untrained convolutional encoder yields cosine similarity scores very close to 1 for each input pair.

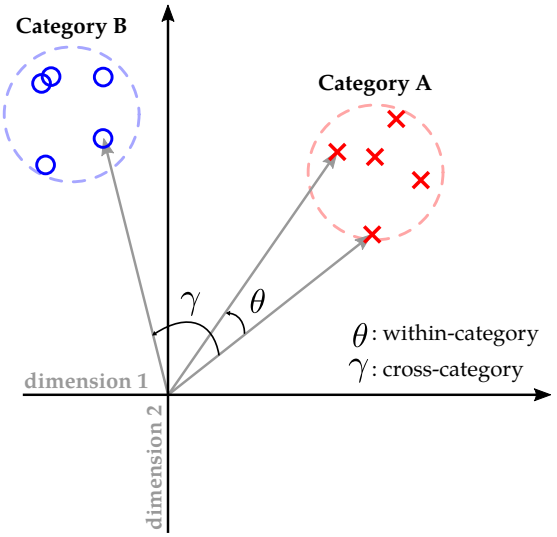


Figure 5.4: A visual illustration of within-category and cross-category cosine similarity in a simplified view of a two-dimensional representation space with two distinct categories.

5.5.2 The Degree of (An)isotropy

Although inspecting the cosine similarity distributions is an insightful analysis, it does not enable us to make well-informed judgments about the uniformity of the representation spaces. Here, we ask two questions: (1) do AWE models utilize all dimensions of the vector space to represent the speech samples and separate the categories? and (2) how do architecture and learning objective affect the distributivity of information in the embedding space? To answer these questions, we inspect the degree of **isotropy** in the representation spaces. A representation space is said to be **maximally isotropic** if the variance is uniformly distributed across all dimensions. That is, the data points are geometrically organized in a uniform matter in all directions from the origin, with no preferred direction of variance. On the other hand, a representation space is said to be **minimally isotropic** if the data points vary along a single dimension. An anisotropic representation space can be anywhere between maximally isotropic and minimally isotropic depending on the degree of the variance uniformity. Isotropy is often a desired property in the representational spaces of deep neural networks. The concept of isotropy is illustrated in Figure 5.6.

Prior work in NLP has found that semantic word embeddings tend to be anisotropic since they only utilize a few dimensions of the vector space—an effect that has been observed for word embeddings that are static (Mimno and Thompson, 2017; Mu and Viswanath, 2018) as well as contextualized (Cai et al.,

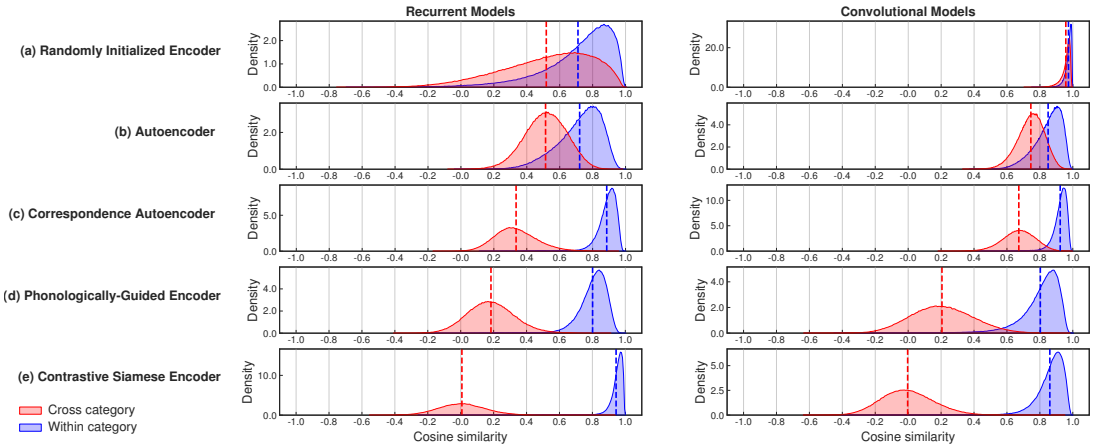


Figure 5.5: Distribution of cosine similarity across different AWE models for within category samples (i.e., exemplar pairs of the same word type) and cross-category samples (i.e., sample pairs that correspond to different word types). Each row in the figure corresponds to one learning objective and each column corresponds to one architecture.

2020; Ethayarajh, 2019; Rudman et al., 2022). The degree of isotropy in acoustic word embeddings, however, remains so far unknown.

To inspect the degree of isotropy of the AWE vector spaces, we use the IsoScore metric recently proposed by Rudman et al. (2022), which is—to the best of our knowledge—the only metric in the literature that is grounded on the mathematical definition of isotropy. The IsoScore metric operates on the covariance matrix of the embedding dimensions and returns values between 0 (minimally isotropic) and 1 (maximally isotropic).[2] We quantify the degree of isotropy using IsoScore for each model type and show the result in Figure 5.7(a). We observe that IsoScore returns values that are within the range $[0.002, 0.095]$, which indicates that embedding spaces for all models tend towards being minimally isotropic. However, the embeddings of untrained, randomly initialized encoders (RIE) tend to be extremely anisotropic (i.e., IsoScore values close to 0). This observation suggests that the anisotropic space does not "emerge" during the model training but rather that it is an inherent property of the encoder architecture. We are not aware of prior work in NLP that has studied the degree of isotropy in untrained NLP models to investigate whether anisotropic spaces are an emergent or inherent feature. In our case, training with a learning objective that encourages the model to separate word categories moves the representation space more towards utilizing more dimensions, therefore resulting in a higher degree of isotropy. Moreover,

2 The detailed mathematical definition of the IsoScore is beyond the scope of this chapter. We refer the reader to the work of Rudman et al. (2022) for a detailed formal description of this metric.

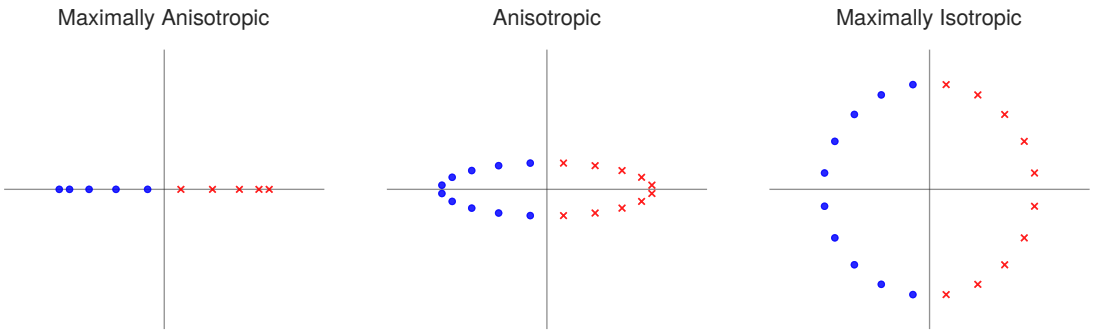


Figure 5.6: A visual illustration of the isotropy concept in a two-dimensional representation space with two distinct categories. In a maximally anisotropic space, the variance in the data is encoded along a single dimension, or a line (left). In general, the variance in an anisotropic representation space is not uniformly distributed across all dimensions (middle). In a maximally isotropic space, the variance is uniformly distributed (right).

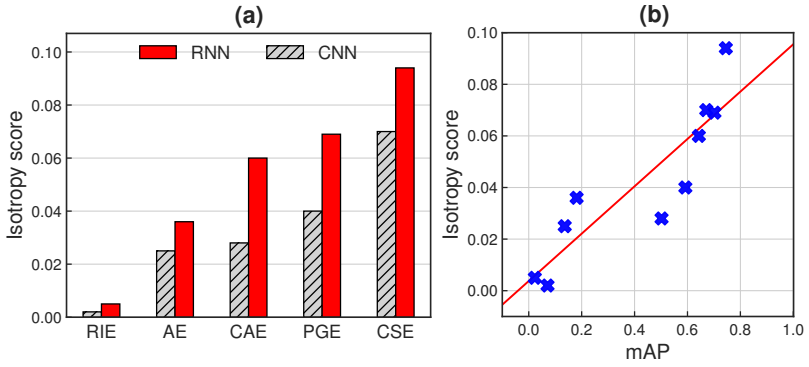


Figure 5.7: (a) The degree of isotropy of AWE for each model. (b) Correlation between the word discrimination performance measured by mAP and isotropy score (Pearson $r = 0.89, p < 0.001$).

recurrent encoders tend to be more isotropic than their convolutional counterparts within the same learning objective.

Despite the tendency of all models to be anisotropic, we find a strong positive correlation between the degree of isotropy and the performance on word discrimination—see Figure 5.7(b). That is, the more dimensions the model utilizes in the representation space, the better it performs on the intrinsic evaluation task.

5.6 Analysis 2: Word Category Discriminability

Ideally, AWE models should project exemplars of the same word category onto the same point in the embedding space. However, there are no strong constraints during training to encourage maximal separability between different word categories. In

this analysis, we seek to answer two questions: (1) how well-separated are the word categories of the training samples? and (2) to what degree do lexical properties predict the discriminability of word categories?

5.6.1 Category Discriminability Index

In order to investigate the geometric density of each word category in the representation space, we need to measure within-category compactness and cross-category separability. Inspired by the exemplar discriminability index proposed in the neuroscience literature (Nili et al., 2020), we define **category discriminability index** (CDI) as a metric that operates on within-category and cross-category distances. If we consider each word category in the training set as a set of its exemplar embeddings $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{C}|}\}$, CDI is defined for a single word category $\mathcal{C}$ as

$$\text{CDI}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\forall \mathbf{x}_i \in \mathcal{C}} \Big(\sum_{\forall \mathbf{x}_j \sim \mathcal{C} | j \neq i} d(\mathbf{x}_i, \tilde{\mathbf{x}}_j) - d(\mathbf{x}_i, \mathbf{x}_j) \Big) \tag{5.6}$$

where $d(.,.)$ is the cosine distance and $\{\mathbf{x}_i, \mathbf{x}_j\}$ is a pair of within-category samples while $\tilde{\mathbf{x}}_j$ is an embedding sampled from a different word category. In this metric, higher CDI values indicate higher word discriminability. We compute CDI for each word category in the training set and take the average over categories to estimate how well the categories are separated in the embedding space of each model type. The result of this analysis is shown in Figure 5.8(a). For each learning objective, we observe that word discriminability is higher in the recurrent encoders compared to their convolutional counterparts. Besides that, the contrastive objective yields encoders with a higher word discriminability index regardless of the architecture type (i.e., recurrent vs. convolutional). Furthermore, we observe a strong positive correlation between average CDI and the performance on the evaluation task—see Figure 5.8(b). This finding indicates that word discrimination performance on future, held-out samples can be predicted based on the CDI computed on the training samples.

5.6.2 Effect of Frequency and Distinctiveness

The proposed CDI in the previous section quantifies the separability and compactness for each word category in the representation space. Next, we aim to identify the factors that could make a word category compact and well-separable.

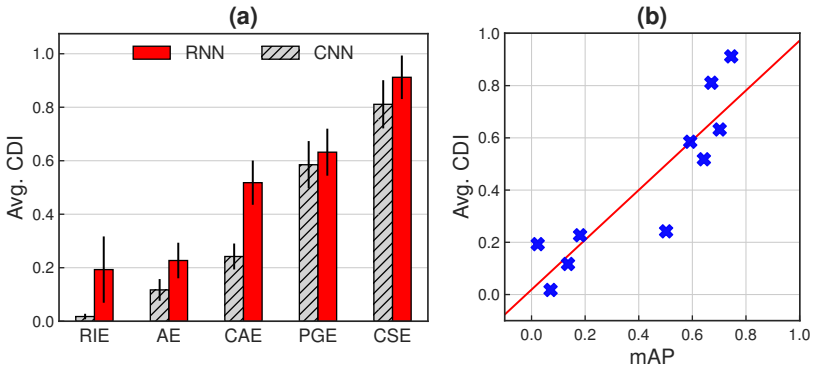


Figure 5.8: Averaged Category Discriminability Index (CDI) for each AWE model with error bars showing standard deviation over word categories. (b) Correlation between the word discrimination performance measured by mAP and averaged CDI (Pearson $r = 0.90, p < 0.001$).

In this analysis, we study the effect of two lexical properties that could be quantified in a data-driven approach: **word frequency** and **acoustic distinctiveness**. Our initial hypothesis is that a word category with many training exemplars becomes more discriminable in the embedding space. This is because the repeated exposure to several within-category samples that exhibit acoustic-phonetic variability should ideally enable the model to learn compact and precise representations for categories with high exemplar frequency. In additions, words that are acoustically distinct have fewer competitors in the perceptual space, thus they should be more separable than words with many phonological neighbours that sound similar. Therefore, we expect word acoustic distinctiveness (WAD) to positively correlate with CDI. In this analysis, we operationalize WAD using two metrics: **word length** (i.e., the number of phonemes) and **phonological distinctiveness**. Word length contributes to WAD since word formation in natural languages is a combinatorial process. That is, increasing the number of phonemes in a word-form decreases the likelihood of encountering a similarly sounding word-form, which makes it less confusable. However, the word formation process is governed by language-specific phonotactic rules which makes some sound combinations more probable than others. To capture the probabilistic nature of sound sequences, we employ **phonological information content** (PIC), an information-theoretic metric that estimates WAD based on its phoneme-to-phoneme transition probabilities (Meylan and Griffiths, 2017). Given a word-form as a sequence of phonemes $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_\tau)$, PIC is defined as

$$\mathrm{PIC}(\boldsymbol{\varphi}) = -\sum_{i=1}^{\tau} \log p_{\boldsymbol{\theta}}(\varphi_i | \boldsymbol{\varphi}_{<i}) \tag{5.7}$$

Learning objective	Architecture	Frequency	Length	PIC
Autoencoder (AE)	CNN	-0.081†	0.315†	0.263†
	RNN	-0.087†	**0.357**†	0.306†
Correspondence Autoencoder (CAE)	CNN	0.021	0.376†	0.274†
	RNN	0.077†	**0.447**†	0.359†
Phonologically Guided Encoder (PGE)	CNN	0.035	0.039*	-0.011
	RNN	-0.043*	**0.325**†	0.263†
Contrastive Siamese Encoder (CSE)	CNN	0.131†	0.075†	0.031
	RNN	0.109†	**0.100**†	0.030

Table 5.1: Pearson correlation (r) between word category discriminability index (CDI) and three lexical properties: frequency, length, and phonological information content (PIC). Statistical significance is marked with * and † for $p < 0.05$ and $p < 0.001$, respectively.

where $p_{\boldsymbol{\theta}}$ is a probabilistic phoneme-level language model (PLM). We estimate $p_{\boldsymbol{\theta}}$ using a trigram PLM with the counts of the phonemes in the training word categories. Higher values of PIC indicate less probable phoneme sequences thus more distinct word-forms. Note that PIC is not length normalized and therefore shorter words tend to have lower PIC.

Next, we conduct a correlation analysis between word CDI and the three predictors: frequency, length, and PIC. The result of this analysis is shown in Table 5.1. Surprisingly, our correlation analysis shows that lexical frequency is a poor predictor of CDI. Although in five out of eight models the frequency positively correlates with CDI, the correlation is rather weak. However, measures of acoustic distinctiveness have a stronger correlation with CDI compared to frequency, and the strength of the correlation is more noticeable in all decoding-based models—except the convolutional PGE—compared to contrastive models. We also find it surprising that PIC is not a better predictor of CDI than word length. However, it has been shown in a prior related work that autoencoder-based AWEs encode duration as an acoustic feature (Matusevych, H. Kamper, et al., 2021). Taken together with our findings, this suggests that the models exploit and rely on acoustic word length as a feature to discriminate between the word categories. Arguably, word length is a more accessible feature to learn from the acoustic signal compared to structural phonological regularities in the training data.

Learning objective	Architecture	mean	max	min	std
Autoencoder (AE)	CNN	0.137	0.141	0.133	0.0026
	RNN	0.183	0.186	0.179	0.0024
Correspondence Autoencoder (CAE)	CNN	0.505	0.510	0.500	0.0040
	RNN	0.646	0.650	0.643	0.0029
Phonologically Guided Encoder (PGE)	CNN	0.595	0.599	0.592	0.0033
	RNN	0.704	0.710	0.687	0.1000
Contrastive Siamese Encoder (CSE)	CNN	0.676	0.680	0.674	0.0023
	RNN	0.742	0.745	0.739	0.0027

Table 5.2: mAP statistics across six different runs for each model type.

5.7 Analysis 3: Network Representational Consistency

Suppose we train two instances of the same architecture and learning objective on the same training samples, but each with a different random initialization. Do these two neural network instances exhibit differences in their representational geometries? In this section, we shed light on the representational discrepancies caused by different initializations. In other words, we are interested in quantifying the degree to which variability in the initial conditions affects the way two models separate the same set of speech samples.

5.7.1 Performance Stability

First, we quantify the effect of the initial weights on the evaluation task performance. To this end, we train six model instances—in identical setup but with different initializations—for each architecture and each learning objective, which yields 48 model instances in total (6×4 RNN runs and 6×4 CNN runs). We evaluate each model instance on the acoustic word discrimination task while observing the result variation per model type. The result of the performance stability analysis is shown in Table 5.2. We observe that all instances have converged and the performance is fairly stable across different runs. Therefore, there are no notable qualitative differences in the performance of models due to variability in the initial conditions of the model in the representation space.

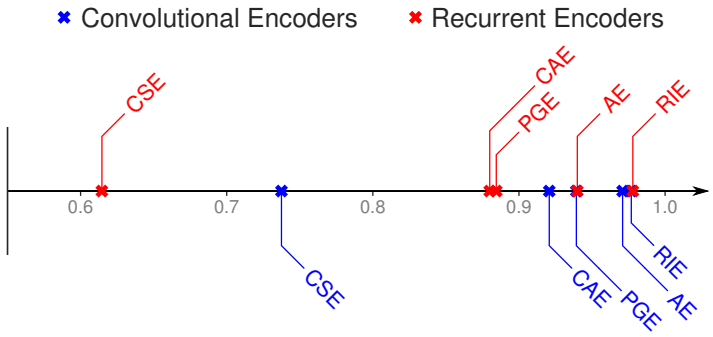


Figure 5.9: Network representational consistency (RC): (top) recurrent encoders and (bottom) convolutional encoders. Values closer to 1 indicates higher RC.

5.7.2 Representational Discrepancies

Our previous performance stability analysis has demonstrated that different DNN instances exhibit only trivial quantitative differences. However, a stable performance on the evaluation task does not necessarily entail an identical representational geometry across different instances. That is, two network instances could have have an identical performance on the evaluation task while each having a distinct representational geometry. To closely investigate representational discrepancies between network instances, we employ the **representational consistency** (RC) analysis (Mehrer et al., 2020), which is a neuroscience-inspired technique based on the representational similarity analysis (RSA) framework (Kriegeskorte et al., 2008). For our analysis, we operationalize the RC using linear Centered Kernal Alignment (CKA) as a representational similarity measure of two views of the same input samples (Kornblith et al., 2019). CKA abstracts away from the embeddings themselves and operates on pairwise distances between the sample representations. Concretely, given K spoken-word samples $\boldsymbol{A}_1^K = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K\}$, we embed the samples using two encoder instances to obtain two different views of the samples $\mathbf{X} \in \mathbb{R}^{K \times D}$ and $\mathbf{Y} \in \mathbb{R}^{K \times D}$. Then, the representational similarity of the two views is computed using CKA as

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\langle \text{vec}(\mathbf{X}), \text{vec}(\mathbf{Y}) \rangle}{||\mathbf{X}||_F ||\mathbf{Y}||_F} \tag{5.8}$$

where the function vec($\mathbf{X}$) reshapes the matrix $\mathbf{X}$ into a vector, $\langle ., . \rangle$ is the inner product, and $||.||_F$ is the Frobenius norm to ensure that CKA $\in [0, 1]$. CKA values close to 1 indicate that the two instances are highly consistent, while values close to 0 indicate low consistency. A more precise mathematical definition of CKA was introduced in the preliminaries chapter.

Using CKA, we conduct pairwise similarity analysis across all six instances which yields 15 comparisons for each model type. We report the mean of the resulting CKA values for each model type in Figure 5.9. First, we observe that randomly initialized encoders (RIE) are highly consistent for both architectures (mean $\text{CKA}_{RIE/RNN} \approx$ mean $\text{CKA}_{RIE/CNN} = 0.98$). However, after training the encoder instances, convolutional networks are more consistent than their recurrent counterparts. Note that this behaviour cannot be attributed to a difference in the number of trainable parameters between the two architectures since they are comparable. Moreover, all decoding-based learning objectives (i.e., AE, CAE, and PGE) return mean CKA values above 0.87, which indicates that their representational profiles are similar despite some noticeable differences especially among the recurrent encoders. The only exception to this trend are model instances trained with contrastive loss since they are significantly less consistent compared to the other learning objectives (mean $\text{CKA}_{CSE/RNN} = 0.61$ and mean $\text{CKA}_{CSE/CNN} = 0.74$). We emphasize that CKA is a second-order isomorphismic approach that operates on the similarity of the pairwise similarity matrices across different views. Therefore, the anisotropic nature of AWEs reported in §5.5 cannot explain their similarity-based representational profiles, and by implication, their representational consistency. Furthermore, and given that different initializations have only resulted in only trivial differences on the evaluation task metric (mAP), we conclude that the network representational consistency cannot be explained by quantitative differences, but rather by representational discrepancies due to disagreement in the geometric arrangement of the speech samples in the embedding space.

5.8 Analysis 4: Qualitative Evaluation

To further inspect the representation space and its neighborhood structure, we conduct a qualitative analysis by querying the representation space with a few word samples. In this analysis, we compute word category centroids by averaging the word embeddings of the training samples, then we use a word centroid as a query and obtain the top-10 ranked nearest neighbors. The result of this analysis is shown in Figure 5.3. For the majority of the examples in Figure 5.3, we observe that there is a strong word onset bias where the most similar words are those that begin with a similar sounding prefix as the query word.

Query (↓)	Convolutional Encoders (CNN)				Recurrent Encoders (RNN)			
	AE	CAE	PGE	CSE	AE	CAE	PGE	CSE
mentioned	mention	mention	mention	mention	mention	mention	mention	mention
	wretched	mansion	mansion	mansion	wretched	mansion	mansion	mansion
	nation	motion	legends	merchant	nation	merchant	merchant	merchant
	midst	merchant	management	mission	merchant	motion	legends	mission
	merchants	making	merchants	mental	motion	nation	mountain	pinching
	motion	wilson	magic	vincent	merchants	making	merchants	massive
	merchant	nation	matrons	pinching	midst	vincent	mission	mental
	message	midst	mission	medicine	milking	nineteen	magician	transient
	regiment	missing	merchant	crouching	winter	nature	motion	motion
	winter	nature	magician	midst	vessel	rachel	wretched	hudson
intellectual	individual	introduction	intellect	intellect	individual	individual	individual	introduction
	interesting	individual	individual	adjoining	interesting	introduction	intelligence	individual
	indifferent	interrupted	introduction	recollection	neglected	uncomfortable	introduction	immature
	newton	indifference	intelligent	delightful	petition	intelligent	intellect	objection
	institution	attraction	encouragement	individual	magician	intelligence	uncomfortable	implacable
	departure	intellect	interrupted	employing	hokosa	interesting	intelligent	delightful
	imitation	immature	intelligence	impetuous	compassion	invisible	interpretation	theatrical
	hokosa	indifferent	indifferently	employed	departure	interrupted	industrial	thoughtful
	encountered	encouragingly	unconditional	natural	convention	imperfectly	incapable	industrial
	neglected	implacable	impetuous	accumulated	consulted	incredible	insensible	election
maker	labor	naked	naked	baker	labor	nature	baker	liquor
	liquor	natured	liquor	naked	nature	local	nature	negro
	labored	nature	natured	negro	walker	naked	liquor	eaten
	labour	local	nature	liquor	local	labour	labour	baker
	wicker	labour	baker	local	naked	labor	labor	nature
	leaping	major	major	native	labour	major	major	labor
	lifted	labor	negro	nature	rachel	natured	negro	naked
	walker	native	native	major	liquor	baker	neighbors	newspaper
	local	making	wicker	matrons	labored	liquor	vapor	mink
	nature	navy	labor	vigor	leaping	negro	labors	vigour
profession	position	procession	procession	professor	position	procession	procession	professor
	proceed	professor	professor	sufficient	professors	possession	proportion	procession
	positions	position	position	procession	possessions	position	perfection	perfection
	physician	possession	petition	professors	proceeded	professor	possession	sufficient
	proceeded	professors	pushing	efficiency	physician	possessions	protection	proposition
	possessed	pushing	professors	efficient	condition	permission	proportions	proportion
	prison	perfection	possession	petition	procession	discussion	position	production
	possessions	positions	physician	prevent	presumption	positions	possessions	petition
	perfect	discussion	positions	position	protested	commission	professor	compassion
	discussion	preferred	precious	physician	proceed	physician	petition	pushing
seized	ceased	ceased	ceased	thieves	ceased	ceased	ceased	thieves
	freedom	season	seizing	ceased	faded	feasts	thieves	ceased
	seated	thieves	season	season	cities	scenes	saves	fuse
	faded	saves	thieves	feast	singing	thieves	seats	jesus
	singing	seems	saves	seizing	scenes	saves	seems	spheres
	scenes	scenes	ceasing	feared	feeding	seems	scenes	feels
	season	ceasing	seems	ceasing	season	feast	seemed	cities
	cities	saints	feast	saves	sweetest	saints	feast	season
	field	feast	seats	species	seated	faced	saved	seats
	seeming	sins	seemed	speed	saying	seemed	seizing	scenes
experiments	experiment	experiment	experiment	experiment	experiment	experiment	experiment	experiment
	experience	experienced	experience	experienced	experienced	experienced	experienced	attendants
	experienced	experience	experienced	garments	experience	experience	experience	extremities
	experiences	experiences	experiences	extermination	extinguished	experiences	experiences	islands
	extinguished	extremities	expense	expense	experiences	exposed	expressions	experienced
	exchange	established	embarrassment	experience	expected	extremities	extermination	prominence
	extremities	extraordinary	expanse	aramis	exchange	expense	extremities	edmunds
	expressions	extinguished	extraordinary	disturbance	expressions	expanse	extremity	instruments
	extremely	extremity	extremities	examined	extremities	extinguished	expression	attendance
	extremity	expanse	expressions	vanished	extent	exclusion	expensive	commons

Table 5.3: Top-10 nearest word embedding centroids for a word sample.

5.9 Discussion of Main Findings

Acoustic word embeddings (AWEs) are vector representations that encode the sound structure and acoustic-phonetic features of spoken words. AWEs are induced from actual acoustic realizations of speech, and therefore AWE models have to abstract away from non-linguistic dimensions of variability in speech signals (e.g., speaker characteristics, speech rate, recording conditions, etc). While analyzing the representational geometry of semantic word embeddings is a topic that has received a substantial attention in the NLP research community, the interpretability of AWEs remains an under-explored topic and we are aware of a few prior studies in this direction (e.g., Matusevych, Herman Kamper, et al., 2020a). In this study, we made a number contributions in analyzing the representational geometry of AWEs and obtained research findings which we discuss and summarize in this section.

Learning objective affects the geometry more than architecture. Our three analyses in this study have shown that the learning objective shapes the representational geometry of the AWE encoders more than their underlying architectures. This finding suggests that recurrent and convolutional encoders may exhibit similar inductive biases while the learning process is mainly guided by the loss function.

AWE models are anisotropic. Our analysis in Section 5.5 demonstrated that AWEs tend towards being minimally isotropic, or anisotropic. This implies that within-category and cross-category word variability are encoded by neural networks in a small subspace, while the majority of dimensions exhibit no significant variance. However, it's important to note that the anisotropic nature of AWE models is not an emergent property of the training process. Instead, we found that isotropy is an inherent property of the underlying neural network architecture itself. Moreover, we found a positive correlation between the degree of isotropy after model training and performance on the acoustic word discrimination evaluation task. Interestingly, different models exhibited varying degrees of isotropy, indicating that the variance is not uniform spoken-word representations from a geometric point of view. As a result, we conclude that making comparisons between different models based solely on absolute distance metrics, such as cosine distance, could lead to inaccurate conclusions.

Word distinctiveness, but not frequency, predicts category discriminability. As found in Section 5.6, word acoustic distinctiveness has been shown to be a good predictor of the extent to which a word category is compact and well-separated in the representation space. However, word frequency does not

correlate with category discriminability. In retrospect, this finding shouldn't be surprising, as frequent words tend to be shorter in length. Shorter words tend to have more phonological neighbors that are perceptually similar in form, thus making them more confusable with other words. Future work can employ linear mixed effects models to analyze the interaction between different lexical properties such as frequency, phonological neighborhood density, and word length, and their impact on word category discriminability.

AWE models exhibit individual differences. Though AWE model instances trained with different random initializations are stable with respect to the performance of the evaluation task, they exhibit individual differences in their representational profiles, as shown in Section 5.7. However, the degree of the network's representational consistency across different initializations depends on both the architecture and the learning objective. Contrastive objectives are less consistent than decoding-based objectives, while recurrent encoders are less consistent than their convolutional counterparts. These findings are relevant when models of AWEs are adopted as cognitive models of human speech processing. For example, it would be important to establish a reasonable upper bound of model similarity when training on one language, before conducting comparisons across models trained on different datasets or languages.

Contrastive models have distinct representational profiles. In the analyses we presented in this study, we observed that the contrastive encoders behave differently than other encoders trained with non-contrastive losses. For example, word distinctiveness has been found to be a weak predictor of category discriminability in the embedding spaces of the contrastive encoders. Recall that our contrastive encoders have a stronger constraint in grouping exemplars of the same category closer in the embedding space guided by the margin hyperparameter, while decoding-based model lack this constraint. We hypothesize that this constraint forces the models to emphasize the separability of the word categories in the embedding space. Therefore, a stronger constraint seems to make contrastive encoders different compared to other learning objectives and different instances of the same contrastive encoder are less consistent in their representational geometry.

5.10 Summary

Computational and conceptual modeling of spoken-word recognition is a well-researched area in cognitive science (Dahan and Magnuson, 2006; Luce and McLennan, 2005; Scharenborg and Boves, 2010; Weber and Scharenborg, 2012,

inter alia). On the other hand, nueral models of spoken-word representations have been independently developed by researcher in speech technology research. However, we still do not know much about how these neural network models encode spoken-word variability in their representational geometry. In this study, we have taken a closer, analytical look at the representational geometry of acoustic word embeddings (AWEs) from three different, but complementary perspectives: (1) representational space uniformity, (2) word category discriminability, and (3) network representational consistency. We have shown that the representational spaces of AWEs tend towards being minimally isotropic, or in other words, they utilize only a few dimensions of the representation space to encode spoken-word variability. Another finding was that most AWE models rely on word length as a feature to discriminate between word categories since the word discriminability index positively correlates with the number of phonemes in a word. Furthermore, our representational consistency analysis have shown that AWE models exhibit individual differences in their representational profiles, with the contrastive encoders being the most inconsistent across different random initializations.

Even though we focused on acoustic word embeddings in this study, our analytic methodology can also be employed for the interpretability of spoken-word representations in self-supervised speech models such as contrastive predictive coding (A. v. d. Oord et al., 2018) and wav2vec (Schneider et al., 2019b). Also, the emergent representations of sub-lexical units such phonemes and syllables in speech neural networks can be analyzed using the our proposed methodology in this study.

6

The Role of Linguistic Experience in Intercomprehension

Closely related languages are often mutually intelligible to various degrees. Cross-linguistic intelligibility is mainly driven by language similarity across different levels of the linguistic hierarchy. Nevertheless, the contribution of lower levels of language processing (i.e., acoustic-phonetic and phonological) to this phenomenon remains unclear. In this paper, we develop a data-driven approach based on computational modeling and the representational similarity analysis framework to quantify the contribution of low-level speech processing to cross-linguistic intelligibility. Our approach quantifies the representational similarity between native and non-native spoken-word representations using acoustic models trained on naturalistic speech data. Therefore, our proposed approach does not require parallel (spoken) word lists which are usually difficult to obtain in a cross-linguistically comparable format. Using our approach, we present a case study on the related Slavic languages and we demonstrate that representational similarity not only captures language similarity in the broad sense, but also predicts the degree of cross-linguistic intelligibility between closely related languages.

6.1 Introduction

Successful comprehension of spoken language requires the processing of the incoming acoustic signal to activate and retrieve the lexical categories intended by the speaker. Despite the highly variable nature of speech, human listeners exhibit a high degree of robustness in recognizing spoken words in their native language (L1), enabling them to resolve inherent ambiguities in speech communication (Luce and McLennan, 2005). However, listening to a non-native (L2) speech is a

completely different experience. Since the human auditory processing system is shaped by the exposure to one's native language, listeners experience perceptual difficulties along dimensions that are not informative for decoding speech in their L1 or when L2 speech deviate from the statistical regularities of their L1 (Pallier et al., 1997). Many of these difficulties can be explained by the misperception of L2 phonological contrasts, both segmental and suprasegmental. For example, L1 Japanese listeners fail to discriminate between English minimal pairs such as *long-wrong* due to the perceptual assimilation of the non-native phonetic categories [ɹ]-[l] onto a single Japanese category [ɾ], thus the two word forms are perceived as homophones (Goto, 1971; MacKain et al., 1981; Miyawaki et al., 1975). Similar auditory processing difficulties have been reported due to inability to recognize suprasegmental contrasts such as variable lexical stress (Peperkamp et al., 2010) and contrastive tones (Y. Wang et al., 1999). However, not all cross-language perceptual difficulties can be explained by misperception of L2 phonological contrasts (Amengual, 2016). For example, L2 Russian learners (native speakers of English) confuse words that are phonological similar (e.g. *molotok-moloko*), despite the absence of unfamiliar L2 phonological contrasts to English listeners (Cook et al., 2016).

On the other hand, the listener's **linguistic experience** can in some cases have a **facilitative** effect on cross-language speech processing. For example, speakers of closely-related languages can comprehend each other's speech to an extent that enables a form of communication known in the sociolinguistics literature as **intercomprehension** or **receptive multilingualism** (refer to Van Heuven (2008), Gooskens (2017), and Gooskens (2019) for an overview). Even in the absence of prior familiarity with each other languages, **L1/L2 structural similarity** across the different levels of the linguistic hierarchy enables interlocutors to decode the incoming L2 speech using their L1 competence. The factors that contribute to intercomprehension and mutual intelligibility can be categorized as either linguistic (e.g., inherent cross-linguistic similarities) or extra-linguistic (e.g., listener's attitude). However, isolating the effects of linguistic versus extra-linguistic factors remains a challenge in experimental studies involving human participants..

Recently, acoustic models based on deep neural networks (DNNs) have been analyzed from a cognitively motivated angle to investigate the degree to which they exhibit human-like behavior on a variety of speech processing tasks (e.g., Matusevych, H. Kamper, et al., 2021; Matusevych, Herman Kamper, et al., 2020a; Millet et al., 2021; Schatz and Feldman, 2018). Furthermore, it has been shown that neural networks reflect human perception of cross-linguistic and dialectal

variation of spoken language (Bartelds, de Vries, et al., 2022; Bartelds and Wieling, 2022). In this study, we build on this line of research and investigate a class of acoustic models that project a spoken-word stimulus of an arbitrary length onto a fixed-dimensionality representation (e.g., Herman Kamper, W. Wang, et al., 2016; Levin et al., 2013; Settle and Livescu, 2016b). In speech technology research, these representations are known as acoustic word embeddings (AWEs) and their underlying acoustic models are employed in voice-based applications such as as query-by-example spoken term discovery (Jansen and Durme, 2012; Metze et al., 2013; Yaodong Zhang and James R Glass, 2009) and keyword spotting (Myers et al., 1980; Rohlicek, 1995). AWE models are trained in a way such that different acoustic exemplars of the same word category, or word type, are ideally projected onto the same point in the representation space. From the cognitive perspective, these models simulate auditory-lexical processing during language comprehension.

In this chapter, we present a computational framework to study the impact of linguistic factors on mutual intelligibility. Our objective is to study how lower-levels of language processing (i.e., acoustic-phonetic processing and phonological decoding) contribute to the facilitating effect of language similarity on cross-language speech processing. Concretely, our study makes the following contributions:

1. We develop a data-driven framework based on representation similarity analysis (RSA) to study the role of linguistic experience in non-native spoken-word representations (§ 6.3). Our framework is visually illustrated in Figure 6.1.

2. Using our framework, we present a case study on the related Slavic languages and demonstrate that cross-lingual representational similarity not only predicts language similarity in the broad sense, but also the degree of mutual intelligibility among related langauges (§6.6).

3. We conduct a qualitative analysis of the model representations and shed light on the sources of cross-linguistic differences among the models (§6.7 and §6.8).

6.2 Background

In this section, we discuss some context that situates our work within the broader literature. In section 6.2.1, we discuss the relevant sociolinguistic research on the empirical testing of mutual intelligibility among closely related languages. In section 6.2.2, we highlight research efforts on modeling human speech processing

using deep neural networks. Finally, in Section 6.2.3, we briefly discuss the neuroscientific framework of representation similarity analysis and its application in interpreting and analyzing neural networks in natural language processing (NLP) and automatic speech recognition (ASR).

6.2.1 Cross-Linguistic Intelligibility

In speech communication, the term *intelligibility* refers to the quality of information decoding at the listener's side when the speech is transmitted under adverse conditions (e.g. noise) or when it deviates from the norm (e.g., foreign-accented speech). Even though speech is highly variable, human listeners are usually able to reliably understand utterances in adverse conditions, and in some cases, comprehend what is being communicated in a different, but related language. A host of studies in the sociolinguistics literature reported varying degrees of **mutual intelligibility** between related languages across different language families and dialect continua including Romance, Germanic, and Slavic languages (e.g., Golubović and Gooskens, 2015; Gooskens and Heuven, 2017; Gooskens, Heuven, et al., 2018). It has been observed that the objective measures of linguistic distance, such as lexical distance, are strong predictors of cross-linguistic intelligibility (Gooskens, 2007).

In this study, we focus on the cross-linguistic intelligibility of Slavic languages and propose a computational, analytic approach to analyze the contributions of lower levels of language processing to mutual intelligibility. Despite their intriguing linguistic features and remarkable similarities, Slavic languages remain understudied within the NLP and speech processing literature. We ground our research on findings from the sociolinguistics literature regarding the effect linguistic distance on mutual intelligibility. For example, Golubovic (2016) shown that languages in a geographically connected area that are members of the same sub-family (e.g., West Slavic languages such as Czech and Polish) exhibit a higher cross-linguistic intelligibility compared to languages that belong to different sub-families (e.g., Bulgarian, a South Slavic language, and Czech). In their study, Golubovic (2016) used a **spoken-word translation** task to test mutual intelligibility among speakers of different South Slavic and West Slavic languages. Human participants in this task listened to L2 spoken-word stimuli and were asked to translate what they heard into their native language. When translating Czech word stimuli, Polish-speaking participants performed better on average (54.3%) compared to Croatian-speaking (43.0%) and Bulgarian-speaking (41.6%) participants. Similarly, when translating Polish word stimuli, Czech-speaking participants performed better on average

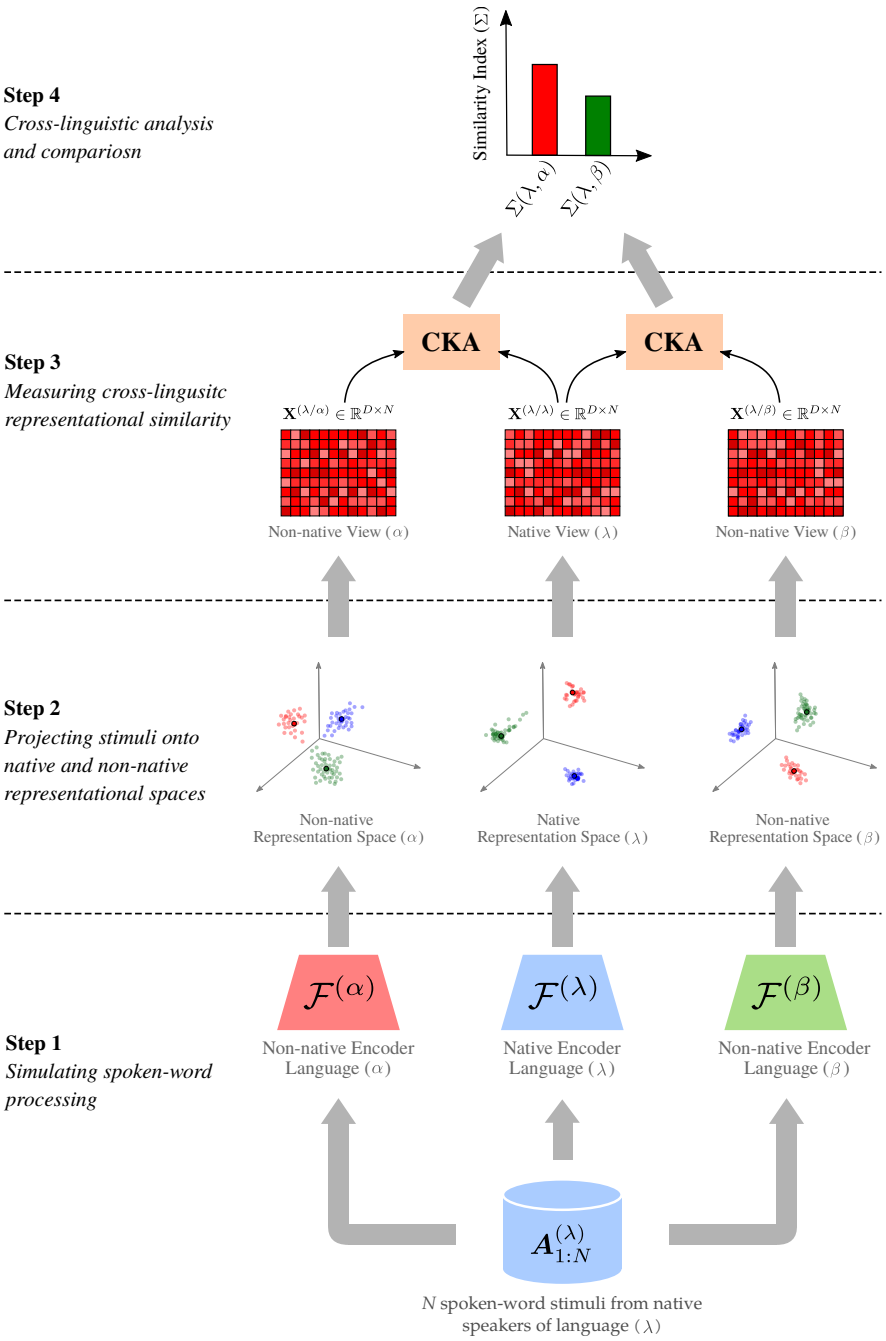


Figure 6.1: A schematic view of our experimental pipeline whereby we quantify the extent to which non-native models produce native-like representations using representational similarity analysis. A set of N spoken-word stimuli from language λ are represented using the encoder $\mathcal{F}^{(\lambda)}$ which was trained on language λ to obtain a *native view* of the data: $\mathbf{X}^{(\lambda/\lambda)} \in \mathbb{R}^{D \times N}$. Simultaneously, the same stimuli are represented using encoders trained on other languages, namely $\mathcal{F}^{(\alpha)}$ and $\mathcal{F}^{(\beta)}$, to obtain two different *non-native views* of the data: $\mathbf{X}^{(\lambda/\alpha)}$ and $\mathbf{X}^{(\lambda/\beta)}$.

(63.2%) compared to Croatian-speaking (37.5%) and Bulgarian-speaking (43.3%) participants. These findings are not surprising given that Slavic languages share a large number of cognate words—words with a common etymological origin that the same meaning and similar form. Nevertheless, the contributions of phonetic perception and phonological decoding to cross-language spoken-word recognition are difficult to isolate in experimental studies with human participants. To address this challenge, our study makes use of computational models that have access solely to the acoustic instances of word-forms, but not their semantic content.

6.2.2 Neural Networks as Models of Human Speech Processing

Recent developments in representation learning have made it possible to develop computational models that simulate language learning from raw, continuous speech input. Working with speech representation learning as a cognitive framework enables researchers to address questions about language acquisition and speech processing without making strong assumptions about the (emergent) categorical nature of speech (Alishahi et al., 2017b; Dupoux, 2018; Gelderloos et al., 2020; Magnuson et al., 2020; Matusevych, Schatz, et al., 2020; Räsänen et al., 2016a; Scharenborg, Gouw, et al., 2019). Complementing this line of research, DNN-based models for automatic speech recognition (ASR) have been adopted as cognitive models of speech processing in several prior studies, with a strong focus on modeling cross-linguistic effects. For example, Schatz and Feldman (2018) have shown that an ASR predicts cross-linguistic perceptual effects due to misperception of non-native phonological contrasts—e.g., L1 Japanese listeners' difficulty with the English [l]-[ɹ] contrast. Matusevych, H. Kamper, et al. (2021) have shown that lexically-constrained acoustic models predict non-native lexical processing difficulties that cannot be explained by phonetic categorization. These studies demonstrate the value of computational modeling to shed light on an important question: how does linguistic experience shape our speech processing system?

6.2.3 Representational Similarity Analysis

Representational Similarity Analysis (RSA) is a data-analytical framework developed in the neuroscience community to enable comparison of neural activity patterns across brain regions and computational models of information processing (Kriegeskorte et al., 2008). The RSA framework abstracts away from the activity patterns themselves and operates on the geometry of the representation or feature space, which makes it applicable for interpretability and analysis of neural

networks where the correspondence between neurons across different layers or architectures is unknown. In the NLP and speech processing community, RSA has previously been employed to study the correlation between neural network and symbolic representations of language (Chrupała and Alishahi, 2019), analyze word representations in language models (Abdou et al., 2019; Abnar et al., 2019; Beinborn and Choenni, 2020; Lepori and McCoy, 2020; J. Wu et al., 2020), and analyze the representations of speech recognition models (Chrupała, Higy, et al., 2020; Chung, Belinkov, et al., 2021). In this study, we build on prior work and employ the RSA framework to quantify the impact of linguistic experience, characterized by the language of exposure, on non-native spoken-word processing. We do so by building spoken-word representation models based on deep neural networks in a controlled setting where we keep all dimensions of variability fixed except the language of exposure (i.e., language of the training samples) to examine cross-linguistic effects. To the best of our knowledge, our study is the first to employ the RSA framework to analyze the impact of linguistic experience on the degree to which non-native speech models exhibit native-like representations.

6.3 Research Methodology

A neural spoken-word representation model can be formally described as a mapping, or an encoder function, $\mathcal{F} : \mathcal{A} \rightarrow \mathbb{R}^D$, where $\mathcal{A}$ is the (continuous) space of acoustic sequences and D is the dimensionality of the representation space. Given an acoustic word signal represented as a temporal sequence of T acoustic events $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T)$, where $\boldsymbol{a}_t \in \mathbb{R}^k$ is a spectral vector of k coefficients, a word representation is computed as

$$\mathbf{x} = \mathcal{F}(\boldsymbol{A}; \boldsymbol{\theta}) \in \mathbb{R}^D \tag{6.1}$$

Here, $\boldsymbol{\theta}$ are the parameters of the encoder, which are learned by training the model in a monolingual supervised setting. That is, the training word segments are speech intervals that are sampled from utterances produced by native speakers where the word identity of each acoustic segment is known. To encourage the model to abstract away from speaker variability, the training samples are obtained from multiple speakers, while the resulting representations are evaluated on a held-out set of speakers.

Step 1: Simulating spoken-word processing

Since our goal is to characterize how the linguistic experience shapes speech processing, the first step in our analytic approach is to simulate spoken-word processing in (adult) listeners who have been exposed to a single language. To achieve this goal, we train monolingual, word-level acoustic models on different languages where the training data and conditions are cross-linguistically comparable with respect to the number of word stimuli, genre, and speaker variability. We therefore have access to several models $\{\mathcal{F}^{(\alpha)}, \mathcal{F}^{(\beta)}, \ldots, , \mathcal{F}^{(\omega)}\}$, where the superscripts $\{\alpha, \beta, \ldots, \omega\}$ denote the language of the training samples. In this study, we focus on models that are based on unidirectional recurrent neural networks because they are better at capturing the sequential incremental nature of speech in comparison to other architectures (i.e., convolutional and transformer networks).

Step 2: Projecting word stimuli onto native vs. non-native representation spaces

Next, we obtain a set of experimental conditions in the form of N held-out spoken-word stimuli produced by native speakers of language λ: $\boldsymbol{A}^{(\lambda)}_{1:N} = \{\boldsymbol{A}^{(\lambda)}_1, \ldots, \boldsymbol{A}^{(\lambda)}_N\}$, where $\boldsymbol{A}_i$ is a spectral representation of a spoken-word instance. Then, each acoustic word stimulus in this set is mapped onto a representation using the model $\mathcal{F}^{(\lambda)}$, which yields a matrix $\mathbf{X}^{(\lambda/\lambda)} \in \mathbb{R}^{D\times N}$. Since the model $\mathcal{F}^{(\lambda)}$ was trained on language λ, we refer to it as the *native model* and consider the matrix $\mathbf{X}^{(\lambda/\lambda)}$ as the *native view* of the stimuli. To obtain a *non-native view* of the same experimental conditions, we project each of sample in the stimuli $\boldsymbol{A}^{(\lambda)}_{1:N}$ onto the representation space of the model $\mathcal{F}^{(\alpha)}$, which was trained on a different language α. Therefore, we have a second matrix $\mathbf{X}^{(\lambda/\alpha)} \in \mathbb{R}^{D\times N}$ for the non-native representations. Here, we read the superscript notation (λ/α) as word stimuli of language λ represented by a model trained on language α.

Step 3: Measuring representational similarity

In this step, we aim to measure the degree to which the non-native model produces native-like representations. This goal can be achieved by quantifying the structural correspondence and alignment between the representational spaces of the models $\mathcal{F}^{(\lambda)}$ and $\mathcal{F}^{(\alpha)}$ in response to the stimuli $\boldsymbol{A}^{(\lambda)}_{1:N}$. To this end, we measure the representation similarity between the two matrices $\mathbf{X}^{(\lambda/\lambda)}$ and $\mathbf{X}^{(\lambda/\alpha)}$ using linear Centered Kernel Alignment (CKA) as

$$\Sigma(\lambda, \alpha) := \mathsf{CKA}(\mathbf{X}^{(\lambda/\lambda)}, \mathbf{X}^{(\lambda/\alpha)}) \tag{6.2}$$

Here, $\Sigma(\lambda, \alpha) \in [0, 1]$ is a scalar that quantifies the agreement between the responses of the two models, i.e., native $\mathcal{F}^{(\lambda)}$ and non-native $\mathcal{F}^{(\alpha)}$, when tested with spoken-word stimuli $\boldsymbol{A}^{(\lambda)}_{1:N}$. If the two models separate the experimental conditions with a similar geometry in their representational spaces, Σ will be close to 1. On the other hand, values close to 0 indicate different representational geometries and a difficulty in establishing an alignment. Note that CKA is a neuroscience-inspired measure that emphasizes the distributivity of information in neural activity patterns and abstracts away from the roles of individual neurons. In addition, CKA is invariant to orthogonal transformation and isotropic scaling, which are desirable properties for our analysis.

Step 4: Analyzing cross-linguistic differences

To analyze cross-linguistic differences in spoken-word processing and representation, we obtain another non-native view $\mathbf{X}^{(\lambda/\beta)}$ of the stimuli $\boldsymbol{A}^{(\lambda)}_{1:N}$ using the non-native model $\mathcal{F}^{(\beta)}$. We then quantify the representational similarity between the models $\mathcal{F}^{(\lambda)}$ and $\mathcal{F}^{(\beta)}$ as

$$\Sigma(\lambda, \beta) := \mathsf{CKA}(\mathbf{X}^{(\lambda/\lambda)}, \mathbf{X}^{(\lambda/\beta)}) \tag{6.3}$$

If $\Sigma(\lambda, \alpha) > \Sigma(\lambda, \beta)$, we can conclude that the representations of the model $\mathcal{F}^{(\alpha)}$ are more native-like compared to the representations of $\mathcal{F}^{(\beta)}$. Note that while $\mathsf{CKA}(.,.)$ is a symmetric metric—i.e., $\mathsf{CKA}(\mathbf{X}, \mathbf{Y}) = \mathsf{CKA}(\mathbf{Y}, \mathbf{X})$—our established similarity metric $\Sigma(.,.)$ is not symmetric—i.e., $\Sigma(\lambda, \alpha) :\neq \Sigma(\alpha, \lambda)$. To compute $\Sigma(\alpha, \lambda)$, we use word stimuli of language α and collect the matrices $\mathbf{X}^{(\alpha/\alpha)}$ and $\mathbf{X}^{(\alpha/\lambda)}$. Then we compute

$$\Sigma(\alpha, \lambda) := \mathsf{CKA}(\mathbf{X}^{(\alpha/\alpha)}, \mathbf{X}^{(\alpha/\lambda)}) \tag{6.4}$$

When we apply our proposed experimental pipeline across L different languages, the effect of language similarity can be characterized by constructing a cross-lingual representational similarity matrix (xRSM) which is an asymmetric $L \times L$ matrix where each cell represents the structural correspondence between two representational spaces.

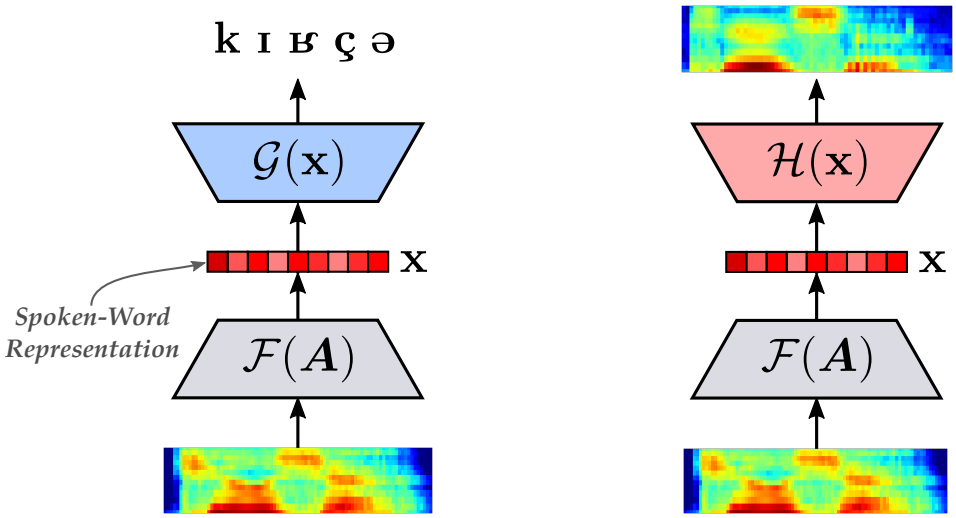


Figure 6.2: A visual illustration of the models in our study: (left) Phonologically Guided Encoder (PGE) (left) and (right) Correspondence Autoencoder (CAE).

6.4 Spoken-Word Representation Models

In this section, we describe the two spoken-word representation models that we employ for our study. Both models are supervised, utilizing top-down lexical constraints to direct the learning process. Training the models requires a dataset $\mathcal{D} = \{(\boldsymbol{A}_i, w_i)\}_{i=0}^{M}$ of M spoken word instances where w_i is the category (or word type) of ith instance and $\boldsymbol{A}_i$ is an exemplar of this word category. Note that the two models, presented in the following two subsections, were introduced in greater depth in the previous chapter, and briefly presented here to contextualize their application in the current study.

6.4.1 Phonologically Guided Encoder

The first model we experiment with is the phonologically guided encoder (PGE), which is a sequence-to-sequence model that is trained with explicit phonological supervision. Given an acoustic word sequence $\boldsymbol{A}$ and its corresponding phonemic transcription $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_\tau)$, a recurrent encoder $\mathcal{F}$ takes $\boldsymbol{A}$ as input and produces a single representation $\mathbf{x}$ (i.e., the last hidden state of the recurrent cell). Then, the phonological decoder $\mathcal{G}$ aims to decode $\boldsymbol{\varphi}$ from $\mathbf{x}$. The objective is to minimize a categorical cross-entropy loss at each timestep in the decoder, which is equivalent to

$$\mathcal{L} = -\sum_{i=1}^{\tau} \log p(\varphi_i | \boldsymbol{\varphi}_{<i}, \mathbf{x}) \tag{6.5}$$

where p is the probability of the phone φ_i at timestep i, conditioned on the previous phone sequence $\boldsymbol{\varphi}_{<i}$ and the embedding $\mathbf{x}$. The intuition of this learning objective is the following: although their acoustic realizations might vary due to speaker and context variability, different exemplars of the same word category have identical phonemic transcriptions. Therefore, the model is expected to project exemplars of the same category nearby in the representation space and the representational distance should ideally reflect phonological (dis)similarity.

6.4.2 Correspondence Autoencoder

The second model we experiment with is the correspondence autoencoder (CAE), which is also sequence-to-sequence model with an identical encoder architecture to the PGE model (Herman Kamper, 2019). However, the CAE differs from the PGE in the nature of the supervision signal. To train the CAE, two acoustic word exemplars of the same category are paired to make a tuple $(\boldsymbol{A}, \boldsymbol{A}^+)$. Then, the encoder $\mathcal{F}$ takes $\boldsymbol{A}$ as input and builds up a representation $\mathbf{x}$, which is then fed into an acoustic decoder $\mathcal{H}$ that aims to (sequentially) reconstruct the corresponding acoustic sequence $\boldsymbol{A}^+$. The objective is to minimize the L_2 distance at each timestep in the decoder, which is equivalent to

$$\mathcal{L} = \sum_{t=1}^{T^+} ||\boldsymbol{A}_t^+ - \mathcal{H}(\mathbf{x})_t||_2 \tag{6.6}$$

where $\boldsymbol{A}_t^+$ is the ground-truth spectral vector at timestep t and $\mathcal{H}(\mathbf{x})_t$ is the reconstructed spectral vector at timestep i as a function of the computed representation $\mathbf{x}$.

Despite the difference in the learning objectives between the PGE and CAE, both models implement a memory bottleneck since the decoders have access only to the last hidden state of the encoder without an attention mechanism. Therefore, both models are encouraged to learn high-level abstractions of the detailed acoustic input in their bottleneck representations. We hypothesize that this constraint also encourages the models to uncover phonological regularities in the speech data in order to efficiently compress the acoustic input into a single bottleneck representation.

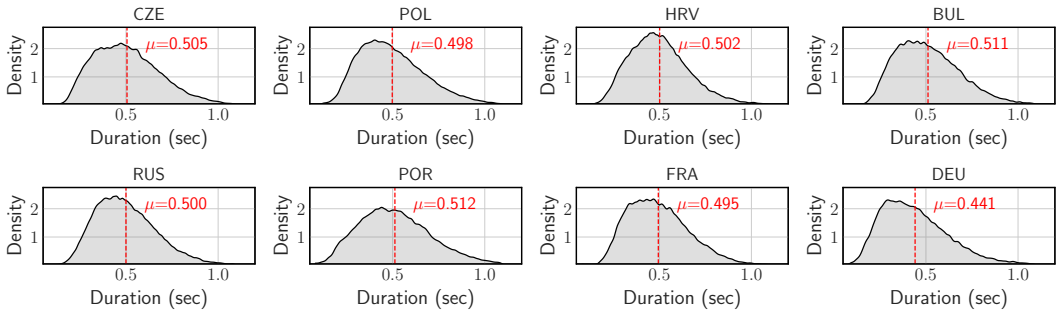


Figure 6.3: Word duration distributions across the languages in our study.

6.5 Data and Experimental Setup

6.5.1 Experimental Data

The data in our study is a subset of the GlobalPhone speech (GPS) database (Schultz et al., 2013) which is a multilingual read speech resource containing utterances recorded by native speakers (who self-reported their gender in the metadata) in a controlled recording environment with minimal noise. Therefore, the recording conditions are comparable across languages which enables us to conduct a cross-linguistic comparison. To access the data, we obtained the license for the languages in our study from the copyright holder, XLingual LLC. We experiment with five Slavic languages: Czech (CZE), Polish (POL), Russian (RUS), Bulgarian (BUL), and Croatian (HRV), and additionally with three Indo-European languages outside the Slavic language family as control languages: Brazilian Portuguese (POR), French (FRA), and German (DEU). To train the models, a set of 42 speakers of balanced gender was sampled from each language to make a training dataset consisting of roughly the same number of training samples for each language (˜28k spoken-word instances each). Spoken-word alignments were produced using the Montreal forced aligner (McAuliffe et al., 2017). Each acoustic word signal is parametrized as a sequence of 39-dimensional Mel-frequency spectral coefficients where frames are extracted over intervals of 25ms with 15ms overlap. đigure 6.3 shows word duration distributions across for each language.

6.5.2 Architecture and Hyperparameters

Acoustic Encoder $\mathcal{F}(.;\boldsymbol{\theta}_{\mathcal{F}})$. We employ a 3-layer recurrent neural network with a unidirectional Gated Recurrent Unit (BGRU) of hidden state dimension of 512,

which yields a 512-dimensional representation. The encoder is identical in both the PGE and CAE model.

PGE–Phonological Decoder $\mathcal{G}(.;\boldsymbol{\theta}_{\mathcal{G}})$. We employ a 1-layer GRU of 512 units hidden state that takes the 512-dimensional bottleneck representation from the encoder as the initial hidden state and decodes the corresponding phonological sequence without teacher forcing.

CAE–Acoustic Decoder $\mathcal{H}(.;\boldsymbol{\theta}_{\mathcal{H}})$. Similar to the phonological decoder, but it decodes the corresponding acoustic exemplar thus the inputs and outputs of this decoder are 39-dimensional vectors, which is the dimensionality of the spectral dimension of our acoustic stimuli.

Training Details. All models in this study are trained for 100 epochs with a batch size of 256 using the ADAM optimizer (Kingma and Ba, 2015) and an initial learning rate (LR) of 0.001.

Implementation. Our models are developed using PyTorch (Paszke et al., 2019) and FAISS (J. Johnson et al., 2017) for efficient similarity search during evaluation. Our code, data statistics, configuration files of hyperparameters, and speaker splits will be publicly available on a public GitHub repository upon publication.

6.5.3 Quantitative Evaluation

To keep track of the learning progress during the models' training, we monitor the performance of the models at each epoch on the exemplar retrieval task. In the speech technology literature, this task is also known as the same-different acoustic word discrimination task and it mainly assesses the model's ability to separate distinct lexical categories and discriminate between word exemplars of different categories, which is quantified using the mean averaged precision metric (mAP). The task can be verbally described as follows: given each acoustic word instance in the evaluation set as a query, the goal is to retrieve all instances that are exemplars of the same category as the query word. The similarity search takes places in the model representation space with cosine distance as the ranking criterion. The word discrimination task and the mAP metric were introduced in greater detail in the previous chapter.

Figure 6.4 shows the results of the quantitative evaluation using the mAP metric during training. It can be observed that the PGE model converges faster than its CAE counterpart and requires fewer epochs to reach a plateau. This behavior is expected if we consider the objective function of the two models since the supervision signal is stronger in the PGE (i.e., deterministic phonological target).

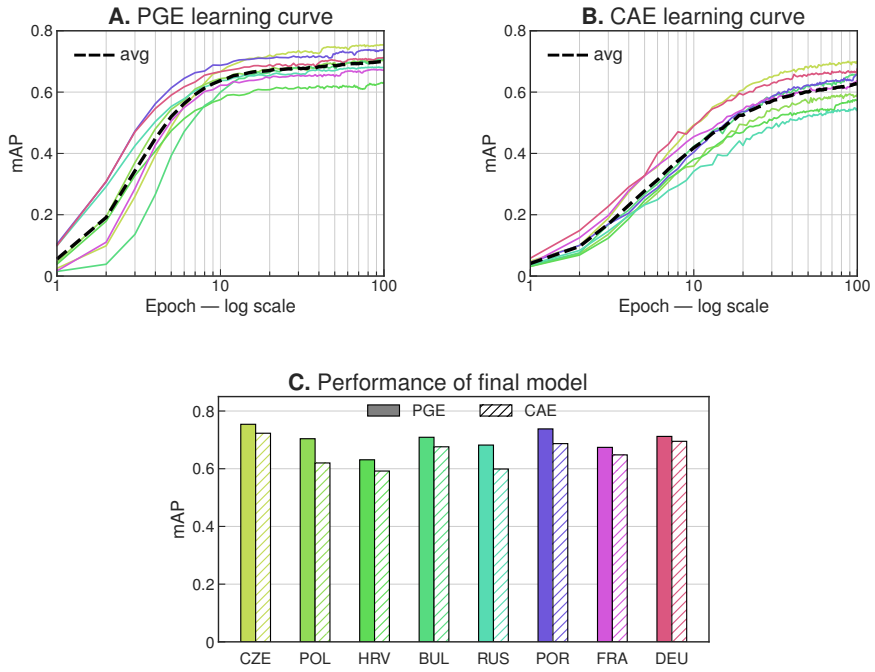


Figure 6.4: (A-B) Learning curves of the PGE model (A) amd CAE model (B) during the first 100 epochs of training, quantified by the exemplar retrieval performance (measured by mAP) on the validation set for each language. The black dashed line is the mean across languages at each epoch. (C) Performance of the converged model measured by mAP.

6.6 Similarity Analysis

6.6.1 Quantifying Within-Language Variation

To investigate the effect of non-language factors of variation, we first quantify the model representational consistency while varying two training conditions: (1) initial weights, and (2) speaker sample. Our aim is to obtain a reasonable upper bound on how (dis)similar two (native) models would be if they vary along a dimension that is not the language of exposure. To this end, we train 10 different model instances for Czech and Polish for each learning objective (PGE vs. CAE) and for each training condition (different seeds, same speakers vs. same seed, different speakers), which yield 80 different instances in total. Then, we measure the representation similarity between native models across each varying training condition. The results of this analysis are shown in Fig 6.5. We observe that both factors of variability affect the (within-language) representational similarity. However, model variation due to exposure to different speakers during training

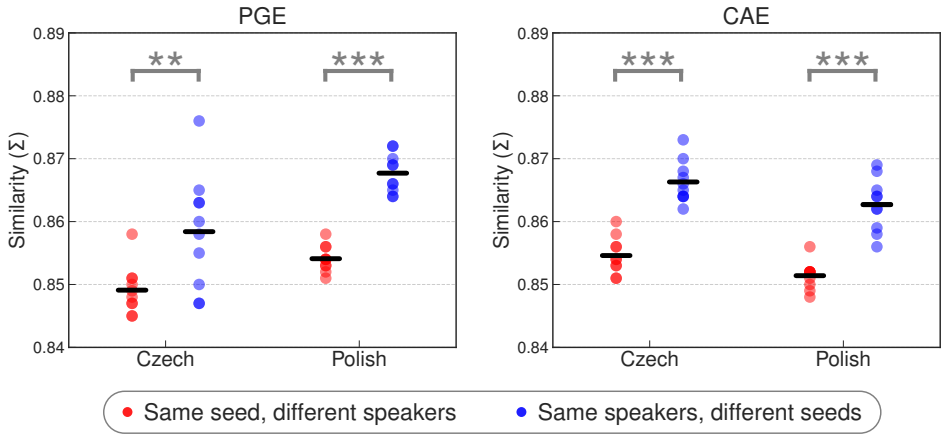


Figure 6.5: Impact of initial weights (●) and speaker sample (●) on the representational similarity of models trained on the same language. Statistical significance is marked with ** and *** for $p < 0.01$ and $p < 0.001$, respectively.

has a greater impact on the extent to which two native models agree on the arrangement of the test samples.

6.6.2 Cross-Linguistic Similarity

For our cross-linguistic RSA experiments, we focus on the similarity of the non-native models to native West Slavic models, namely Czech and Polish. The motivation to focus on these two languages stems from the fact that both languages are within the West Slavic dialect continuum—a geographically connected linguistic area with closely-related languages that exhibit high cross-linguistic intelligibility. Figure 6.7 shows the results of this analysis where each data point corresponds to the similarity index between one of the native models from the previous section (i.e., speaker variability experiment) and a non-native model. We observe that the extent to which both the PGE and CAE models exhibit a native-like representation is largely conditioned on the language of exposure during training. That is, the non-native representations of Czech and Polish test stimuli are more native-like in the Slavic models compared to the non-Slavic models. Furthermore, both PGE and CAE predict the West Slavic advantage since the Polish view of the Czech stimuli is the most native-like and vice versa. In the PGE view of the Czech stimuli, mean similarity scores to the native model are $\mu_{\text{Polish}} = 0.761 > \mu_{\text{Slavic}} = 0.749 > \mu_{\text{non-Slavic}} = 0.702$, while for Polish stimuli $\mu_{\text{Czech}} = 0.743 > \mu_{\text{Slavic}} = 0.733 > \mu_{\text{non-Slavic}} = 0.667$. A similar trend can be observed in the CAE model as it predicts both the Slavic advantage as well as the West Slavic advantage when processing Czech and Polish word stimuli.

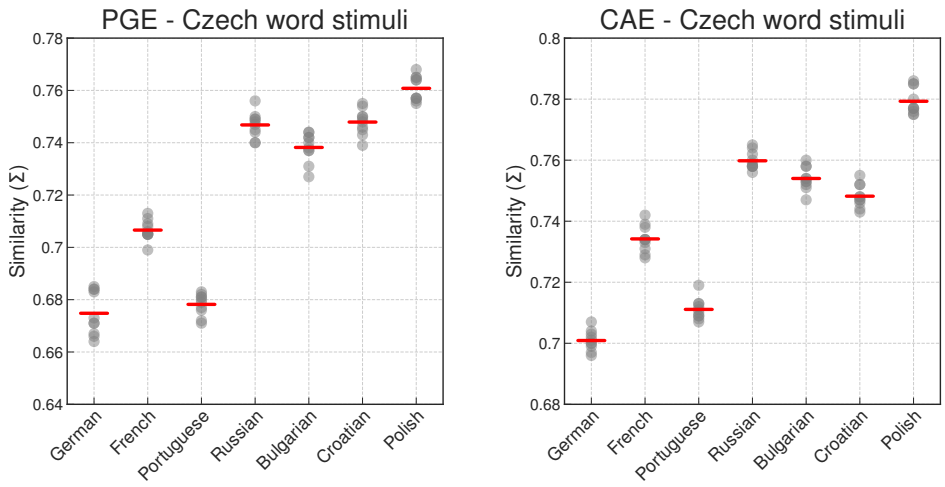


Figure 6.6: Cross-linguistic RSA of the Czech stimuli which quantifies the extent to which L2 models produce native-like (here, Czech) representations. Each point in the figure corresponds to one comparison between a native Czech model and an L2 model. Red dashes represent the means over 10 native models.

It is worth pointing out that the PGE and CAE models also predict the L1 advantage, since all cross-linguistic similarity scores in Figure 6.7 are significantly below those reported in the within-language variation experiments in Figure 6.5. Also, similarity scores between native models and an untrained model are always below 0.180 for both encoders, which demonstrates that similarities can only be attributed to the linguistic experience and not to architectural inductive biases.

6.6.3 Clustering Analysis

To get further insights into the impact of the linguistic experience on representational similarity, we constructed a cross-lingual representational similarity matrix (xRSM) using the similarity scores (8×8 matrix, where 8 is the number of languages in our study). The construct matrix for each model (PGE vs. CAE) is illustrated as a heatmap in Figure 6.8, where warmer colors indicate higher representational similarity. We then applied hierarchical clustering with the Ward algorithm (Ward, 1963) on each matrix, which is also illustrated in the dendrograms in Figure 6.8. One can observe that the Slavic languages form a single cluster in both dendrograms. Likewise, the Romance languages in our study, French and Portuguese, are grouped together in each dendrogram. This grouping demonstrates that the representational geometries of two models exhibit higher agreement if their training languages are structurally similar

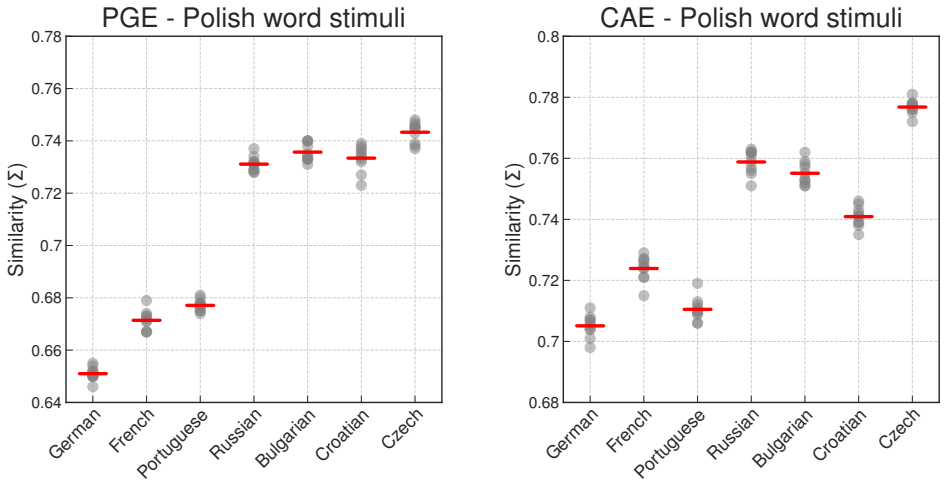


Figure 6.7: Cross-linguistic RSA of the Polish stimuli which quantifies the extent to which L2 models produce native-like (here, Polish) representations. Each point in the figure corresponds to one comparison between a native Polish model and an L2 model. Red dashes represent the means over 10 native models.

However, the internal grouping of the Slavic languages varies between the two models. While both models are consistent with the grouping of West Slavic languages (i.e., Czech and Polish) into one sub-cluster, the grouping of South Slavic (i.e., Croatian and Bulgarian) languages differs. The PGE clustering analysis groups Croatian and Bulgarian together, while the CAE groups Russian and Bulgarian first. In addition, the CAE clustering analysis shows that Russian and Bulgarian is the most similar language pair, even though Russian belongs to the East Slavic branch. Note that the key difference between the two models is the nature of the supervision signal, and this difference could explain the inconsistency between the two models regarding the grouping of South Slavic languages. The supervision

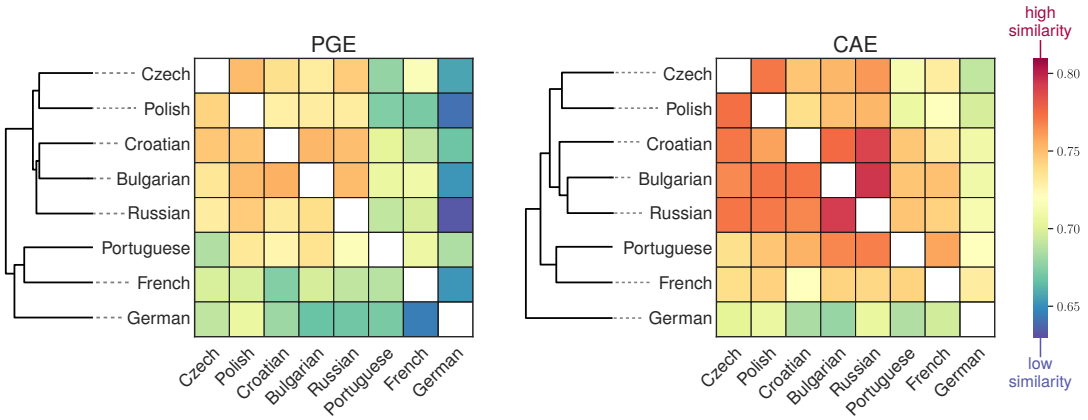


Figure 6.8: Cross-lingual representational similarity matrix of the two models: PGE (left) and CAE (right).

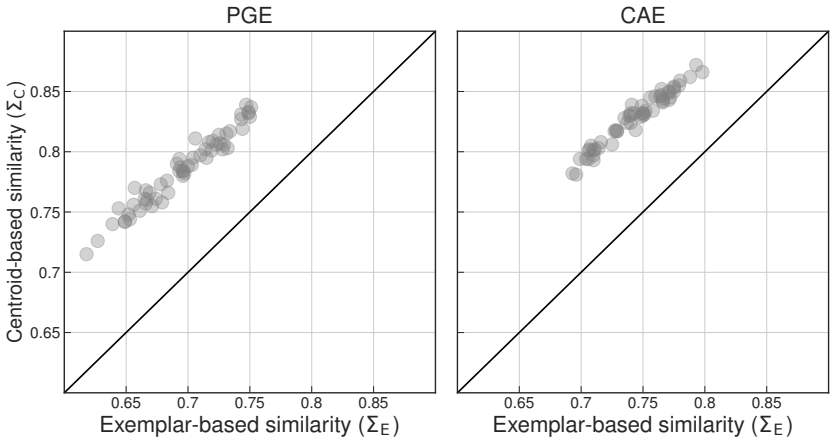


Figure 6.9: Exemplar-based RSA (x-axis) vs. centroid-based RSA (y-axis) of the two models.

signal in the PGE is symbolic in the form of discrete phonological sequences, thus it is more likely to preserve historical relatedness and reflect the phylogenetic signal. In contrary, the CAE does not have access to symbolic units during training since the supervision signal is purely acoustic based on correspondence learning between exemplars of the same word category. Therefore, the CAE is more likely to be sensitive to contemporary phonetic and word-internal prosodic similarities that are encoded in the acoustic signals. As an example, Bulgarian has variable (or free) word stress and strong phonetic reduction patterns, which are common features in Russian. Due to their variable word stress, Bulgarian and Russian are characterized by a stronger contrast between stressed and unstressed syllables, and vowels in unstressed syllables undergo a process that is known as vowel-quality alternation, or lexicalized vowel reduction (Barry and Andreeva, 2001). This finding suggests that the CAE model is more sensitive to fine-grained details in the acoustic realizations of vocalic segments. We suggest further in-depth studies analyzing the differences of vowel representations between the two models as future work.

6.7 Exemplar vs. Centroid Similarity

One of our key findings from the previous section is that L1/L2 representational similarity in the acoustic models not only predicts language similarity in the broad sense but also the degree of mutual intelligibility within a set of closely-related languages. In this section, we investigate potential factors that may contribute to the cross-linguistic differences among the models. Both models we consider in this study (PGE and CAE) are trained with decoding objectives that do not explicitly

aim to separate word categories in the representation space. That is, there are no strong constraints on the arrangement of categories and category exemplars in the high-dimensional representation space. Therefore, differences between native and non-native representations can be driven by two factors: (1) differences in the arrangement of individual exemplars within their respective word category cluster, and/or (2) differences in the placement of overall word category clusters in the representation space.

To investigate the sources of differences among the native and non-native views of the stimuli, we first obtained the centroid for each word category by computing the arithmetic mean of all exemplar representations within a category. Then, we apply a statistical approach based on bootstrap re-sampling to estimate the consistency of category centroids compared to category exemplars in the representation space. To control for the difference between the overall number of exemplars and number of categories, we sample exactly 1750 word categories for each language and then sample one exemplar per category in each bootstrapped sample. For each pair of samples (centroids vs. exemplars), we compute the representational similarity between each native model and each of the non-native models using our proposed similarity score, which yields a set of pairwise comparisons between centroid-based RSA and exemplar-based RSA. The results of this analysis is depicted in Figure 6.9. We observe that centroid-based RSA scores are always higher than their exemplar-based RSA counterparts, which can be visually inferred from the positions of comparison data points above the identity line. This finding suggests that the relative positions of category centroids are more consistent between the native and non-native representations, and the relative positions of individual exemplars within lexical clusters seem to have a stronger effect in making L2 representations dissimilar than the native ones.

6.8 Analyzing Exemplar Representations

To further investigate the discrepancy between exemplar arrangements within lexical clusters among L1 vs. L2 models, we conduct a qualitative evaluation of the representation space by inspecting the distances between held-out exemplars of word categories that the native model observed during training. We conduct this analysis on the representations of Czech word stimuli from the CAE model using the L1 Czech model and two non-native models; the Polish model (most similar) and the German model (least similar).

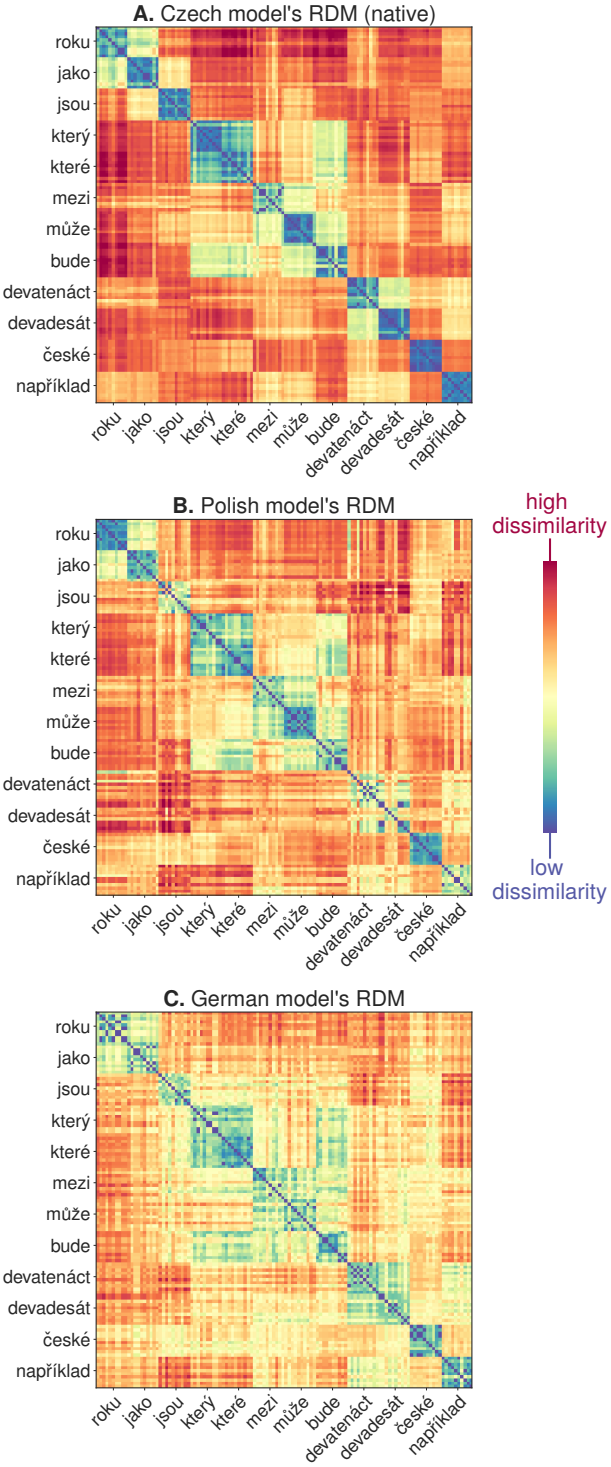


Figure 6.10: Representation dissimilarity matrix (RDM) of Czech stimuli of 12 word categories, each with 10 exemplars. Warmer colors indicate higher dissimilarity (large distances), while cooler colors indicate higher similarity (small distances). Note that each cell contains the cosine distance between two representations associated with two different stimuli. The RDM is symmetric along a diagonal of zeros.

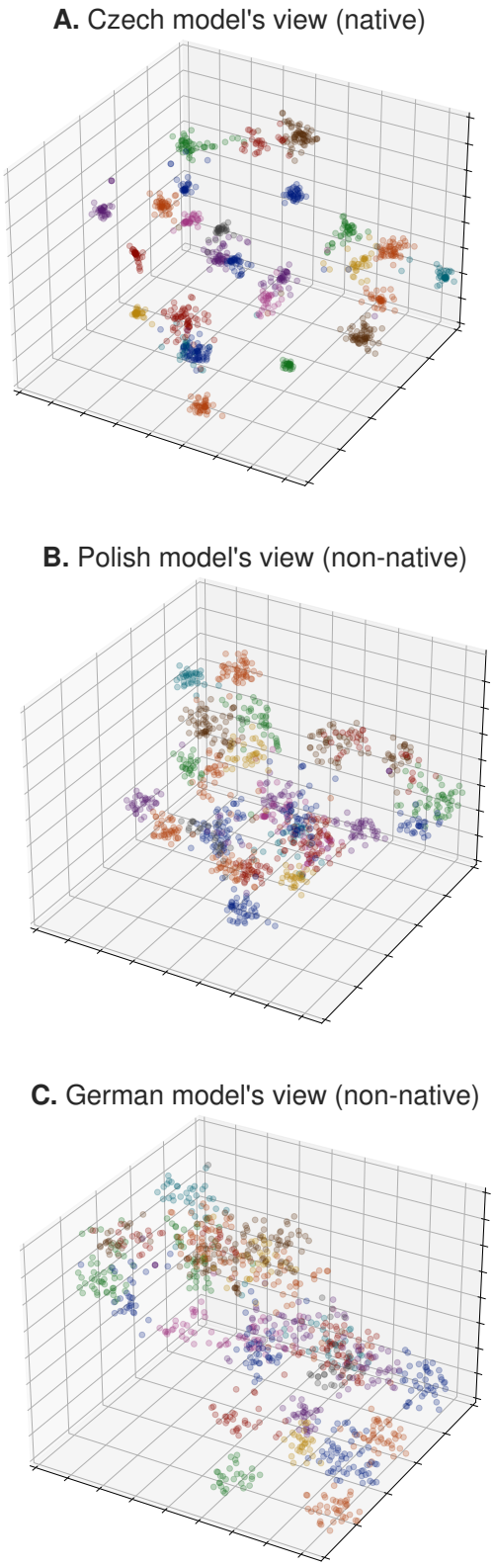


Figure 6.11: Three-dimensional t-SNE visualizations of a sample of Czech spoken word stimuli. Data points that are close in the representation space and have the same color are exemplars of the same word category.

Our first analysis is based on visualizing the representational dissimilarity matrix (RDM) which characterizes the pairwise distances between the word stimuli in the representation space (Figure 6.10). We analyze 12 word categories where sample 10 held-out exemplars (unheard speakers) for each category. From the L1 Czech representations in Figure 6.10–A, we observe a block-like structure centered along the diagonal which is visually recognized by cooler colors, revealing small distances between exemplars of the same category. This block-like structure is less obvious in the L2 representations of the Czech stimuli produced by the Polish model in Figure 6.10–B and the German model in Figure 6.10–C. Therefore, our analysis reveals that the native exemplar representations are tightly clustered around their centroid while their non-native counterparts are loosely tied to their centroid. Moreover, the RDMs reveal intriguing cross-linguistic differences between the Polish and German views of the Czech stimuli. One can observe that distinct Czech word categories are better separated by the Polish model compared to the German model (characterized by the presence of warmer colors in the Figure 6.10–B compared to Figure 6.10–C). This behaviour can be explained by the fact that Czech and Polish exhibit similar phonological regularities, which enables the Polish model to produce more native-like representations, thus better category separability, compared to the German model.

Finally, we inspect the representation space by means of visualization and dimensionality reduction using the the t-SNE algorithm (Van der Maaten and Hinton, 2008). Figure 6.11 shows 3-dimensional projections of Czech held-out stimuli by the native Czech model and non-native Polish and German models. The geometric structures of the exemplar placement in Figure 6.11 confirms our observations from the RDMs. That is, category exemplars are tightly clustered around their centroid in the native view, while the non-native views exhibit a looser geometric arrangement. The effect of loosely clustered exemplars is more prominent in the German view compared to the Polish view of the Czech stimuli.

6.9 Discussion and Summary

In the sociolinguistics literature, it has has been reported that the objective measures of linguistic distance (e.g., lexical distance) is a strong predictor of mutual intelligibility among related languages (Gooskens, 2007). However, it remains a challenge to disentangle the contributions of language similarity at different levels (i.e., phonetic, lexical, syntactic, semantic, etc.) to cross-linguistic intelligibility. In addition, the contributions of lower levels of the linguistic hierarchy (e.g., phonetic, phonological, and prosodic) to mutual intelligibility are difficult to

study and quantify due to confounding experimental variables (i.e., extra-linguistic factors such as listeners' prior exposure to different languages besides their mother tongue). The key contribution of our study is a computational, analytic approach for quantifying the effect of cross-language similarity at lower levels of the linguistic hierarchy on L2 speech processing using a neural acoustic modeling and the RSA framework. By quantifying the extent to which non-native models exhibit native-like behavior—operationalized as representational similarity—our approach does not require a parallel list of word stimuli. Therefore, it is applicable to any sample experimental stimuli as long as the training conditions of the models are cross-linguistically comparable.

The computational models we investigate in this study simulate the first stage of lexical access during speech comprehension, namely the acoustic-phonetic analysis and the phonological decoding of the incoming speech signal. Since our acoustic models operate at the word level, we isolate the effects of sentence-level contextual cues during language processing, and focus on word recognition. Furthermore, the models have access only to word-form information, but not the semantic content of the lexical data. Therefore, the observed effects of the linguistic experience in our presented analysis cannot be attributed to the transparency of form-to-meaning mapping of word cognates, but rather to similarity in phonological regularities. In other words, since our models have no access to information beyond the word-form level, our analysis shows that the observed cross-linguistic intelligibility among the Slavic languages reported in sociolinguistic studies can be partly attributed to word-internal acoustic-phonetic, phonological, and prosodic similarities which enable the listener to activate the correct word-form mental representation prior to lexical access.

In this study, we aim to bridge between several directions of (computational) linguistics research that have been so far unconnected. Therefore, our work avoids the "Square One Bias" (Ruder et al., 2022) in language and speech processing research by connecting the speech modality, neural networks interpretability, and multilinguality research in a single study. Furthermore, our study exemplifies how an interdisciplinary perspective can enable us to pose novel research questions that are grounded on sociolinguistic studies of cross-language speech processing.

7

Semantically-Enriched Spoken-word Representations

The presented representation models of acoustic word embeddings (AWEs) in the previous chapters learn to map variable-length spoken-word segments onto fixed-dimensionality vector representations. These models are trained in a bottom-up approach that integrates acoustic cues to build up a word representation given an acoustic or symbolic supervision signal. Therefore, these models do not leverage or capture high-level lexical knowledge during the learning process. In this study, we propose a multi-task learning model that incorporates top-down lexical knowledge into the training procedure of AWEs. Our model learns a mapping between the acoustic input and a lexical representation that encodes high-level information such as word semantics in addition to bottom-up form-based supervision. We experiment with three languages and demonstrate that incorporating lexical knowledge improves the embedding space discriminability and encourages the model to better separate word categories.

7.1 Introduction

The development of robust automatic speech recognition (ASR) systems requires large collections of high-quality transcribed speech, which are only available for a small subset of the world languages. To facilitate access to spoken content for language varieties that are not yet supported by conventional ASR systems, researchers have developed voice-based search applications such as query-by-example (QbE) search (e.g., Jansen and Durme, 2012; Metze et al., 2013; Yaodong Zhang and James R Glass, 2009). These systems rely on vector-space acoustic models that map variable-length spoken-word segments onto fixed-size vector

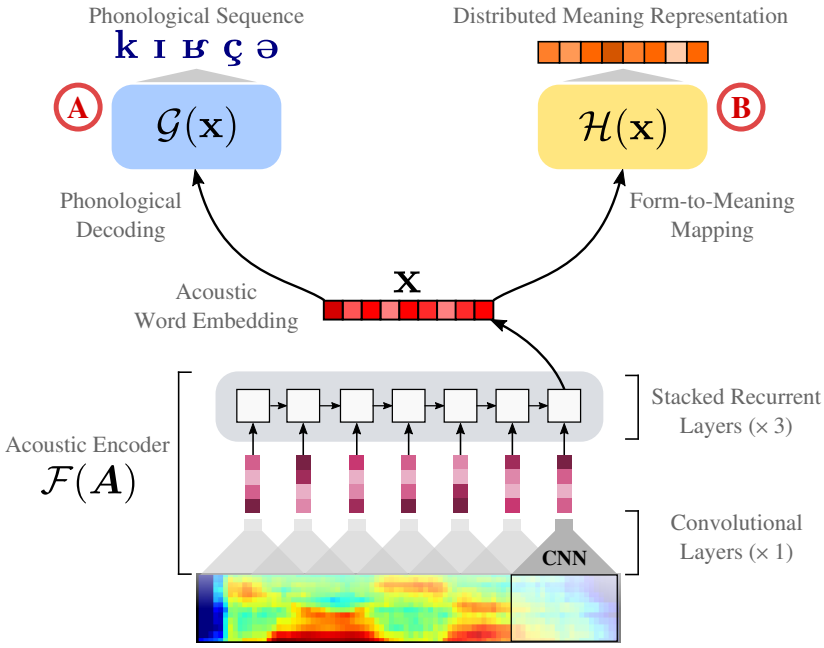


Figure 7.1: A schematic view of our proposed model.

representations such that exemplars of the same word are (ideally) projected onto the same vector (S. Bengio and Heigold, 2014; Herman Kamper, W. Wang, et al., 2016; Levin et al., 2013; Settle, Levin, et al., 2017; Settle and Livescu, 2016b, *inter alia*). In the speech technology literature, these fixed-dimensionality vector representations are known as acoustic word embeddings (AWEs). Currently, the top performing and the most efficient models of AWEs are based on deep neural networks (DNNs, He et al., 2017; H. Kamper et al., 2016; Herman Kamper, 2019; Herman Kamper, Elsner, et al., 2015). Due to the ubiquity of computers that support DNNs coupled with highly-optimized vector-space search algorithms (i.e., FAISS (J. Johnson et al., 2017)), AWEs enable efficient indexing and retrieval of spoken content at an unprecedented scale.

In addition to their applications in speech technology, DNN-based models of AWEs have been adopted as models of human speech processing and analyzed from a cognitively motivated angle in recent studies. It has been shown that AWEs predict non-native perceptual difficulties in phonetic categorization (Matusevych, Schatz, et al., 2020), cross-linguistic effects in auditory lexical processing (Matusevych, H. Kamper, et al., 2021), and non-native lexical production patterns of second language (L2) learners (Ando et al., 2021). These empirical findings from cognitively motivated, computational word perception and production studies encourage further integration between speech technology and cognitive science.

Nevertheless, the majority of existing AWEs rely on supervision signals that only capture low-level, form-based information about the word. That is, AWEs are learned in a bottom-up approach whereby acoustic-phonetic cues are integrated in the model to build up a word-form representation that encodes its phonetic features and phonological structure. However, a host of psycholinguistic studies with human listeners have shown that top-down, **high-level lexical properties**—such as **word semantics**—not only interact with the word recognition process but also facilitate discrimination between word competitors (Cortese et al., 1997; Hino and Lupker, 1996; Mirman and Magnuson, 2009; Strain et al., 1995; Zhuang et al., 2011). We take inspiration from these experimental findings and introduce an AWE model based on the **multi-task learning** framework that integrates **form-based** and **meaning-based** supervision signals into a single model (Figure 1). Contrary to prior work that aims to learn the semantic content directly using a very large speech corpus (Chung and James R. Glass, 2020), our model incorporates word semantics as an additional supervision signal, thus requiring only a few hours of speech and being more applicable in low-resource settings. We experiment with read speech corpora for three languages and empirically demonstrate that integrating high-level lexical knowledge into training AWEs improves the ability of the model to discriminate between word categories.

7.2 AWEs via Multi-Task Learning

Given an acoustic signal that corresponds to a spoken-word represented as a temporal sequence of T spectral vectors, i.e., $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T)$, the goal of an AWE model is to transform $\boldsymbol{A}$ into a D-dimensional vector representation $\mathbf{x}$. This task corresponds to learning an encoder function $\mathcal{F}_{\boldsymbol{\theta}} : \mathcal{A} \rightarrow \mathbb{R}^D$, where $\mathcal{A}$ is the (continuous) space of acoustic sequences, D is the embedding dimensionality, and $\boldsymbol{\theta}$ are the parameters of the function. Sequences in $\mathcal{A}$ can vary in length, thus the function $\mathcal{F}_{\boldsymbol{\theta}}$ should be modeled with a suitable neural architecture such as recurrent networks (RNNs). Therefore, transforming a variable-length acoustic input into a D-dimensional AWE can be described as

$$\mathbf{x} = \mathcal{F}(\boldsymbol{A}; \boldsymbol{\theta}_{\mathcal{F}}) \in \mathbb{R}^D \tag{7.1}$$

Different approaches in the literature have been proposed for modeling the function $\mathcal{F}(.; \boldsymbol{\theta}_{\mathcal{F}})$, which can be characterized as either architectural innovations or introducing new loss functions. In the approach we propose in this study, which is inspired by a classical connectionist model of spoken-word processing (Gaskell

and Marslen-Wilson, 1997), our goal is to integrate two sources of supervision signals—namely phonological form and lexical semantics—into the training procedure. To this end, we assume a dataset $\mathcal{D} = \{(\boldsymbol{A}^1, w^1), \ldots, (\boldsymbol{A}^N, w^N)\}$ of N spoken words where w^i is the written form of the i^{th} word. Such a dataset can be automatically obtained using a forced alignment tool on a transcribed speech dataset. Furthermore, we assume the availability of two look-up dictionaries: (1) a dictionary that maps each written word onto its phonemic transcription as $\Phi(w) = \boldsymbol{\varphi}_{1:\tau} = (\varphi_1, \ldots, \varphi_\tau)$, which can be automatically created using a grapheme-to-phoneme (G2P) tool, and (2) a lookup dictionary that maps each word into a distributed word representation as $\Lambda(w) = \mathbf{w} \in \mathbb{R}^K$. The distributed word representation ideally encodes high-level lexical knowledge about the word—such as its semantic and syntactic properties—and can be obtained independently using a large text corpus or from a public repository of semantic word embeddings such as ***Glove*** (Pennington et al., 2014) or ***fasttext*** (Mikolov, Grave, et al., 2018).

7.2.1 Form-based Phonological Supervision

Our first learning objective is based on the sequence-to-sequence learning framework in which the network is trained as a word-level acoustic model (Figure7.1, branch [A]). Given the output of acoustic encoder $\mathbf{x}$, a phonological decoder $\mathcal{G}(.;\boldsymbol{\theta}_\mathcal{G})$ aims to decode the corresponding phonological sequence $\boldsymbol{\varphi}_{1:\tau}$ of the word-form $\mathbf{x}$. The objective is to minimize a categorical cross-entropy loss at each timestep in the decoder, which is equivalent to minimizing the term

$$\begin{aligned}\mathcal{L}^\phi(\boldsymbol{\theta}_\mathcal{F}, \boldsymbol{\theta}_\mathcal{G}) &= -\sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \log \mathbf{P}\big(\Phi(w^i) \,|\, \mathcal{F}(\boldsymbol{A}^i; \boldsymbol{\theta}_\mathcal{F});\, \boldsymbol{\theta}_\mathcal{G}\big) \\ &= -\sum_{(\boldsymbol{A}^i, w^i)\in\mathcal{D}} \sum_{t=1}^{\tau} \log \mathbf{P}\big(\varphi_t \,|\, t,\, \mathbf{x}^i;\, \boldsymbol{\theta}_\mathcal{G}\big)\end{aligned} \tag{7.2}$$

where $\mathbf{P}\big(\varphi_t \,|\, t,\, \mathbf{x}^i;\, \boldsymbol{\theta}_\mathcal{G}\big)$ is the probability of the phoneme φ_t at the t^{th} timestep, conditioned on the previous phoneme sequence $\boldsymbol{\varphi}_{1:t-1}$ and the AWE $\mathbf{x}$, and $\boldsymbol{\theta}_\mathcal{G}$ are the parameters of the decoder. The learning objective is based on the idea that while acoustic realizations of words vary across speakers and contexts, all exemplars of a specific word should have the same phonemic transcription. Thus, the model should map exemplars of the same word close together in the embedding space, where the distance should ideally reflect phonological (dis)similarity.

7.2.2 Meaning-based Lexical Supervision

Our second learning objective aims to map the acoustic input $\boldsymbol{A}$ onto a high-level lexical representation (Figure7.1, branch [B]). The goal here is to incorporate a supervision signal from a level that is higher in the linguistic hierarchy compared to form-based phonological supervision. Inspired by Maas et al. (2012), we model this task as a vector regression problem. The output of the acoustic encoder $\mathbf{x}$ is transformed via a feed-forward network into a semantic vector as $\boldsymbol{v} = \mathcal{H}(\mathbf{x}; \boldsymbol{\theta}_{\mathcal{H}}) \in \mathbb{R}^K$. Thus, the objective is to minimize the term

$$\begin{aligned} \mathcal{L}^{\lambda}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{H}}) &= \sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} || \, \mathcal{H}\big(\mathcal{F}(\boldsymbol{A}^i; \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{H}}\big) - \Lambda(w^i) \, ||_2 \\ &= \sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} || \, \boldsymbol{v}^i - \Lambda(w^i) \, ||_2 \end{aligned} \tag{7.3}$$

where $\Lambda(w^i) \in \mathbb{R}^K$ is the ground-truth distributed representation, or semantic word embedding, of the ith sample. We assume that continuous, distributed word representations are available to the model during training. Given the ubiquity of word embeddings in the NLP research and the availability of text corpora for many languages, we believe that our assumption is reasonable. Since all exemplars of a word category are associated with the same semantic representation, we expect this objective to separate word categories in the representation space.

7.2.3 Integrating Form and Meaning Supervision

To integrate the two sources of supervision when training the model, we jointly minimize the term

$$\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{H}}) = \alpha \cdot \mathcal{L}^{\phi}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{G}}) + \beta \cdot \mathcal{L}^{\lambda}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{H}}) \tag{7.4}$$

Here, α and β are trade-off hyperparameters (i.e., scalars) that control the contribution of each term to the overall loss.

7.3 Baseline: Contrastive Acoustic Model

We compare the performance of our proposed model to a strong baseline that explicitly minimizes the distance between exemplars of the same word category. The baseline model employs a contrastive triplet loss that has been extensively explored in the AWEs literature with different underlying architectures and has

shown strong discriminative performance (Abdullah et al., 2021; Jacobs and Herman Kamper, 2021b; H. Kamper et al., 2016; Settle and Livescu, 2016a). Given a matching pair of AWEs $(\mathbf{x}^a, \mathbf{x}^+)$—i.e., embeddings of two exemplars of the same word type—the objective is then to minimize a triplet margin loss

$$\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}) = \sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} \max\left[0, m + d(\mathbf{x}^i, \mathbf{x}^+) - d(\mathbf{x}^i, \mathbf{x}^-)\right] \tag{7.5}$$

where $\mathbf{x}^-$ is an AWE that corresponds to a word other than w^i, and $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, 1]$ is the cosine distance. This objective aims to map acoustic exemplars of the same word closer in the embedding space while pushing away segments of different word types by a distance defined by the margin hyperparameter m. To obtain negative samples, we create mismatching pairs from the mini-batch such that $d(\mathbf{x}^i, \mathbf{x}^-)$ is minimized (Jansen, Plakal, et al., 2018).

7.4 Experiments

7.4.1 Experimental Data

The data in our study is drawn from the GlobalPhone multilingual speech database (Schultz et al., 2013) for Portuguese, German, and Polish (see Table 7.1). We sample 42 speakers from each language for training and obtain spoken-word segments using the Montreal Forced Aligner (McAuliffe et al., 2017). It is worth pointing out that the speakers in the validation and test splits are held-out and not used while training. The phonemic transcription for each word is produced using the *eSpeak* G2P tool. Then, each acoustic segment is parametrized as a sequence of 39-dimensional Mel-frequency spectral coefficients of 25ms frames with 10ms overlap.

Table 7.1: Word-level statistics of our experimental data.

	# segments per split			duration	Type-Token Ratio
	train	valid	test	(*mean* ± *std*)	
Portuguese	28810	9029	9580	0.51 ± 0.19	0.147
German	28914	9683	9372	0.44 ± 0.18	0.193
Polish	27979	9656	9089	0.50 ± 0.18	0.267

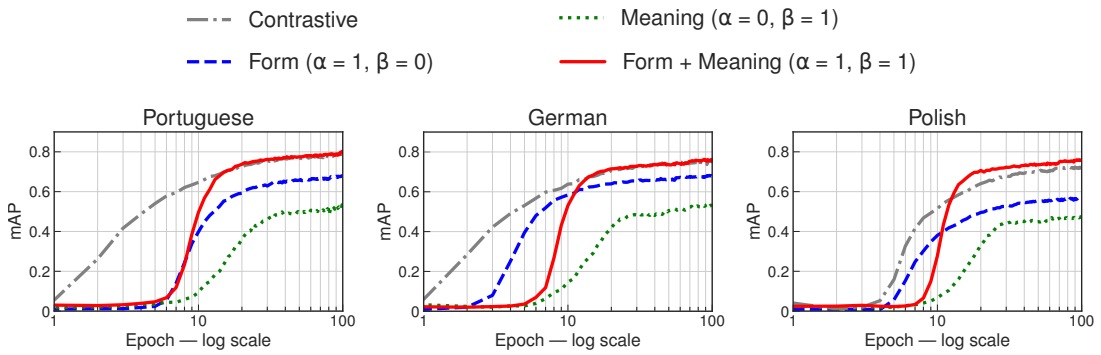


Figure 7.2: Learning curves of the models for 100 training epochs, quantified by the word discrimination task and the mAP metric.

7.4.2 Architecture and Hyperparameters

Acoustic Encoder $\mathcal{F}(.;\boldsymbol{\theta}_{\mathcal{F}})$. Our acoustic encoder consists of a hybrid, convolutional-recurrent neural network architecture. The frond-end consists of a 1D convolutional layer of 64 filters with a kernel size of 5 spectral vectors and stride of 2. Then, the output of the convolutional layer is fed sequentially into a recurrent block that consists of a 3-layer unidirectional Gated Recurrent Unit (GRU) with a hidden state of 512 units, which yields a 512-dimensional AWE as the last hidden state of the GRU. We apply layer-wise dropout with a probability of 0.2. Bidirectional GRUs did not yield further improvements.

Phonological Decoder $\mathcal{G}(.;\boldsymbol{\theta}_{\mathcal{G}})$. We employ a 1-layer GRU of 512 units hidden state that takes the 512-dimensional AWE as the initial hidden state and decodes the corresponding phonological sequence without teacher forcing.

Form-to-Meaning Regressor $\mathcal{H}(.;\boldsymbol{\theta}_{\mathcal{H}})$. We employ a linear layer ($512 \rightarrow 300$) followed by a `Tanh` non-linearity to project the AWE $\mathbf{x}$ onto the corresponding distributed word representation. We use pre-trained 300-dimensional ***fasttext*** embeddings as distributed word representations. Deeper feed-forward networks did not yield further improvements.

Contrastive Loss. For the baseline model with the contrastive loss, we experiment with different values of the margin hyperparameter $m = \{0.2, 0.3, 0.4, 0.5\}$, out of which 0.4 yields the best performance on the validation set.

Training Details. We train all models in this study for 100 epochs with batches of 256 samples using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate (LR) of 0.001. The LR is reduced by a factor of 0.5 if the performance on the validation set does not improve for 10 epochs. The epoch with the best validation performance is used for evaluation on the test set.

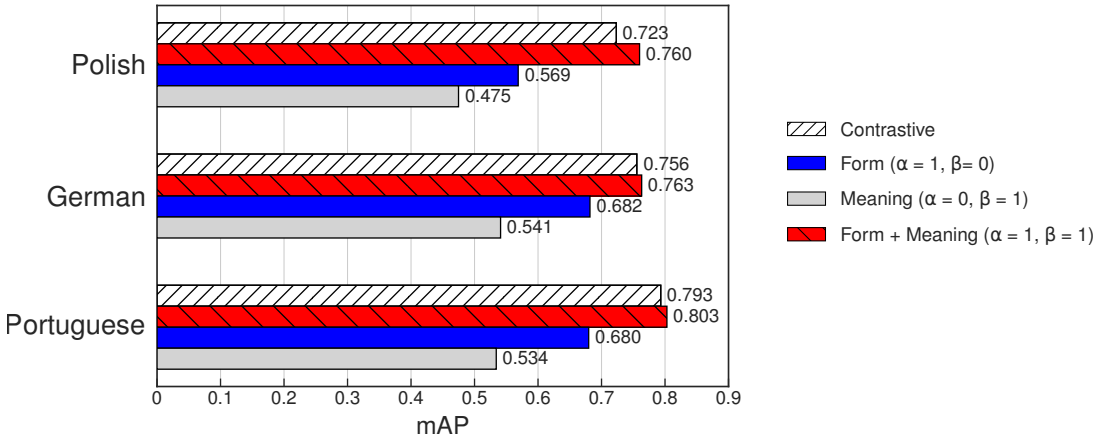


Figure 7.3: Word discrimination performance (mAP) on test set.

7.4.3 Experimental Results

We conduct an intrinsic evaluation for the AWEs to assess the performance of our models using the same-different acoustic word discrimination task with the mean average precision (mAP) metric, which was defined in Chapter 5 (§5.4). Prior work has shown that performance on this task positively correlates with improvement on downstream QbE speech search (Jacobs and Herman Kamper, 2021b). This task evaluates the ability of the model to determine whether two given lexical segments correspond to the same word type—that is, whether or not two acoustic instances are exemplars of the same category.

Figure 7.2 shows the learning curves for the models during 100 epochs of training quantified by the performance on the validation set. Contrary to the other models, we observe that the contrastive baseline model reaches a reasonable performance before the 10th epoch, which we attribute to the fact that the evaluation task (word discrimination) and the learning objective (contrastive triple loss) are analogous. Figure 7.3 shows the final performance on the test set. We observe that both the form-only model ($\alpha = 1, \beta = 0$) and the meaning-only model ($\alpha = 0, \beta = 1$) perform poorly compared to the contrastive baseline. However, integrating the two sources of supervision in the form + meaning setting ($\alpha = 1, \beta = 1$) enables the model to outperform the contrastive baseline for the three languages in our study.

The gain in performance is more prominent in the Polish language (relative mAP gain by 5.06%), which is the most morphologically complex language in our study due to its rich inflection system. The Polish morphological complexity is also reflected in its relatively high type-to-token ratio (TTR) in Table 7.1. These findings show that integrating high-level linguistic knowledge in training acoustic

models improves the discriminability of the embeddings space, and the effect seems to be more prominent on a language with a rich morphological system.

7.4.4 Embedding Visualization

To analyze our multi-task learning model and gain further insights into its embedding space, we use the t-SNE (Van der Maaten and Hinton, 2008) dimensionality reduction algorithm on the AWEs from the setting where $\alpha = 1$ and $\beta = 1$ (form + meaning supervision). We visualize the embeddings of acoustic word samples from held-out speakers (i.e., a set of speakers the models were not trained on). Note that the t-SNE objective aims to preserve the local structure within the higher-dimensional space when reducing the dimensionality. Therefore, the local distance between the two-dimensional projections of the embeddings mainly reflects the cluster structure within the embedding space, which enables us to visually inspect the emergent clusters and investigate whether or not they correspond to distinct word categories.

The t-SNE visualizations for the three languages in our study are illustrated in Figure 7.4 for Portuguese and German, and in Figure 7.5 for Polish. We observe a clear tendency for exemplars of the same word category to closely cluster in the embedding space, despite the lack of an explicit clustering objective in the learning procedure. One notable exception we observe in Figure 7.5 for the Polish language is the two nearby, nearly overlapping clusters that correspond to the word forms [tysięcy] and [tysiące]. Note that these two word forms are two morphological variants of the same lemma (i.e., [tysiąc], the Polish word for *thousand*). Given their semantic and phonological similarity, a small distance between the centroids of their clusters is expected.

7.5 Summary

AWEs are vector representations of spoken words that encode their acoustic-phonetic features and phonological structures. In addition to their utility in speech technology applications, models of AWEs have shown to produce human-like behavior in various auditory lexical processing tasks. Existing methods for learning AWEs from speech corpora employ training strategies with acoustic, phonological feature-based, or symbolic form-based supervision. These learning strategies correspond to the bottom-up integration of acoustic-phonetic cues to build up a word-form representation. In this study, we have introduced a methodology

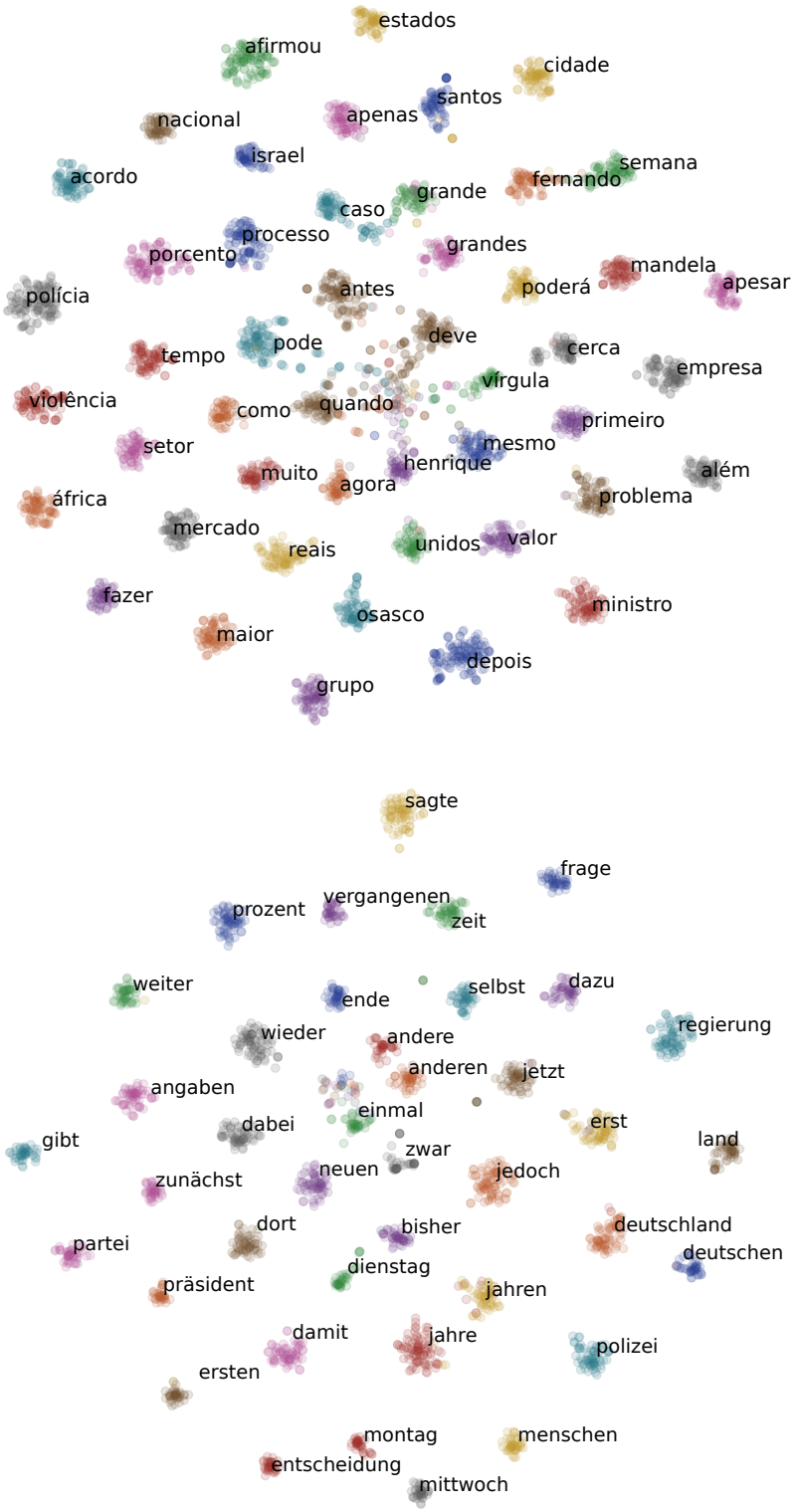


Figure 7.4: Two-dimensional visualization of the word embedding space using the t-SNE algorithm for Portuguese (Top) and German (Bottom). Figure is best viewed in color.

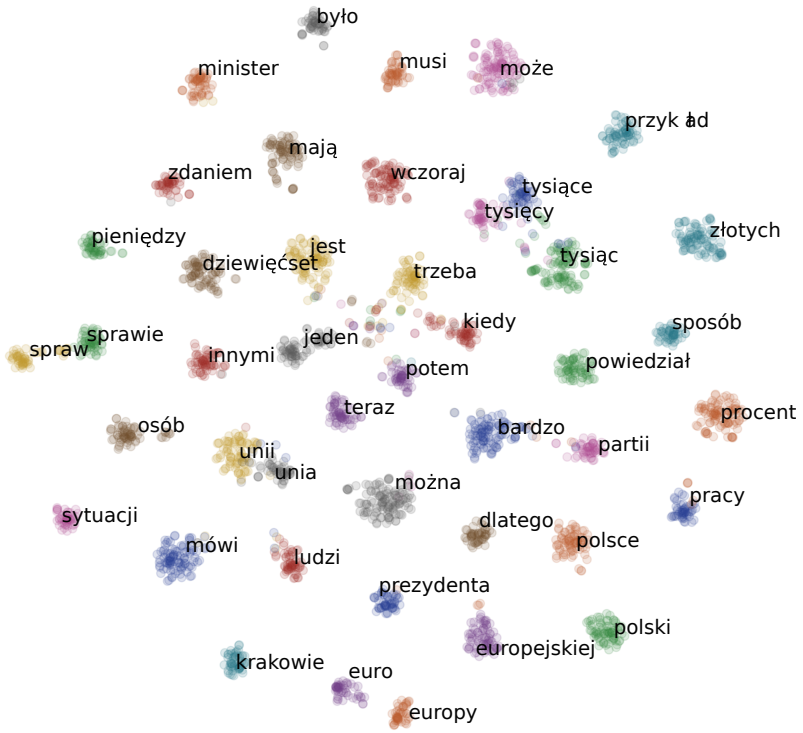


Figure 7.5: Two-dimensional visualization of the word embedding space using the t-SNE algorithm for Polish. Figure is best viewed in color.

based on the multi-task learning framework that leverages top-down, high-level lexical knowledge to learn semantically-enriched AWEs. We have experimented with semantic word embeddings as distributed meaning representations that guide the learning process in addition to form-based phonological supervision. Our experiments have demonstrated that integrating the two sources of supervision (i.e., phonological form and lexical semantics) improves the discriminability of the embeddings space—for the three languages in our study—as evidenced by the competent performance of our model compared to a strong contrastive AWE model. Furthermore, the t-SNE visualization analysis has supported our experimental findings in the word discrimination evaluation and provided further evidence that incorporating top-down lexical knowledge encourages the model to better separate the lexical categories in the embedding space without explicit supervision that directs the network to minimize the distance in the embedding space or a clustering

Part IV

DISCRETE SPEECH REPRESENTATIONS

8

Discrete Representations of Speech and Phonetic Variability

Self-supervised representation learning for speech often involves a quantization step that transforms the acoustic input into discrete units. However, it remains unclear how to characterize the relationship between these discrete units and abstract phonetic categories such as phonemes. In this chapter, we propose an information-theoretic characterization whereby we represent each phonetic category as a distribution over discrete units. We then apply our framework to two different self-supervised models (namely English wav2vec 2.0 and Multilingual XLSR-53) and use American English speech as a case study. Our study demonstrates that the entropy of phonetic distributions reflects the variability of the underlying speech sounds, with phonetically similar sounds exhibiting similar distributions. While our study confirms the lack of direct, one-to-one correspondence, we find an intriguing, indirect relationship between phonetic categories and discrete units.

8.1 Introduction

Self-supervised learning (SSL) for the speech modality is an active area of research that aims to develop models that build meaningful speech representations from raw audio without any explicit labels or transcriptions (see A. Mohamed et al. (2022) for an overview). These models can be further adapted for downstream tasks such as automatic speech recognition and speaker identification, and have become the state-of-the-art approach even when limited labeled data are available (Baevski et al., 2020; Hsu et al., 2021; A. v. d. Oord et al., 2018; Schneider et al., 2019a). Recently, it has become a common practice to include a quantization module within the architecture of SSL speech models that transforms the acoustic input into a

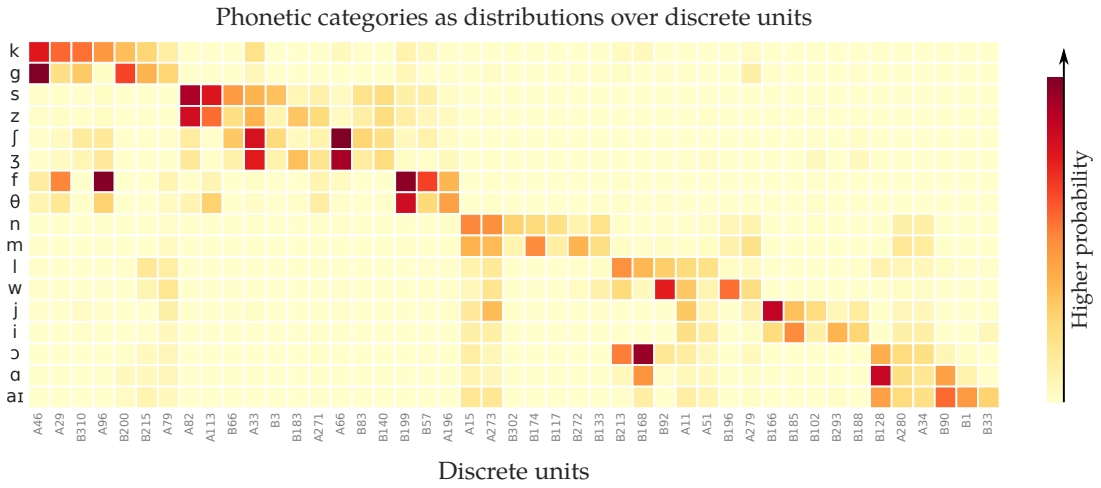


Figure 8.1: Phonetic categories as empirical probability distributions over highly responsive discrete units in the multilingual wave2vec2.0-XLSR-53 model.

sequence of discrete entities. Besides representing the complex acoustic signal in a compact and computationally efficient manner, learning discrete representations of speech can also facilitate training large SSL speech models using a masked language modeling objective similar to those employed in natural language processing (e.g., BERT (Devlin et al., 2019)).

Nevertheless, the nature of the discrete units learned via self-supervision remains an under-explored area of research. A key question is whether these discrete representations correspond to abstract phonetic categories such as phonemes. A few recent studies have investigated the discrete units from a neural network interpretability point of view (e.g., Higy et al., 2021; T. A. Nguyen et al., 2022; Sicherman and Adi, 2023; Wells et al., 2022). The phonetic analysis of Wells et al. (2022) have shown that discrete representations of speech correspond to low-level "sub-phonetic" events—rather than high-level phonetic categories—since they are sensitive to context-dependent and non-phonemic variations in speech. In Sicherman and Adi (2023), the authors concluded that there exists a strong correspondence between discrete units and phonemes, and attributed the lack of consistent phoneme-to-unit mapping to variations in phonological contexts. These findings seem to be contradictory and rely on different definitions of the term "phoneme", and thus remain inconclusive.

Although information theory was initially proposed as a mathematical theory of communication (Shannon, 2001), it also provides a quantitative framework for measuring the amount of information conveyed by linguistic units, such as words or sounds. Information theory has been adopted as a framework to study various aspects of linguistic structure, including phonology (Pimentel, Meister,

et al., 2021; Pimentel, Roark, et al., 2020), morphology (Rathi et al., 2021; S. Wu et al., 2019), and syntax (Futrell, Mahowald, et al., 2015; Hahn et al., 2018). This study builds on this line of research and develop information-theoretic metrics to characterize the relationship between phonetic categories and discrete units. Concretely, we make the following contributions:

1. We develop an empirical approach to represent each phonetic category as a probability distribution over discrete units using two self-supervised pre-trained models: English wav2vec 2.0 (henceforth W2V2) and multilingual wav2vec-XLSR (henceforth XLSR) (§ 8.2).

2. We characterize each phonetic category using the notion of information entropy and demonstrate that entropy quantifies acoustic-phonetic variability (§ 8.4).

3. We quantify the dissimilarity between phonetic distributions using Jensen-Shannon divergence and illustrate that this metric highly reflects feature-based phonetic similarity (§ 8.5).

8.2 Research methodology

8.2.1 Speech quantization via self-supervised learning

Consider a continuous acoustic signal represented as a sequence of T acoustic frames $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Here, each $\mathbf{x}_t$ is a short temporal window over the raw waveform with a stride. Given a pre-trained speech encoder, the signal $\mathbf{x}$ is first transformed via a local, temporal convolutional encoder $\mathcal{F} : \mathcal{X} \mapsto \mathcal{Z}$ into a sequence of latent speech representations in a continuous space as $\mathcal{F}(\mathbf{x}_1, \ldots, \mathbf{x}_T) = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$, where $\mathbf{z}_t \in \mathbb{R}^f$. As a part of the quantization step, the sequence of continuous representations gets discretized to produce a sequence of discrete units $\mathcal{D}(\mathbf{z}_1, \ldots, \mathbf{z}_T) = (\omega_1, \ldots, \omega_T)$, where $\mathcal{D} : \mathcal{Z} \mapsto \Omega$ can be viewed as a vector-to-centroid mapping and $\omega_t \in \Omega$ is the index of the centroid. Here, we use Ω to denote the finite set of discrete units within the model codebook. During pre-training using masked learning objectives, the corresponding quantized representations of these discrete units become the targets of the model prediction.

Figure 8.2 illustrates the computational flow within the quantization module in the wav2vec 2.0 model. Here, we give a concise formal description of the quantization process. Consider the output of the convolutional encoder at a single

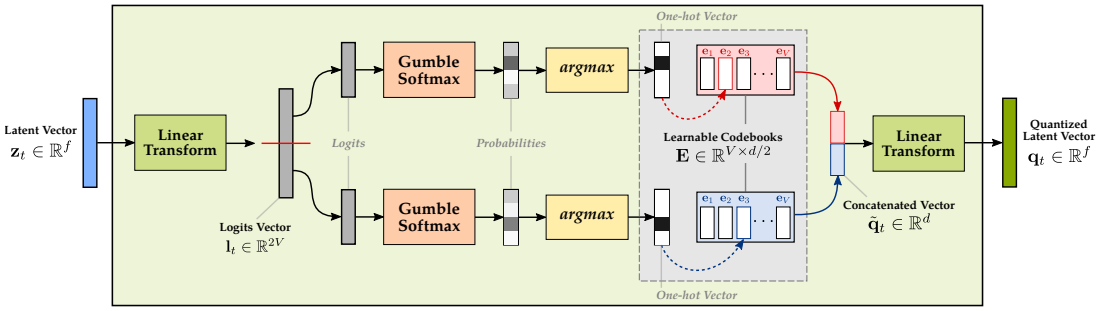


Figure 8.2: A visual illustration of the quantization module in wav2vec 2.0. In this study, we analyze the emergent discrete units within the quantization module. In particular, we look into the correspondence between the indices of the codebook and phonetic categories.

type step $\mathbf{z}_t$. Within the quantization module, a linear transformation is first applied to map the vector $\mathbf{z}_t$ is into a vector as follows

$$\mathbf{l}_t = \mathbf{W}_q \, \mathbf{z}_t \in \mathbb{R}^{(G \cdot V)} \tag{8.1}$$

where G is the number of codebooks within the quantization module and V is the number of entries in each codebook. Given that the developers of the wav2vec 2.0 model have made a design choice of incorporating two independent codebooks, the resulting vector of this transformation is $\mathbf{l}_t \in \mathbb{R}^{2V}$. Then, the vector $\mathbf{l}_t$ is organized into two groups such that each group goes through a computation that eventually activates a specific entry in the codebook. This computation takes a logit vector as input and produces a one-hot encoding vector as output through a differential `Gumble-softmax` function followed by an `argmax` function. Each codebook in the quantization module resembles an embedding lookup in NLP models, which can be formally described as a matrix $\mathbf{E} \in \mathbb{R}^{V \times d/2}$. Each entry in a codebook can be characterized by an index i and a corresponding vector $\mathbf{E}_i \in \mathbb{R}^{d/2}$. Let i and j be the two selected indices based on the resulting one-hot vectors, the corresponding codebook entries are concatenated to yield the vector

$$\tilde{\mathbf{q}}_t = \mathbf{E}_i^{(A)} \oplus \mathbf{E}_j^{(B)} \in \mathbb{R}^d \tag{8.2}$$

Here, $\mathbf{E}_i^{(A)}$and $\mathbf{E}_j^{(B)}$ are the selected vectors from codebooks A and B, respectively, and $\oplus$ is the concatenation operation. Finally, a linear transformation is applied to map the resulting vector into a vector of the same dimensionality as the

contextualized representation of the current speech frame within the Transformer network as follows

$$\mathbf{q}_t = \mathbf{W}_p \, \tilde{\mathbf{q}}_t \in \mathbb{R}^f \tag{8.3}$$

During pre-training, the learning objective is to minimize the distance between the contextualized representation of the final Transformer layer $\mathbf{c}_t$ and the output of the quantization module $\mathbf{q}_t$. Although many the design choices made by wav2vec 2.0 developers are not thoroughly justified and the motivation behind this particular process of vector quantization remains unclear, the model has been influential and effective for all speech processing tasks.

8.2.2 Phonetic categories as distributions over discrete units

Consider a speech corpus that is transcribed and aligned to phonetic segments given an inventory of phonetic categories Φ. In this scenario, a phonetic category can be considered as a set of K different acoustic exemplars obtained from the corpus, $\boldsymbol{\varphi} = \{\varphi^1, \ldots, \varphi^K\}$. These exemplars represent different acoustic realizations of the underlying phonetic category, and should optimally be produced by various speakers in diverse phonological contexts. Using the feature encoder and quantization module of a self-supervised speech model, we transform the associated acoustic segments of all exemplars $\{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$ into a discrete representation to obtain a collection of discrete sequences $\{(\omega_1^1, \ldots, \omega_{\tau_1}^1), \ldots, (\omega_1^k, \ldots, \omega_{\tau_k}^k)\}$ for each phonetic category. We then discard the exemplar identity as well as the sequential nature of each discrete sequence and view each phonetic category as a bag of discrete units. In this approach, each phonetic category can be described as a frequency distribution over the units in Ω. To facilitate our information-theoretic analysis, we turn the frequency distribution into a probability distribution where the probability of observing a discrete unit ω under a phonetic category $\boldsymbol{\varphi}$ is calculated using maximum likelihood estimation as follows

$$\boldsymbol{p}_\varphi(\omega_i) = \frac{N_\varphi(\omega_i)}{\sum_{\pi \in \Omega} N_\varphi(\pi)} \tag{8.4}$$

Here, $N_\varphi : \Omega \mapsto \mathbb{Z}^+$ is a function that returns the number of occurrences of a discrete unit under the phonetic category $\boldsymbol{\varphi}$, and therefore $\boldsymbol{p}_\varphi : \Omega \mapsto [0, 1]$ is a probability mass function defined over Ω such that $\sum_{\omega \in \Omega} \boldsymbol{p}_\varphi(\omega) = 1$. Note that each phonetic category in our analysis has its own $\boldsymbol{p}_\varphi$ and N_φ functions. For example, the vowels /æ/ and /ɔ/ are represented as two empirical distributions

Articulatory Class	Phonetic Categories (IPA)
Vowels	i ɪ eɪ ɛ ə æ aɪ aʊ ɑ ɔ ʌ oʊ ɔʊ ʊ u ɝ
Approximants	j w r l
Nasals	m n ŋ
Fricatives	f v θ ð s z ʃ ʒ h
Affricates	tʃ dʒ
Plosives	p b t d k g

Table 8.1: Phonetic categories in the TIMIT speech corpus of American English, grouped by several articulatory classes.

$\boldsymbol{p}_{/æ/}$ and $\boldsymbol{p}_{/ɔ/}$, respectively. Given our representation of a phonetic category as a distribution over discrete units $\boldsymbol{p}_{\varphi}$, we can employ information-theoretic metrics to characterize each phonetic distribution. For simplicity, we henceforth omit the subscript notation in $\boldsymbol{p}_{\varphi}$ and use $\boldsymbol{p}$ to denote a distribution associated with a single phonetic category.

8.3 Experimental setup

Experimental Speech data. We use the TIMIT speech corpus which consists of recordings from 630 American English speakers each speaking 10 different sentences, for a total of 6,300 sentences covering a diverse range of ages, genders, and regional accents from across the United States (Garofolo, 1993). Following (Räsänen et al., 2016b), the original phonetic categories of TIMIT annotation are mapped to the reduced set of 40 categories. We exclude silences and closures from our analysis.

Self-supervised speech models. We conduct our analysis using two SSL speech models that are publicly available via the HuggingFace Model Hub: (1) monolingual English wav2vec 2.0-BASE (Baevski et al., 2020), which is a 12-layer transformer model, and (2) multilingual wav2vec XLSR-53-LARGE (Conneau et al., 2020), which is a 24-layer transformer model trained on different languages. Both models employ two codebooks with 320 discrete units each, for a total of 640 units in each model. We consider the concatenation of the two codebooks as the set of discrete units in our analysis, thus $|\Omega| = 640$.

8.4 Analysis 1: Phonetic variability as information entropy

8.4.1 Information content and entropy

For any discrete unit within the codebook $\omega \in \Omega$, we measure its information content, or surprisal under a specific phonetic category as

$$\eta(\omega) = -\log_2 \boldsymbol{p}(\omega) \tag{8.5}$$

which quantifies the unexpectedness of the discrete unit to be observed under the phonetic category associated with the distribution $\boldsymbol{p}$. It is measured in bits. The uncertainty or "randomness" of the distribution $\boldsymbol{p}$ can be quantified as the average surprisal, or entropy

$$H(\boldsymbol{p}) = \sum_{\omega \in \Omega} \boldsymbol{p}(\omega)\, \eta(\omega) \tag{8.6}$$

where $0 \leq H(\boldsymbol{p}) \leq \log_2|\Omega|$. If all acoustic realizations of a phonetic category are associated with a single discrete unit, then its entropy is minimal $H(\boldsymbol{p}) = 0$. On the other hand, a distribution of a phonetic category is maximally entropic (i.e., $H(\boldsymbol{p}) = \log_2|\Omega|$) when all discrete units are equally likely to be aligned to this category. Therefore, entropy can be viewed as a measure of (within-category) acoustic-phonetic variability in our case. That is, the more entropic a phonetic category is, the higher the difficulty of predicting its alignment to discrete units. Note that our measure of variability is similar to the measure of diversity (i.e., the unit purity measure) introduced in (Hsu et al., 2021), but we express the variability of phonetic distributions using information-theoretic metrics.

8.4.2 Entropy per phonetic category

We compute the entropy of each phonetic category using Eq. 8.6. First, we find that phonetic categories are more entropic on average under W2V2 (mean $H = 3.97$) compared to XLSR (mean $H = 3.52$). After inspecting the phone-to-unit alignment of the TIMIT corpus, we attribute this behavior to different utilization of the codebooks across the two models. While there are 56.6% of the discrete units under W2V2 with non-zero counts across all phonetic categories, only 24.2% of the units have non-zero counts under XLSR. This difference gets reflected in lower entropy values in XLSR compared to W2V2.

Fig. 8.3 illustrates the results of our analysis with entropy as a measure of phonetic variability. Fig. 8.3A shows the entropy of each phonetic category in

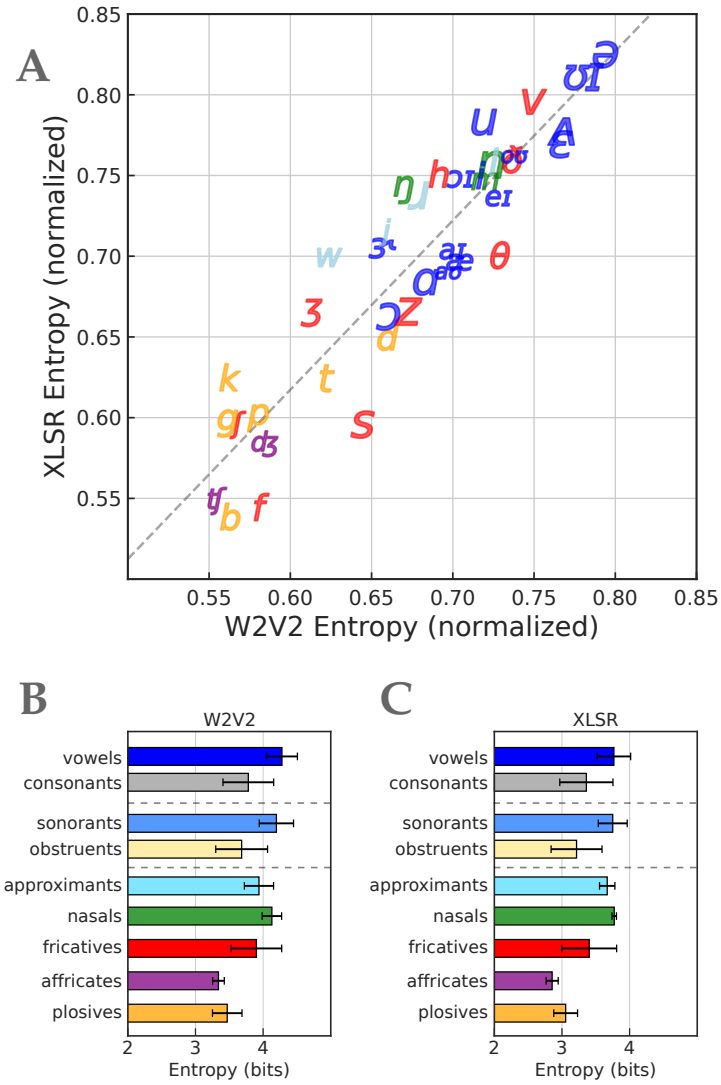


Figure 8.3: (A) The (normalized) entropy of each phonetic category in W2V2 (x-axis) vs XLSR (y-axis). (B-C) The entropy of several selected articulatory classes in W2V2 (B) and XLSR (C).

W2V2 (x-axis) and XLSR (y-axis). We report the normalized entropy in Fig. 8.3A to account for differences in entropy values between the two models. In addition, we group phonetic categories according to several articulatory classes, average the entropy over the categories within each class, and depict the result for W2V2 (Fig. 8.3B) and XLSR (Fig. 8.3C). From Fig. 8.3A, we observe a strong correlation between the two models (Pearson's $r = 0.92, p \ll 0.001$). When considering entropy values, we see that none of the phonetic categories is minimally entropic (i.e, $H(\boldsymbol{p}) = 0$), which confirms the findings in the literature about the lack of one-to-one correspondence between high-level abstract phonetic categories and discrete units in self-supervised speech models.

Regarding the variation of entropy across different phonetic categories, we observe that vowels tend to be more entropic than consonants in W2V2 ($H_V = 4.28 > H_C = 3.78$) and XLSR ($H_V = 3.77 > H_C = 3.36$). This reflects a higher variability in the acoustic realizations of vowels compared to consonants, since vowels are subject to a higher degree of variation due to vowel reduction in unstressed syllables and co-articulation, as well as other factors such as cross-speaker and dialect variability (Hagiwara, 1997; Hillenbrand et al., 1995; G. E. Peterson and Barney, 1952). For consonants, the nasal sounds (i.e., /n, m, ŋ/) are the most entropic consonant group, followed by the approximant sounds (i.e., /l, j, w, ɹ/), and then by the fricative sounds (i.e., /ð, z, ʒ, v, θ, s, ʃ, f, h/). We also observe that resonating consonants (i.e., nasals and approximants) exhibit higher variability on average than obstruents (i.e., plosives, fricatives, and affricates). Furthermore, we find an effect of voicing on variability since the voiced fricatives (i.e., /ð, z, ʒ, v/) are more entropic than their voiceless counterparts (i.e., /θ, s, ʃ, f/). For example, consider the voiceless-voiced contrast /f-v/ where /v/ is substantially more entropic than /f/ under W2V2 ($H(/\text{v}/) = 4.40 > H(/\text{f}/) = 3.41$) and XLSR ($H(/\text{v}/) = 4.01 > H(/\text{f}/) = 2.75$). This effect of voicing can be explained by the presence of low-frequency voicing energy in voiced fricatives which is likely to vary due to cross-speaker variability. Finally, the affricates (i.e., /dʒ, tʃ/) are found to be the least entropic consonant category under both W2V2 ($H = 3.33$) and XLSR ($H = 2.86$).

8.5 Analysis 2: Phonetic dissimilarity as Jensen-Shannon divergence

8.5.1 Relative entropy and divergence

Consider two phonetic distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ that are defined over the same set of discrete units Ω. To quantify how different $\boldsymbol{p}$ is from $\boldsymbol{q}$, we measure the expected surprisal from using $\boldsymbol{q}$ as a model distribution when the true distribution is $\boldsymbol{p}$. This quantity is known as the relative entropy or Kullback–Leibler divergence

$$D_{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) = -\sum_{\omega \in \Omega} \boldsymbol{p}(\omega) \log_2 \frac{\boldsymbol{q}(\omega)}{\boldsymbol{p}(\omega)} \tag{8.7}$$

Here, $D_{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) \geq 0$, with $D_{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) = 0$ only if $\boldsymbol{p} = \boldsymbol{q}$. Note that relative entropy is not symmetric, that is, $D_{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) \neq D_{KL}(\boldsymbol{q} \,||\, \boldsymbol{p})$. Since a symmetric

metric is more suitable for our analysis, we therefore measure the distance between two probability distributions using Jensen-Shannon divergence (JSD) as

$$D_{JS}(\boldsymbol{p}\,||\,\boldsymbol{q}) = \frac{1}{2}D_{KL}(\boldsymbol{p}\,||\,\boldsymbol{m}) + \frac{1}{2}D_{KL}(\boldsymbol{q}\,||\,\boldsymbol{m}) \tag{8.8}$$

where $\boldsymbol{m} = \frac{1}{2}\,\boldsymbol{p} + \boldsymbol{q}$ and $0 \leq D_{JS}(\boldsymbol{p}\,||\,\boldsymbol{q}) \leq 1$. Here, our goal is to investigate the degree to which the distance between distributions reflects phonetic similarity. Therefore, we use JSD as a measure of phonetic (dis)similarity in our analysis.

8.5.2 Exploratory similarity analysis

Table 8.2 presents a qualitative similarity analysis for a few selected phonetic categories under both models we analyze in this study. Concretely, we retrieve five phonetic categories that exhibit the lowest JSD scores (and by implication the highest similarity) for each of the categories in the set /w, ɛ, ʃ, g/. We then provide a ranking in the table from the most similar to the least. In the case of the approximant or semivowel /w/, we observe that the approximant sound /l/ exhibits the highest similarity under both models, but four vowels appear in ranks $2-5$. This indicates a high similarity in phonetic distributions between the approximant /w/ and vowels, which we further study in the clustering analysis below. For the front vowel /ɛ/, the top-5 similar categories are all vowels under both models, although no strong preference for other front vowels can be observed since similar vowels are a mixture of front and central vowels. The two models exhibit the highest agreement in the case of the unvoiced post-alveolar fricative /ʃ/, since both models have identical ranks that include the voiced post-alveolar

Table 8.2: Top-5 most similar phonetic categories to each of the categories /w, ɛ, ʃ, g/ in both W2V2 (W) and XLSR (X).

	/w/		/ɛ/		/ʃ/		/g/	
	W	X	W	X	W	X	W	X
1	/l/	/l/	/æ/	/æ/	/tʃ/	/tʃ/	/k/	/k/
2	/u/	/u/	/ʌ/	/ɪ/	/ʒ/	/ʒ/	/b/	/b/
3	/ʊ/	/ʊ/	/ɪ/	/ʌ/	/dʒ/	/dʒ/	/d/	/d/
4	/ɔ/	/ə/	/eɪ/	/eɪ/	/s/	/s/	/p/	/p/
5	/ɔɪ/	/oʊ/	/aʊ/	/aɪ/	/z/	/z/	/ð/	/h/

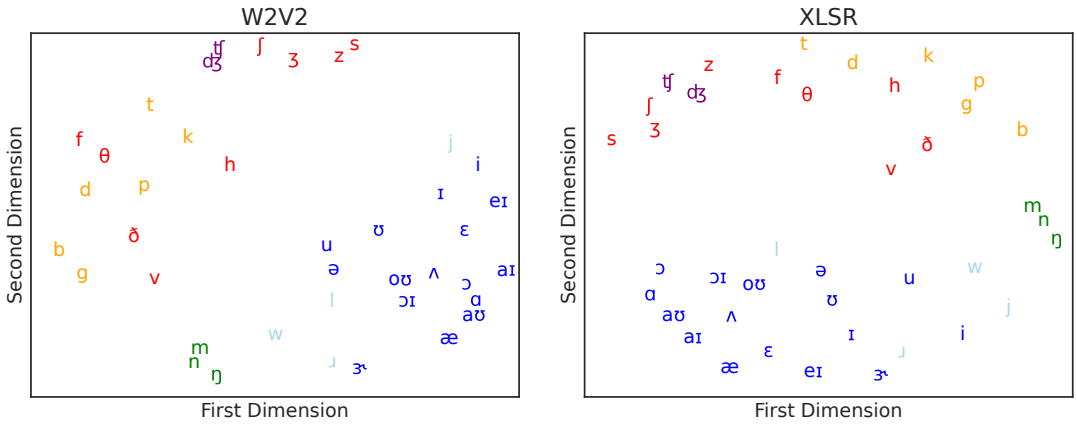


Figure 8.4: A multidimensional scaling (MDS) plot illustrating the distances between phonetic distributions in W2V2 (left) and XLSR (right).

fricative /ʒ/ and the affricates /tʃ, dʒ/ among the most similar. For the voiced velar plosive /g/, the unvoiced velar plosive /k/ is the most similar, as expected.

Furthermore, we apply the multidimensional scaling (MDS) to visualize the view of the pairwise distances on a two-dimensional plan. The resulting plots of the MDS are shown in Figure 8.4, where we can identify four different clusters: (1) vowels and approximants, (2) nasals, (3) sibilant fricatives, and (4) non-sibilant obstruents.

8.5.3 Hierarchical clustering

To study the similarity patterns among the phonetic categories, we apply agglomerative hierarchical clustering with the Ward algorithm (Ward, 1963) over the distance matrix generated by category-wise JSD values. The result of this clustering is illustrated in Fig. 8.5, where each phonetic category is colored by the manner of articulation. We observe that the clustering analysis yields a similar high-level grouping between W2V2 and XLSR, except for the placement of nasals which differs across the two models. For W2V2 in Fig. 8.5A, the highest level of organization divides the phonetic categories into two groups: a group that represents obstruent sounds (i.e., plosives, fricatives, and affricates) as well as nasals, and another group that represents vowels and approximants. On the other hand, the highest level of organization in XLSR in Fig. 8.5B reveals a pure obstruent vs. sonorant division, since both approximants and nasals exhibit a higher similarity to vowels than to other consonants. The consistent grouping of approximant sounds with vowels is not surprising given their acoustic-phonetic properties. Even though approximant sounds are considered consonants from a

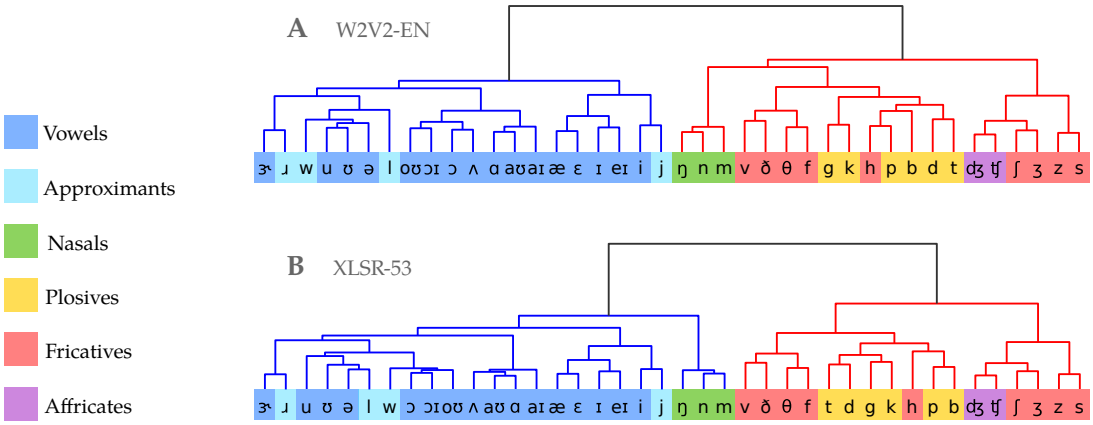


Figure 8.5: The resulting clusters from applying agglomerative hierarchical clustering over the distance matrix, where our measure of the distance is the Jensen-Shannon divergence between phonetic distributions: (A) w2v2 and (B) xlsr.

phonological point of view, they are produced with a (relatively) unconstricted articulation and exhibit a formant structure similar to vowels (Raphael, 2021).

Considering lower-level grouping for obstruent consonants, labio-dental and dental fricatives /f, v, θ, ð/ exhibit a higher similarity to plosive sounds /p, b, t, d, k, g/ than alveolar and postalveolar fricatives /s, z, ʃ, ʒ/ in both models. The affricates /ʤ, ʧ/ are grouped together with alveolar and postalveolar fricatives under both models, indicating the prominence of the fricative component of affricates in their underlying distributions over the discrete units. The only phonetic category that exhibits unexpected behavior in this analysis is the glottal fricative /h/, which is grouped within plosives under both models. However, the placement of the fricative /h/ among plosives should not be surprising given that the voiceless plosives /p, t, k/ are typically aspirated in syllable-initial position before a stressed vowel. Plosive aspiration is acoustically realized as a friction noise following the release of the plosive, similar to the friction of the sound /h/. Furthermore, the lowest level of grouping reflects the high similarity of phonetic minimal pairs (i.e., voicing contrasts) among all plosive contrasts (i.e., /t, d/, /p, b/, and /k, g/), but only two fricative contrasts (i.e., /s, z/ and /ʃ, ʒ/). As for the vowels, the lower-level grouping seems to reflect vowel backness more than vowel height in both models, although only a slight tendency to separate front vowels from back vowels can be observed.

8.5.4 Correlation with feature-based phonetic distance

To study the degree to which our measure of (dis)similarity (JSD) reflects phonetic distance, we correlate the distance among phonetic distributions over discrete units against a measure of feature-based phonetic distance. To this end, we map each phonetic category in the TIMIT inventory onto a discrete, multi-valued feature vector based on the PHOIBLE feature set (Moran and McCloy, 2019). We then compute the feature-based distance as the Hamming distance between their feature vectors. When we consider all phonetic categories, we find a strong positive correlation between the JSD and feature-based phonetic distance in W2V2 ($r = 0.63$) and XLSR ($r = 0.61$). Surprisingly, the correlation becomes stronger when we consider only the vowels in our analysis for both W2V2 ($r = 0.77$) and XLSR ($r = 0.80$), while it becomes weaker for consonants in W2V2 ($r = 0.47$) and XLSR ($r = 0.43$). The weaker correlation among the consonants could be attributed to the high similarity between the phonetic distributions of vowels and approximants in both W2V2 and XLSR, and vowels and nasals in XLSR. The correlation coefficients reported in this section are all Pearson's r and significant with $p \ll 0.001$.

8.6 Summary

We presented an information-theoretic framework for characterizing the relationship between phonetic categories and discrete units in self-supervised speech models. By representing each phonetic category as a distribution over discrete units, we have shown that the distribution entropy reflects the acoustic-phonetic variability of the underlying speech sounds, with vowels being more entropic on average than consonants. Moreover, phonetically similar sounds have been found to exhibit similar distributions, with the highest level of division separating obstruents and sonorants. Our findings confirm the characterization of discrete units as sub-phonemic events, rather than high-level categories such as phonemes, which is consistent with the findings of Wells et al. (2022). Given that speech sounds are dynamic acoustic signals that vary considerably due to many factors such as context and speaker, we argue that the characterization of phonetic categories as distributions over sub-phonemic events allows for a more nuanced understanding of the relationships between phonetic categories and discrete units in self-supervised speech models.

9

Conclusion and Future Outlook

In this chapter we summarize the thesis and highlight its contributions. Furthermore, we discuss perspectives and directions for future research that can build on the work and methodolody presented in this thesis.

9.1 Thesis Summary

One of the keys attributes that distinguishes humans from other intelligent species is our remarkable ability to communicate through language. Language serves as a powerful vehicle that enables us to exchange ideas, thoughts, and emotions across diverse cultures and communities. Through language, we can transmit and preserve knowledge to ensure our continuity and evolution across generations.

While human language is often described as a linguistic system using symbolic and discrete elements, the actual acoustic realization of language is continuous and dynamic. This dynamic nature of language is subject to various sources of variability in speech communication. Factors such as individual speaking styles, accents, dialects, and contextual influences all contribute to the unique use of language by individuals and communities. Despite the inherent variability in spoken language, speakers communicate ideas to listeners who can almost effortlessly decode the intended message.

Significant advancements in representation learning have allowed us to develop computational models that can effectively process and decode human language. While these models have empowered many successful speech and language technology applications, we have limited knowledge regarding how they encode, process, and represent different dimensions of speech variability. This thesis argues that a variability-centric perspective enables us to ask novel research questions and

conduct better-informed explorations of neural speech representations. To this end, each chapter in this thesis presented a case study that analyzes a dimension of speech variability and investigate how it shapes the representation profile of neural networks. In the following section, the thesis contributions are summarized.

9.2 Contributions

- A novel approach that improves the adaptability of neural networks to domain variability. The approach is based on unsupervised adversarial adaptation and requires only unlabeled samples from the unseen, target domain. We experimented with neural models of spoken language identification and we have demonstrated that adversarial training improves the model robustness against variability in recording conditions. Further analysis has revealed that adversarial training prevents neural networks from exploiting dataset-specific artefacts and spurious correlations as shortcuts for to predict the language.

- A linguistically-informed exploration of spoken language representations to analyze whether and to what extent they capture cross-linguistic variation and similarity. This exploration has revealed that neural networks uncover phylogenetic relations among the related Slavic languages, while representational similarity positively correlates with geographic proximity.

- An exploration of acoustic word embeddings as models of spoken-word representations from a neural network interpretability perspective. This exploration was mainly concerned with the encoding of variability in spoken-word representations and has revealed that spoken-word variability is encoded in a small fraction of the representation space. Our analysis has also shown that acoustically distinct words are easier to discriminate for the models, while word categories with many exemplars are not necessarily easier to discriminate. Furthermore, we have found that variability in initial conditions of the models lead to substantial individual differences in their representational geometry.

- A computational framework to study the role of linguistic experience, characterized by the first language of exposure (L1), on non-native spoken-word representations. The framework is based on representational similarity analysis and it aims to shed light on how variability in linguistic experience shapes speech representations of a non-native language (L2). With our framework,

we presented a case study on several Indo-European languages with various degrees of similarity and demonstrated that representational similarity predicts cross-linguistic intelligibility.

- A novel neural model that integrates form-based and meaning-based supervision signals for spoken-word representations. The two sources of supervision are integrated using the multi-task learning framework. Our model exemplifies how variability in lexical semantic content can encourage the neural network to better separate perceptually similar word categories.
- An information-theoretic exploration of the encoding of phonetic structure in discrete representations of speech that emerge in Transformer-based models via self-supervision. The exploration has revealed that entropy of phonetic distributions reflects the acoustic-phonetic variability of the underlying speech sounds, with sonorants being more entropic on average than obstruents. In addition, phonetically similar sounds are found to exhibit similar distributions while a clustering analysis has shown that phonetic categories are mainly grouped by manner of articulation.

9.3 Future Work

Embracing an interdisciplinary perspective, which resonates with the spirit of this thesis, opens up avenues for addressing additional research questions in future work. Moreover, the methodologies developed in this thesis can be adapted for various research directions that are relevant to the studies presented. The following subsections present and discuss some of these research questions.

9.3.1 Linguistically-informed Cross-lingual Transfer Learning

Previous studies have shown that cross-lingual transfer learning is most effective when transferring from a source training language that is linguistically similar to the target language. This language similarity effect has been observed in natural language processing (e.g., Lauscher et al., 2020; Pires et al., 2019) as well as speech processing (e.g., Jacobs and Herman Kamper, 2021a; Żelasko et al., 2020). Our analytical studies have demonstrated that similar languages exhibit higher representational similarity, indicating that the geometry of neural network representations are strongly shaped by the language(s) of exposure. These findings shed light on the role of cross-linguistic similarity in cross-lingual learning. However, in this thesis, cross-linguistic similarity was operationalized based on

mutual intelligibility and phylogenetic relatedness within a language family, which represents a coarse-grained approach. An alternative approach at the level of typological similarity would provide a more fine-grained analysis of cross-linguistic similarity and its impact on cross-lingual transfer across different tasks. In light of this, the following research question is posed:

What is the role of typological similarity in cross-lingual transfer learning? Which level of typological similarity is most relevant for a given task?

We hypothesize that typological similarity at the level of phonetic and phonological structure is more relevant for tasks such as automatic speech recognition, while higher linguistic levels (e.g., syntactic similarity) are more relevant for tasks such as language modeling (See also Papadimitriou and Jurafsky, 2020). This perspective will also inform language sampling when training multilingual models and benefit the resource construction process for under-resource languages.

9.3.2 Encoding of Indexical Properties in Multilingual Speech models

The development of multilingual, self-supervised models has opened up possibilities for speech technology applications in less-resourced languages (e.g., Babu et al., 2021; Conneau et al., 2020). However, the representation of multilinguality in these models remains an under-explored research topic. While it has been demonstrated that monolingual speech models encode indexical properties (e.g., speaker identity), in their representations (S. Li et al., 2022; Liu et al., 2023), it remains unknown how multilingual speech models disentangle speaker and language information. Therefore, the following research question is posed:

Is the representation of speaker information shared across languages in multilingual speech models?

This question is relevant to speech variability, as previous research in speech perception has shown that listeners can identify voices of their native language more accurately than those of an unfamiliar language (Goggin et al., 1991; E. K. Johnson et al., 2018). These findings allow us to generate research hypotheses about the encoding of speaker identity in multilingual speech representations. The probing method, or diagnostic classifiers, can be employed to extract speaker information from the representation of one language and evaluate the performance of the probe on speech samples from another language. If the performance remains similar to that of the probing language, it indicates that the speaker representation

subspace is shared across languages. Furthermore, we hypothesize that language similarity will have a strong effect on the probing results.

9.3.3 Language Representations in Multilingual Transformers for Speech Translation

The task of direct speech-to-text translation has greatly benefited from the use of Transformer-based speech models. In this task, an end-to-end encoder-decoder model is trained to take unsegmented acoustic input in the source language (speech encoder) and generate a text translation in the target language (text decoder). These models can also be trained in multilingual settings, where the encoder can process speech utterances from various languages and produce text translations in the target language. The encoder component of these speech translation models acts as a multilingual processor, dealing with significant cross-linguistic variability in speech sounds. However, it remains unclear how the encoder handles the multilingual nature of the input to generate accurate translations. To gain further insights into the behavior of these models, the following research question is posed:

Does the encoder of multilingual Transformers for speech translation learn a universal semantic space that is shared across languages?

Answering this question requires analyzing a multilingual speech translation model trained on typologically diverse languages. We hypothesize that the lower layers of the encoder perform language-specific auditory processing tasks, such as noise filtering, speaker normalization, and acoustic-phonetic processing. In contrast, the deeper layers are expected to uncover a universal semantic space that is shared across languages, where utterances that encode the same message are nearby in the representation space.

List of Figures

List of Tables

Bibliography

Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu (2014). "Convolutional neural networks for speech recognition." In: *IEEE/ACM Transactions on audio, speech, and language processing* 22.10, pp. 1533–1545 (cit. on p. 23).

Abdou, Mostafa, Artur Kulmizev, Felix Hill, Daniel M. Low, and Anders Søgaard (Nov. 2019). "Higher-order Comparisons of Sentence Encoder Representations." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5838–5845. DOI: 10.18653/v1/D19-1593. URL: https://aclanthology.org/D19-1593 (cit. on pp. 31, 101).

Abdullah, Badr M., Marius Mosbach, Iuliia Zaitova, Bernd Mobius, and Dietrich Klakow (2021). "Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study." In: *Interspeech* (cit. on p. 124).

Abnar, Samira, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema (Aug. 2019). "Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 191–203. DOI: 10.18653/v1/W19-4820. URL: https://aclanthology.org/W19-4820 (cit. on pp. 31, 101).

Algayres, Robin, Mohamed Salah Zaiem, Benoît Sagot, and Emmanuel Dupoux (2020). "Evaluating the Reliability of Acoustic Speech Embeddings." In: *Proc. Interspeech*. DOI: 10.21437/Interspeech.2020-2362 (cit. on p. 78).

Alishahi, Afra, Marie Barking, and Grzegorz Chrupała (2017a). "Encoding of phonology in a recurrent neural model of grounded speech." In: *arXiv preprint arXiv:1706.03815* (cit. on p. 5).

– (Aug. 2017b). "Encoding of phonology in a recurrent neural model of grounded speech." In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Compu-

tational Linguistics, pp. 368–378. DOI: 10.18653/v1/K17-1037. URL: https://aclanthology.org/K17-1037 (cit. on p. 100).

Amengual, Mark (2016). "The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals." In: *Applied Psycholinguistics* 37.5, pp. 1221–1251 (cit. on p. 96).

Ando, Shintaro, Nobuaki Minematsu, and Daisuke Saito (2021). "Lexical Density Analysis of Word Productions in Japanese English Using Acoustic Word Embeddings." In: *Proc. Interspeech 2021*, pp. 4433–4437. DOI: 10.21437/Interspeech.2021-853 (cit. on pp. 73, 120).

Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. (2021). "XLS-R: Self-supervised cross-lingual speech representation learning at scale." In: *arXiv preprint arXiv:2111.09296* (cit. on p. 150).

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations." In: *Advances in neural information processing systems* 33, pp. 12449–12460 (cit. on pp. 27, 133, 138).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate." In: *arXiv preprint arXiv:1409.0473* (cit. on p. 26).

Bangalore, Srinivas, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez (2012). "Real-time incremental speech-to-speech translation of dialogs." In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 437–445 (cit. on p. 37).

Barry, William and Bistra Andreeva (2001). "Cross-language similarities and differences in spontaneous speech patterns." In: *Journal of the International Phonetic Association* 31.1, pp. 51–66 (cit. on p. 112).

Bartelds, Martijn, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling (2022). "Neural representations for modeling variation in speech." In: *Journal of Phonetics* 92, p. 101137. ISSN: 0095-4470. DOI: https://doi.org/10.1016/j.wocn.2022.101137. URL: https://www.sciencedirect.com/science/article/pii/S0095447022000122 (cit. on p. 97).

Bartelds, Martijn and Martijn Wieling (2022). "Quantifying language variation acoustically with few resources." In: *arXiv preprint arXiv:2205.02694* (cit. on p. 97).

Bednarczuk, Leszek (2018). *Poczkatki i pogranicza polszczyzny*. Lexis (cit. on p. 55).

Beery, Sara, Grant Van Horn, and Pietro Perona (2018). "Recognition in terra incognita." In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473 (cit. on p. 39).

Beinborn, Lisa and Rochelle Choenni (2020). "Semantic drift in multilingual representations." In: *Computational Linguistics* 46.3, pp. 571–603 (cit. on pp. 31, 101).

Belinkov, Yonatan and James Glass (2017). "Analyzing hidden representations in end-to-end automatic speech recognition systems." In: *Advances in Neural Information Processing Systems* 30 (cit. on p. 5).

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan (2010). "A theory of learning from different domains." In: *Machine learning* 79.1, pp. 151–175 (cit. on p. 38).

Bengio, Samy and Georg Heigold (2014). "Word Embeddings for Speech Recognition." In: *Proc. Interspeech* (cit. on pp. 72, 120).

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult." In: *IEEE transactions on neural networks* 5.2, pp. 157–166 (cit. on p. 26).

Bent, Tessa and Rachael F Holt (2017). "Representation of speech variability." In: *Wiley Interdisciplinary Reviews: Cognitive Science* 8.4, e1434 (cit. on p. 4).

Benzeghiba, Mohamed, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. (2007). "Automatic speech recognition and speech variability: A review." In: *Speech communication* 49.10-11, pp. 763–786 (cit. on p. 4).

Bérard, Alexandre, Olivier Pietquin, Laurent Besacier, and Christophe Servan (2016). "Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation." In: *NIPS Workshop on end-to-end learning for speech and audio processing* (cit. on p. 4).

Bird, Steven (2021). "Sparse transcription." In: *Computational Linguistics* 46.4, pp. 713–744 (cit. on p. 72).

Bjerva, Johannes, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein (2019). "What do language representations really represent?" In: *Computational Linguistics* 45.2, pp. 381–389 (cit. on pp. 57, 62, 65, 66).

Brodersen, Kay Henning, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann (2010). "The balanced accuracy and its posterior distribution." In: *2010 20th International Conference on Pattern Recognition*. IEEE, pp. 3121–3124 (cit. on p. 46).

Bulut, Ahmet E, Qian Zhang, Chunlei Zhang, Fahimeh Bahmaninezhad, and John HL Hansen (2017). "UTD-CRSS submission for MGB-3 Arabic dialect identification: Front-end and back-end advancements on broadcast speech." In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 360–367 (cit. on p. 38).

Cai, Xingyu, Jiaji Huang, Yuchen Bian, and Kenneth Church (2020). "Isotropy in the contextual embedding space: Clusters and manifolds." In: *International Conference on Learning Representations* (cit. on p. 81).

Carlin, Michael A, Samuel Thomas, Aren Jansen, and Hynek Hermansky (2011). "Rapid evaluation of speech representations for spoken term discovery." In: *Proc. Interspeech* (cit. on p. 78).

Cathcart, Chundra and Florian Wandl (July 2020). "In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology." In: *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics (cit. on pp. 58, 65, 66).

Chan, William, Navdeep Jaitly, Quoc Le, and Oriol Vinyals (2016). "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4960–4964 (cit. on p. 4).

Chelba, Ciprian, Timothy J Hazen, and Murat Saraclar (2008). "Retrieval and browsing of spoken content." In: *IEEE Signal Processing Magazine* 25.3, pp. 39–49 (cit. on p. 37).

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation." In: *arXiv preprint arXiv:1406.1078* (cit. on p. 26).

Chrupała, Grzegorz and Afra Alishahi (July 2019). "Correlating Neural and Symbolic Representations of Language." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2952–2962. DOI: 10.18653/v1/P19-1283. URL: https://aclanthology.org/P19-1283 (cit. on pp. 31, 101).

Chrupała, Grzegorz, Bertrand Higy, and Afra Alishahi (July 2020). "Analyzing analytical methods: The case of phonology in neural models of spoken language." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4146–4156. DOI: 10.18653/v1/2020.acl-main.381. URL: https://aclanthology.org/2020.acl-main.381 (cit. on pp. 31, 101).

Chung, Yu-An, Yonatan Belinkov, and James Glass (2021). "Similarity Analysis of Self-Supervised Speech Representations." In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3040–3044 (cit. on pp. 5, 31, 101).

Chung, Yu-An and James R. Glass (2020). "Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech." In: *IProc. Interspeech* (cit. on p. 121).

Clopper, Cynthia G and David B Pisoni (2005). "Speech Perception, Hearing Impairment and Linguistic Variation." In: *Clinical Sociolinguistics*, pp. 207–218 (cit. on p. 18).

– (2021). "Perception of dialect variation." In: *The handbook of speech perception*, pp. 333–364 (cit. on pp. 4, 19).

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural language processing (almost) from scratch." In: *Journal of Machine Learning Research* 12.Aug, pp. 2493–2537 (cit. on p. 23).

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2020). "Unsupervised cross-lingual representation learning for speech recognition." In: *arXiv preprint arXiv:2006.13979* (cit. on pp. 138, 150).

Cook, Svetlana V, Nick B Pandža, Alia K Lancaster, and Kira Gor (2016). "Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings." In: *Frontiers in Psychology* 7, p. 1345 (cit. on p. 96).

Cortese, Michael J, Greg B Simpson, and Steph Woolsey (1997). "Effects of association and imageability on phonological mapping." In: *Psychonomic Bulletin & Review* 4.2, pp. 226–231 (cit. on p. 121).

Dahan, Delphine and James S Magnuson (2006). "Spoken word recognition." In: *Handbook of psycholinguistics*. Elsevier, pp. 249–283 (cit. on p. 92).

Dalewska-Greń, Hanna (2020). *Języki słowiańskie*. second. Wydawnictwo Naukowe PWN, Warszawa (cit. on p. 55).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on p. 134).

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/ (cit. on p. 57).

Dupoux, Emmanuel (2018). "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner." In: *Cognition* 173, pp. 43–59. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2017.11.008. URL: https://www.sciencedirect.com/science/article/pii/S0010027717303013 (cit. on p. 100).

Ethayarajh, Kawin (Nov. 2019). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 55–65. DOI: 10.18653/v1/D19-1006. URL: https://aclanthology.org/D19-1006 (cit. on p. 82).

Fügen, Christian, Alex Waibel, and Muntsin Kolss (2007). "Simultaneous translation of lectures and speeches." In: *Machine translation* 21.4, pp. 209–252 (cit. on p. 37).

Futrell, Richard and Michael Hahn (2022). "Information theory as a bridge between language function and language form." In: *Frontiers in Communication* 7, p. 657725 (cit. on p. 30).

Futrell, Richard, Kyle Mahowald, and Edward Gibson (Aug. 2015). "Quantifying Word Order Freedom in Dependency Corpora." In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala,

Sweden: Uppsala University, Uppsala, Sweden, pp. 91–100. URL: https://aclanthology.org/W15-2112 (cit. on p. 135).

Ganin, Yaroslav and Victor Lempitsky (2015). "Unsupervised Domain Adaptation by Backpropagation." In: *International Conference on Machine Learning*, pp. 1180–1189 (cit. on pp. 40, 42, 44, 48).

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016). "Domain-adversarial training of neural networks." In: *The journal of machine learning research* 17.1, pp. 2096–2030 (cit. on p. 38).

Garcia-Romero, Daniel and Carol Y Espy-Wilson (2011). "Analysis of i-vector length normalization in speaker recognition systems." In: *Twelfth annual conference of the international speech communication association* (cit. on p. 38).

Garofolo, John S (1993). "TIMIT acoustic phonetic continuous speech corpus." In: *Linguistic Data Consortium, 1993* (cit. on p. 138).

Gaskell, M Gareth and William D Marslen-Wilson (1997). "Integrating form and meaning: A distributed model of speech perception." In: *Language and cognitive Processes* 12.5-6, pp. 613–656 (cit. on p. 121).

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann (2020). "Shortcut learning in deep neural networks." In: *Nature Machine Intelligence* 2.11, pp. 665–673 (cit. on p. 39).

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In: *arXiv preprint arXiv:1811.12231* (cit. on p. 39).

Gelderloos, Lieke, Grzegorz Chrupała, and Afra Alishahi (July 2020). "Learning to Understand Child-directed and Adult-directed Speech." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/2020.acl-main.1. URL: https://aclanthology.org/2020.acl-main.1 (cit. on p. 100).

Gelly, Gregory and Jean-Luc Gauvain (2017). "Spoken Language Identification Using LSTM-Based Angular Proximity." In: *INTERSPEECH*, pp. 2566–2570 (cit. on p. 56).

Gelly, Grégory, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet Bac Le, and Abdel Messaoudi (2016). "Language Recognition for Dialects and Closely Related Languages." In: *Odyssey*, pp. 124–131 (cit. on pp. 38, 54, 57).

Glass, James Robert (1988). "Finding acoustic regularities in speech: applications to phonetic recognition." In: (cit. on p. 17).

Glockner, Max, Vered Shwartz, and Yoav Goldberg (2018). "Breaking NLI systems with sentences that require simple lexical inferences." In: *arXiv preprint arXiv:1805.02266* (cit. on p. 39).

Goggin, Judith P, Charles P Thompson, Gerhard Strube, and Liza R Simental (1991). "The role of language familiarity in voice identification." In: *Memory & cognition* 19.5, pp. 448–458 (cit. on p. 150).

Golubovic, Jelena (2016). "Mutual intelligibility in the Slavic language area." In: *Groningen: Center for Language and Cognition* (cit. on p. 98).

Golubović, Jelena and Charlotte Gooskens (2015). "Mutual intelligibility between West and South Slavic languages." In: *Russian linguistics* 39.3, pp. 351–373 (cit. on p. 98).

Gonzalez-Dominguez, Javier, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez, and Pedro J Moreno (2014). "Automatic language identification using long short-term memory recurrent neural networks." In: *Fifteenth Annual Conference of the International Speech Communication Association* (cit. on pp. 38, 56).

Gooskens, Charlotte (2005). "Travel time as a predictor of linguistic distance." In: (cit. on p. 62).

– (2007). "The contribution of linguistic factors to the intelligibility of closely related languages." In: *Journal of Multilingual and multicultural development* 28.6, pp. 445–467 (cit. on pp. 54, 98, 116).

– (2017). "Dialect intelligibility." In: *The handbook of dialectology*, pp. 204–218 (cit. on pp. 19, 96).

– (2019). "Receptive multilingualism." In: *Multidisciplinary perspectives on multilingualism: The fundamentals*, pp. 149–174 (cit. on p. 96).

Gooskens, Charlotte and Wilbert Heeringa (2004). "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data." In: *Language variation and change* 16.3, pp. 189–207 (cit. on p. 60).

Gooskens, Charlotte, Wilbert Heeringa, and Karin Beijering (2008). "Phonetic and lexical predictors of intelligibility." In: *International journal of humanities and arts computing* 2.1-2, pp. 63–81 (cit. on p. 54).

Gooskens, Charlotte and Vincent J van Heuven (2017). "Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups." In: *Speech Communication* 89, pp. 25–36 (cit. on p. 98).

Gooskens, Charlotte, Vincent J van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt (2018). "Mutual intelligibility between closely related languages in Europe." In: *International Journal of Multilingualism* 15.2, pp. 169–193 (cit. on p. 98).

Goto, Hiromu (1971). "Auditory perception by normal Japanese adults of the sounds" L" and" R."" In: *Neuropsychologia* (cit. on p. 96).

Goyal, Yash, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh (2017). "Making the v in vqa matter: Elevating the role of image understanding in visual question answering." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913 (cit. on p. 39).

Grand, Gabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko (2022). "Semantic projection recovers rich human knowledge of multiple object features from word embeddings." In: *Nature Human Behaviour*, pp. 1–13 (cit. on p. 71).

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In: *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376 (cit. on p. 4).

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks." In: *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, pp. 6645–6649 (cit. on p. 26).

Gutkin, Alexander, Tatiana Merkulova, and Martin Jansche (2018). "Predicting the Features of World Atlas of Language Structures from Speech." In: *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 248–252 (cit. on p. 57).

Hagiwara, Robert (1997). "Dialect variation and formant frequency: The American English vowels revisited." In: *The Journal of the Acoustical Society of America* 102.1, pp. 655–658 (cit. on p. 141).

Hahn, Michael, Judith Degen, Noah D. Goodman, Dan Jurafsky, and Richard Futrell (2018). "An Information-Theoretic Explanation of Adjective Ordering Preferences." In: *Cognitive Science* (cit. on p. 135).

Hansen, John HL and Taufiq Hasan (2015). "Speaker recognition by machines and humans: A tutorial review." In: *IEEE Signal processing magazine* 32.6, pp. 74–99 (cit. on p. 5).

Hawkins, Sarah (2003). "Roles and representations of systematic fine phonetic detail in speech understanding." In: *Journal of phonetics* 31.3-4, pp. 373–405 (cit. on p. 19).

Hazen, Timothy J and James R Glass (1997). "A comparison of novel techniques for instantaneous speaker adaptation." In: *Fifth European Conference on Speech Communication and Technology* (cit. on p. 5).

He, Wanjia, Weiran Wang, and Karen Livescu (2017). "Multi-view Recurrent Neural Acoustic Word Embeddings." In: *Proc. ICLR* (cit. on p. 120).

Heeringa, Wilbert, Keith Johnson, and Charlotte Gooskens (2009). "Measuring Norwegian dialect distances using acoustic features." In: *Speech Communication* 51.2, pp. 167–183 (cit. on p. 54).

Heeringa, Wilbert Jan (2004). "Measuring dialect pronunciation differences using Levenshtein distance." PhD thesis. University Library Groningen][Host] (cit. on p. 60).

Higy, Bertrand, Lieke Gelderloos, Afra Alishahi, and Grzegorz Chrupała (2021). "Discrete representations in neural models of spoken language." In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 163–176 (cit. on p. 134).

Hillenbrand, James, Laura A Getty, Michael J Clark, and Kimberlee Wheeler (1995). "Acoustic characteristics of American English vowels." In: *The Journal of the Acoustical society of America* 97.5, pp. 3099–3111 (cit. on p. 141).

Hino, Yasushi and Stephen J Lupker (1996). "Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts." In: *Journal of Experimental Psychology: Human Perception and Performance* 22.6, p. 1331 (cit. on p. 121).

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. (2012). "Deep neural networks for acoustic modeling in speech

recognition: The shared views of four research groups." In: *IEEE Signal processing magazine* 29.6, pp. 82–97 (cit. on p. 3).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780 (cit. on p. 26).

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed (2021). "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3451–3460 (cit. on pp. 133, 139).

Jacobs, Christiaan and Herman Kamper (2021a). "Multilingual Transfer of Acoustic Word Embeddings Improves When Training on Languages Related to the Target Zero-Resource Language." In: *Proc. Interspeech*, pp. 1549–1553. DOI: 10.21437/Interspeech.2021-461 (cit. on p. 149).

– (2021b). "Multilingual transfer of acoustic word embeddings improves when training on languages related to the target zero-resource language." In: *Interspeech* (cit. on pp. 124, 126).

Jacobs, Christiaan, Yevgen Matusevych, and Herman Kamper (2021). "Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation." In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 919–926 (cit. on p. 76).

Jansen, Aren and Benjamin Van Durme (2012). "Indexing raw acoustic features for scalable zero resource search." In: *Proc. Interspeech* (cit. on pp. 72, 97, 119).

Jansen, Aren, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous (2018). "Unsupervised learning of semantic audio representations." In: *Proc. ICASSP* (cit. on p. 124).

Johnson, Elizabeth K, Laurence Bruggeman, and Anne Cutler (2018). "Abstraction and the (misnamed) language familiarity effect." In: *Cognitive Science* 42.2, pp. 633–645 (cit. on p. 150).

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2017). "Billion-scale similarity search with GPUs." In: *IEEE Transactions on Big Data* (cit. on pp. 78, 107, 120).

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. (2017). "Google's multilingual neural machine translation system: Enabling

zero-shot translation." In: *Transactions of the Association for Computational Linguistics* 5, pp. 339–351 (cit. on p. 57).

Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever (2015). "An empirical exploration of recurrent network architectures." In: *International Conference on Machine Learning*, pp. 2342–2350 (cit. on p. 26).

Jurafsky, Dan (2003). "Probabilistic modeling in psycholinguistics: Linguistic comprehension and production." In: *Probabilistic linguistics* 21 (cit. on p. 73).

Jurafsky, Dan and James H. Martin (2000). "Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition." In: *Prentice Hall series in artificial intelligence* (cit. on pp. 21, 153).

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom (2014). "A convolutional neural network for modelling sentences." In: *arXiv preprint arXiv:1404.2188* (cit. on p. 23).

Kamper, H., W. Wang, and Karen Livescu (2016). "Deep convolutional acoustic word embeddings using word-pair side information." In: *Proc. ICASSP* (cit. on pp. 77, 120, 124).

Kamper, Herman (2019). "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models." In: *Proc. ICASSP* (cit. on pp. 75, 79, 105, 120).

Kamper, Herman, Micha Elsner, Aren Jansen, and Sharon Goldwater (2015). "Unsupervised neural network based feature extraction using weak top-down constraints." In: *Proc. ICASSP* (cit. on pp. 78, 120).

Kamper, Herman, Weiran Wang, and Karen Livescu (2016). "Deep convolutional acoustic word embeddings using word-pair side information." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4950–4954 (cit. on pp. 72, 97, 120).

Kavumba, Pride, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui (2019). "When choosing plausible alternatives, clever hans can be clever." In: *arXiv preprint arXiv:1911.00225* (cit. on p. 39).

Kenny, Patrick (2010). "Bayesian speaker verification with heavy-tailed priors." In: *Odyssey*, p. 14 (cit. on p. 38).

Kim, Yoon (2014). "Convolutional neural networks for sentence classification." In: *arXiv preprint arXiv:1408.5882* (cit. on p. 23).

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization." In: *Proc. ICLR* (cit. on pp. 78, 107, 125).

Klatt, Dennis H (1979). "Speech perception: A model of acoustic–phonetic analysis and lexical access." In: *Journal of phonetics* 7.3, pp. 279–312 (cit. on p. 4).

– (1989). "Review of selected models of speech perception." In: (cit. on p. 4).

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton (2019). "Similarity of neural network representations revisited." In: *International Conference on Machine Learning*. PMLR, pp. 3519–3529 (cit. on pp. 32, 88).

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience." In: *Frontiers in systems neuroscience* 2, p. 4 (cit. on pp. 31, 88, 100).

Ladd, D Robert (2011). "Phonetics in phonology." In: *The handbook of phonological theory*, pp. 348–373 (cit. on p. 17).

Lamel, Lori F and Jean-Luc Gauvain (1994). "Language identification using phone-based acoustic likelihoods." In: *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 1. IEEE, pp. I–293 (cit. on p. 38).

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (2020). "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499 (cit. on p. 149).

Lee, Leo Jingyu (2004). *Hidden dynamic models for speech processing applications*. Citeseer (cit. on p. 17).

Lehr-Spławiński, Tadeusz, Władysław Kuraszkiewicz, and Franciszek Sławski (1954). *Przegląd i charakterystyka języków słowiańskich*. Państwowe Wydawn. Naukowe (cit. on p. 55).

Lepori, Michael and R. Thomas McCoy (Dec. 2020). "Picking BERT's Brain: Probing for Linguistic Dependencies in Contextualized Embeddings Using Representational Similarity Analysis." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3637–3651. DOI: 10.18653/v1/2020.coling-main.325. URL: https://aclanthology.org/2020.coling-main.325 (cit. on pp. 31, 101).

Levin, Keith, Katharine Henry, Aren Jansen, and Karen Livescu (2013). "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings." In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (cit. on pp. 72, 97, 120).

Levow, Gina-Anne, Emily P Ahn, and Emily M Bender (2021). "Developing a Shared Task for Speech Processing on Endangered Languages." In: *Proceedings of the Workshop on Computational Methods for Endangered Languages*. Vol. 1, pp. 96–106 (cit. on p. 37).

Levow, Gina-Anne, Emily M Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, et al. (2017). "STREAMLInED challenges: Aligning research interests with shared tasks." In: *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 39–47 (cit. on p. 37).

Li, Haizhou and Bin Ma (2005). "A phonotactic language model for spoken language identification." In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 515–522 (cit. on p. 38).

Li, Haizhou, Bin Ma, and Kong Aik Lee (2013). "Spoken language recognition: from fundamentals to practice." In: *Proceedings of the IEEE* 101.5, pp. 1136–1159 (cit. on pp. 37, 39, 54, 56).

Li, Sirui, Qinya Zhang, Yunpeng Li, Guanyu Li, Senyan Li, and Shaoxuan Wang (2022). "Analyzing speaker information in self-supervised models to improve unsupervised speech recognition." In: *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*, pp. 1300–1305 (cit. on p. 150).

Liao, Hank (2013). "Speaker adaptation of context dependent deep neural networks." In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7947–7951 (cit. on p. 5).

Liu, Ollie, Hao Tang, and Sharon Goldwater (2023). "Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces." In: *ArXiv* abs/2305.12464 (cit. on p. 150).

Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno (2014). "Automatic language identification using deep neural networks." In: *2014 IEEE interna-*

tional conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 5337–5341 (cit. on pp. 38, 56).

Luce, Paul A. and Conor T. McLennan (2005). "Spoken Word Recognition: The Challenge of Variation." In: pp. 590–609 (cit. on pp. 4, 92, 95).

Maas, Andrew L, Stephen D Miller, Tyler M O'neil, Andrew Y Ng, and Patrick Nguyen (2012). "Word-level acoustic modeling with convolutional vector regression." In: *Proc. ICML Workshop Representation Learn* (cit. on p. 123).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov, pp. 2579–2605 (cit. on pp. 51, 60).

MacKain, Kristine S, Catherine T Best, and Winifred Strange (1981). "Categorical perception of English/r/and/l/by Japanese bilinguals." In: *Applied psycholinguistics* 2.4, pp. 369–390 (cit. on p. 96).

MacKay, David JC (2003). *Information theory, inference and learning algorithms*. Cambridge university press (cit. on p. 30).

Magnuson, James S, Heejo You, Sahil Luthra, Monica Li, Hosung Nam, Monty Escabi, Kevin Brown, Paul D Allopenna, Rachel M Theodore, Nicholas Monto, et al. (2020). "EARSHOT: A minimal neural network model of incremental human speech recognition." In: *Cognitive science* 44.4, e12823 (cit. on pp. 20, 31, 100).

Mahowald, Kyle, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson (2013). "Info/information theory: Speakers choose shorter words in predictive contexts." In: *Cognition* 126.2, pp. 313–318 (cit. on p. 30).

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell (Sept. 2017). "Learning Language Representations for Typology Prediction." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics (cit. on p. 57).

Mańczak, Witold (2004). *Przedhistoryczne migracje Słowian i pochodzenie języka staro-cerkiewno-słowiańskiego*. Nakładem Polskiej Akad. Umiejętności (cit. on p. 55).

Martinez, David, Oldřich Plchot, Lukáš Burget, Ondřej Glembek, and Pavel Matějka (2011). "Language recognition in ivectors space." In: *Twelfth annual conference of the international speech communication association* (cit. on p. 38).

Mateju, Lukas, Petr Cerva, Jindrich Zdánskỳ, and Radek Safarik (2018). "Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal." In: *Interspeech*, pp. 1803–1807 (cit. on pp. 38, 40, 44, 54, 57).

Matusevych, Yevgen, H. Kamper, Thomas Schatz, Naomi H Feldman, and S. Goldwater (2021). "A phonetic model of non-native spoken word processing." In: *Proc. EACL* (cit. on pp. 20, 73, 86, 96, 100, 120).

Matusevych, Yevgen, Herman Kamper, and Sharon Goldwater (2020a). "Analyzing Autoencoder-Based Acoustic Word Embeddings." In: *Bridging AI and Cognitive Science Workshop, ICLR* (cit. on pp. 76, 91, 96).

– (2020b). "Analyzing autoencoder-based acoustic word embeddings." In: *BAICS Workshop ICLR* (cit. on p. 73).

Matusevych, Yevgen, Thomas Schatz, Herman Kamper, Naomi Feldman, and Sharon Goldwater (2020). "Evaluating computational models of infant phonetic learning across languages." In: *Proc. CogSci* (cit. on pp. 73, 100, 120).

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech* (cit. on pp. 77, 106, 124).

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger (2018). "UMAP: Uniform Manifold Approximation and Projection." In: *Journal of Open Source Software* 3.29, p. 861 (cit. on p. 60).

Mehrer, Johannes, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann (2020). "Individual differences among deep neural network models." In: *Nature communications* 11.1, pp. 1–12 (cit. on p. 88).

Meng, Zhong, Zhuo Chen, Vadim Mazalov, Jinyu Li, and Yifan Gong (2017). "Unsupervised adaptation with domain separation networks for robust speech recognition." In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 214–221 (cit. on p. 42).

Meng, Zhong, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang (2018). "Speaker-invariant training via adversarial learning." In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5969–5973 (cit. on p. 5).

Merkx, Danny and Odette Scharenborg (2018). "Articulatory feature classification using convolutional neural networks." In: *Interspeech 2018*, pp. 2142–2146 (cit. on p. 23).

Metze, Florian, Xavier Anguera, Etienne Barnard, Marelie Davel, and Guillaume Gravier (2013). "The spoken web search task at MediaEval 2012." In: *Proc. ICASSP* (cit. on pp. 72, 97, 119).

Meylan, Stephan C. and Thomas L. Griffiths (2017). "Word forms - not just their lengths- are optimized for efficient communication." In: *ArXiv* abs/1703.01694 (cit. on p. 85).

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin (May 2018). "Advances in Pre-Training Distributed Word Representations." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1008 (cit. on p. 122).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems* 26 (cit. on p. 71).

Millet, Juliette, Ioana Chitoran, and Ewan Dunbar (2021). "Predicting non-native speech perception using the Perceptual Assimilation Model and state-of-the-art acoustic models." In: *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 661–673 (cit. on p. 96).

Mimno, David and Laure Thompson (Sept. 2017). "The strange geometry of skip-gram with negative sampling." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2873–2878. DOI: 10.18653/v1/D17-1308. URL: https://aclanthology.org/D17-1308 (cit. on p. 81).

Mirman, Daniel and James S Magnuson (2009). "Dynamics of activation of semantically similar concepts during spoken word recognition." In: *Memory & cognition* 37.7, pp. 1026–1039 (cit. on p. 121).

Miyawaki, Kuniko, James J Jenkins, Winifred Strange, Alvin M Liberman, Robert Verbrugge, and Osamu Fujimura (1975). "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English." In: *Perception & Psychophysics* 18.5, pp. 331–340 (cit. on p. 96).

Mohamed, Abdelrahman (2014). "Deep Neural Network Acoustic Models for ASR." PhD thesis. University of Toronto (cit. on p. 45).

Mohamed, Abdelrahman et al. (2022). "Self-Supervised Speech Representation Learning: A Review." In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1179–1210. DOI: 10.1109/JSTSP.2022.3207050 (cit. on p. 133).

Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. URL: https://phoible.org/ (cit. on p. 145).

Mu, Jiaqi and Pramod Viswanath (2018). "All-but-the-Top: Simple and Effective Postprocessing for Word Representations." In: *International Conference on Learning Representations* (cit. on p. 81).

Müller, Meinard (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer (cit. on p. 79).

Myers, Cory, Lawrence Rabiner, and Andrew Rosenberg (1980). "An investigation of the use of dynamic time warping for word spotting and connected speech recognition." In: *ICASSP'80. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5. IEEE, pp. 173–177 (cit. on pp. 72, 97).

Nalepa, Jerzy (1968). *Słowiańszczyzna północno-zachodnia: podstawy jedności i jej rozpad*. Vol. 25. Państwowe Wydawn. Naukowe; Oddz. w Poznaniu (cit. on p. 55).

Nematzadeh, Aida, Stephan C. Meylan, and Thomas L. Griffiths (2017). "Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words." In: *Cognitive Science* (cit. on p. 71).

Nguyen, Tu Anh, Benoit Sagot, and Emmanuel Dupoux (2022). "Are discrete units necessary for spoken language modeling?" In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1415–1423 (cit. on p. 134).

Nili, Hamed, Alexander Walther, Arjen Alink, and Nikolaus Kriegeskorte (2020). "Inferring exemplar discriminability in brain representations." In: *Plos one* 15.6, e0232551 (cit. on p. 84).

Nouza, Jan, Radek Safarik, and Petr Cerva (2016). "ASR for South Slavic Languages Developed in Almost Automated Way." In: *Interspeech 2016*, pp. 3868–3872 (cit. on p. 44).

Novotná, Petra and Václav Blažek (2007). "Glottochronology and its application to the Balto-Slavic languages." In: *Baltistica* 42.2, pp. 185–210 (cit. on p. 65).

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding." In: *arXiv preprint arXiv:1807.03748* (cit. on p. 133).

– (2018). "Representation learning with contrastive predictive coding." In: *arXiv preprint arXiv:1807.03748* (cit. on p. 93).

Östling, Robert and Jörg Tiedemann (Apr. 2017). "Continuous multilinguality with language vectors." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics (cit. on p. 57).

Palaz, Dimitri, Mathew Magimai- Doss, and Ronan Collobert (2015). "Convolutional neural networks-based continuous speech recognition using raw speech signal." In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4295–4299 (cit. on p. 23).

Pallier, Christophe, Anne Christophe, and Jacques Mehler (1997). "Language-specific listening." In: *Trends in Cognitive Sciences* 1.4. Dolphin Communication and Cognition, pp. 129–132. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(97)01044-9. URL: https://www.sciencedirect.com/science/article/pii/S1364661397010449 (cit. on p. 96).

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015). "Librispeech: an asr corpus based on public domain audio books." In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5206–5210 (cit. on p. 77).

Papadimitriou, Isabel and Dan Jurafsky (2020). "Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6829–6839 (cit. on p. 150).

Pasad, Ankita, Ju-Chieh Chou, and Karen Livescu (2021). "Layer-wise Analysis of a Self-supervised Speech Representation Model." In: *arXiv e-prints*, arXiv–2107 (cit. on pp. 5, 28).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. (2019). "Pytorch: An imperative style, high-performance deep learning library." In: *Proc. NeuRIPS* (cit. on pp. 78, 107).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (cit. on pp. 71, 122).

Peperkamp, Sharon, Inga Vendelin, and Emmanuel Dupoux (2010). "Perception of predictable stress: A cross-linguistic investigation." In: *Journal of Phonetics* 38.3, pp. 422–430 (cit. on p. 96).

Pereira, Francisco, Samuel Gershman, Samuel Ritter, and Matthew Botvinick (2016). "A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data." In: *Cognitive neuropsychology* 33.3-4, pp. 175–190 (cit. on p. 71).

Peters, Ben, Jon Dehdari, and Josef van Genabith (Sept. 2017). "Massively Multilingual Neural Grapheme-to-Phoneme Conversion." In: *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Copenhagen, Denmark: Association for Computational Linguistics (cit. on p. 57).

Peterson, Gordon E and Harold L Barney (1952). "Control methods used in a study of the vowels." In: *The Journal of the acoustical society of America* 24.2, pp. 175–184 (cit. on p. 141).

Peterson, Joshua C, Jordan W Suchow, Krisha Aghi, Alexander Y Ku, and Thomas L Griffiths (2018). "Capturing human category representations by sampling in deep feature spaces." In: Proceedings of the 40th Annual Conference of the Cognitive Science Society. (cit. on p. 65).

Pianka, Włodzimierz and Emil Tokarz (2000). *Gramatyka konfrontatywna języków słowiańskich. 1 (2000)*. Śląsk (cit. on p. 55).

Piantadosi, Steven T, Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication." In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529 (cit. on p. 30).

Pierrehumbert, Janet B (2002). "Word-specific phonetics." In: *Laboratory Phonology 7* 4.1, p. 101 (cit. on p. 19).

Pimentel, Tiago, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell (Nov. 2021). "A surprisal–duration trade-off across and within the world's languages." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 949–962. DOI:

10.18653/v1/2021.emnlp-main.73. URL: https://aclanthology.org/2021.emnlp-main.73 (cit. on p. 134).

Pimentel, Tiago, Brian Roark, and Ryan Cotterell (Jan. 2020). "Phonotactic Complexity and Its Trade-offs." In: *Transactions of the Association for Computational Linguistics* 8, pp. 1–18. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00296. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00296/1923388/tacl_a_00296.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00296 (cit. on p. 135).

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001 (cit. on p. 149).

Pisoni, David B (1993). "Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning." In: *Speech communication* 13.1-2, pp. 109–125 (cit. on p. 4).

Pisoni, David B and Susannah V Levi (2012). "Representations and representational specificity in speech perception and spoken word recognition." In: *The Oxford Handbook of Psycholinguistics*. Oxford University Press (cit. on p. 19).

Port, Robert (2007). "How are words stored in memory? Beyond phones and phonemes." In: *New ideas in psychology* 25.2, pp. 143–170 (cit. on p. 19).

Rabinovich, Ella, Noam Ordan, and Shuly Wintner (July 2017). "Found in Translation: Reconstructing Phylogenetic Language Trees from Translations." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics (cit. on pp. 57, 66).

Raphael, Lawrence J (2021). "Acoustic cues to the perception of segmental phonemes." In: *The handbook of speech perception*, pp. 603–631 (cit. on p. 144).

Räsänen, Okko, Tasha Nagamine, and Nima Mesgarani (2016a). "Analyzing distributional learning of phonemic categories in unsupervised deep neural networks." In: *Annual Conference of the Cognitive Science Society (CogSci). Cognitive Science Society (U.S.). Conference* 2016, pp. 1757–1762 (cit. on p. 100).

– (2016b). "Analyzing distributional learning of phonemic categories in unsupervised deep neural networks." In: *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*. Vol. 2016. NIH Public Access, p. 1757 (cit. on p. 138).

Rathi, Neil, Michael Hahn, and Richard Futrell (Nov. 2021). "An Information-Theoretic Characterization of Morphological Fusion." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10115–10120. DOI: 10.18653/v1/2021.emnlp-main.793. URL: https://aclanthology.org/2021.emnlp-main.793 (cit. on p. 135).

Rohlicek, Jan Robin (1995). "Word spotting." In: *Modern Methods of Speech Processing*. Springer, pp. 123–157 (cit. on pp. 72, 97).

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2022). "Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold." In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2340–2354 (cit. on p. 117).

Rudman, William, Nate Gillman, Taylor Rayne, and Carsten Eickhoff (May 2022). "IsoScore: Measuring the Uniformity of Embedding Space Utilization." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 3325–3339. DOI: 10.18653/v1/2022.findings-acl.262. URL: https://aclanthology.org/2022.findings-acl.262 (cit. on p. 82).

Sainath, Tara N, Brian Kingsbury, George Saon, Hagen Soltau, Abdel-rahman Mohamed, George Dahl, and Bhuvana Ramabhadran (2015). "Deep convolutional neural networks for large-scale speech tasks." In: *Neural networks* 64, pp. 39–48 (cit. on p. 23).

San, Nay, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. (2021). "Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages." In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 1094–1101 (cit. on p. 72).

Saon, George, Hagen Soltau, David Nahamoo, and Michael Picheny (2013). "Speaker adaptation of neural network acoustic models using i-vectors." In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, pp. 55–59 (cit. on p. 5).

Sawicka, Irena (1991). "Problems of the phonetic typology of the Slavic languages." In: *Studies in the Phonetic Typology of the Slavic Languages. Warszawa: Slaw-*

istyczny Osrodek Wydawniczy przy Instytucie Slowianoznawstwa PAN (cit. on p. 55).

Scharenborg, Odette and Lou Boves (2010). "Computational modelling of spoken-word recognition processes: Design choices and evaluation." In: *Pragmatics & Cognition* 18.1, pp. 136–164 (cit. on p. 92).

Scharenborg, Odette, Nikki van der Gouw, Martha Larson, and Elena Marchiori (2019). "The representation of speech in deep neural networks." In: *International Conference on Multimedia Modeling*. Springer, pp. 194–205 (cit. on pp. 5, 100).

Schatz, Thomas and Naomi H Feldman (2018). "Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception." In: *Proceedings of the Conference on Cognitive Computational Neuroscience* (cit. on pp. 66, 96, 100).

Schneider, Steffen, Alexei Baevski, Ronan Collobert, and Michael Auli (2019a). "wav2vec: Unsupervised Pre-training for Speech Recognition." In: *Interspeech* (cit. on p. 133).

– (2019b). "wav2vec: Unsupervised pre-training for speech recognition." In: *arXiv preprint arXiv:1904.05862* (cit. on p. 93).

Schultz, Tanja, Ngoc Thang Vu, and Tim Schlippe (2013). "Globalphone: A multilingual text & speech database in 20 languages." In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 8126–8130 (cit. on pp. 40, 44, 106, 124).

Serva, Maurizio and Filippo Petroni (2008). "Indo-European languages tree by Levenshtein distance." In: *EPL (Europhysics Letters)* 81.6, p. 68005 (cit. on p. 66).

Settle, Shane, Kartik Audhkhasi, Karen Livescu, and Michael Picheny (2019). "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition." In: *Proc. ICASSP* (cit. on p. 78).

Settle, Shane, Keith Levin, Herman Kamper, and Karen Livescu (2017). "Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings." In: *Proc. Interspeech* (cit. on p. 120).

Settle, Shane and Karen Livescu (2016a). "Discriminative Acoustic Word Embeddings: Recurrent Neural Network-Based Approaches." In: *Proc. IEEE Spoken Language Technology Workshop (SLT)* (cit. on pp. 77, 124).

Settle, Shane and Karen Livescu (2016b). "Discriminative acoustic word embeddings: Recurrent neural network-based approaches." In: *IEEE Spoken Language Technology Workshop (SLT)* (cit. on pp. 72, 97, 120).

Shah, Jui, Yaman Kumar Singla, Changyou Chen, and Rajiv Ratn Shah (2021). "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure." In: *arXiv preprint arXiv:2101.00387* (cit. on p. 5).

Shannon, Claude Elwood (2001). "A mathematical theory of communication." In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1, pp. 3–55 (cit. on p. 134).

Shen, Gaofei, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała (2023). "Wave to Syntax: Probing spoken language models for syntax." In: *arXiv preprint arXiv:2305.18957* (cit. on p. 5).

Shen, Peng, Xugang Lu, Sheng Li, and Hisashi Kawai (2018). "Feature Representation of Short Utterances Based on Knowledge Distillation for Spoken Language Identification." In: *Interspeech*, pp. 1813–1817 (cit. on p. 38).

Shinohara, Yusuke (2016). "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition." In: *Interspeech*. San Francisco, CA, USA, pp. 2369–2372 (cit. on p. 42).

Shon, Suwon, Ahmed Ali, and James Glass (2018). "Convolutional Neural Network and Language Embeddings for End-to-End Dialect Recognition." In: *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pp. 98–104 (cit. on pp. 38, 45, 54, 57).

Sicherman, Amitay and Yossi Adi (2023). "Analysing Discrete Self Supervised Speech Representation for Spoken Language Modeling." In: *arXiv preprint arXiv:2301.00591* (cit. on p. 134).

Skirgård, Hedvig, Seán G Roberts, and Lars Yencken (2017). "Why are some languages confused for others? Investigating data from the Great Language Game." In: *PloS one* 12.4 (cit. on pp. 38, 53, 54, 62).

Sriram, Anuroop, Heewoo Jun, Yashesh Gaur, and Sanjeev Satheesh (2018). "Robust speech recognition using generative adversarial networks." In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5639–5643 (cit. on p. 4).

Strain, Eamon, Karalyn Patterson, and Mark S Seidenberg (1995). "Semantic effects in single-word naming." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21.5, p. 1140 (cit. on p. 121).

Strunk, Jan, Florian Schiel, and Frank Seifart (2014). "Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3940–3947 (cit. on p. 77).

Su, Hang and Steven Wegmann (2016). "Factor Analysis Based Speaker Verification Using ASR." In: *Interspeech*, pp. 2223–2227 (cit. on p. 38).

Tachbelie, M. Y., A. Abulimiti, S. T. Abate, and T. Schultz (2020). "DNN-Based Speech Recognition for Globalphone Languages." In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8269–8273 (cit. on p. 44).

Tatman, Rachael and Conner Kasten (2017). "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions." In: *Interspeech*, pp. 934–938 (cit. on p. 4).

Titus, Andrew, Jan Silovsky, Nanxin Chen, Roger Hsiao, Mary Young, and Arnab Ghoshal (2020). "Improving Language Identification for Multilingual Speakers." In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8284–8288 (cit. on p. 38).

Toshniwal, Shubham, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao (2018). "Multilingual speech recognition with a single end-to-end model." In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 4904–4908 (cit. on p. 4).

Tripathi, Aditay, Aanchan Mohan, Saket Anand, and Maneesh Singh (2018). "Adversarial learning of raw speech features for domain invariant speech recognition." In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5959–5963 (cit. on p. 4).

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (cit. on pp. 116, 127).

Van Heuven, Vincent J (2008). "Making sense of strange sounds:(Mutual) intelligibility of related language varieties. A review." In: *International journal of humanities and arts computing* 2.1-2, pp. 39–62 (cit. on pp. 20, 66, 96).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in neural information processing systems* 30 (cit. on pp. 27, 29).

Waibel, Alex, Petra Geutner, L Mayfield Tomokiyo, Tanja Schultz, and Monika Woszczyna (2000). "Multilinguality in speech and spoken language systems." In: *Proceedings of the IEEE* 88.8, pp. 1297–1313 (cit. on p. 37).

Wang, Yue, Michelle M Spence, Allard Jongman, and Joan A Sereno (1999). "Training American listeners to perceive Mandarin tones." In: *The Journal of the acoustical society of America* 106.6, pp. 3649–3658 (cit. on p. 96).

Wang, Zhirong, Tanja Schultz, and Alex Waibel (2003). "Comparison of acoustic model adaptation techniques on non-native speech." In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* Vol. 1. IEEE, pp. I–I (cit. on p. 4).

Ward, Joe H (1963). "Hierarchical grouping to optimize an objective function." In: *Journal of the American statistical association* 58.301, pp. 236–244 (cit. on pp. 65, 110, 143).

Weber, Andrea and Odette Scharenborg (2012). "Models of spoken-word recognition." In: *Wiley Interdisciplinary Reviews: Cognitive Science* 3.3, pp. 387–401 (cit. on p. 92).

Wells, Dan, Hao Tang, and Korin Richmond (2022). "Phonetic Analysis of Self-supervised Representations of English Speech." In: *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*. ISCA, pp. 3583–3587 (cit. on pp. 134, 145).

Wu, John, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass (2020). "Similarity Analysis of Contextual Word Representation Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4638–4655 (cit. on pp. 31, 101).

Wu, Shijie, Ryan Cotterell, and Timothy O'Donnell (July 2019). "Morphological Irregularity Correlates with Frequency." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5117–5126. DOI: 10.18653/v1/P19-1505. URL: https://aclanthology.org/P19-1505 (cit. on p. 135).

Żelasko, Piotr, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak (2020). "That sounds familiar: an analysis of phonetic

representations transfer across languages." In: *arXiv preprint arXiv:2005.08118* (cit. on p. 149).

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang (2018). "The unreasonable effectiveness of deep features as a perceptual metric." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595 (cit. on p. 65).

Zhang, Yaodong and James R Glass (2009). "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams." In: *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (cit. on pp. 72, 97, 119).

Zhang, Ye and Byron Wallace (2015). "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." In: *arXiv preprint arXiv:1510.03820* (cit. on p. 23).

Zhang, Ying, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, and Aaron Courville (2016). "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks." In: *Interspeech 2016*, pp. 410–414 (cit. on p. 4).

Zhuang, Jie, Billi Randall, Emmanuel A Stamatakis, William D Marslen-Wilson, and Lorraine K Tyler (2011). "The interaction of lexical semantics and cohort competition in spoken word recognition: an fMRI study." In: *Journal of Cognitive Neuroscience* 23.12, pp. 3778–3790 (cit. on p. 121).

Declaration

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, June 25, 2024

Badr Mohammed Badr Abdullah