IT-UNIVERSITETET I KØBENHAVN

Department of Computer Science

Ph.D. Thesis

Marija Stepanović

Phonetic Vowel Representations for Cross-Lingual Automatic Speech Recognition

This thesis has been submitted to the Ph.D. School of the IT University of Copenhagen on September 30, 2024.

Committee

Advisor	
Dr. Christian Hardmeier	IT-Universitetet i København
Former Advisor	
Prof. Dr. Barbara Plank	Ludwig-Maximilians-Universität München, IT-Universitetet i København
Members	
Dr. Anna Rogers	IT-Universitetet i København
Prof. Dr. Laurent Besacier	Naver Labs Europe
Prof. Dr. Torbjørn Karl Svendsen	Norges teknisk-naturvitenskapelige universitet

Abstract

Multilingual automatic phone recognition models can learn language-independent pronunciation patterns from large volumes of spoken data and recognize them across languages. This potential can be harnessed to improve speech technologies for under-resourced languages. However, these models are typically trained on phonological representations of speech sounds, which do not necessarily reflect the phonetic realization of speech. A mismatch between a phonological symbol and its phonetic realizations can lead to phone confusions and reduce performance.

This thesis introduces a formant-based vowel categorization method aimed at improving cross-lingual vowel recognition by uncovering a vowel's phonetic quality from its formant frequencies, and reorganizing the vowel categories in a multilingual speech corpus to increase their consistency across languages. The work investigates vowel categories obtained from a trilingual speech corpus of Danish, Norwegian, and Swedish using four categorization techniques. Cross-lingual phone recognition experiments reveal that uniting the vowel categories of different languages into a shared set of formant-based categories can improve cross-lingual recognition of the shared vowels, but also interfere with recognition of vowels not present in one or more training languages. Nevertheless, improved recognition of individual vowels can translate to improvements in overall phone recognition on languages unseen during training.

To assess their wider applicability in automatic speech recognition (ASR), the investigated vowel representations are also evaluated as part of pronunciation lexicons used in hybrid ASR systems. These experiments, however, do not reveal many conclusive patterns, which demonstrates that hybrid systems are more robust to divergence in pronunciation from the phonological norm. Nonetheless, a qualitative analysis of phone predictions shows that the models trained on formant-based vowel representations can infer the distinctive vowel qualities of an unseen language, especially when their vowel set and training data align with the vowel system of the target language. This indicates that formant-based vowel representations could provide useful information for tasks where phonological description is preferred.

Resumé

Flersproglige modeller til automatisk genkendelse af sproglyde kan lære fonetiske mønstre fra store mængder taledata og genkende dem på tværs af sprog. Dette kan bruges til at forbedre sprogteknologier for sprog der kun har få dataressourcer. Disse modeller er imidlertidig trænet på fonologiske repræsentationer af sproglyde, som ikke nødvendigvis afspejler den fonetiske realisering. En uoverensstemmelse mellem fonologiske symboler og den fonetiske realisering kan medføre en sammenblanding af sproglyde og forringe modellen.

Denne afhandling præsenterer en formant-baseret kategorisering af vokallyde med det formål at forbedre vokallydsgenkendelse på tværs af sprog. Dette sker ved at undersøge vokallydenes fonetiske karaktertræk ud fra formantfrekvenser og derudfra omorganisere vokallydskategorier i et flersprogligt talekorpus for at forøge ensartetheden af vokallydene på tværs af sprog. Denne afhandling undersøger fire forskellige måder at kategorisere vokallyde på i et tresprogligt korpus med dansk, norsk og svensk. I eksperimenter med fongenkendelse på tværs af sprog kombineres vokalydene fra de forskellige sprog til et fælles sæt af formant-baserede kategorier. Disse eksperimenter viser, at man kan forbedre genkendelsen af sprogenes fælles vokaler, men det indvirker også negativt på hvor godt vokallyde, der ikke findes på alle sprogene, bliver genkendt. Ikke desto mindre bliver resultatet en forbedret fongenkendelse på sprog, som modellen ikke er trænet på.

De undersøgte vokallydsrepræsentationer bliver også evalueret som del af et udtaleleksikon brugt i hydridsystemer til automatisk talegenkendelse (ASR) for at vurdere deres anvendelighed for ASR-systemer. Disse eksperimenter viser at imidlertid ikke mange tydelige mønstre, hvilket betyder at hybride systemer er mere robuste over for udtaleafvigelser i forhold til den fonologiske norm. Ikke desto mindre viser en kvalitativ analyse af modellens sproglydsgenkendelse, at modeller trænet på formant-baserede vokallydsrepræsentationer kan udlede distinkte vokallydskaraktertræk fra et nyt sprog, særligt hvis modellens vokallydssystem og træningsdata stemmer overens med det nye sprog. Dette indikerer, at formant-baserede vokallydsrepræsentationer kunne give nyttig information til opgaver, hvor fonologisk beskrivelse foretrækkes.

Acknowledgements

[To be added after defense]

Declaration of Work

I, Marija Stepanović, declare that this thesis – submitted in partial fulfillment of the requirements for the conferral of Ph.D. from the IT University of Copenhagen – is solely my own work unless otherwise referenced or attributed. Neither the thesis nor its content have been submitted (or published) for qualifications at another academic institution.

– Marija Stepanović

Table of Contents

Part I

Introduction

Motivation 1

In recent years, the advancements made in automatic speech recognition (ASR) and speech technologies centered on ASR have been nothing short of remarkable. In particular, self-supervised learning has enabled general speech representation learning without the need for large-scale manually transcribed data. This has led to substantial progress in many of the essential aspects of ASR, including increased accuracy, robustness, language diversity, contextual language understanding, and integration with other technologies. These developments have, in turn, given rise to a wide range of practical applications, which have significantly impacted our lives, from improving accessibility to revolutionizing industries.

However, these improvements in performance have mostly been restricted to the languages for which large amounts of annotated data are available. On the other hand, for low-resource languages, such as indigenous and minority languages, regional varieties, and unwritten languages, ASR performance generally remains much lower (Scharenborg et al., 2020). Unequal access to effective ASR technologies can have adverse effects on the speakers of low-resource languages and their communities by worsening social inequalities, economic disadvantages, and cultural marginalization, as well as limiting their opportunities for language preservation. It is, thus, crucial to prioritize the development of ASR technologies for low-resource languages in order to bridge the digital and cultural divide.

1.1 Increasing the Linguistic Diversity of ASR Systems

The advent of self-supervised learning and large pre-trained language and speech models has undeniably made it easier to expand the linguistic diversity of ASR technologies. This is because large pre-trained models can be fine-tuned on the target task with target domain data.

For ASR, fine-tuning approaches generally require less data, but training ASR systems that perform well for languages with very little data remains challenging (Bartelds et al., 2023).

1.1.1 Challenges

The main obstacle to developing accurate and robust ASR models for low-resource languages is the scarcity of linguistic resources in the target language. Examples of such resources include annotated speech corpora, which consist of audio recordings paired with their corresponding orthographic or phonetic transcriptions, pronunciation lexicons, language models, and natural language processing tools, such as tokenizers and grapheme-to-phoneme converters. In particular, transcribing speech data can be a complex and arduous task. Depending on the speech domain, audio quality, and type of transcription, it could take over 50 work hours to manually transcribe an hour of recorded speech (Strik and Cucchiarini, 2014). Therefore, collecting and manually transcribing large volumes of speech data, at the scale required to power modern ASR systems and match their performance on high-resource languages, is no longer feasible. To address this challenge, ASR researchers and developers are exploring techniques such as data collection and automatic creation of speech corpora from public data sources, automatic phonetic transcription, and transfer learning from related languages to improve ASR performance for low-resource languages.

Namely, when developing ASR models for under-resourced languages, multilingual and cross-lingual models that leverage phonological representations of speech sounds have shown promise in learning language-independent pronunciation patterns from higher-resource languages and recognizing them cross-linguistically in low- and zero-resource scenarios (Żelasko et al., 2020; Feng et al., 2021; Li et al., 2020c; Żelasko et al., 2022; Xu et al., 2022). However, their phone recognition rates on unseen languages are still far from those achieved on languages seen during training (Gao et al., 2021; Xu et al., 2022).

Apart from the lack of labeled speech data in the target language, another possible explanation for the relatively poor results of such models could be the lack of multilingual speech data transcribed

phonetically using a unified cross-linguistically consistent system, such as the one provided by the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999), in which phone symbols correspond to their articulatory description. As a result, these models have usually relied on combined monolingual phonological systems (Żelasko et al., 2020; Xu et al., 2022), which are rarely consistent across languages due to differences in phonological inventories, phonological notation, and transcription conventions (Laver, 1994, p. 549). Moreover, even when designed to be consistent with the IPA, monolingual phonological systems are typically based on the canonical pronunciation forms from a dominant language variety, which means they do not take account of all the variation in speech, such as allophonic, regional, or socioeconomic variation (Laver, 1994, p. 551).

Finally, evaluating multilingual phone recognition models presents unique challenges due to the aforementioned diversity of languages, notational and phonetic variation, and the potential for cross-lingual interference. For example, pooling training languages and their different phonological systems together might result in a model that predicts tones or vowel length in a target language that does not distinguish these features (Żelasko et al., 2020). Likewise, it might result in a model that fails to recognize target phones that were not found in its training data. Ultimately, if we do not have ground-truth phonetic annotations for the evaluated language, how can we assess whether the predicted phones are phonetically and phonologically relevant?

1.1.2 Applications

1.1.2.1 Inclusivity

Multilingual and cross-lingual ASR play a crucial role in promoting inclusivity and fostering a just and equitable society. When people are able to interact with technologies in their native language, they can access relevant information, be more efficient in their work and daily lives, feel empowered to participate in important activities and conversations, and in general connect with the world and community around them.

1.1.2.2 Accessibility

ASR has a crucial role in making technology more accessible for people with disabilities, particularly those who have difficulty speaking, typing, or reading. Multilingual and cross-lingual ASR can make assistive technologies more effective and tailor them to the specific needs of individuals from diverse linguistic and cultural backgrounds, ensuring that everyone has equal access to the tools and resources they need to thrive. This is particularly important for people with disabilities who already face challenges due to language or social isolation.

1.1.2.3 Scientific Research and Innovation

Multilingual and cross-lingual ASR play a vital role in advancing scientific research and innovation. By breaking down language barriers and enabling researchers from diverse linguistic backgrounds to collaborate more effectively, ASR facilitates the exchange of ideas and information across cultures. This, in turn, leads to the development of more innovative and comprehensive research projects. Additionally, increased diversity and accessibility provide opportunities to analyze large-scale datasets in multiple languages, providing valuable insights and supporting interdisciplinary research.

1.1.2.4 Language and Cultural Preservation

Multilingual and cross-lingual ASR are important for language and cultural preservation. Being able to use speech technologies in one's native language facilitates access to information and education, which are crucial for language preservation efforts. Furthermore, by documenting endangered languages and preserving spoken language data, ASR helps to ensure that these languages and their associated cultures are not lost to time. Additionally, ASR can be used to analyze and understand the nuances of different languages, providing valuable insights into their history, structure, and cultural significance. This helps to preserve the richness and diversity of human language and culture, and potentially reveal general insights into language and human nature.

1.1.2.5 Economic Development

Multilingual and cross-lingual ASR are beneficial for economic development. By enabling access to information and education in one's native language, multilingual speech technologies can help individuals learn new job skills, find employment, and improve financial stability. Additionally, by breaking down language barriers and facilitating communication between people from diverse backgrounds, multilingual ASR can foster international trade, attract foreign investment, and promote tourism.

1.2 Focus Areas, Objectives, and Scope

Due to the high complexity and diversity of methods that have been developed over the years to increase the linguistic diversity of ASR systems and improve their cross-lingual transfer while limiting interference, it would be impossible to encompass all of them in this thesis. We have, therefore, singled out three specific focus areas which are centered on the challenges outlined in the previous section. In this section, we define each focus area, describe how we approach them in our research, and list the research questions we seek to answer.

1.2.1 Phonetic Vowel Representations

The creation of phonetic vowel representations that can be recognized in unseen languages is the main focus area of the thesis and one that is explored in the most depth. This is because vowels are particularly prone to phonetic variation and notational inconsistencies (Labov et al., 2005; Tanner et al., 2022). As a result, phone errors involving vowels constitute a large portion of the phone error rates of multilingual and cross-lingual phone recognition models.

Different languages have different vowel inventories where each vowel category is represented as a discrete phonological symbol (Ladefoged and Maddieson, 1990). Different vowel categories can have wide and often overlapping ranges of realizations, which can result in two languages or dialects using the same phonological symbol for

two phonetically distant sets of vowel realizations,[1] or, vice versa, using multiple different symbols to denote overlapping ranges of vowel realizations.[2] Moreover, vowels can exhibit various additional features, such as nasalization, rhotacization, lengthening, tone, stress, etc., which might be contrastive in one language, but not another. Determining which of these features are language-independent, and how they can be captured and transferred to unseen languages is crucial for minimizing cross-lingual interference.

Fortunately, the periodicity and resonance of vowels, which can be measured reliably from the speech signal as vowel formants (Catford, 2001, p. 153), make them amenable to comparative cross-linguistic studies that could be used to both improve notational consistency and incorporate phonetic variation. Leveraging these characteristics of vowels, we propose a formant-based vowel categorization method for increasing the consistency of phonetic vowel representations used in multi- and cross-lingual ASR. We hypothesize that the new formant-based vowel categories would reduce cross-lingual vowel confusions that stem from a mismatch between a vowel's phonological symbol and its phonetic manifestations. We believe that this could lead to lower phone error rates on unseen languages, including their non-standard regional dialects.

More specifically, we investigate two approaches to formant-based vowel categorization: *formant-based vowel categorization with language- specific vowel sets*, and *formant-based vowel categorization with a language-universal vowel set.* In both approaches, the investigated vowel categories are obtained from a trilingual speech corpus of Scandinavian languages: Danish, Norwegian, and Swedish.

1.2.1.1 Formant-Based Vowel Categorization with Language- Specific Vowel Sets

Formant-based vowel categorization with language-specific vowel sets reorganizes the vowel categories of the source languages based on their

[1] For example, Danish /a/ (Grønnum, 1998) vs. Bosnian-Croatian-Montenegrin-Serbian (BCMS) /a/ (Landau et al., 1995)

[2] For example, Danish /e, ɛ, a/ (Grønnum, 1998) vs. BCMS /e/ (Landau et al., 1995)

formant frequencies while preserving their original language-specific vowel sets. This means that, for example, if a source language has 14 distinctive vowel categories, after formant-based categorization, it will still distinguish the same 14 vowel categories, but the distribution of vowel tokens across the categories will change.

We present the rationale and methodology behind the creation of language-specific formant-based vowel categories in Chapter 10. The effect of this vowel categorization approach on the cross-lingual phone recognition on the three Scandinavian languages is presented in Chapter 11. Additionally, we break down the trained cross-lingual models' performance by dialect region and examine how our vowel categorization methods affect cross-lingual phone recognition on non-standard regional dialects. This will highlight whether our approach is particularly effective for under-resourced speech varieties. Finally, we investigate for which vowels our approach is most effective. Although no longer considered low-resource, the three Scandinavian languages comprise a diverse trilingual corpus suitable for experiments in multilingual and cross-lingual phonetic transfer.

Here, we ask the following main research question:

RQ1 Can we derive phonetic vowel representations, which are consistent across languages, from the measurements of vowel formant frequencies using *language-specific* vowel categories?

We decompose this larger question into:

RQ1.1 Can formant-based vowel categorization with language-specific vowel sets improve cross-lingual phone recognition on Danish, Norwegian, and Swedish?

RQ1.2 Does adding more fine-tuning data further improve the cross-lingual phone recognition?

RQ1.3 Can formant-based vowel categorization improve cross-lingual phone recognition on under-represented language varieties, such as regional dialects?

RQ1.4 How does formant-based vowel categorization affect individual vowel predictions in cross-lingual phone recognition? Does it reduce some vowel confusions and which ones?

1.2.1.2 Formant-Based Vowel Categorization with a Language-Universal Vowel Set

Formant-based vowel categorization with a language-universal vowel set converts the language-specific vowel sets of the different training languages into a single unified set of vowel categories shared universally by all training languages. Unlike language-specific categorization, language-universal categorization changes both the size of the vowel sets of the source languages and the distribution of vowel tokens across vowel categories. This means that, after the categorization, all source languages will have the same vowel set, regardless of their originally distinctive vowels categories.

We present the rationale and methodology behind the creation of language-universal formant-based vowel categories in Chapter 13. The effect of this vowel categorization approach on the cross-lingual phone recognition on various evaluation languages and speech domains is presented in Chapter 14. Moreover, we examine the cross-lingual models' predictions on individual vowels and how they relate to the vowel systems of the different evaluation languages. This will demonstrate whether formant-based vowel representations can transfer to unseen languages, such as low-resource and typologically distant languages. It will also tell us which languages and vowel systems they are more aligned with.

Here, we ask the following main research question:

RQ2 Can we derive phonetic vowel representations, which are consistent across languages, from the measurements of vowel formant frequencies using *language-universal* vowel categories?

We decompose this larger question into:

RQ2.1 Can formant-based vowel categorization with a language-universal vowel set improve cross-lingual phone recognition on a diverse set of languages and speech domains?

RQ2.2 Can formant-based vowel categorization with a language-universal vowel set improve cross-lingual vowel recognition on different languages, including low-resource and typologically distant languages?

RQ2.3 How do the formant-based vowel representations obtained from a corpus of Scandinavian languages relate to the vowel systems of the different evaluation languages?

RQ2.4 Can we use phone recognition models trained on language-universal formant-based vowel categories to infer the vowel inventory of an unseen language?

Both approaches to vowel categorization entail the estimation and normalization of vowel formants from a phonemically transcribed and aligned speech corpus, followed by new categorizations of vowel phones based on their location in the vowel space. The obtained vowel phones are then inserted into the original phonetic utterance transcripts in place of their canonical pronunciations. Finally, we evaluate the new representations on a cross-lingual phone recognition task and investigate their potential for cross-lingual transfer to languages and dialects unseen during training.

It should be noted that our study is focused exclusively on the major features of the phonetic quality of monophthong vowels, i.e. height, backness, and lip rounding, as these are distinctive in most

languages, including all languages studied in this thesis, and can be directly associated with vowel formants (Ladefoged and Maddieson, 1990). Minor features of vowel quality, such as nasalization, pharyngealization, rhotacization, phonation, length, diphthongization, and tone, are not clearly distinctive in all three languages in this study, and thus ignored. For example, Danish and Swedish do not have a clear-cut distinction between diphthongs and monophthong vowel-vowel or vowel-consonant sequences (Grønnum, 1998; Riad, 2014).

1.2.2 Extrinsic Evaluation of Phonetic Vowel Representations

Extrinsic evaluation of phonetic vowel representations is a secondary focus area in the thesis. It aims to investigate whether the vowel representations obtained using the formant-based vowel categorization methods developed in the thesis can be used to recognize words in downstream ASR tasks. Namely, the phone recognition models trained on formant-based vowel representations are used to create pronunciation lexicons for word-based ASR systems, which are then trained and evaluated on corpora from different languages and speech domains. This evaluation procedure is introduced in Section 13.6 and demonstrated in Section 14.5.

Here, we ask the following main research question:

RQ3 Can we show that formant-based vowel representations are useful in word-based speech recognition?

We decompose this larger question into:

RQ3.1 Can we use hybrid ASR systems to evaluate whether formant-based vowel representations can be phonologically relevant, i.e. can be used to recognize words?

RQ3.2 Can cross-lingual lexicons created using phone recognition models trained on formant-based vowel representations improve the performance of monolingual hybrid ASR systems?

RQ3.3 How do phone predictions on individual vowels in the cross-lingual lexicons affect the performance of the monolingual hybrid ASR systems?

1.2.3 Corpus Creation for Lesser-Resourced Languages

The creation of *FT Speech*: Danish Parliament Speech Corpus is a secondary focus area in the thesis, intended to provide additional speech data for Danish at the time when Danish language resources were more scarce. With over 1,800 hours of speech, it remains the largest publicly available speech corpus for Danish to date. The details on how the corpus was created and evaluated are presented in Chapter 7. It is used as one of the evaluation corpora in our ASR experiments in Chapter 14.

Here, we ask the following main research question:

RQ4 Can we use Danish parliamentary data to create a large speech corpus that will significantly expand publicly available ASR resources for Danish?

We decompose this larger question into:

RQ4.1 Can we create and release an ASR corpus from the the recorded meetings of the Danish Parliament?

RQ4.2 Can the newly created ASR corpus be used to train general-purpose hybrid ASR systems?

RQ4.3 How does this ASR corpus change the landscape of existing resources for Danish and the status of Danish as a medium-resource language?

1.3 Specific Contributions

The main contributions of this thesis are summarized below.

1. ***FT Speech*: Danish Parliament Speech Corpus**, the largest public speech corpus for Danish to date, created from the recorded meetings of the Danish Parliament,

2. **Methodology for large-scale speech corpus creation**, which can be used to extend the FT Speech corpus in the future and applied to similar corpus creation efforts based on forced alignment,

3. **Formant-based vowel categorization**: methodology for the creation of formant-based vowel representations, which can be expanded and applied to any spoken language or speech corpus,

4. **Application of formant-based vowel representations to unseen languages and speech domains**, including real-world spontaneous speech and low-resource and typologically diverse languages

5. **Downstream application of formant-based vowel representations**: creation of pronunciation lexicons for hybrid ASR systems

6. **Extensive and rigorous evaluation pipeline** for assessing the applicability of phonetic vowel representations to additional languages, domains, and downstream tasks

Key Terms and Concepts

2

2.1 Speech and Speech Sounds

Speech is a form of linguistic communication produced with the human vocal system and perceived with the auditory system. The basic units of speech are called *speech sounds*, *speech segments*, or *phones*. A stretch of speech (series of phones) by a single speaker bounded by silence is referred to as an *utterance*. Individual speech sounds can be described in terms of *segmental features*, which tell us how the sounds are produced and how to categorize them based on their place and manner of articulation.

However, speech is a complex signal, both acoustically and linguistically, and often carries more information than the sum of its constituent segments. For example, it could reveal personal information about the speaker, such as their age and gender, or convey their emotional state, tone, intent, and many other pragmatic functions. Rather than being found at the level of individual segments, this information is superimposed over an entire utterance and can be described in terms of *suprasegmental* or *prosodic features*.

According to Laver (1994, p. 26), speech can be analyzed at multiple levels ranging from the underlying abstract representation to its physical realization. The initial level of analysis closest to the physical realization is the *acoustic level*. At this level, two separate speech events will almost never be acoustically the same. Even when they have the same underlying representation and are produced by the same speaker, there will almost certainly be a measurable acoustic difference between them. The next level of analysis is the *perceptual level*, which deals with how speech is registered by the listener. At this level, two speech events are said to be different if there is an audible difference between them, i.e. a difference that can be registered by the human auditory system. Then comes the *organic level* of analysis,

which takes into account the anatomical and physiological characteristics of the speaker. Finally, the last two and most abstract levels of analysis are the *phonetic* and the *phonological level*.

The phonetic and phonological levels deal with the learnable linguistic aspects of speech, which allow us to perceive, produce, and recognize vocal sounds as potentially meaningful speech. The distinction between the phonetic and phonological level is not always easy to make, but it is important to define. It will help us understand how the relatively small set of relevant speech sounds of a particular language relates to our ability to learn to recognize and produce a wide array of possible speech sounds, not just in our language but any spoken language in general. It will also form the foundation of this thesis and our attempts to create automatic speech recognition models that transcribe the pronunciation of any speech regardless of its language.

2.1.1 Phonetic Analysis

The phonetic level of speech analysis refers to the learnable aspects of the use of the vocal apparatus, which allow us to discern and learn any speech sound regardless of the language it belongs to. Phonetic analysis is based on the assumption that we can analyze a speech event phonetically without knowing what linguistic value it might have in some particular language. In other words, phonetic description of a given utterance is held to be independent of the phonological description of the language involved. From this perspective, descriptive phonetic theory can be regarded as a general theory applicable to the sounds of any language in the world (Laver, 1994, p. 29).

Phonetic theory allows us to study speech and speech sounds systematically from a language-neutral point of view in terms of their phonetic features. The most widely used system for the notation of speech sounds, known as the International Phonetic Alphabet (IPA), categorizes the sounds of the world's languages according to their articulatory and acoustic features, assigning the same written symbol to the characteristics of speech from different languages which can be described as having the same phonetic quality (International Phonetic Association, 1999).

However, even though, in theory, there should be only one pho-

netic system that can be applied universally to all languages, phonetic analysis has not evolved independently of typical correspondences between phonological units and their phonetic manifestations in languages. For this reason, general phonetic theory is inevitably colored by general phonological considerations and contrasts (Laver, 1994, p. 29), (Ladefoged, 1990).

2.1.2 Phonological Analysis

While phonetics is the study of all possible speech sounds, phonology studies the ways in which speakers of one language "systematically use a selection of these sounds in order to express meaning" (Crystal, 2010, p. 168). Due to anatomical and physiological differences, every speaker's pronunciation is different. Even the pronunciation of a single speaker can exhibit significant variation. Nevertheless, to use language efficiently, we are able to disregard the irregularities and focus only on the sounds and features relevant to the communication of meaning. The function of phonology, therefore, is to find general patterns and principles underlying the phonetic layer of speech and relate them to higher levels of linguistic analysis, notably morphology, syntax, and semantics (Laver, 1994, p. 30). This means that there is only one phonetic system, which is, in theory, applicable to all languages, but every language (or language variety) has its own phonological system, as an abstraction of the universal phonetic system.

The phonological abstractions of speech sounds (phones) are called *phonemes*. According to the traditional definitions, the term phoneme denotes a speech sound which brings about a difference in meaning between a pair (or set) of words in a given language, whereas the phone is any distinct speech sound regardless of the language in which it occurs or how it affects the meaning of an utterance.

What complicates this distinction is the fact that it is often impossible to find a minimal pair that distinguishes a given pair of phonemes in a language. A minimal pair is a pair of words that differ in only one sound, and substituting this sound for the other causes the difference in meaning (e.g. *read* /riːd/ vs. *lead* /liːd/). For example, there are no minimal pairs that differentiate the sounds [ʃ] and [ʒ], and yet, the native speakers of English perceive these two sounds as different

enough to constitute two distinct phonemes.

On the other hand, two phonetically different sounds may be perceived as the same phoneme when their different phonetic realizations, called *allophones*, are in complementary distribution (i.e. they are mutually exclusive and never appear in the same phonetic environment), or free variation (i.e. when their use cannot be reliably predicted from the phonetic context). Therefore, modern phonological theories define the phoneme in rather abstract and subjective terms: as a set of phonetically similar speech sounds which the speakers of a particular language perceive as a single distinctive unit of sound in the language.

2.2 Phonetic Representations, Variation, and Notation

In this thesis, we are mainly concerned with phones as we try to transcribe speech in general language-independent terms. We use the term *phonetic transcription* (conventionally written between square brackets, [ ]) to refer to a detailed language-independent written representation of speech, and *phonological transcription* (written between slanted brackets, / /) to indicate a simplified language-specific written representation of speech which ignores the details whose omission or mispronunciation does not obstruct communication. We use the term *phonetic representations* to refer to the individual phone tokens specified using the symbols of the IPA that we use to transcribe speech data phonetically.

A potential source of confusion comes from the fact that, when designing a system for the phonological transcription of a language, phonologists often use IPA symbols to represent phonemes, but simplify the symbols to make the transcription easier to read, usually by dropping inessential diacritical marks or replacing uncommon characters with more convenient ones. As an illustration, compare the transcriptions of the English word *preach* using IPA phonetic transcription: [pʰɹ̥itʃ] and simplified IPA-based phonological notation used by the Oxford dictionary: /priːtʃ/. To those unfamiliar with the notation, this kind of transcription system may seem indistinguishable from the IPA, but is, nonetheless, language-specific and thereby not necessarily comparable to either the IPA or similar IPA-based systems used in the

phonological analysis of other languages.

For instance, in the shown example, we can see that Oxford uses the phoneme /r/ to denote the English r-sound. In standard British English, this sound is pronounced as an alveolar approximant, which is represented phonetically as [ɹ]. The symbol that Oxford uses actually represents an alveolar trill in the IPA, which is considered a different phone. As most English varieties have only one type of /r/-phoneme, we can simplify the notation and use /r/. Indeed, there are many varieties of English where the /r/-phoneme is realized phonetically as a trill ([ɾ]). These speakers will have no problem being understood by the speakers of standard British English as both of their pronunciations fall under the same phoneme. Nevertheless, they are considered different phones as there are numerous languages where at least one of them is distinctive. Therefore, if we want to create a model that can differentiate these two sounds, it would be beneficial to use a consistent system of phonetic notation to transcribe our training data.

Part II

Theoretical Framework

Linguistic Background

3

3.1 Introduction

Since our vowel categorization experiments are performed on a trilingual corpus consisting of Danish, Norwegian, and Swedish speech, we provide a brief introduction to the phonology and vowel systems of these three closely related languages. Danish, Norwegian, and Swedish belong to the North Germanic language group, a branch of the Indo-European family, together with Icelandic and Faroese. Although they are thought to descend from distinct branches of North Germanic, modern Danish, Norwegian, and Swedish are now considered part of the same continuum of dialects with varying degrees of mutual intelligibility, commonly referred to as the *Continental North Germanic* or *Scandinavian* dialect continuum (Gooskens, 2020). According to a survey of studies on the mutual intelligibility of Scandinavian languages, Norwegian and Swedish have the highest degree of mutual intelligibility in spoken communication in the Scandinavian group, while Danish and Swedish have the lowest. However, the mutual intelligibility is asymmetrical and depends on various factors including amount of exposure to the other language, geographical distance from the border, attitude toward regional variation, and historical political influences. For example, Norwegian speakers understand to a relatively high degree both spoken and written Swedish and Danish (Gooskens, 2020), while spoken Danish seems to be the most difficult to understand for both Norwegian and Swedish speakers (Grønnum, 2003), (Basbøll, 2005, p. 7).

3.2 Phonological Characteristics of Danish, Norwegian, and Swedish

All three languages have complex phonological systems with particularly large vowel inventories. There are many parallels among their phonological systems, especially between those of Norwegian and Swedish, such as their phoneme sets and certain patterns of allophonic variation. There are also a number of differences, especially in Danish, which might explain why speakers of Norwegian and Swedish find Danish more difficult to understand. Namely, Danish exhibits several radical reduction processes, such as lenition of obstruents in syllable-final positions and assimilation and deletion of post-tonic syllables (Grønnum, 1998; Grønnum, 2003). Another distinguishing feature of Scandinavian languages is the contrastive use of pitch with two distinct pitch patterns, often termed *tonal accents*, which are found in most varieties of Norwegian and Swedish, as well as some southern dialects of Danish, and may vary considerably across regions (Wetterlin, 2010, p. 2-4). On the other hand, most Danish dialects do not feature tonal accents and instead use *stød*, typically described as a form of creaky voice, whose distribution often corresponds to the distribution of tonal accent 1 in Norwegian and Swedish (for more on stød, see, e.g., Fischer-Jørgensen (1989) and Grønnum (2023)). These prosodic differences further reduce the mutual intelligibility of Norwegian and Swedish with Danish (Grønnum, 2003).

3.3 Comparing Vowel Systems Across Scandinavian Languages

The phonological systems of Danish, Norwegian, and Swedish presented here belong to the varieties spoken in and around the capital regions. Although none of the languages have a mandated spoken standard, the capital regions enjoy a relatively high level of cultural influence.[1] The consonant sets of Norwegian and Swedish have sig-

[1]The spoken variety of the capital region has a weaker status in Norwegian compared to its counterparts in Denmark and Sweden. This is a result of historical circumstances, as well as strong social policies in favor of dialect use and preservation.

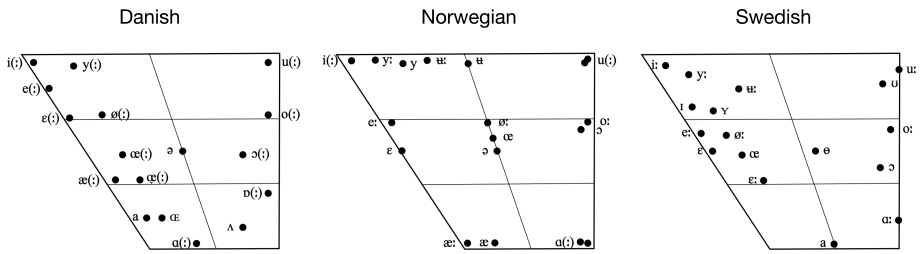


Figure 3.1: Abstract vowel spaces of Danish, Norwegian, and Swedish based on the varieties spoken in their respective capital regions.

nificant overlap and include 18 consonant phonemes each, i.e. 23 if retroflex allophones are counted. Comparatively, Danish has 15 consonant phonemes, or 19 if the most common allophones are included (Grønnum, 1998). When it comes to their vowel systems, Danish has 10 vowel phonemes, Norwegian 8, and Swedish 9. When allophonic variation relating to length, stress, and phonetic context is taken into account, the Norwegian and Swedish vowel sets increase to 19 and 21 vowel categories respectively, while the Danish one increases to 30 Grønnum (1998, 1996). Figure 3.1 shows a side-by-side comparison of the monophthong vowel systems of Danish (Grønnum, 1998), Norwegian (Kristoffersen, 2000, p. 11), and Swedish (Engstrand, 1990) plotted on the cardinal vowel quadrilateral (reproduced here with permission).

We can see that most of the vowels occur in pairs of short and long vowels, and that within some pairs there is also a qualitative difference (e.g. [ɪ, ʏ, ʊ] vs [iː, yː, uː] in Swedish). In some cases, the decision to denote short and long vowels in a pair with different symbols is a matter of convention (Kristoffersen, 2000, p. 11). Another unusual characteristic of these vowel systems is the large number of rounded front vowels, whose formant values might overlap with not only the surrounding rounded vowels but also their unrounded counterparts. While most of these vowels can be distinguished via minimal pairs, it should be noted that their number and symbols are not definitive and might vary across speakers and phonological interpretations (Grønnum, 1996). In addition to monophthongs, all three Scandinavian languages have a number of diphthongs, which will not be explored here. As mentioned before, they are frequently

analyzed as consonant-vowel and vowel-consonant sequences, both in literature and the trilingual corpus used in our experiments Grønnum (1998); Riad (2014). As a result, they are excluded from this study, as well.

3.4 Dialectal Variation

When it comes to dialectal variation in Scandinavia, traditional regional dialects, with their own phonological and morphological systems, have largely disappeared in the past century due to industrialization, urbanization, and migration (Gooskens, 2020). Especially in Denmark and Sweden, where the national standard has held a dominant role, many traditional dialects have been replaced by varieties of the national standard, often called *regional standards* (Basbøll, 2005, p. 13), (Riad, 2014, p. 7).[2] The perceived differences among the present-day regional standards can be explained, to a large extent, by differences in prosody and phonetic quality, while morphological, syntactic, and lexical variation across regions has decreased significantly (Leinonen, 2011). On the other hand, regional dialects have a much stronger position in Norwegian, where the official language policy is that all spoken varieties are to be considered equal. Nevertheless, Norwegian dialects have also undergone regionalization and leveling to the extent that most regional dialects today are mutually intelligible (Kristoffersen, 2000, p. 7). Like in Danish and Swedish, phonetic and prosodic features play an important role in perceived and measured dialect distances (Gooskens and Heeringa, 2004; Heeringa et al., 2009).

[2] Local dialects that significantly differ from the standards still exist, mostly in peripheral areas, e.g. South Jutland and the island of Bornholm in Denmark (Pedersen, 2003), and Jämtland and the island of Gotland in Sweden (Riad, 2014, p. 9).

Acoustic Analysis of Vowels 4

4.1 Introduction

Vowels are speech sounds produced without any obstructions in the vocal tract. Traditionally, they have been described by specifying the position of the tongue and lips during their articulation, namely, in terms of three parameters: vertical tongue position (vowel height), horizontal tongue position (backness), and lip shape (rounding) (Catford, 2001, p. 120). The articulatory vowel space is thus commonly defined as a quadrilateral whose points are vowels produced with the tongue in an extreme position, as far front, back, high, or low as possible without creating friction. These four points delimiting the vowel space together with the intermediate points along the edges and inside of the quadrilateral form a system of reference vowels known as *cardinal vowels* (Figure 4.1). Although the vowel space is continuous, the cardinal vowel system allows us to describe any vowel in any spoken language based on its position within the vowel quadrilateral (International Phonetic Association, 1999, p. 13).

However, the articulatory basis of the vowel space has long been disputed as the positions of the cardinal vowels do not accurately reflect their corresponding tongue positions (Ladefoged and Disner, 2012, p. 131). In fact, it has been demonstrated that vowel quality is more accurately characterized in acoustic terms, using *formants*, which represent spectral prominences computed from the speech signal that correspond to the acoustic resonances of the human vocal tract and depend on the size, shape, and position of the speech organs during speech production (Joos, 1948; Lindau, 1978; Ladefoged and Maddieson, 1990). This means that the posited vowel space is more indicative of our perception of the acoustic properties of vowels than it is of their articulation. It should thus be viewed as an abstraction rather than a direct mapping of tongue position (International

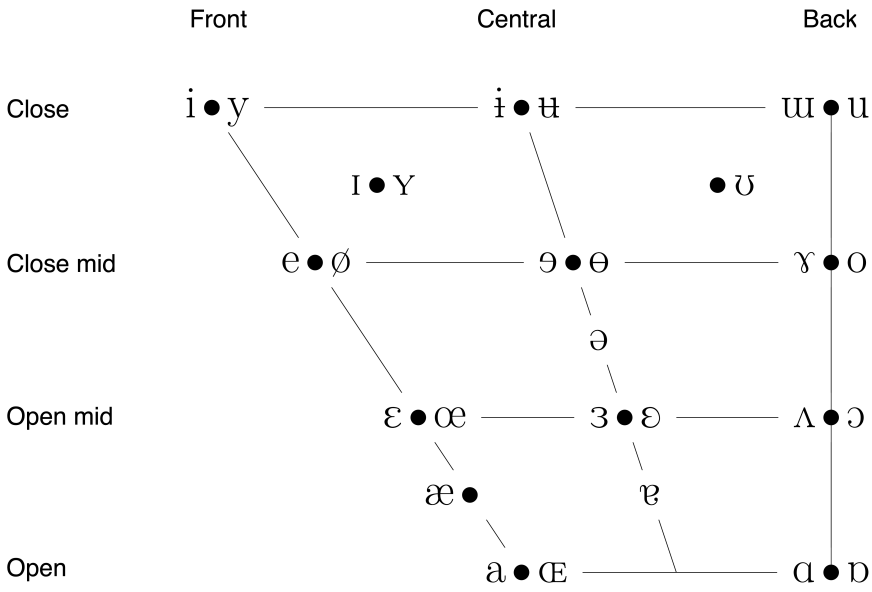


Figure 4.1: IPA vowel quadrilateral with cardinal vowels. Symbols on the left side of the dots indicate unrounded vowels, and symbols on the right indicate rounded vowels (International Phonetic Association, 1999).

Phonetic Association, 1999, p. 12).

Apart from the three traditional dimensions, height, backness, and rounding, there are additional vowel features which may be used to form phonological contrasts within a language. They include, but are not limited to, nasalization, advanced tongue root, pharyngealization, rhotacization, phonation, length, diphthongization, and tone (Ladefoged and Maddieson, 1990). Moreover, vowels play a vital role in stress and prosody. They may take on extra features such as pitch or loudness, or undergo lengthening or reduction, to convey syllabic and prosodic prominence, intonation, and rhythm. These features can interact with each other in complex ways to produce a wide range of vowel sounds in different languages. In this work, we only consider vowel height and backness, as well as rounding to a certain degree, as we would like to find general patterns that capture the three most important vowel features. Nevertheless, we should always keep in mind that vowels are combinations of complex and dynamic factors

that are not easily separable.

4.2 Vowel Formants

In acoustic studies of vowels, the first two formants (F_1, F_2), which correspond to the two lowest resonant frequencies of the vocal tract, are typically used to characterize vowels (Ladefoged and Disner, 2012, p. 39). More specifically, F_1 has been found to correlate with vowel height and F_2 with vowel backness and lip rounding (Johnson, 2011, p. 144), (Ladefoged and Johnson, 2015, p. 208). For this reason, plotting vowels in terms of F_1 and F_2 allows us to locate them within the abstract cardinal vowel quadrilateral. As the last major feature of vowel quality, lip rounding has been proposed as the third dimension in a 3D representation of the vowel quadrilateral (Ladefoged and Maddieson, 1990). However, since rounding has an effect on all formants (Fant, 1960, p. 64), the third dimension of the vowel space cannot be independently interpreted as the degree of rounding. Furthermore, while F_1 and F_2 have often proved sufficient for vowel identification in studies on the perception and discrimination of natural and synthetic vowels (Fry et al., 1962), F_3 and higher formants might be required to distinguish features such as rounding and rhoticity. However, we restrict our study to the first two formants to be able to visualize our results in two dimensions and compare them to existing studies of Danish, Norwegian, and Swedish vowel spaces.

4.3 Cross-Lingual Vowel Normalization

Formant values cannot be directly compared across different speakers, as they also encode information about the physiological characteristics of a speaker's vocal tract (Ladefoged and Broadbent, 1957). As a result, any comparison of vowels produced by different people, including those who differ by dialect or language, requires a vowel normalization procedure in order to reduce the confounding effects of individual speaker differences on the formants (Disner, 1980). This procedure is designed to minimize the acoustic overlap among vowel categories. This is believed to simulate the ability of human listeners

to deal with acoustic variability of vowels in speech recognition (Reetz and Jongman, 2020, p. 285).

A number of different vowel normalization techniques have been developed and applied to various languages. They are typically described as *vowel-intrinsic* or *vowel-extrinsic* depending on the type of information they use to transform the raw formant frequencies. Vowel-intrinsic procedures have been developed with the aim of modeling human speech perception. In order to normalize a given vowel, they rely solely on the acoustic information present in that single vowel token. They include transformations into log, bark (Zwicker, 1961; Zwicker and Terhardt, 1980; Traunmüller, 1990), ERB (Glasberg and Moore, 1990), or mel (Stevens and Volkmann, 1940) frequency scales, as well as procedures that adjust each formant value based on the values of the other formants in the same vowel, such as Miller's formant-ratio method (Miller, 1989). On the other hand, vowel-extrinsic normalization requires external knowledge about the speaker, and typically describes vowels in relation to the other vowels in the speakers vowel space. The most widely used vowel-extrinsic procedures include Nearey1 and Nearey2 (Nearey, 1978; Adank et al., 2004), which center the formant values around a speaker's mean, and Lobanov (Lobanov, 1971), which further standardizes the centered values to unit standard deviation. Previous studies comparing normalization procedures have found the vowel-extrinsic methods that involve speaker-specific centering and standardization to be the best at separating vowel categories while preserving socio-linguistic variation (Lobanov, 1971; Disner, 1980; Carpenter and Govindarajan, 1993; Adank et al., 2004; Kohn and Farrington, 2012; Richter et al., 2017; Persson and Jaeger, 2023). However, few of these methods allow direct comparisons of vowel systems across different languages, as the systems may not be comparable on the basis of their mean vowels (Disner, 1980).

The normalization technique we employ in this paper was devised for a cross-lingual study of vowel spaces by Chung et al. (2012). It is a modification of the Nearey1 method (as defined in Adank et al. (2004)), which makes it more robust to cross-lingual differences in vowel systems. The study demonstrates that this technique is effective at reducing the variation in formant frequencies due to speakers' gender and age while maintaining cross-lingual variation. It is performed

using the following equation:

$$F_{i,s}^{Norm} = F_{i,s}^{L} - \bar{F}_{i,s}^{L},$$

where $F_{i,s}^{L}$ is the log-transformed value of F_i for speaker s, and $\bar{F}_{i,s}^{L}$ is the weighted mean of the mean log-transformed F_i values of each of the point vowels [i, a, ɑ, u] for speaker s. The mean is weighted by the number of tokens in each vowel category to account for the different number of tokens available for each speaker. Intuitively, this procedure converts all formant frequencies into log space where each vowel is represented in terms of its distance from the speaker-specific centroid vowel, i.e. the weighted mean of a speaker's mean point vowels ($\bar{F}_{i,s}^{L}$).

Automatic Speech Recognition 5

5.1 Introduction

Automatic Speech Recognition (ASR) is the task of transcribing a digital recording of a speech signal, commonly referred to as *utterance*, into its corresponding textual representation. The textual representation is usually orthographic, but could also be phonemic or phonetic. In recent years, monolingual ASR, in which an ASR system is trained, optimized, and deployed on a single high-resource language or dialect, has advanced to the point where it can be successfully applied to a number of practical tasks, such as human-computer communication, dictation, and automatic caption generation (Jurafsky and Martin, 2020). Their importance for multilingual and cross-lingual ASR is paramount as most of the multi- and cross-lingual methods were first developed and perfected in a monolingual setting in line with one of these two approaches.

At present, there are two main types of ASR architectures in active development: traditional modular HMM-based ASR systems and modern neural end-to-end models. Traditional HMM-based ASR systems range from monophone, simple context-independent phone models, to triphone, context-dependent phone models, to hybrid Hidden Markov Model / Deep Neural Network (HMM/DNN) models. Their advantages include high computational efficiency and ability to learn from a relatively limited number of training samples. But their disadvantages are highly complicated architectures, lack of easily extensible frameworks for prototype development, and the requirement for pronunciation dictionaries, which are very time-consuming to create and maintain. On the other hand, neural end-to-end models are easier to protype and train, but they require large amounts of training data and greater computational power.

5.2 Modular Systems

A conventional HMM-based ASR system consists of three separate modules: an acoustic model (AM), which estimates the observation likelihoods of the acoustic feature vectors at each time frame of the input sequence, a pronunciation model, which is a pronunciation dictionary providing phonemic representations for each word in the vocabulary, and a language model (LM) which estimates the a priori probability for each word in the output sequence (Jurafsky and Martin, 2009, p. 321-329). Inferring the transcription of a given utterance requires integrating all three modules and implementing highly optimized finite state transducers that can search efficiently over all possible word sequences in order to find the most likely one (Jelinek, 1998, p. 5-9), (Mohri et al., 2002, 2008).

Today, the mainstream traditional ASR systems have a hybrid acoustic model architecture combining HMMs with deep neural networks (HMM/DNN), in which a feed-forward neural network acts as a phonetic classifier instead of the previously used Gaussian mixture models (GMMs) (Jurafsky and Martin, 2020). The most successful hybrid AMs include time-delay neural networks (TDNNs) (Peddinti et al., 2015; Povey et al., 2016) and long short-term memory networks (LSTMs) (Sak et al., 2014; Peddinti et al., 2018). The main advantage of modular ASR systems is that they can achieve state-of-the-art performance using relatively small amounts of spoken training data. However, their major disadvantages include the need for meticulously handcrafted pronunciation dictionaries, which exist for only the best-resourced languages, and a complicated overall architecture (Watanabe et al., 2017b).

HMM-based ASR systems view the task of speech recognition through the metaphor of the noisy channel, which treats the acoustic signal as a noisy version of the underlying string of words. The goal of an HMM speech recognizer is to build a model of this channel which allows us to understand how the channel distorts the underlying sentences and thus recover them by searching through the huge space of potential source sentences and choosing the one with the highest probability of generating the “noisy” sentence (Jurafsky and Martin, 2009, p. 321).

Since probability is used as a performance metric, the problem of speech recognition can be described as a special case of Bayesian inference. As a result, the probabilistic noisy-channel ASR architecture tries to answer the following question: *What is the most likely sentence $\hat{W}$ out of all sentences W in the language $\mathscr{L}$ given some acoustic input O?* This can be expressed as:

$$\hat{W} = \underset{W \in \mathscr{L}}{\operatorname{argmax}} P(W|O)$$

To make it easier to compute, the above equation is expanded using Bayes' rule into the following:

$$\begin{aligned} \hat{W} &= \underset{W \in \mathscr{L}}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)} \\ &= \underset{W \in \mathscr{L}}{\operatorname{argmax}} P(O|W)P(W) \end{aligned}$$

where the *observation likelihood* $P(O|W)$ is computed by the acoustic model, the *prior probability* $P(W)$ is estimated by the language model, and $P(O)$ can be ignored because it is a constant term for each sentence.

5.2.1 Acoustic Model

A general Hidden Markov Model is characterized by the following five components (Jurafsky and Martin, 2009, p. 325):

In HMMs used for speech recognition, the states of the HMM are phonetic units of speech called *subphones*. Subphones are parts of a larger phonetic unit known as *phone*. In ASR, each phone is modeled as consisting of three subphones: beginning, middle, and end subphone, as shown in Figure 5.1. The three-subphone model of a phone is made to take into account the temporal variation of acoustic features throughout a phone. Figure 5.2 shows a full HMM for the example word *six* which consists of four phones: /s ɪ k s/.

Acoustic observations used as input to a speech recognizer are derived from digital audio recordings of speech, termed *waveforms* when plotted as a function of time (Figure 5.3a). The waveforms are

$Q = q_1, q_2, ..., q_N$	a set of *states*
$A = a_{01}, a_{02}, ..., a_{n1}, ..., a_{nm}$	a *transition probability matrix* A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{i=1}^{n} a_{ij} = 1 \, \forall \, i$
$O = o_1, o_2, ..., o_N$	a set of acoustic *observations*
$B = b_i(o_t)$	a set of *observation likelihoods* (emission probabilities), each expressing the probability of an observation o_t being generated from state i
q_0, q_{end}	special *start and end states*, not associated with observations

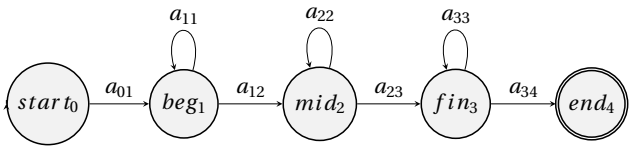


Figure 5.1: The standard five-state HMM for a phone (based on (Jurafsky and Martin, 2009, 2, p. 328)).

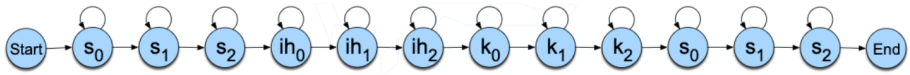


Figure 5.2: A full HMM for word *six* /s ɪ k s/, formed by joining four phone models (Jurafsky and Martin, 2009, p. 326).

first converted to a time-frequency representation called *spectrogram* using the short-time Fourier transform (STFT), which segments the input signal into overlapping frames and computes the discrete Fourier transform for the individual frames:

$$X[k,\lambda] = STFT\{x[n]\} = \sum_{n=0}^{M-1} \tilde{x}[n+\lambda R] \mathrm{e}^{-j\frac{2\pi}{M}kn}$$

$$\tilde{x}[n+\lambda R] = \begin{cases} x[n+\lambda R]\, w[n]; & n = 0, 1, ..., N-1 \\ 0; & n = N, N+1, ..., M-1 \end{cases}$$

where x is a discrete-time signal of length N_x, w an analysis window

of length N, usually the Hamming window, N the window size, M the size of the discrete Fourier transform after zero-padding, R the shift between adjacent windows, Z the overlap between adjacent windows $Z = N - R$, λ the frame index $\lambda \in \{0, 1, ..., L-1\}$, k the frequency bin index $k \in \{0, 1, ..., M-1\}$, and L the number of frames $L = \lceil (N_x - Z)/R \rceil$. The spectrogram is then converted to the mel scale, which gives higher resolution to frequencies below 1000 Hz to which the human ear is most sensitive, resulting in a *mel spectrogram* (Figure 5.3b). The mel frequency can be computed from the raw acoustic frequency as follows (Jurafsky and Martin, 2009, p. 333):

$$mel(f) = 1127 \ln(1 + \frac{1}{700})$$

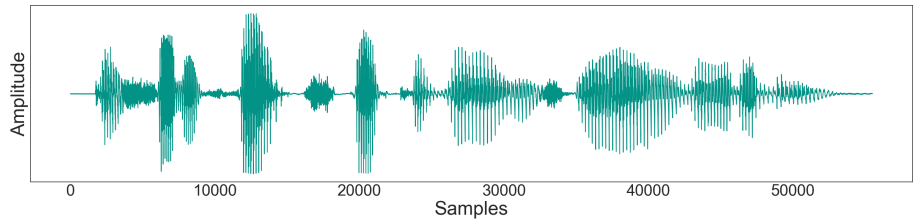
The final step in the feature extraction procedure transforms the mel spectrogram into *mel-frequency cepstral coefficients* (MFCCs) using the inverse short-time Fourier transform. The goal of this is to decorrelate the cepstral features in order to reduce the number of parameters needed to estimate the emission probabilities (Jurafsky and Martin, 2009, p. 395). For the purposes of ASR, it is common to take only the first 13 MFCC values (Figure 5.3c), which are described as providing the most information about the vocal tract.

The observation likelihoods, or emission probabilities, are modeled using multivariate Gaussian mixture models (GMMs), which constitute a weighted sum of multivariate Gaussians, which are parametrized by the mean vector μ, covariance matrix Σ, and mixture weights c (Jurafsky and Martin, 2009, p. 346).

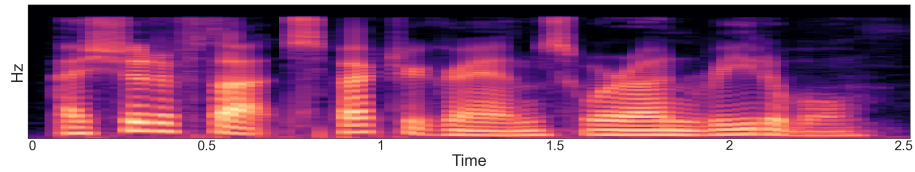
$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{2\pi|\Sigma_{jm}|}} \exp\left[(x - \mu_{jm})^\top \sigma_{jm}^{-1} (o_t - \mu_{jm})\right]$$

5.2.2 Pronunciation Lexicon

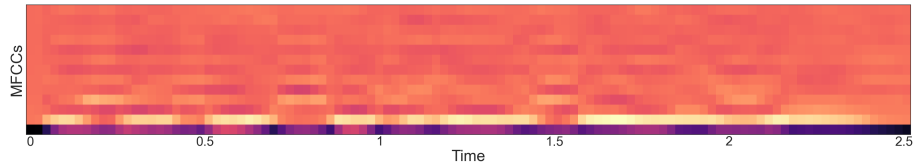
The pronunciation lexicon, or dictionary, contains phonemic pronunciations for each word in the vocabulary of the training data. The pronunciation is given in terms of *phonemes* (sometimes referred to as *phones*), the basic units of speech in a particular language or dialect, but may also include suprasegmental markings, such as word stress



(a) Waveform of a speech signal.



(b) Mel spectrogram of the above waveform.



(c) First 13 MFCC features of the above mel spectrogram.

Figure 5.3: MFCC feature extraction for ASR

or tone. The lexicon is used to estimate the transition probabilities between within-word HMM states. Lexicons are notoriously difficult to compile, especially for languages with low grapheme-to-phoneme correspondence and/or high lexical productivity, where each word in the vocabulary needs to be manually transcribed phonemically by expert phoneticians. For this reason, a more common approach to compiling pronunciation lexicons is by means of grapheme-to-phoneme conversion tools, such as espeak-ng (eSpeak NG, 2016), Phonetisaurus (Novak et al., 2012) and its pre-trained models LanguageNet (Hasegawa-Johnson et al., 2020), Sequitur G2P (Bisani and Ney, 2008), or Epitran (Mortensen et al., 2018).

5.2.3 Language Model

Language models are used in ASR to obtain transition probabilities between individual words in a word sequence. They are typically N-gram

language models trained on large amounts of text data. Depending on the amount of training data, the size of the N-gram usually ranges from trigram to 5-gram. N-gram LMs are N^{th}-order Markov chains that approximate the conditional probability of the next word in a sequence by just the previous $N-1$ words. Given this approximation of the probability of an individual word, the probability of an entire word sequence can be found by (Jurafsky and Martin, 2009, p. 122):

$$P(W) \approx \prod_{k=1}^{N} P(w_k | w_{k-N+1}^{k-1})$$

5.2.4 Search and Decoding

Integrating all the HMM probability estimators to retrieve the most probable word sequence is called decoding. This is achieved using the graph search algorithm named Viterbi, which takes as input an observation sequence and a trained HMM and returns the probability of the state path through the HMM trellis that assigns maximum likelihood to the observation sequence as well as the state path itself.

The value of each cell in the trellis is computed recursively by taking the most probable path that could lead to this cell. This is done by first computing the probability of being in every state at time $t-1$. Then, the Viterbi probability is computed by taking the most probable of the extensions of the paths that lead to the current cell. For a given state q_j at time t, the value $v_t(j)$ is computed as (Jurafsky and Martin, 2009, p. 351-358):

$$v_t(j) = \max_{i-1}^{N} v_{t-1}(i) a_{aj} b_j(o_t)$$

where $v_{t-1}(i)$ is the previous Viterbi path probability from the previous time step, a_{ij} the transition probability from previous state q_i to current state q_j, and $b_j(o_t)$ the state observation likelihood of the observation symbol o_t given the current state j.

5.2.5 Embedded Training

ASR systems train each phone HMM embedded in an entire sentence. The segmentation and phone alignment are performed automatically during training without the need for manually time-coded transcripts. The transition matrix A and the emission probability estimator B for the HMMs are conventionally trained using the Baum-Welch algorithm, also known as the forward-backward algorithm, which sums over all possible segmentations of words and phones using the probability of being in a certain state and a certain time step and generating a particular observation. However, since the execution of the Baum-Welch is very time-consuming in practice, it is common to approximate it using the Viterbi algorithm. In Viterbi training, instead of accumulating counts by summing over all paths passing through a given state at a given time, we approximate this by choosing the most probable (Viterbi) path. Thus, instead of running the Baum-Welch at every step of the training, we repeatedly run the Viterbi. This is called *forced Viterbi alignment* or just *forced alignment* (Jurafsky and Martin, 2009, p. 361).

Therefore, we can summarize the embedded training procedure as shown below. Given a phone set, pronunciation lexicon, and transcribed wavefiles (Jurafsky and Martin, 2009, p. 361):

1. Build a whole sentence HMM for each sentence
2. Initialize A probabilities to 0.5 (for loop-backs and correct next subphone) or to 0 (for other transitions)
3. Initialize B probabilities by setting the mean and variance for each Gaussian to the global mean and variance for the entire training set
4. Run multiple iterations of the Viterbi algorithm (Viterbi forced alignment)

5.2.6 Discriminative Training

In traditional HMM-based recognizers, the acoustic model relies on the *maximum likelihood estimation* (MLE) where the model parame-

ters are trained to maximize the likelihood of the training data. For a particular observation sequence O and a particular HMM model M_i corresponding to a word sequence W_i out of all possible sentences W', the MLE criterion maximizes:

$$\mathscr{F}_{MLE}(\lambda) = \sum_{i=1}^{R} \log P_\lambda(O_i|M_i)$$

However, the goal in speech recognition is to have the correct transcription for the largest number of sentences, which means that the probability of the correct word sequence should be high while the probability of all the wrong sequences should be low. Therefore, more recent ASR systems feature an acoustic model that relies on the criterion that directly maximizes the probability of a word sequence given an acoustic observation. This criterion is called *conditional maximum likelihood estimation* (CMLE) and the models that utilize it are known as *discriminative models* (Jurafsky and Martin, 2009, p. 384). The expression below can be used to describe the CMLE mathematically. The first line is expanded using Bayes' rule, and the second one expands $P_\lambda(O)$ by summing over all sequences that could have produced the given observation.

$$\begin{aligned}\mathscr{F}_{CMLE}(\lambda) = \sum_{i=1}^{R} \log P_\lambda(M_i|O_i) &= \sum_{i=1}^{R} \log \frac{P_\lambda(O_i|M_i)P(W_i)}{P_\lambda(O_i)} \\ &= \sum_{i=1}^{R} \log \frac{P_\lambda(O_i|M_i)P(W_i)}{\sum_{W' \in L} P_\lambda(O_i|M_{W'})P(W')}\end{aligned}$$

However, in literature, the CMLE criterion is typically referred to as *maximum mutual information estimation* (MMIE). The reason for this is that maximizing $P(W|O)$ is actually equivalent to maximizing the mutual information between the word sequence and the acoustic observation. As shown below maximizing MMI becomes equivalent to minimizing conditional entropy because the entropy of a word sequence is difficult to maximize and considered constant for a given language model.

$$
\begin{aligned}
I(O,W) &= \sum_{O,W} P(O,W) \log \frac{P(O,W)}{P(O)P(W)} \\
&= \sum_{O,W} P(O,W) \log \frac{P(W|O)}{P(W)} \\
&= H(W) - H(W|O) \\
&\equiv H(W|O)
\end{aligned}
$$

$$
\begin{aligned}
H(W|O) &= -\sum_{W,O} P(W,O) \log P(W|O) \\
&= -\sum_{W,O} P(O|W)P(W) \log \frac{P(O|W)P(W)}{\sum_{W' \in L} P_\lambda(O|W')P(W')} \\
&\Rightarrow \mathscr{F}_{MMIE}(\lambda) = \sum_{i=1}^{R} \log \frac{P_\lambda(O_i|M_i)P(W_i)}{\sum_{W' \in L} P_\lambda(O_i|M_{W'})P(W')}
\end{aligned}
$$

5.2.7 Hybrid HMM/DNN ASR Systems

Hybrid ASR systems combine traditional statistical models with neural networks to leverage the strengths of both approaches. Neural networks, particularly Time-Delay Neural Networks (TDNNs) and Recurrent Neural Networks (RNNs), have significantly improved the performance of modular ASR systems. In hybrid ASR systems, neural networks primarily serve as the acoustic model. They can extract discriminative features from the input speech signal, which are then used to represent the acoustic properties of the speech. Additionally, their output layer predicts the posterior probabilities of each acoustic unit at each time step. These probabilities are then used by the language model to generate the final transcription.

RNNs using a dynamically changing contextual window over all of the sequence history have been shown to achieve state-of-the-art performance on large-vocabulary ASR tasks (Sak et al., 2014). However due to their recurrent connections, parallelization during training cannot be exploited to the same extent as in plain feed-forward neural networks. Unlike traditional RNNs, TDNNs use a delay mechanism

to access previous time steps of the input sequence, making them particularly effective for capturing temporal dependencies in speech. Some of the most effective neural networks for ASR, which are still in use today, include those described in Peddinti et al. (2015), Povey et al. (2018), and Peddinti et al. (2018).

5.3 End-to-End Systems

Neural end-to-end models were designed with the intention to eliminate the need for handcrafted lexical resources and simplify the module-based architecture into a single sequence-to-sequence network that can map the acoustic features of an input utterance directly into the graphemes of its corresponding transcript.

Until recently, the leading types of end-to-end architectures were recurrent neural networks (RNNs) based on connectionist temporal classification (CTC) (Graves et al., 2006) and the ones based on the attention mechanism (Bahdanau et al., 2015; Cho et al., 2014). Loosely speaking, CTC outputs a single character for every input frame and then collapses the sequences of identical characters to obtain the final output (Jurafsky and Martin, 2020; Hannun, 2017). Although it has led to numerous state-of-the-art models, such as (Graves and Jaitly, 2014; Hannun et al., 2014; Amodei et al., 2016), CTC often requires large amounts of training data, a separate language model, and graph-based decoding (Watanabe et al., 2017b). On the other hand, attention-based ASR models use an encoderdecoder architecture to map speech feature sequences to text, as well as an attention mechanism that aligns each element of the output sequence with the hidden states generated by the encoder network for each frame of the acoustic input. These models have also led to breakthroughs in end-to-end ASR (Chorowski et al., 2014, 2015; Chan et al., 2016; Lu et al., 2016; Park et al., 2019), however, the attention mechanism has been found to result in non-sequential alignments, making it too flexible for speech recognition where the acoustic inputs and corresponding orthographic outputs typically proceed in the same order (Watanabe et al., 2017b). For these reasons, a joint CTC-attention architecture was proposed (Kim et al., 2017; Hori et al., 2017; Watanabe et al., 2017b), which leverages the advantages of both methods by combining attention-based and CTC

scores in a rescoring beam search algorithm, thereby significantly reducing the number of irregular alignments (Watanabe et al., 2017b).

However, in recent years, the transformer has become the most dominant model architecture in both ASR and wider speech and language processing. Introduced in 2017, the transformer was designed to reduce the computational cost of RNNs by eliminating the recurrent connections and relying solely on attention mechanisms (Vaswani et al., 2017). Some of the first transformer models successfully applied to speech were developed by Dong et al. (2018), Karita et al. (2019), Karita et al. (2019), and Li et al. (2020a).

A transformer model consists of an encoder and a decoder, both of which are composed of a stack of identical layers. Each layer in the encoder and decoder contains two sub-layers: multi-head attention and a feed-forward neural network. Multi-head attention allows the model to capture dependencies between different parts of the input sequence. It consists of multiple attention heads, each of which is a scaled dot-product attention mechanism. The attention mechanism is the core component of transformer models. It allows the model to focus on different parts of the input sequence at different times, capturing dependencies between words or sub-words. Since transformers process input sequences in parallel, positional encoding is used to provide information about the order of elements in the sequence. This is typically a sinusoidal function that varies with the position in the sequence, thus helping the model capture positional relationships.

5.4 Large Pre-Trained Speech Models

The gains in computational efficiency brought on by transformer-based models have ushered in a new era of self-supervised learning, in which models are trained on massive amounts of unlabeled data. This allows them to learn to identify general patterns and relationships in speech, which can be transferred to various downstream tasks via fine-tuning on a relatively small amount of target domain data.

One of the first large pre-trained speech models was wav2vec 2.0 proposed in Baevski et al. (2020b). It uses a transformer-based encoder to learn contextualized representations of the input audio, which allow the model to capture long-range dependencies in the speech signal. It

was pre-trained on 53,000 hours of untranscribed speech and able to achieve near state-of-the-art results on a downstream ASR task after fine-tuning on only 10 minutes of target domain data. Other first-generation large pre-trained speech models include HuBERT (Hsu et al., 2021) and SpeechT5 (Ao et al., 2022).

The latest generation of large speech models are pre-trained on even more data. For example, Meta's Seamless is a multi-modal streaming translation model supporting over 140 languages (Communication et al., 2023). It was pre-trained on over 4.5 million hours of unlabeled speech (approximately 513 years) and fine-tuned on 125,000 hours. OpenAI's Whisper is a transformer-based attention encoder-decoder model with ASR capabilities in 96 languages (Radford et al., 2022). It was trained initially on 680,000 hours of data which was later extended to 5 million hours (approximately 570 years). Google's Universal Speech Model was trained on 12 million hours of speech (approximately 1370 years), spanning over 300 languages (Zhang et al., 2023). It uses a Conformer-based encoder, which combines convolutional layers and transformer blocks, making them particularly well-suited for speech recognition tasks (Gulati et al., 2020). These models exhibit impressive ASR capabilities and portability to a large number of languages. However, they come with huge computational costs and are not easily extensible to additional languages and speech domains.

5.5 Multilingual and Cross-Lingual Speech Recognition

With the global expansion of speech technologies came the need to make ASR systems equally usable in all languages, which often means overcoming a serious lack of resources. Fortunately, it has been shown that training and optimizing ASR models on multiple languages simultaneously can improve performance on minority languages from the training data when compared to monolingual models trained and evaluated on the same monolingual data set (Pratap et al., 2020a). This approach has become known as multilingual ASR. Multilingual models are usually trained end-to-end, using methods similar to those used in monolingual end-to-end ASR, as this paradigm does not require explicit pronunciation modeling. Among the first to make inroads into large-scale multilingual ASR were private companies using large

amounts of internal proprietary data (Huang et al., 2013; Heigold et al., 2013; Li et al., 2018; Toshniwal et al., 2018; Kim and Seltzer, 2018; Pratap et al., 2020a), though, with the improvements in end-to-end modeling and rise in open-access multilingual speech corpora, open-source models were able to follow suit (Watanabe et al., 2017a; Cho et al., 2018; Karafiát et al., 2019; Adams et al., 2019; Hou et al., 2020).

The most popular public multilingual data sets in ASR literature are *GlobalPhone* (Schultz, 2002; Schultz and Schlippe, 2014), *Babel* (Chan et al., 1995), *Common Voice* (Ardila et al., 2020), *CMU Wilderness* (Black, 2019), and *Multilingual LibriSpeech* (MLS) (Pratap et al., 2020b). However, despite their size and linguistic diversity, these data sets still cover only a small percentage of the world's languages, and even among the languages they do cover, there is often great imbalance in favor of already high-resource languages. For example, the English portion of MLS is seven times larger than the remaining seven languages put together. Meanwhile, research has shown that the performance of multilingual models on individual low-resource languages improves with the amount of data belonging to the target language present in the training set (Karafiát et al., 2019).

These limitations of multilingual ASR have spurred interest in zero-shot cross-lingual ASR in which models are trained to recognize speech in languages not seen during training. However, this task faces several challenges. First, for most languages, there is little to no overlap among their orthographic systems or grapheme-to-phoneme correspondence. In such scenarios, transfer to unseen languages is poor without fine-tuning or adaptation to the target data (Cho et al., 2018; Karafiát et al., 2019; Pratap et al., 2020a). Second, many languages or dialects have no resources whatsoever, or even a writing system to begin with. For these languages, no pronunciation or language models can be made without considerable human effort.

HMM-based systems are relatively amenable to cross-lingual adaptation as they use a separate acoustic model, but their adaptation requires both a pronunciation dictionary and language model of the target language (Wiesner et al., 2019). For example, Wiesner et al. try to create zero-shot pronunciation lexicons for cross-lingual adaptation by transcribing a small set of words from an unseen language using grapheme-to-phoneme (G2P) models of higher-resource lan-

guages with the most similar orthography, which they use to train a new G2P transducer that can automatically extend the pronunciation dictionary for the unseen language. However, they assume that languages with similar orthography will also have similar pronunciation, which is not necessarily true (one example being Danish and Norwegian). Moreover, they find that incorrectly transcribed pronunciations adversely affect their performance, especially when they occur in frequent words (Wiesner et al., 2019).

As introduced in Section 5.4, fine-tuning large speech models to target languages and domains has become the dominant paradigm in low-resource ASR. Models such as Whisper (Radford et al., 2022), Seamless (Communication et al., 2023), and USM (Zhang et al., 2023) can be fine-tuned to a large number of languages using a relatively small amount labeled speech. However, they are not easily extensible to new languages and their performance on unseen languages is still not on par with that of the fine-tuned models (Bartelds et al., 2023).

5.6 Automatic Phonetic Transcription

To deal with cross-lingual differences in orthography, phonetic representations have been proposed as a potentially more viable channel for transferring ASR models to very low-resource languages. Notable past attempts include the bottleneck approach, which aimed to extract language-independent phonetic knowledge from a bottleneck layer of a multilingual model and use it as additional input features when training an acoustic model of a target language (Veselý et al., 2012; Thomas et al., 2012; Knill et al., 2013; Grézl et al., 2014), and the phoneme-based cross-lingual CTC approach, where a CTC model trained on multilingual data was ported to an unseen language by means of a new output layer which was trained on a small amount of target language data (Tong et al., 2018a,b; Dalmia et al., 2018; Dalmia et al., 2019; Li et al., 2020c,b). Cross-lingual phonetic transcription has also been attempted with large pre-trained speech models, such as wav2vec 2.0 with modest improvements in cross-lingual transfer (Gao et al., 2021; Conneau et al., 2020; Xu et al., 2022).

Our method is based on the investigation of cross-lingual transfer of phonetic representations performed by Żelasko et al. (2020)

who train end-to-end transformers with CTC-attention to phonetically transcribe speech in different languages. Namely, they perform three types of experiments on 13 languages from the GlobalPhone and Babel data sets: monolingual, multilingual, and cross-lingual, in which they find that all languages benefit from a multilingual training regime (in comparison to monolingual), but observe significant performance degradation in the cross-lingual regime, i.e. when attempting to transcribe an unseen language. Nevertheless, they notice that the prediction accuracy of certain phones which are shared by most of the studied languages is relatively stable across all experiments, including the cross-lingual ones. In their follow-up work, presented in Feng et al. (2021), where they train modular ASR systems with a separate hybrid HMM/DNN AM and phone-level n-gram LM, they observe similar performance degradation in the cross-lingual evaluation scheme, but contrary to their prior findings in Żelasko et al. (2020), the performance also decreases in the multilingual setting in comparison to monolingual, which suggests that transformers are better at leveraging multilingual data to gain performance improvements on low-resource languages.

However, despite trying to model language-universal representations, in both papers, they employ LanguageNet G2P models to transcribe the training data (Hasegawa-Johnson et al., 2020), which are often trained on language-specific representations, so their notation is not always consistent across languages. Moreover, in some cases, when no training data is available, LanguageNet G2P models are only rule-based, which can result in inaccurate transcriptions for languages with highly irregular spelling. They also do not analyze the performance of their cross-lingual models on out-of-vocabulary (OOV) tokens, which are phonetic symbols unique to the unseen languages. It is, therefore, not certain whether the cross-lingual models could learn broader phonetic categories which, although not identical to the ground-truth OOV phones, might still be close enough to prove useful in downstream word recognition. Another potentially confounding factor could be the fact that the audio recordings in the Babel data sets have lower sound quality, as the authors themselves note, which makes it difficult to judge whether the poor cross-lingual performance is due to a lack of cross-lingual phonetic transfer or

the presence of background noise. Finally, neither paper provides a downstream evaluation of the phonetic transcription models, so it is unclear how they could be applied to predict conventional orthographic transcriptions. In this thesis, we address these issues by proposing a method for uncovering more cross-linguistically consistent phonetic representations through an acoustic-phonetic corpus analysis of vowels in multilingual ASR data, and evaluating the obtained representations both intrinsically on APT and extrinsically on downstream ASR tasks.

5.7 Evaluation of ASR Systems

To measure their accuracy and performance, ASR systems are typically evaluated in terms of word error rate (WER) or character error rate (CER). WER and CER are computed using the same formula. They only differ in how the transcripts are tokenized and what is considered a token. First, the hypothesized (predicted) transcripts are paired and aligned with their corresponding reference (ground-truth) transcripts. The alignment is performed using a dynamic programming algorithm that performs a global minimization of a Levenshtein distance function. It allows us to count the number of correctly recognized words, as well as the number of insertions, deletions, and substitutions, which are considered errors. The overall error rate is obtained as the sum of insertions, deletions, and substitutions divided by the total number of tokens in the reference transcript. These metrics are usually interpreted as the percentage of tokens that are incorrectly recognized in a transcription. However, since the error counts include inserted tokens, which are not part of the reference transcript, the error rates can actually go over 100%.

The modular (including hybrid) ASR systems are typically evaluated in terms of WER, because they predict word-level tokens from the pronunciation lexicon, whereas neural end-to-end systems are usually evaluated in terms of CER, because they predict character tokens. When the target task is phone recognition or phonetic transcription, these metrics can also be used to measure phone, phoneme, or phonetic symbol error rate, depending on the relevant token.

Although error rate is a widely used metric for evaluating ASR systems, it is not the only measure of performance to consider. Other evaluation methods include human evaluation, where human experts assess the quality of the transcriptions based on factors such as accuracy, fluency, and coherence, and domain-specific metrics tailored to specific applications, such as medical, legal, or call center transcription.

Part III

Data Description

Nordic Language Technology Corpus 6

6.1 Introduction

The central data used for the creation of formant-based vowel representations are the ASR databases for Danish (Språkbanken: The Norwegian Language Bank, 2003a), Norwegian (Språkbanken: The Norwegian Language Bank, 2003c), and Swedish (Språkbanken: The Norwegian Language Bank, 2003g) created in 1990's by the company Nordisk Språkteknologi (NST). NST went bankrupt in 2003, but was soon acquired collectively by a group of Norwegian universities, Language Council of Norway, and IBM, to ensure that the linguistic resources NST had developed were preserved. In 2011, these resources were transferred to the National Library of Norway, where they were made publicly available as part of the Norwegian Language Bank (Språkbanken). Since all three ASR datasets were part of the same resource creation effort, we refer to them collectively as the *NST corpus*, and its individual monolingual subsets as *NST subcorpora*.

6.2 Corpus Description

The subcorpora consist of a number of short but phonetically diverse read-aloud sentences and phrases recorded in a quiet office environment using high-quality recording equipment. The recordings have high signal-to-noise ratio, consistent annotations, and speaker metadata, which includes speakers gender, age, and regional dialect. All utterances in the datasets are recorded in the uncompressed WAV format containing 16-bit linear PCM audio sampled at 16 kHz, and paired with their corresponding orthographic transcripts.

Furthermore, each subcorpus is accompanied by its corresponding pronunciation lexicon, namely, Danish (Språkbanken: The Norwegian Language Bank, 2003d), Norwegian (Språkbanken: The Norwegian

Language Bank, 2003e), and Swedish NST lexicon (Språkbanken: The Norwegian Language Bank, 2003f). The lexicons provide canonical pronunciations of the most frequent lexical items in the three languages, including all words and multi-word expressions from the NST corpus, manually transcribed in X-SAMPA, an ASCII-based encoding of the IPA. Furthermore, they come with detailed guides on their respective transcription conventions, which include X-SAMPA-to-IPA conversion tables and a cross-lingual comparison chart of the three phonological inventories. These guides will be used to convert the utterance transcripts to the IPA, to be able to compare vowel categories across languages and to the theoretical cardinal vowels.

Since the subcorpora do not come with gender and regionally balanced validation and evaluation partitions, we will perform the splits ourselves. Further steps on subcorpus partitioning, speech data and lexicon pre-processing, as well as corpus statistics will be presented in Section 10.2.

FT Speech: Danish Parliament Speech Corpus

7

7.1 Introduction

In this chapter, we describe the development of a new ASR resource for Danish, *FT Speech*. It represents the biggest speech corpus for Danish spanning nine years of source material (2010–2019) and advancing Danish from a medium-resource to a high-resource language with respect to open-access speech data (Kirkedal et al., 2019). We evaluate baselines for the new corpus and compare them to the ones trained on existing resources. Since parliamentary recordings are released on an ongoing basis, our plan is to update the corpus as more source data becomes available.

To ensure replicability, we release the code required to reproduce the corpus creation and evaluation from scratch. At the same time, to ensure accessibility, we also provide the resulting timestamps and transcripts that can be used to extract the corpus utterances directly from the meeting recordings. All materials we provide are freely available for research purposes only. The data, license, and terms of use can be obtained via `https://ftspeech.github.io`.[1] This chapter was adapted from the publication: Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish Parliament Speech Corpus. In *Proc. Interspeech 2020*, pages 442–446.

7.2 Related Work

Research into ASR for Danish has been rather limited owing to the scarcity of publicly available speech corpora (Kirkedal et al., 2019). At present, there are only two public Danish speech corpora: Danish subset of the NST corpus and DanPASS, which were developed under different research questions and objectives.

[1] We thank the Danish Parliament for making their data available.

The more comprehensive of the two, Danish NST subcorpus, is included in Språkbanken (Norwegian Language Bank), a collection of open-access and open-source language resources for Norwegian, Swedish, and Danish compiled by Nordisk Språkteknologi (NST). It contains two subsets designed specifically for the development of ASR systems: NST Danish ASR Database (NST-Read) (Språkbanken: The Norwegian Language Bank, 2003a), which is used in our experiments and introduced in Section 6, and NST Danish Dictation (NST-Dictate) (Språkbanken: The Norwegian Language Bank, 2003b).

NST-Read is the only public data set suitable for training ASR systems. It contains around 320 hours of phonetically balanced read-aloud speech by a total of 748 speakers, as well as general metadata about the speakers. A standardized version of this data set was released with a recipe to train ASR systems in the Kaldi repository (`sprakbanken`) (Kirkedal, 2016). On the other hand, NST-Dictate is a smaller data set with roughly 54 hours of speech by 151 speakers aimed at acoustic modeling of automatic dictation. However, despite their size and speaker variety, both of these data sets are limited by their highly contrived nature. Namely, the utterances in these data sets constitute read-aloud sentences or phrases such as personal names, place names, acronyms, numerals, spelled out letters, etc.

DanPASS is a phonetically annotated speech corpus primarily intended for acoustic and auditory phonetic analyses (Grønnum, 2006). It contains about 9 hours of speech by 27 speakers recorded in a studio using professional recording equipment. Although this corpus may offer a theoretical basis for the development of speech technologies, it is not particularly suitable for ASR due to its small size and artificial setting.

Currently, the Kaldi recipe `sprakbanken` represents the state-of-the-art on the NST-Read test set and DanPASS (Kirkedal, 2018). However, for reasons outlined above, the performance of these models degrades sharply when they are confronted with large-vocabulary spontaneous speech, as we will show in Section 7.5.

In order to compile a corpus of utterances more akin to rapid spontaneous speech, we follow the recent trend of converting open parliamentary data into ASR speech corpora. This has been accomplished for languages such as Icelandic (Helgadóttir et al., 2017), Finnish (Man-

sikkaniemi et al., 2017), and Bulgarian (Geneva et al., 2019). In addition, a multilingual speech corpus has been constructed from the debates held in the European Parliament (Iranzo-Sánchez et al., 2020). Meanwhile, the official reports of Folketing meetings have already been used to create a text corpus released within CLARIN (Hansen, 2018) and proved invaluable for multiple research disciplines (Hansen et al., 2018; Pedersen et al., 2016; Hansen et al., 2019).

In constructing *FT Speech*, we follow a procedure similar to the one used to create LibriSpeech (Panayotov et al., 2015). Other influential work on automatic alignment methods in the creation or correction of speech corpora includes Hazen (2006); Haubold and Kender (2007); Anguera et al. (2014); McAuliffe et al. (2017).

7.3 Corpus Preparation and Alignment

This section presents the corpus preparation and alignment procedures, including the description of the raw source data, audio and text preprocessing, lexicon creation, and alignment.

7.3.1 Source Data Description

FT Speech was created from the recordings of Danish parliamentary sessions and their annotated reports, which are freely available on the Folketing's official website: `ft.dk`. The audio recordings are available in two formats: `MP3` (audio only) and `AAC` (as part of the audio and video stream container `MP4`).

The sessions used to create the corpus include 1,003 meetings of the Parliament recorded in the period from October 5, 2010 (first video broadcast) until December 20, 2019 (last meeting in 2019). This amounted to about 4,960 hours of recorded audio material featuring 447 different speakers.

The reports of the parliamentary meetings are transcribed and published by the Office of the Folketing Hansard. Each report contains a comprehensive account of all parliamentary activities in the course of one meeting, including near-verbatim transcripts of the speeches by Members of Parliament (MPs) accompanied by their corresponding

metadata, such as the speaker's name, role, and political affiliation, as well as the approximate start and end timestamps of the speech.

The reports are released online as `XML` and `PDF` documents. Initially, only a preliminary version is released while the report is still subject to revision. From this point forward, it can take up to several months until the final version is published. During this period, the reports, and, in particular, the speech transcripts may undergo a number of modifications to ensure adherence to the formal guidelines established by the Presidium of the Danish Parliament. Therefore, the speeches are not transcribed strictly verbatim, but are instead adapted into standard written form by omitting speech disfluencies, correcting factual errors and slips of the tongue, and adding context to ensure the transcripts reflect the intentions of the speaker clearly and accurately (Folketingstidende).

In addition to prescribing documentation guidelines, the Presidium also enforces observance of parliamentary etiquette, which mandates decorum and the use of formal and respectful language in the Parliament. Some of the official rules state that the MPs must be addressed as either *Mr.* or *Ms.* followed by their full name, while the Ministers must be addressed with their minister titles. Furthermore, the MPs may not interject, applaud, express disapproval, or otherwise make noise during speeches and debates[2] (Hansen et al., 2018). This makes the Folketing meetings well-suited for the extraction of speaker-annotated monologues used in ASR research and development. However, the recordings still occasionally contain an audible level of spoken background noise.

The speakers come from different administrative regions of Denmark, as well as the two autonomous territories within the Kingdom of Denmark: Greenland and the Faroe Islands. Although some of the speakers may be native speakers of other local languages or dialects, the official language of the Parliament is Danish. In particular, since the linguistic register is strictly formal, while the topics discussed are primarily concerned with social, political, economic, and legal matters, the idiolects used in the Parliament converge on Standard Danish. The manner of delivery ranges from read or rehearsed to spontaneous

[2]A parliamentary debate is a sequence of monologues on the same topic.

speech.

The main challenges of converting this kind of raw data into a corpus suitable for ASR stem from: 1) the inaccuracy of the timestamps indicating the beginning and end of speeches in the reports by up to 30 seconds, 2) discrepancy between the written transcripts and the actual speeches, 3) presence of background noise in the audio data, and 4) use of lossy compression formats (MP3 and AAC) to encode the audio data.

7.3.2 Audio and Text Preprocessing

First, we downloaded the audio recordings of all Folketing meetings available on the official website up to and including December 2019.[3] All the recordings were in the MP3 format with a bitrate of 128 kbit/s. Their duration ranged from 5 minutes to 16 hours (mean ≈ 5 h, SD ≈ 3 h).

We began the audio processing by extracting the left channel stream from the stereo recordings. This was an arbitrary decision since the two channels were identical. The mono recordings were left unchanged at this stage. Next, we converted the selected single-channel MP3 recordings to WAV using a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 16 kHz. Finally, to extract speeches by single speakers, we segmented the obtained WAV files according to the timestamps and speaker names provided in the annotated meeting reports in the XML format.[4] To ensure the speaker names in the annotations referred to unique individuals, we cross-checked them with the biographies of past and present MPs available on the official website.[5] This procedure resulted in 414K speeches whose duration ranged from less than 1 second to 15 minutes (mean ≈ 40.1 s, SD ≈ 63.68 s). However, most speeches did not perfectly align with their

[3] URL to the video and audio recordings:
ft.dk/da/dokumenter/dokumentlister/referater

[4] URL to XML transcripts of the proceedings:
ftp://oda.ft.dk/ODAXML/Referat/samling

[5] URLs to the biographies of the Folketing MPs in Danish and English:
https://www.ft.dk/da/medlemmer
https://www.thedanishparliament.dk/en/members

corresponding transcripts due to the inaccuracy of the timestamps in the XML annotations, which is one of the issues we try to overcome with the alignment procedure outlined in Section 7.3.4

As stated earlier, the textual transcripts of the speeches were extracted from the XML documents containing the reports of the Folketing meetings. Their preprocessing involved expanding all common abbreviations, numbers, dates, and symbols, as well as removing all punctuation, capitalization, and unspoken parenthetical remarks and references.

7.3.3 Alignment Lexicon

The lexicon used for alignment was produced by concatenating the vocabulary created from the preprocessed transcripts of the Folketing speeches with the `sprakbanken` lexicon containing all words from the NST-Read train set. This yielded around 233K unique words (types). Their pronunciations were generated using eSpeak NG,[6] a multilingual rule-based grapheme-to-phoneme converter and speech synthesizer. We stripped these pronunciations off all stress, vowel length, and stød markers. We also made the pronunciations of foreign words consistent with the Danish phonetic alphabet in eSpeak, and manually added the unstressed forms of common function words.

7.3.4 Alignment Model

Because the Folketing meeting reports are not transcribed verbatim, we expect a large proportion of words in the aligned utterances to be misaligned. For instance, if a speaker mistakenly stated, *My uh party is against tax- taxation of the air used to to create soft ice*, but the party were, in fact, in favor of such taxation, the transcript would be edited by changing *against* to *for* and removing filler words (*uh*), restarts (*tax-*), and repetitions (*to*). In this example, if *against* were correctly recognized by an ASR system, the word would be misaligned because it did not match the transcript. For this reason, we need to extract verbatim transcriptions while allowing for word repetitions, restarts,

[6] `https://github.com/espeak-ng/espeak-ng`

and filler words that occur frequently in spontaneous speech. In our example, we would extract two segments: *My party is* and *taxation of the air used to create soft ice*.

First, we create a word alignment with a procedure similar to the one used in LibriSpeech.[7] To compare ASR hypotheses to transcripts, we decode the timestamp-segmented Folketing data with a speaker-independent GMM AM trained with boosted MMI on training data from `sprakbanken`. We use standard MFCC features and a GMM AM because we expect better performance on data from a mismatched domain than with a DNN AM. The LMs used to generate the word alignment are trained on groups of utterances and the most frequent words in the Folketing data. We want to bias the LMs to the training utterances to find a trade-off between accurate decoding of the verbatim segments and robustness to the text mismatch.

Next, we segment the utterances again and keep segments that start with a correctly recognized word and end with a misrecognized word (*against* in our example). If the segmentation algorithm classifies a misrecognized word as a word repetition such as *to*, the word is inserted into the reference and we do not split the segment. If segments contain silences longer than 0.3 seconds, they are split again. All ends of segments are extended in the audio and transcript to include the next word and to reduce edge effects in feature extraction from the core segment. The transcript is padded with a silence token such as `<UNK>` unless the segment is at the utterance boundaries.

We restrict the utterances we include in the corpus to utterances with a duration ranging from 2 to 60 seconds to ensure they can be used to train AMs. As a result, we discard 487,938 utterances of which 1,320 were longer than 1 minute.

7.4 Corpus Description and Organization

Following the alignment, segmentation, and elimination of overly long or short utterances, the finished corpus, termed *FT Speech*, contains

[7] Implemented in the bash script `https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/clean_and_segment_data.sh` from the Kaldi GitHub repository.

1,017,244 utterances with a total duration of 1,857 hours produced by 434 speakers.

We partition this corpus into a training, development, and test set with no speaker overlap. To create the development and test set, we select the same number of male and female speakers with at least 150 utterances and 900 seconds per speaker, while trying to minimize training data loss.

Since the total duration per speaker varies significantly, it was impossible to create a speaker-balanced development or test set. Therefore, we decided to further partition both of these sets into two subsets: *balanced* and *other*. The balanced portions (*dev-balanced* and *test-balanced*) contain approximately equal amounts of speech per speaker. They were created by choosing a random sample of utterances from each speaker such that the total duration per speaker was 900 seconds and that the difference in the number of utterances per speaker was kept low. The other portions (*dev-other* and *test-other*) consist of the remaining utterances by the same speakers which had to be removed from the training data to avoid speaker overlap. Detailed statistics of the corpus and each of the partitions are shown in Table 7.1.

7.5 Speech Recognition Experiments

This section describes the ASR experiments conducted for the purpose of evaluating the new resource, *FT Speech*. We build two acoustic models: one trained on *FT Speech* train (FT AM) and the other on NST-Read train data (SB AM), as well as two language models: one trained on the Folketing text data (FT LMs) and the other on NST-Read training transcripts (SB LMs), and subsequently evaluate their in-domain and out-of-domain combinations on three different test sets. Since NST-Read is an established ASR corpus, we use the acoustic and language models trained on it as a reference point in our performance evaluation.

Table 7.1: Corpus partitions and their size in hours, total number of utterances, tokens, types, OOV tokens, and speakers. The speaker counts by gender are marked as **F** (female) and **M** (male). Out-of-vocabulary tokens (OOVs) comprise all tokens whose types are not part of our ASR lexicon.

Subset	Hours	Utterances	Tokens	Types	OOVs	Speakers (F+M)
train	1,816.29	995,677	18,865,071	147,326	0	374 (146+228)
dev-balanced	5.03	2,601	51,497	6,888	2,846	20 (10+10)
dev-other	14.96	7,595	152,225	12,701	8,248	
test-balanced	10.05	5,534	103,439	10,050	5,871	40 (20+20)
test-other	10.88	5,837	111,818	10,491	6,145	
total	1,857.21	1,017,244	19,284,050	149,239	23,110	434 (176+258)

7.5.1 Acoustic Models

We follow the `sprakbanken` recipe[8] to train monophone and triphone segmentation GMM AMs (Bing-Hwang Juang et al., 1986) from scratch on *FT Speech* and generate an alignment to train an iVector model (Dehak et al., 2011) for speaker adaptation of a Time-Delay Neural Network (TDNN) (Peddinti et al., 2015) AM. The only modification compared to the `sprakbanken` recipe is that we do not perform data augmentation with speed-perturbation on *FT Speech* because the size of the training data is larger than the size of NST-Read augmented with speed-perturbation. For NST-Read, we use the training, development, and test split introduced in (Kirkedal, 2018).

We train so-called *chain* TDNN AMs with the LF-MMI objective on *FT Speech* and NST-Read. LF-MMI is a sequence discriminative training criterion that maximizes the log probability of the correct phone sequence (Povey et al., 2016). We train the AMs for 4 epochs on minibatches of 128 chunks where each chunk contains 150 frames. The frames are 40-dimensional MFCC features. The feature frames are subsampled so we only train on every third frame, but we create different versions of the training data by shifting the frames by 1 and 2 frames to create 3 versions. The effect is that every training epoch corresponds to 3 epochs. We use HMMs with a single state rather than the classic 3-state topology because of the low frame rate.

The first layer of the TDNN stacks 3 frames and a 100-dimensional iVector and projects the supervector to a 450-dimensional vector with an affine transform. The remaining layers consists of an affine transform, ReLU activation and a *renorm* component which is a layernorm without the mean term. We use a learning rate that decays from 0.001 to 0.0001 during training and we clip parameters at a Frobenius norm of 2.0. All hyperparameters are copied from the `sprakbanken` recipe and are identical for the two AMs. Note that there are several important differences to the AM in (Kirkedal, 2018) (see section 7.1).

[8]The recipe can be found here: `https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/run.sh`

7.5.2 ASR Lexicon and Language Models

To create the lexicon for the ASR experiments, we reuse the alignment lexicon but remove all types that appear only in the preprocessed transcripts of speeches by the speakers placed in the *FT Speech* development and test sets.[9] With SRILM (Stolcke, 2002), we estimated several 3-gram and 4-gram LMs with Witten-Bell or Kneser-Ney smoothing on both text from the Folketing meeting reports and on the NST-Read transcripts. Ultimately, we choose the trigram models with Witten-Bell smoothing, which we will refer to as FT LM and SB LM, for the final evaluation, as they achieve the best performance on their corresponding development sets. Before training, the transcripts were segmented into sentences using the spaCy sentence segmenter for Danish (Honnibal and Montani), and then preprocessed as described in Section 7.3.2.

7.6 Performance Evaluation

We use the standard word error rate metric (WER) to evaluate the performance. Our evaluation spans all four combinations of the two AMs and the two LMs (the lexicon is constant in all cases). We evaluate each of them on three test sets: NST-Read test (introduced in (Kirkedal, 2018) as SPTEST), NST-Dictate test (included in the original NST-Dictate data (Språkbanken: The Norwegian Language Bank, 2003b)), and *FT Speech* test-balanced (Table 7.1). The WER results for all combinations of AMs, LMs, and test data are shown in Table 7.2.

From Table 7.2, we can see that for *FT Speech*, the best WER of 14.01% is obtained with the in-corpus LM and AM. As expected, in-domain AM and LM combinations perform best in all in-domain settings (boldface in Table 7.2).[10] However, going across domains remains a challenge.

When the LM does not match the domain of the test set, WER rises by 5–14% absolute, presumably due to significant lexical differences between NST-Read and *FT Speech*. As stated previously, a large

[9] Note that neither the ASR nor the alignment lexicon contains any NST test or development data.

[10] Our system SB AM+LM achieves a new best result on NST-Read.

Table 7.2: WER performance of all four AM+LM combinations on three different test sets.

AM	**Test set**	SB LM	FT LM
SB	NST-Read	**8.81**	15.98
	NST-Dictate	14.46	19.77
	FT Speech	37.52	23.86
FT	NST-Read	13.07	27.22
	NST-Dictate	20.71	33.73
	FT Speech	24.25	**14.01**

number of NST-Read utterances consist solely of either proper nouns, numerals, spelled out names, or imperative sentences, while some also contain articulated punctuation symbols used for modeling automatic dictation. These kinds of utterances, most of which were devised to increase phonetic diversity and type-to-token ratio, do not occur in *FT Speech* nor general spontaneous speech. For these reasons, NST-Read is a less challenging resource, reflected in the lower WER (8.81).

We see a gap in performance when we fix the test set and LM, but replace the AM. This decrease in performance occurs as a result of the acoustic differences between the *FT Speech* and NST-Read utterances, especially, the differences in the speech genre, recording environment and equipment, and audio encodings. Namely, NST-Read was recorded in a quiet office and encoded in a lossless format, whereas *FT Speech* was recorded in the Folketing meeting chamber and encoded into a lossy format.

Most importantly, however, we observe that the combination of FT AM and SB LM evaluated on NST-Read test achieves a WER of 13.07%, which is comparable to the results previously published on this test set (13.08–13.38% WER, presented in Table 7 in (Kirkedal, 2018)). On NST-Dictate, it achieves a WER of 20.71 with FT AM+SB LM, which is 6.25% absolute WER off from in-domain data results. This means that the FT AM generalizes well to NST-Read data, and it shows the benefit of the new corpus containing more spontaneous speech in a more realistic environment with disfluencies and background noise. Interestingly,

the converse is not the case: the SB AM does not generalize to the *FT Speech* domain, resulting in the worst overall WER. This shows how poorly existing resources generalize, which further underlines the value of the proposed resource. While not strictly comparable, our WER results on FT are in similar ranges to related work on Icelandic, another Northern Germanic language, were a WER of 14.76% was reported on parliamentary speeches (Helgadóttir et al., 2017).

7.7 Conclusion

This work introduces *FT Speech*, a novel corpus for Danish ASR containing more than 1,800 hours of speech. It enriches the limited landscape of existing resources for Danish with a resource containing more spontaneous speech in challenging realistic conditions. Our baseline results show that a combination of *FT Speech* with in-domain language data provides not only comparable results to prior work, but also a more challenging benchmark for future studies. As the source material expands naturally, we will update the corpus with new data.

Other Parliamentary Speech Corpora 8

8.1 Introduction

This section will introduce and briefly describe the remaining four parliamentary speech corpora used in our experiments.

8.2 Althingi: Icelandic Parliament Speech Corpus

Althingi Parliamentary Speech is an Icelandic ASR corpus created from the recorded meetings of the Icelandic Parliament (*Althingi*). It was built in 2016 for the purpose of developing an ASR system for the transcription of parliamentary meetings that would reduce the need for manual speech transcription.

The corpus consists of approximately 542 hours of recorded speech along with corresponding utterance transcripts, a pronunciation dictionary, and two language models. The parliamentary speeches date from 2005-2016. The corpus is publicly available and distributed through Linguistic Data Consortium (LDC) (Helgadóttir et al., 2021). The corpus creation procedure is described in Helgadóttir et al. (2017).

8.3 ParlamentParla: Catalan Parliament Speech Corpus

ParlamentParla is a Catalan ASR corpus created in 2021 from the recorded meetings of the Catalan Parliament (*Parlament de Catalunya*), which took place between 2007 and 2018.

The corpus consists of approximately 611 hours of recorded speech along with corresponding utterance transcripts. It is published with a CC-BY license and fully downloadable from Külebi (2021). The corpus creation procedure is described in Külebi et al. (2022).

8.4 ParlaSpeech-RS: Serbian Parliament Speech Corpus

ParlaSpeech-RS is a Serbian ASR corpus created in 2024 from the recorded proceedings of the National Assembly of Serbia (*Narodna skupština*). The transcripts of the parliamentary proceedings were obtained from the multilingual corpus of parliamentary debates called ParlaMint 4.0 (Erjavec et al., 2023), while the recordings were taken from the Serbian Parliament's YouTube channel.

The corpus consists of almost 900 hours of recorded speech along with corresponding utterance transcripts and speaker metadata. It is published with a CC-BY license and freely downloadable from Ljubešić et al. (2024). The corpus creation procedure is described in Ljubešić et al. (2022). The corpus is not partitioned into training, development, and test subsets. We describe further steps on corpus partitioning, data pre-processing, and corpus statistics in Section 13.2.

8.5 FinParl: Finnish Parliament Speech Corpus

FinParl is a Finnish ASR corpus created from the recorded meetings of the Finnish Parliament (*Suomen eduskunta*) by the Aalto Speech Recognition group. It was first built in 2016 and updated in 2023. It includes recordings which took place between 2008 and 2020.

The corpus consists of approximately 3,130 hours of recorded speech along with corresponding utterance transcripts, as well as a 30-million-word-token text corpus that can be used for language modeling. It is freely available from the website of the Language Bank of Finland Aalto University, Department of Signal Processing and Acoustics (2023). The corpus creation and evaluation procedure is described in Mansikkaniemi et al. (2017) and Virkkunen et al. (2023).

Babel: Low-Resource Noisy Telephone Speech Corpus

9

9.1 Introduction

The IARPA Babel program was a research initiative funded by the Intelligence Advanced Research Projects Activity (IARPA) with the goal of developing ASR systems that can accurately transcribe speech in a variety of languages and dialects, even in challenging acoustic conditions. The main focus of the program was the development of speech recognition technologies for noisy telephone conversations in languages with very little transcribed data. As part of this program, data from 25 low-resource languages was collected and made publicly available via the Linguistic Data Consortium.

We have selected a subset of five of the Babel languages for our experiments. They are Amharic, Javanese, Lao, Mongolian, and Zulu. We aimed to choose languages that were far geographically and typologically from both our training languages and the languages of the parliamentary corpora, as well ones that had little other data publicly available. In the following sections, we will introduce and briefly describe the selected Babel subcorpora, as they are described in their accompanying documentation. However, while the documentation states that over 200 hours of speech were recorded from each language, the actual releases contain only about 50 hours of transcribed speech. More information on Babel data pre-processing and corpus statistics is provided in Section 13.2.

9.2 Amharic Subcorpus

IARPA Babel Amharic Language Pack was developed by Appen for the IARPA Babel program. It contains approximately 204 hours of Amharic conversational and scripted telephone speech collected in 2014 along with corresponding utterance transcripts. The corpus can be found

via (Bills et al., 2019).

The Amharic speech in this release comes from the native speakers of Amharic from the Addis Ababa, Shewa, and Gondar dialect regions of Ethiopia. The gender distribution among speakers is approximately equal, while their age ranges from 16 years to 60 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

Audio data is presented as 8 kHz 8-bit a-law encoded audio in sphere format and 48 kHz 24-bit PCM encoded audio in wav format. Transcripts are encoded in UTF-8 in fidel (Geez/Ethiopic) script and in a romanization scheme developed by Appen. Transcripts are included for approximately 75% of the speech.

9.3 Javanese Subcorpus

IARPA Babel Javanese Language Pack was developed by Appen for the IARPA Babel program. It contains about 204 hours of Javanese conversational and scripted telephone speech collected in 2014 and 2015 along with corresponding utterance transcripts. The corpus can be found via (Bills et al., 2020a).

The Javanese speech in this release represents the Central, Western, and Eastern Javanese dialect regions of Indonesia. The gender distribution among speakers is approximately equal, while their age ranges from 16 years to 65 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

Audio data is presented as 8 kHz 8-bit a-law encoded audio in sphere format and 48kHz 24-bit PCM encoded audio in wav format. Transcripts are encoded in UTF-8 in Latin script for approximately 77% of the speech data.

9.4 Lao Subcorpus

IARPA Babel Lao Language Pack was developed by Appen for the IARPA Babel program. It contains approximately 207 hours of L conversa-

tional and scripted telephone speech collected in 2013 along with corresponding utterance transcripts. The corpus can be found via (Benowitz et al., 2017).

The Lao speech in this release represents that spoken in the Vientiane dialect region in Laos. The gender distribution among speakers is approximately equal, while their age ranges from 16 years to 60 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

Audio data is presented as 8 kHz 8-bit a-law encoded audio in sphere format and 48 kHz 24-bit PCM encoded audio in wav format. Transcripts are encoded in UTF-8. The romanization scheme was developed by Appen and was based on the scheme developed by the American Library Association and Library of Congress.

9.5 Mongolian Subcorpus

IARPA Babel Mongolian Language Pack was developed by Appen for the IARPA Babel program. It contains approximately 204 hours of Mongolian conversational and scripted telephone speech collected in 2014 along with corresponding utterance transcripts. The corpus can be found through Bills et al. (2020b).

The speech in this release represents Halh Mongolian, which is spoken by roughly 3 million speakers living in Mongolia. Only native speakers of Halh Mongolian in Mongolia were recruited for data collection. The gender distribution among speakers is approximately equal, while speakers' ages range from 16 years to 61 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

Most of the audio data is presented as 8 kHz 8-bit a-law encoded audio in the sphere format, with some data also being in the 48 kHz 24-bit PCM encoded wav format.

The utterance transcripts are encoded in UTF-8 in both Mongolian Cyrillic and a romanization scheme developed by Appen. They cover approximately 77% of the speech data. Further information

about transcription methodology is contained in the documentation accompanying the data set.

9.6 Zulu Subcorpus

IARPA Babel Zulu Language Pack was developed by Appen for the IARPA Babel program. It contains approximately 211 hours of Zulu conversational and scripted telephone speech collected in 2012 and 2013 along with corresponding utterance transcripts. The corpus can be found via (Adams et al., 2017).

The Zulu speech in this release represents that spoken in the KwaZulu-Natal urban dialect region of South Africa. The gender distribution among speakers is approximately equal, while their age ranges from 16 years to 70 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

Audio data is presented as 8 kHz 8-bit a-law encoded audio in sphere format and 48 kHz 24-bit PCM encoded audio in wav format. Transcripts are encoded in UTF-8.

Part IV

Formant-Based Vowel Representations

Experimental Setup

10

10.1 Introduction

Since the starting pronunciation transcripts are taken from a pronunciation dictionary, where a given word or phrase will always have the same pronunciation regardless of the speaker or linguistic context, they are neither intra- nor cross-linguistically consistent, because they do not reflect the actual realization of words in connected speech. We try to mitigate these inconsistencies by performing vowel categorizations based on the normalized formant values. Namely, we categorize normalized vowels in three ways using k-means clustering: monolingual language-dependent, multilingual language-dependent, and language-independent categorization. Subsequently, we relabel the vowels depending on which cluster they are assigned to.

Monolingual language-dependent categorization (*mono*) is performed at the level of a monolingual subcorpus by clustering all monophthong tokens in the subcorpus based on their position in the vowel formant space and relabeling them according to their cluster membership. This increases the within-language consistency of vowel representations by allowing vowels to vary in terms of, e.g., their allophonic realization or the speaker's socio-linguistic identity, regional dialect, or emotional state.

Multilingual language-dependent categorization (*multi*) clusters and relabels all monophthong tokens from each language in the corpus based on their position in the vowel formant space of a multilingual corpus. This should increase both the within- and cross-language consistency of phonetic vowel representations, and could, thereby, help improve vowel recognition on non-standard speech, as well as low- and zero-resource languages and language varieties.

Language-independent categorization (*cardinal*) involves vowel categorization with respect to a set of *cardinal vowels*, a system of

reference vowels that allows us to describe any vowel in any spoken language based on the tongue position during articulation (Laver, 1994, p. 276). The hypothesized values of cardinal vowel formants are taken from Catford (2001, p. 154). This form of vowel categorization should increase the cross-lingual consistency of vowel representations since all vowels are categorized into the same set of cardinal vowel categories regardless of the language. Like multilingual clustering, this could also help improve vowel recognition for low- and zero-resource languages and varieties. Additionally, it should generalize to unseen languages better than monolingual and multilingual clustering as the cardinal vowels do not depend on the vowel systems of the training languages. However, since the peripheral cardinal vowels are produced with the tongue in an extreme position, there are few languages with a vowel system that spans the entire range of cardinal vowels (Catford, 2001, p. 133–134). Therefore, this kind of clustering might require a large number of diverse languages to achieve better generalization in cross-lingual vowel recognition.

10.2 Data Preparation

All three NST subcorpora are originally split into a training and test set only. Since there is no validation data and the test sets are significantly larger than typical ASR test sets, we extract smaller tune, development, and test sets from each original test set. The splits are performed manually in order to maintain gender and dialect balance in the tune, development, and test sets, and preclude speaker overlap between any two subsets. In the end, each non-training partition consists of exactly one male and one female speaker from each regional dialect available in the corpus. The leftover data from the original test set is added to the train data. As opposed to the non-training partitions, the new training sets are not regionally balanced. The capital region is the majority dialect region in each subcorpus, and about twice as large as the other regions. The respective sizes of the resulting train, tune, development, and test partitions for Danish, Norwegian, and Swedish are shown in Table 10.1.

To transcribe the speech data phonemically, we use the accompanying NST pronunciation lexicons of Danish, Norwegian, and Swedish

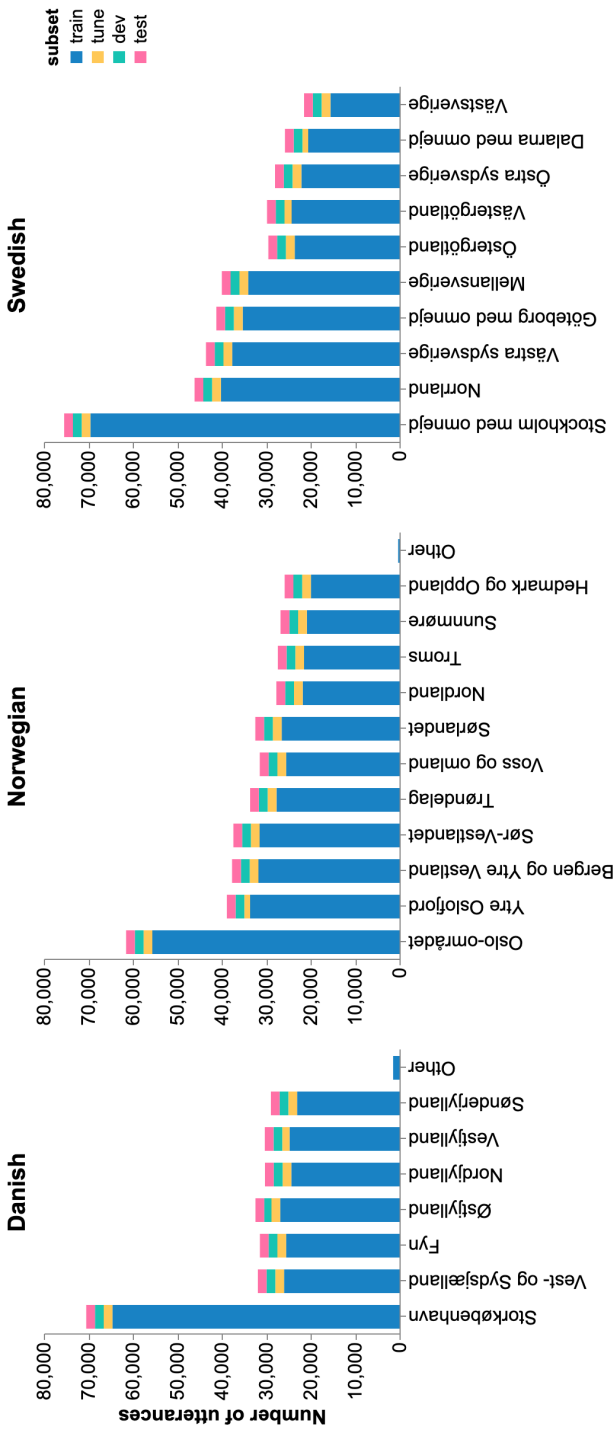


Figure 10.1: The distribution of utterances across NST corpus partitions and regional dialects of Danish, Norwegian, and Swedish, as provided in the corpus speaker metadata.

Table 10.1: NST subcorpus partitions and their size in hours, total number of utterances, tokens, types, and speakers. The speaker counts by gender are marked as **F** (female) and **M** (male). Out-of-vocabulary types (OOVs) comprise all types that are not found in the train set.

	hours	utts	tokens	types	OOVs	speakers (F+M)
Danish						
train	263.9	201,580	1,865,932	62,785	/	644 (335+309)
tune	17.9	13,440	130,604	8,532	1,350	14 (7+7)
dev	18.9	13,459	131,426	8,559	1,466	14 (7+7)
test	20.2	13,804	134,546	8,573	1,481	14 (7+7)
total	320.9	242,283	2,262,508	88,440	3,468	748 (397+351)
Norwegian						
train	428.1	301,168	2,522,539	82,792	/	911 (489+422)
tune	32.5	21,014	183,753	11,159	2,450	22 (11+11)
dev	33.1	21,691	189,245	11,036	2,570	22 (11+11)
test	33.0	21,692	189,853	10,825	2,504	22 (11+11)
total	526.7	365,565	3,085,390	115,812	5,852	977 (522+455)
Swedish						
train	420.2	306,882	2,387,280	87,002	/	954 (526+428)
tune	25.4	18,681	151,288	11,205	2,215	20 (10+10)
dev	28.6	19,720	159,650	10,980	2,315	20 (10+10)
test	27.7	19,720	159,287	11,093	2,350	20 (10+10)
total	501.9	365,003	2,857,505	120,280	5,655	1,014 (556+458)

(Språkbanken: The Norwegian Language Bank, 2003d,e,f). As stated in Chapter 6, the NST lexicons provide canonical pronunciations of the most frequent lexical items in the three languages, including all words and multi-word expressions from the NST corpus, manually transcribed in X-SAMPA, an ASCII-based encoding of the IPA. Furthermore, they come with detailed guides on their respective transcription conventions, which include X-SAMPA-to-IPA conversion tables and a cross-lingual comparison chart of the three phonological inventories. We use these guides to convert all utterance transcripts to the IPA, to be able to compare vowel categories across languages and to the abstract cardinal vowels.

To be able to represent Danish, Norwegian, and Swedish vowel phonetic qualities cross-linguistically, we strip the dictionary phonemic representations of all suprasegmental features, i.e. stress, tone, length, and stød (Danish creaky voice) markers. Table 10.2 shows the number of phone and vowel types for each of the three languages in the NST corpus.

Table 10.2: The number of phone and monophthong vowel types in the phonological inventory of each language in the NST corpus.

Number of types	**Danish**	**Norwegian**	**Swedish**
phones	33	45	41
monophthongs (unround. + rounded)	14 (7 + 7)	17 (7 + 10)	16 (6 + 10)
language unique (of which monophth.)	5 (2)	9 (1)	4 (0)

10.3 Formant-Based Vowel Categorization with Language-Specific Vowel Sets

The vowel categorization pipeline consists of three steps: phonetic corpus alignment, vowel normalization, and vowel clustering and recategorization.

To obtain the start and end times of each vowel in the NST corpus, we segment the speech into phones by force-aligning the speech and the transcriptions of each monolingual subcorpus individually with forced alignment models. To that end, we use Kaldi's `sprakbanken` recipe to train monophone and triphone acoustic models based on hidden Markov models and Gaussian mixture models (HMM-GMM).[1] The resulting segmentation of the speech signal is used to determine the start and end times of all vowels in the data.

[1]The original recipe was created for the Danish NST subcorpus and can be found here: `https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/run.sh`.

The parameters of the acoustic models are estimated by alternating between training and alignment phases where each new training step uses the aligned output from the previous step to refine the model's parameters and improve the alignment between the acoustic data and the reference transcript. For the final alignment, we train a speaker-adapted model with fMLLR transforms estimated at the speaker-level (Gales, 1998). After this stage, we extract phone alignments for each utterance and convert the integer phone labels to their corresponding IPA symbols.

Subsequently, for each vowel, the formant frequencies are estimated using Praat (Boersma and Weenink, 2018) and its Python port Parselmouth (Jadoul et al., 2018). We use standard formant settings in Praat: pre-emphasis from 50 Hz, Gaussian analysis window with window length of 0.025 s, dynamic range of 30 dB, 5 formants per frame, and a formant ceiling of 5500 Hz for female voices and 5000 Hz for male voices. The output of the formant estimation is a sequence of formant values for each vowel formant. Since we are dealing with monophthongs whose formants are relatively constant, we create a single value that represents the formant frequency as accurately as possible. First, we discard outlier values that are more than two standard deviations away from the mean of the sequence, which are assumed to be formant mistrackings. Then, we extract the midpoint value of the resulting sequence, which is a point where the formant is considered the most stable and least affected by adjacent phones (Ladefoged, 2003, p. 104). Finally, the obtained formant midpoints, which we refer to as raw $F_1 - F_2$, are normalized following the procedure explained in Section 4.3.

The effect of this procedure is visualized in Figure 10.2 which shows the female and male vowel categories for each language before and after $F_1 - F_2$ normalization. The plots show that this method greatly reduces both the spread within each vowel and the difference between corresponding male and female categories. This indicates an overall reduction of the effects of individual speaker characteristics on formant values and allows us to compare normalized vowel spaces across languages.

Subsequently, the three formant-based vowel categorizations are carried out (as introduced in Section 10.1). This results in three types

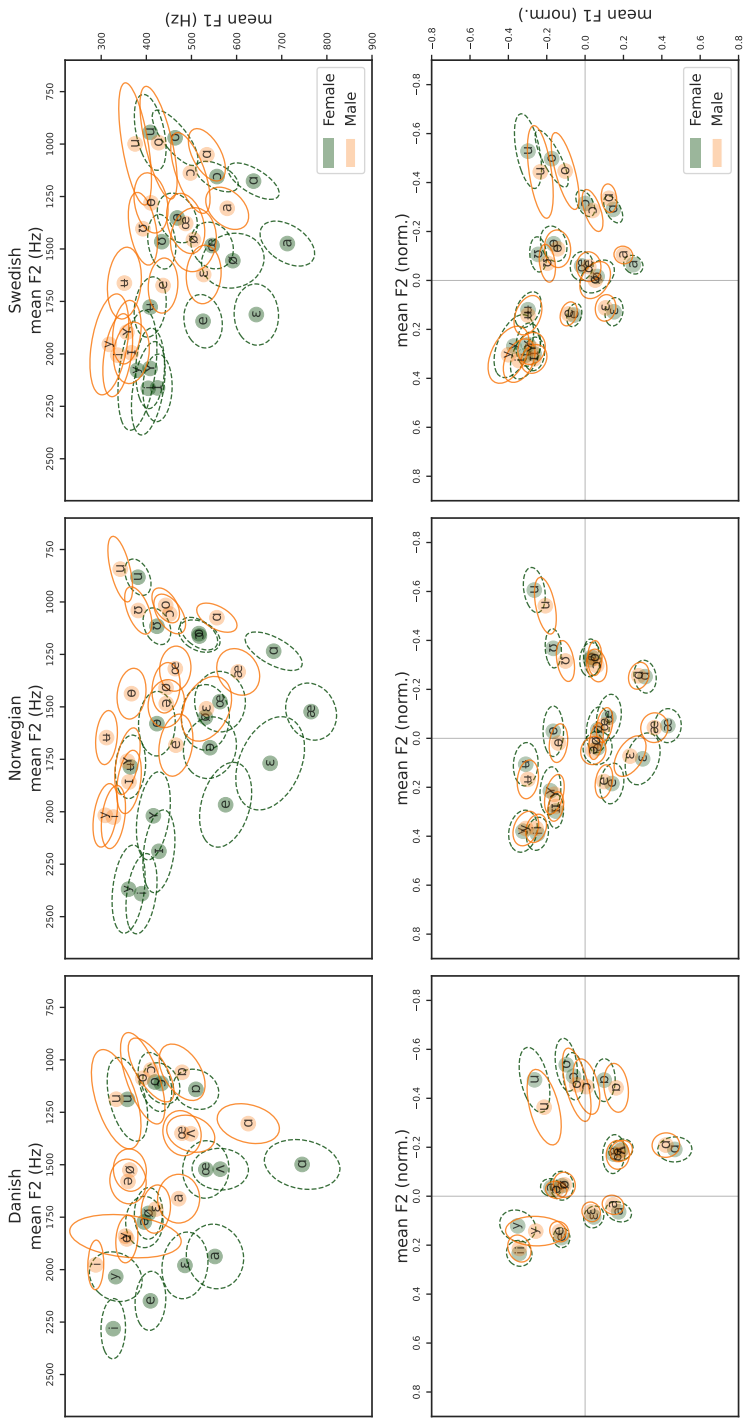


Figure 10.2: Vowel spaces for each language in terms of mean raw F_1 and F_2 in Hz (top) and mean normalized F_1 and F_2 (bottom). Female (dashed green circles) and male (solid orange circles) vowel distributions are plotted separately [color plots online]. The locations of the vowel labels represent the grand means for each vowel category, i.e., the mean of all speakers' means. The ellipses around them correspond to mean vowel spread one standard deviation from the grand mean.

of relabeled phonetic utterance transcripts: *mono*, *multi*, and *cardinal* transcripts. Since all three languages in our corpus are rich in both unrounded and rounded vowels whose first two formant values can overlap considerably (Basbøll, 2005; Kristoffersen, 2000; Riad, 2014), we cluster these two sets of vowels separately. Therefore, for each language and categorization, we first separate unrounded and rounded vowels based on their dictionary IPA symbol, and, then, use k-means to cluster each group into k clusters, where k is the number of vowel types in the given vowel group of a given language.

For language-dependent categorization (*mono* and *multi*), we cluster the unrounded and rounded vowels of a given language into k clusters, where k is the number of unrounded or rounded vowels in its vowel system. For *mono*, the cluster centers are estimated from the vowels of each monolingual subcorpus separately, whereas, for *multi*, they are estimated from the vowels of all three subcorpora together. In each case, the k-means algorithm is initialized with a predefined set of cardinal vowels as cluster centers for the purpose of preserving the vowel cluster labels. To minimize the effects of outlier vowels, which might result from errors in phonetic alignment or formant estimation, for each vowel type, we only cluster the vowel instances whose normalized formant values are within 2 standard deviations (std) from the mean. The outlier vowels over 2 std from the mean, are, therefore, left unchanged.

For language-independent categorization (*cardinal*), we do not learn the clusters from the data, but rather create a trained k-means model using a set of predefined cardinal vowels as cluster centers. We use this model to determine which cardinal vowel cluster each monophthong in the speech corpus belongs to. This is equivalent to classifying each monophthong with a 1-nearest neighbor classifier trained on a set of normalized cardinal vowels.

The outcome of each of the three forms of clustering is a new categorization of monophthong vowels which should more accurately reflect their acoustic realization. These are used to create a new set of utterance transcripts for each language in the corpus by relabeling the monophthong vowel tokens of the original transcripts with the new labels. Figures 10.3-10.5 show the clustering decision boundaries for each of the categorization methods in relation to the original vowel

distributions of Danish, Norwegian, and Swedish respectively, while Figure 10.6 contains all three previous figures for easier cross-lingual comparison.

It should be noted that none of the categorization methods change the phone sets of the source languages provided by the NST pronunciation lexicons. They only change the distribution of monophthong vowel tokens in the utterance transcripts. Preserving the same phone sets across different clustering methods and their resulting utterance transcripts makes it possible to compare our experiment results across the three categorization techniques and the original transcripts.

With *mono* transcripts, about 21% of the original phone tokens have undergone relabeling in each NST subcorpus. With *multi* transcripts, about 22% of the Swedish, 25% of the Norwegian, and 26% of the Danish phone tokens have undergone a label change. Finally, with *cardinal* transcripts, about 22% of the Swedish, 25% of the Norwegian, and 25% of the Danish phone tokens have undergone a label change. About 4% of all monophthong tokens were considered outliers and excluded from any categorization. Tables 10.3 and 10.4 show how each of the categorization methods affects the distribution of unrounded and rounded monophthongs in each subcorpus.

10.4 Intrinsic Evaluation: Cross-Lingual Phone Recognition

The utility of the three vowel categorization approaches is assessed intrinsically in a set of cross-lingual phone recognition experiments. All phone recognition models are created by fine-tuning the pretrained multilingual wav2vec 2.0 model, XLSR-53 (Baevski et al., 2020b; Conneau et al., 2020), on two NST subcorpora (*training languages*). The trained models are then evaluated on the third, unseen NST subcorpus (*evaluation language*). For each evaluation language, we fine-tune three cross-lingual models using the *mono*, *multi*, and *cardinal* transcriptions, individually, and one cross-lingual model using the original dictionary-based pronunciations (*nst*), which is used as the baseline. We furthermore investigate the effect of the number of labeled fine-tuning samples on the cross-lingual (zero-resource) models'

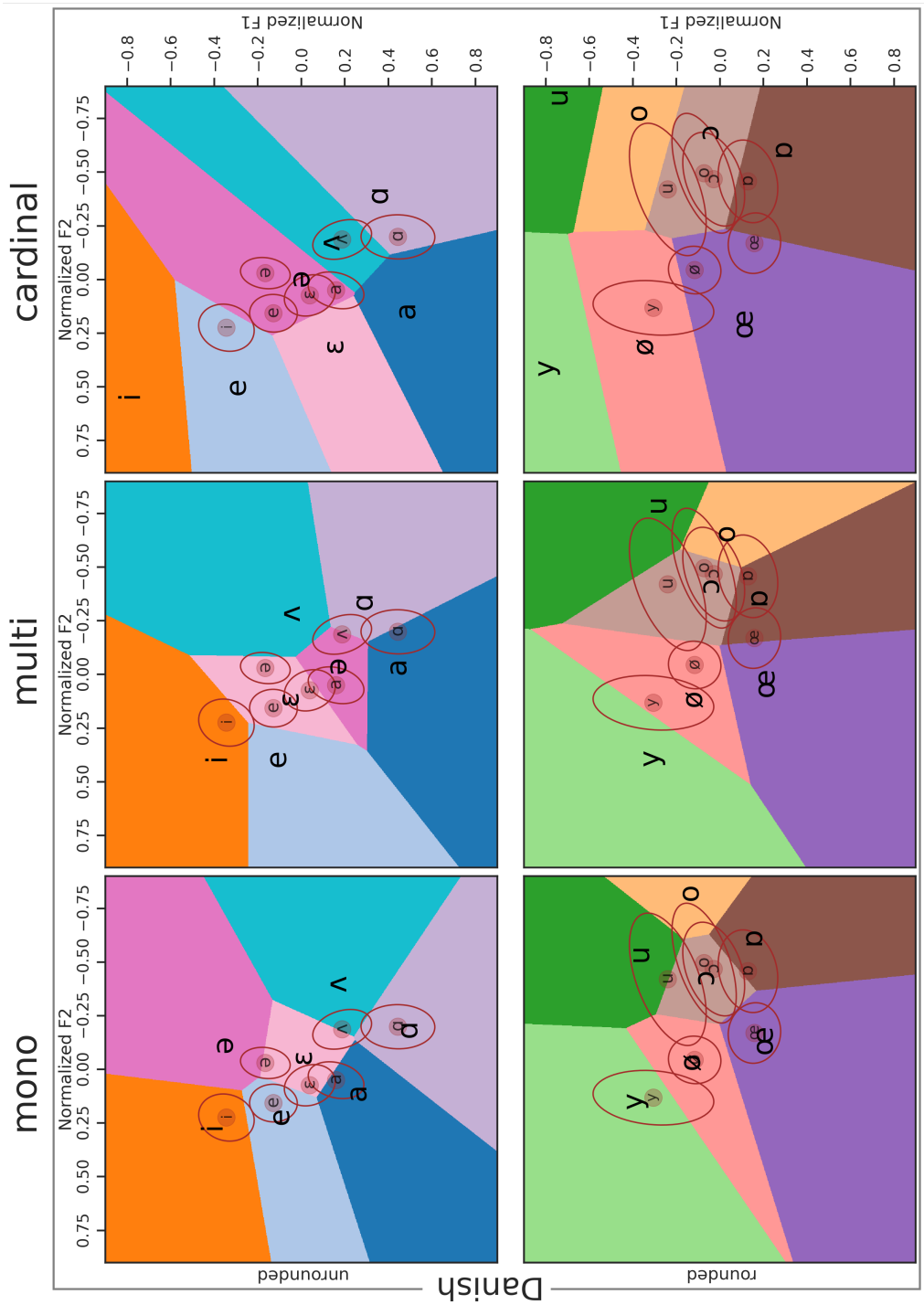


Figure 10.3: The decision boundaries of each vowel category and vowel categorization methods for Danish. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

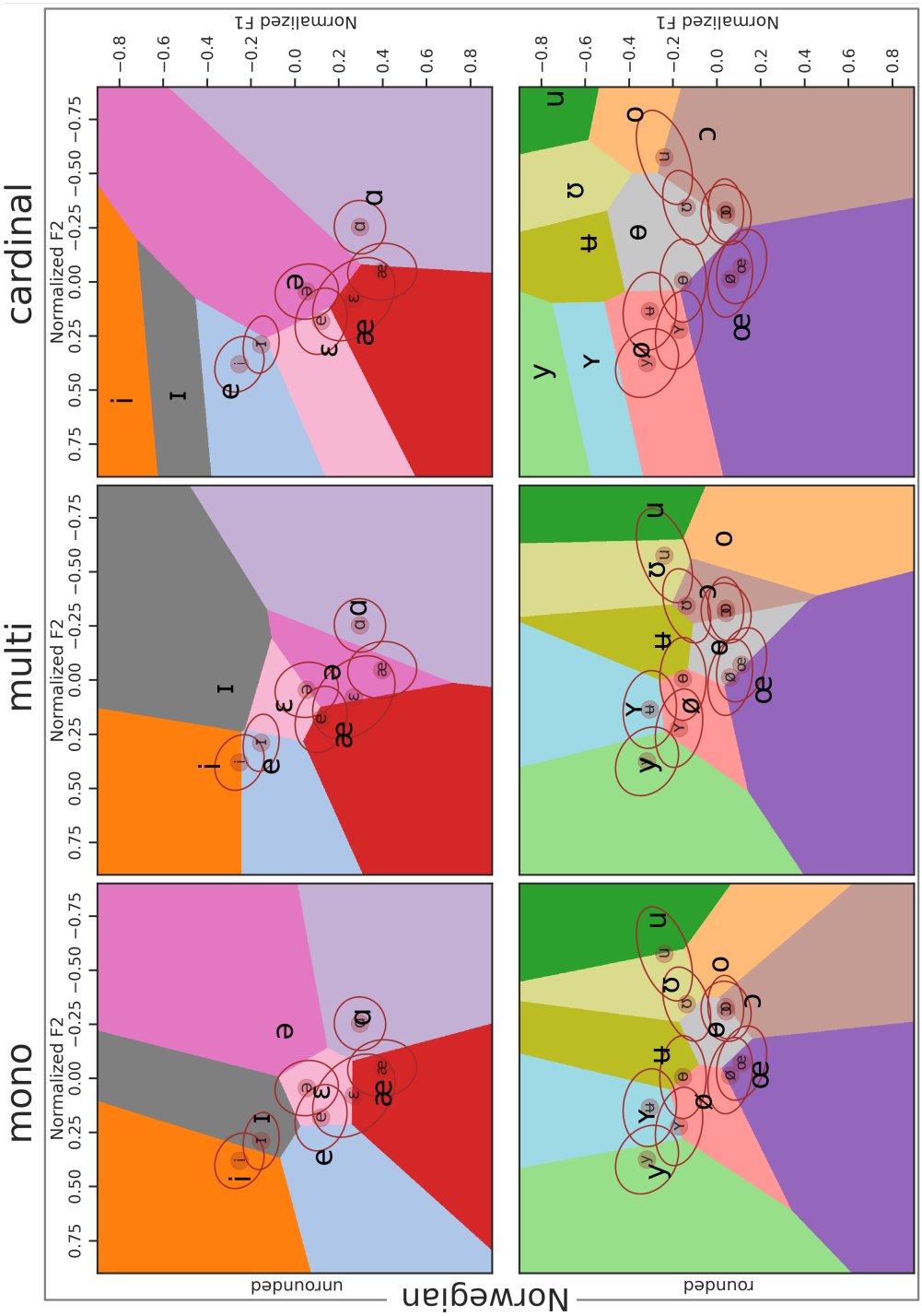


Figure 10.4: The decision boundaries of each vowel category and vowel categorization methods for Norwegian. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

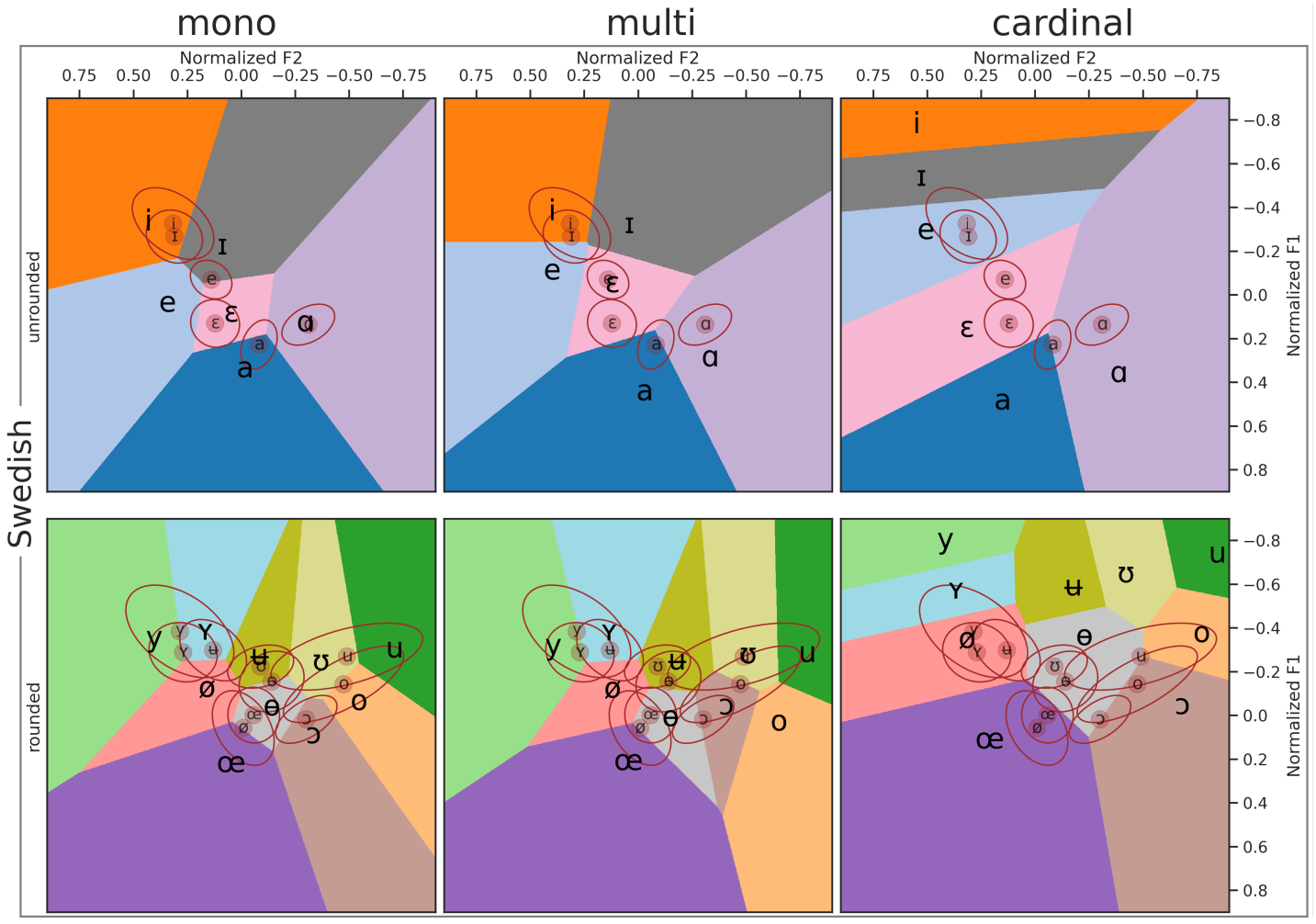


Figure 10.5: The decision boundaries of each vowel category and vowel categorization methods for Swedish. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

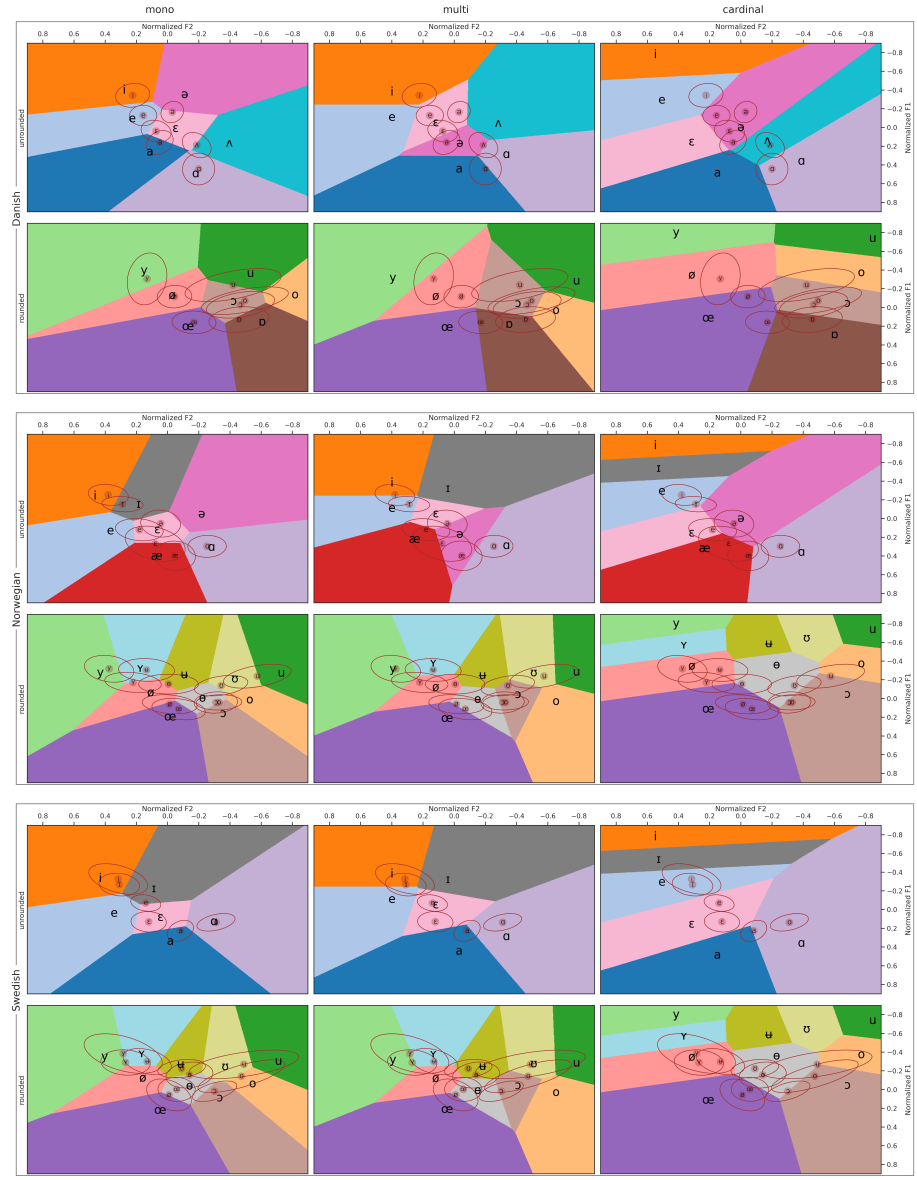


Figure 10.6: The decision boundaries of each vowel category for each of the three categorization methods per language. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

Table 10.3: Unrounded monophthong vowel distribution for each of the categorization methods for each NST subcorpus. The numbers indicate a percentage of total unrounded monophthongs.

vowel	Danish				Norwegian				Swedish			
	nst	*mono*	*multi*	*cardinal*	*nst*	*mono*	*multi*	*cardinal*	*nst*	*mono*	*multi*	*cardinal*
i	11.42	13.44	13.86	2	7.88	15.3	7.62	0	6.15	17.19	12.81	0.26
ɪ	/	/	/	/	15.29	15.12	4.84	1.46	13.59	14.42	7.86	4.18
e	15.29	21.03	8.18	15.43	11.92	13.22	17.1	17.52	29.75	14.24	14.55	22.48
ɛ	13.69	17.88	27.86	10.46	9.64	15.83	13.73	16.83	15.39	20.14	29.85	37.65
æ	/	/	/	/	3.75	13.05	17.66	15.95	/	/	/	/
a	10.71	15.4	11.17	8.5	/	/	/	/	27.52	18	17.74	12.12
ɑ	6.66	10.59	11.07	11.63	18.13	19.02	19.76	20.69	7.6	16.01	17.19	23.3
ʌ	23.83	11.38	11.85	15.5	/	/	/	/	/	/	/	/
	18.41	10.27	16.01	36.47	33.39	8.47	19.3	27.55	/	/	/	/

Table 10.4: Rounded monophthong vowel distribution for each of the categorization methods for each NST subcorpus. The numbers indicate a percentage of total rounded monophthongs.

vowel	Danish				Norwegian				Swedish			
	nst	*mono*	*multi*	*cardinal*	*nst*	*mono*	*multi*	*cardinal*	*nst*	*mono*	*multi*	*cardinal*
y	9.46	10.37	5.28	0.55	3.51	6.66	7.78	0	4.03	7.01	8.36	0.07
ʏ	/	/	/	/	3.5	9.84	6.99	1.26	3.03	10.05	11.12	2.56
ø	8.05	15.19	16.64	14.21	4.1	10.94	12.19	16.26	8.3	10.36	11.59	18.83
œ	4.18	10.29	5.58	16.58	3.04	7.88	7.1	22.56	5.87	6.49	6.4	19.07
ɵ	/	/	/	/	13.24	14.33	16.59	20.52	11.82	13.65	14.51	29.14
ʉ	/	/	/	/	9.85	6.78	6.7	0.52	12.12	9.61	12.42	0.97
ɒ	15.11	14.58	11.12	14.5	/	/	/	/	/	/	/	/
ɔ	20.78	21.43	24.89	40.07	23.7	12.07	22.5	33.05	21.98	11.53	15.17	21.58
o	20.13	15.77	19.55	13.42	22.59	14.11	7.84	5.49	12.12	10.57	4.29	7.08
ʊ	/	/	/	/	6.45	10.2	7.61	0.31	13.14	11.77	10.32	0.62
u	22.28	12.37	16.94	0.67	10.02	7.18	4.7	0.02	7.58	8.97	5.83	0.09

performance by fine-tuning on 1000, 2000, 3000, 4000, 5000, and 10K samples from either training language (so double that number of fine-tuning samples in total). For fine-tuning, we use randomly sampled utterances from the training sets of the NST subcorpora, while the entire tune and development sets are used for validation and evaluation respectively.

The pretrained model is fine-tuned using Connectionist Temporal Classification (CTC) for wav2vec 2.0 (Graves et al., 2006; Baevski et al., 2020a) provided by the Hugging Face Transformers library (Wolf et al., 2020). Each cross-lingual model is fine-tuned on one GPU over 12,000–30,000 training steps. The number of steps is determined as the number of fine-tuning samples + 10,000. We use a train batch size of 4 with 4 gradient accumulation steps, which simulates training on larger batches by accumulating gradients over 4 batches before performing a backward pass. For optimization, we use AdamW with a learning rate of $3 \times 10^{-}5$ with 2,000 warm-up steps and 0.005 weight decay. Once fine-tuning is finished, cross-lingual evaluation is conducted on the entire development set of the evaluation language using the model checkpoint that had the lowest validation loss on the whole tune sets of both training languages. The test sets are not used in this study. They are reserved for prospective follow-up studies that will evaluate the vowel categorization approach extrinsically in downstream tasks.

All fine-tuned models are evaluated in terms of phone error rate (PER) and phone feature Hamming edit distance (PFHED) (Mortensen et al., 2016). PER is computed as the standard word error rate in which each phone token is treated as a word token. It represents the ratio of errors in the hypothesis to the total number of phones in the reference transcript, averaged over all utterances in the evaluation set. However, in this metric, each phone error carries the same weight. For example, for a reference utterance such as [ð ɪ s ɪ z ə kʰ æ t] (*This is a cat.*), the hypotheses such as [d i s ɪ z a k ɛ t] and [o m s ɪ z ɾ v n t] would have the same PER (55.6%), but, all else being equal, a speaker of English is more likely to understand the former whose articulation is closer to the reference than that of the latter. For this reason, we also evaluate the models using PFHED, which takes articulatory/acoustic features (e.g., high, low, front, back, round, etc. for vowels) into account when measuring the error rate. Since both PER and PFHED measure error

rates, lower scores mean better performance. Furthermore, to ensure the observed results are not coincidental, we fine-tune and evaluate each model three times using the same data and hyperparameters, and report the mean error rates and their standard deviation (std) over the three experiment runs.

Results

11

11.1 Introduction

In this chapter, we present and interpret the performance results of the phone and speech recognition models. We start with general corpus-level metrics: phone error rate (PER) and phone feature Hamming edit distance (PFHED), which are calculated on the development partitions of the NST corpora. Subsequently, we break down the corpus-level error rates by dialect region to analyze the models' performance on non-standard regional dialects. Finally, we perform a deeper analysis by looking specifically at phone predictions for each individual vowel in the three evaluation languages. For each language and categorization method, we also interpret the phone prediction results by comparing the prediction rates for each reference vowel and the amount of cross-lingual overlap between the reference and hypothesis vowels in the normalized F_1-F_2 space.

11.2 Cross-Lingual Phone Recognition

To investigate the effect of different types of vowel categorizations on the overall performance on the cross-lingual phone recognition task, we first take a look at the PER and PFHED results of each cross-lingual model.

Table 11.1 shows the mean PERs and Table 11.2 the mean PFHEDs of each cross-lingual model fine-tuned on the different utterance transcripts and different amounts of fine-tuning data. For Danish, when the models are fine-tuned on Norwegian and Swedish, the *multi* and *cardinal* models consistently outperform the baseline by 1.34–1.8 and 3.01–3.47 percentage points on average respectively, with the *cardinal* models achieving the lowest PERs. In terms of PFHED, all models fine-tuned on relabeled transcripts consistently outperform the baselines,

but it is the *multi* models that achieve the lowest edit distances. For Norwegian, when the models are fine-tuned on Danish and Swedish, the *mono* models consistently outperform the baseline in terms of PER, by 1.09–1.64 percentage points on average, while the *cardinal* are mostly below the baseline. In terms of PFHED, however, the baseline models outperform all models fine-tuned on relabeled transcripts despite achieving worse PERs than the *mono* and *cardinal* models. Finally, for Swedish, when the models are fine-tuned on Danish and Norwegian, the PER scores are similar to those for Norwegian. The *mono* models consistently outperform the baseline by 1.13–1.83 percentage points on average, while the *cardinal* models are mostly below the baseline. In terms of PFHED, all models fine-tuned on relabeled transcripts consistently outperform the baselines with the *mono* and *cardinal* achieving the best results (*cardinal* with up to 2000 labeled samples per training language and *mono* with 3000 and more).

Regarding the amount of fine-tuning data, most of the models, irrespective of the transcript type and evaluation language, show little to no improvement with the increase in fine-tuning data past 3000 samples per training language. This indicates that the pre-trained XLSR-53 does not benefit from larger amounts of fine-tuning data when applied to cross-lingual phone recognition on the NST corpus.

11.3 Phone Recognition on Dialect Regions

To examine how the three formant-based vowel categorizations affect cross-lingual phone recognition on non-standard regional dialects, we first select the best cross-lingual models for each evaluation language, and, then, break down their performance by dialect region. The best models have the lowest mean PER + std among the models fine-tuned on different amounts of labeled data and are shown enclosed in a rectangular frame in Table 11.1.

Tables 11.3, 11.4, and 11.5 show the mean PERs and std of the best cross-lingual models on Danish, Norwegian, and Swedish respectively broken down by dialect region. We also computed the vowel distance between a non-standard region and the capital region as the mean Mahalanobis distance (MD) between all vowel points of the non-standard region, expressed in terms of normalized $F_1 - F_2$, and

Table 11.1: Mean PERs and std (%) of all cross-lingual models averaged over three experiment runs. The best results for each number of fine-tuning samples per training language and each evaluation language are shown in bold. The results in a rectangular frame indicate the best models for each categorization type across different sample sizes.

samples per train. lang.		1,000	2,000	3,000	4,000	5,000	10,000
eval. lang.	**transcript**						
Danish	*nst*	53.65±0.20	53.35±0.12	53.17±0.10	52.94±0.19	52.92±0.24	52.98±0.12
	mono	54.00±0.29	54.18±0.05	54.18±0.17	54.40±0.16	54.33±0.07	54.24±0.07
	multi	51.85±0.26	51.85±0.11	51.51±0.07	51.60±0.03	51.36±0.03	51.29±0.14
	cardinal	**50.41±0.32**	**50.29±0.18**	**49.70±0.12**	**49.92±0.11**	**49.75±0.26**	**49.97±0.18**
Norwegian	*nst*	41.85±0.30	41.22±0.16	40.89±0.15	40.92±0.11	40.96±0.28	40.69±0.16
	mono	**40.61±0.23**	**39.65±0.31**	**39.32±0.30**	**39.35±0.29**	**39.32±0.20**	**39.60±0.49**
	multi	42.59±0.31	41.77±0.24	41.91±0.17	41.86±0.23	41.92±0.16	42.22±0.33
	cardinal	40.77±0.45	39.99±0.08	40.03±0.07	40.20±0.20	39.91±0.54	40.82±0.26
Swedish	*nst*	44.72±0.25	43.78±0.05	43.58±0.17	43.31±0.05	43.21±0.12	42.91±0.26
	mono	**43.59±0.17**	**42.26±0.10**	**41.75±0.09**	**41.68±0.06**	**41.46±0.16**	**41.18±0.06**
	multi	46.24±0.05	45.93±0.25	45.74±0.20	45.38±0.12	45.37±0.38	45.00±0.22
	cardinal	44.35±0.20	43.43±0.04	43.16±0.18	43.16±0.04	43.03±0.06	42.86±0.16

Table 11.2: Mean PFHEDs and std of all cross-lingual models averaged over three experiment runs. The best results for each number of fine-tuning samples per training language and each evaluation language are shown in bold.

samples per train. lang.		1,000	2,000	3,000	4,000	5,000	10,000
eval. lang.	**transcript**						
Danish	*nst*	7.83±0.19	7.68±0.08	7.80±0.07	7.64±0.14	7.72±0.09	7.93±0.11
	mono	7.19±0.05	7.22±0.03	7.25±0.01	7.41±0.06	7.33±0.07	7.47±0.02
	multi	**6.92±0.14**	**6.92±0.08**	**6.89±0.04**	**7.05±0.02**	**6.93±0.01**	**7.04±0.04**
	cardinal	7.16±0.19	7.17±0.08	7.19±0.01	7.26±0.02	7.17±0.06	7.29±0.06
Norwegian	*nst*	**6.46±0.02**	**6.14±0.07**	**6.23±0.06**	**6.28±0.06**	**6.28±0.02**	**6.28±0.11**
	mono	6.80±0.09	6.59±0.05	6.50±0.04	6.55±0.04	6.63±0.01	6.62±0.02
	multi	7.09±0.09	6.75±0.19	6.95±0.02	6.93±0.01	6.90±0.06	7.01±0.02
	cardinal	6.98±0.11	6.70±0.01	6.66±0.04	6.64±0.05	6.68±0.04	6.71±0.07
Swedish	*nst*	7.02±0.08	6.84±0.01	6.87±0.05	6.83±0.02	6.82±0.02	6.80±0.09
	mono	6.68±0.02	6.54±0.02	**6.46±0.03**	**6.51±0.08**	**6.52±0.04**	**6.55±0.01**
	multi	6.62±0.03	6.70±0.06	6.69±0.05	6.67±0.07	6.71±0.13	6.76±0.05
	cardinal	**6.57±0.02**	**6.53±0.05**	6.54±0.03	6.55±0.05	6.57±0.06	6.60±0.04

the vowel distributions of the capital region. The MD is chosen because, for each vowel distribution, it takes into account the variance and correlations in the data. Furthermore, it can be interpreted as the number of standard deviations away from the mean of the capital region vowel distributions (Lohninger, 2013).

For Danish, the models fine-tuned on *cardinal* transcripts consistently achieve the best PERs, with some of the (more distant) dialect regions, i.e., West, South, and East Jutland, outperforming the baselines by ≥4% points. In the case of the Norwegian dialect regions, the lowest PERs are achieved alternately by models fine-tuned on *mono* and *cardinal* transcripts. Some of the non-capital regions with particularly better PERs than the baselines (≥2% points of performance gain) include Oslo Outer Fjord, Voss, Hedmark, and Bergen. Finally, for Swedish, the lowest PERs are achieved by the *mono* models. Here, the non-standard regions with particularly better PERs than the baseline (≥2% points of performance gain) are Middle Sweden, Östergötland, Västergötland, West Sweden, and Gothenburg.

To systematically investigate the effect of different vowel categorizations on the cross-lingual performance on non-standard dialect regions, we carry out correlation analyses on the models' performance gain on each dialect region as a function of the regions vowel distance from the capital region. The performance gain is in comparison to the performance of the baseline models. For each non-standard dialect region and categorization method (*mono*, *multi*, *cardinal*), we calculate how much the model's performance differs from the baseline (*nst*) on the same dialect. Then, we plot these performance gains as a function of the region's mean MD from the vowel space of the capital region and measure the correlations for each categorization method and evaluation language (Figure 11.1). The analyses reveal weak and statistically non-significant trends. For *mono*, the Pearson's correlation coefficients (r) are all negative (r={−0.69, −0.42, −0.2} for all three evaluation languages (Danish, Norwegian, and Swedish regions respectively), which makes sense as monolingual clustering is not expected to be helpful for cross-lingual phone recognition. They are close to 0 for *multi* (r={0.03, 0.17, −0.22}) and slightly positive for *cardinal* (r={0.5, 0.19, 0.15}). Though a very weak trend, the performance on dialect regions more distant from the capital seems to improve

Table 11.3: Mean PERs and std (%) of the best cross-lingual models on Danish broken down by dialect region and averaged over the three experiment runs. The non-capital regions are sorted by their mean vowel distance from the capital region. The distance values are shown in parentheses with the region names in the column headers. The best results for each dialect region are shown in bold.

	Copenhagen metro. area	Funen (1.16)	N. Jutland (1.20)	W. Jutland (1.21)	S. Jutland (1.21)	W. and S. Zealand (1.22)	E. Jutland (1.23)	**Total**
nst	53.12 ±0.23	51.56 ±0.13	53.40 ±0.15	53.03 ±0.26	53.89 ±0.19	52.83 ±0.16	52.75 ±0.29	52.94 ±0.19
mono	53.95 ±0.08	52.28 ±0.18	54.39 ±0.03	53.68 ±0.20	55.38 ±0.05	54.78 ±0.07	55.01 ±0.06	54.18 ±0.05
multi	50.53 ±0.09	50.33 ±0.08	52.25 ±0.10	50.55 ±0.13	51.50 ±0.01	53.24 ±0.13	51.07 ±0.21	51.36 ±0.03
cardinal	**49.20 ±0.24**	**49.69 ±0.12**	**50.12 ±0.15**	**48.82 ±0.02**	**49.93 ±0.11**	**51.06 ±0.13**	**48.94 ±0.17**	**49.70 ±0.12**

Table 11.4: Mean PERs and std (%) of the best cross-lingual models on Norwegian broken down by dialect region and averaged over the three experiment runs. The non-capital regions are sorted by their mean vowel distance from the capital region. The distance values are shown in parentheses with the region names in the column headers. The best results for each dialect region and evaluation language are shown in bold.

	Oslo metro. area	Troms (1.17)	Voss and surroundings (1.20)	Oslo Outer Fjord (1.20)	Nordland (1.20)	Hedmark and Oppland (1.21)	Sørlandet (1.23)	Trøndelag (1.25)	Sunnmøre (1.27)	Bergen and Outer Vestland (1.28)	South Vestland (1.30)	**Total**
nst	41.49	39.05	39.42	39.37	39.58	41.09	38.88	40.28	39.78	46.98	41.74	40.69
	±0.11	±0.04	±0.14	±0.19	±0.20	±0.09	±0.13	±0.22	±0.14	±0.78	±0.15	±0.16
mono	**39.82**	**38.10**	**36.42**	**36.78**	38.92	**38.85**	37.81	**39.20**	**39.77**	46.12	40.72	**39.32**
	±0.11	**±0.17**	**±0.22**	**±0.18**	±0.33	**±0.29**	±0.19	**±0.35**	**±0.14**	±0.43	±0.29	**±0.20**
multi	41.28	41.78	40.01	40.28	41.31	40.00	40.40	43.41	42.99	45.49	42.49	41.77
	±0.39	±0.18	±0.23	±0.32	±0.32	±0.25	±0.13	±0.24	±0.14	±0.20	±0.52	±0.24
cardinal	41.00	39.86	37.46	37.53	**38.71**	41.05	**37.76**	40.46	41.61	**44.60**	**39.88**	39.99
	±0.23	±0.09	±0.07	±0.17	**±0.15**	±0.24	**±0.10**	±0.12	±0.11	**±0.32**	**±0.03**	±0.08

Table 11.5: Mean PERs and std (%) of the best cross-lingual models on Swedish broken down by dialect region and averaged over the three experiment runs. The non-capital regions are sorted by their mean vowel distance from the capital region. The distance values are shown in parentheses with the region names in the column headers. The best results for each dialect region and evaluation language are shown in bold.

	Stockholm metro. area	Middle Sweden (1.10)	Östergötland (1.15)	West Sweden (1.17)	Dalarna with surroundings (1.18)	Norrland (1.19)	Eastern South Sweden (1.19)	Gothenburg with surroundings (1.20)	Västergötland (1.21)	Western South Sweden (1.26)	**Total**
nst	42.77 ±0.11	43.66 ±0.24	45.10 ±0.22	42.72 ±0.14	42.15 ±0.07	42.02 ±0.07	43.41 ±0.14	43.07 ±0.07	41.37 ±0.15	45.80 ±0.24	43.21 ±0.12
mono	**40.16 ±0.15**	**40.91 ±0.08**	**42.47 ±0.04**	**40.58 ±0.07**	**41.01 ±0.03**	**41.57 ±0.09**	**42.09 ±0.01**	**40.13 ±0.11**	**39.61 ±0.04**	**43.21 ±0.18**	**41.18 ±0.05**
multi	45.48 ±0.29	45.86 ±0.31	45.51 ±0.14	44.99 ±0.14	43.92 ±0.27	43.94 ±0.24	44.36 ±0.19	43.66 ±0.25	43.89 ±0.16	48.38 ±0.32	45.00 ±0.22
cardinal	44.15 ±0.16	44.11 ±0.09	42.93 ±0.28	42.69 ±0.11	41.92 ±0.14	42.41 ±0.07	41.96 ±0.22	41.44 ±0.26	42.76 ±0.21	44.29 ±0.21	42.86 ±0.16

slightly as vowel categories shift from language-dependent to being more language-independent. However, the correlations are measured on very small samples. Future research should focus on larger sets of speakers and dialects to further investigate these weak trends.

11.4 Phone Prediction Analysis

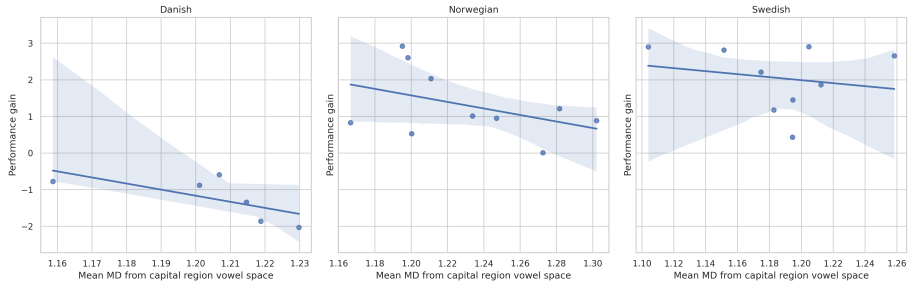
In this section, we analyze individual phone predictions for each of the best cross-lingual models. Specifically, we look at phone confusion matrices normalized over the predicted phones. For each reference phone p_{ref}, the confusion matrix shows the percentage of its tokens that is predicted as each hypothesis phone p_{hyp}. We refer to these percentages as prediction rates.

A review of the full confusion matrices for each evaluation language reveals that all models recognize consonants much better than vowels. Namely, all seen consonants, with the exception of Danish unaspirated stops [b, d, g][1], have relatively high correct recognition rates: over 80% (and most of them over 90%). The recognition rates are similar across models fine-tuned on different relabeled transcript types and on par with those achieved by the baselines. Furthermore, vowel-consonant and consonant-vowel confusions are rare: most of them have prediction rates below 0.5%. On the other hand, the mean and std of the recognition rates of seen vowels are 45.7±14.9%.

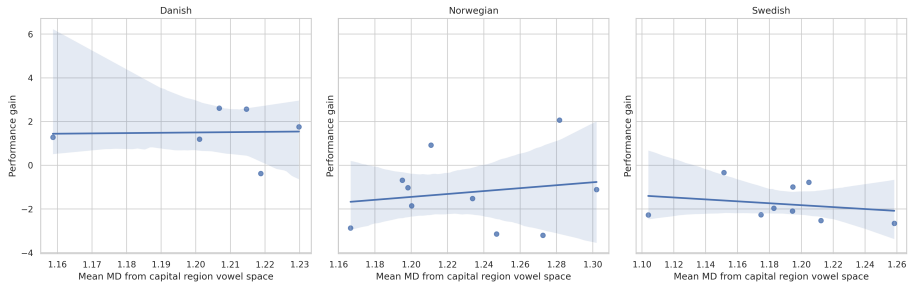
Vowel prediction rates vary substantially across the different transcript types for all three languages. We examined the top 3 predictions for each of the 10 reference vowels that are found in all three languages, which are shown in Table 11.6. Here, we see that the models fine-tuned on any type of relabeled transcripts outperform the baselines on 8 of the 10 shared vowels when evaluated on Danish, 6 out of 10 when evaluated on Norwegian, and 6 out of 10 when evaluated on Swedish.

The biggest improvement over the baseline is seen in the models evaluated on Danish: the *mono* model outperforms the baseline on 6 shared vowels, the *multi* model outperforms the baseline on 8 shared vowels, and the *cardinal* model outperforms all other models on 7 and

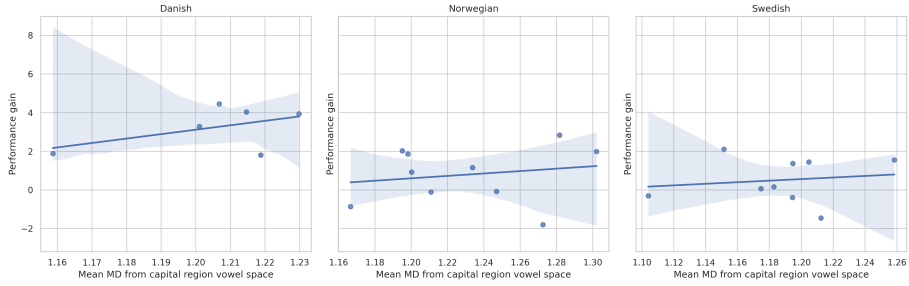
[1] In Danish, [b, d, g] are commonly realized as voiceless unaspirated stops (Grønnum, 1998; Puggaard-Rode et al., 2022).



(a) *mono*



(b) *multi*



(c) *cardinal*

Figure 11.1: Correlation analyses of the cross-lingual models' performance gain on non-standard dialect regions of Danish, Norwegian, and Swedish as a function of the regions vowel distance from the capital region for: a) *mono*, b) *multi*, and c) *cardinal* models. The performance gain is in comparison to the performance of the baseline models (*nst*). The individual plots show the data points fitted with a regression line and a 95% confidence interval.

Table 11.6: Top 3 phone predictions and their prediction rates in % for each reference vowel shared by all three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while *spn* stands for spoken noise and *sil* for silence. Bolded results are correct predictions.

	eval lang			Danish			Norwegian			Swedish
	top hyps	1	2	3	1	2	3	1	2	3
ref	ttype									
i	*nst*	**i: 84.77**	ɪ: 7.38	del: 2.82	**i: 92.35**	ɛ: 1.58	ɪ: 1.41	**i: 95.29**	ɪ: 2.27	del: 0.35
	mono	**i: 89.73**	del: 3.47	ɪ: 2.71	**i: 93.38**	e: 1.96	del: 1.40	**i: 88.88**	ɪ: 4.96	del: 2.34
	multi	**i: 87.79**	del: 4.43	ɪ: 2.05	**i: 95.61**	e: 1.29	del: 1.00	**i: 83.72**	e: 8.43	del: 2.94
	cardinal	**i: 91.16**	ɪ: 2.95	del: 1.93	**i: 97.94**	e: 0.79	ɛ: 0.38	**i: 98.00**	ɪ: 0.59	e: 0.48
e	*nst*	**e: 27.91**	ɪ: 23.07	i: 16.05	**e: 55.17**	ɛ: 17.46	del: 13.48	ə: 38.75	**e: 29.56**	del: 12.45
	mono	i: 29.81	ɪ: 20.34	**e: 18.37**	**e: 60.85**	del: 10.53	ɛ: 7.74	**e: 47.17**	ɛ: 9.13	i: 8.60
	multi	**e: 51.91**	i: 15.36	del: 12.26	**e: 48.32**	i: 22.52	del: 8.41	**e: 49.51**	æ: 19.30	del: 6.96
	cardinal	**e: 60.11**	ɪ: 14.60	del: 11.14	**e: 73.03**	ɪ: 7.75	del: 6.69	**e: 55.39**	ə: 15.60	ɛ: 10.09
ɛ	*nst*	del: 25.65	**ɛ: 23.65**	e: 16.82	**ɛ: 73.27**	a: 8.02	e: 8.00	**ɛ: 44.38**	e: 17.56	del: 8.33
	mono	**ɛ: 28.12**	del: 18.98	ɪ: 17.77	**ɛ: 56.99**	e: 10.16	ɪ: 8.52	**ɛ: 51.56**	æ: 13.39	del: 10.41
	multi	**ɛ: 23.66**	e: 21.83	del: 19.32	**ɛ: 63.11**	ɪ: 8.87	e: 8.85	ə: 24.67	**ɛ: 23.30**	æ: 21.56
	cardinal	**ɛ: 42.63**	e: 18.30	del: 15.73	**ɛ: 71.60**	e: 14.43	del: 5.25	ə: 26.64	**ɛ: 25.27**	æ: 20.02
ɑ	*nst*	**ɑ: 55.81**	del: 18.49	æ: 6.59	**ɑ: 41.34**	a: 35.29	del: 7.34	**ɑ: 85.84**	del: 6.73	o: 3.55
	mono	**ɑ: 36.51**	æ: 19.23	del: 16.48	**ɑ: 63.94**	a: 13.84	del: 6.04	**ɑ: 71.38**	ə: 6.50	del: 5.86
	multi	**ɑ: 42.85**	del: 14.86	ə: 10.31	**ɑ: 68.54**	a: 10.21	del: 5.67	**ɑ: 69.90**	ə: 14.25	del: 4.81
	cardinal	**ɑ: 52.80**	del: 13.43	ɔ: 9.14	**ɑ: 75.79**	a: 6.25	ɛ: 4.87	**ɑ: 69.42**	ə: 10.49	del: 5.22
ɔ	*nst*	ɵ: 32.58	o: 23.47	**ɔ: 12.42**	**ɔ: 56.63**	del: 11.93	œ: 4.34	**ɔ: 58.29**	o: 20.88	del: 5.84
	mono	ɵ: 30.94	ʊ: 18.58	del: 14.87	**ɔ: 49.99**	o: 12.07	del: 9.89	**ɔ: 50.72**	o: 16.67	ɵ: 9.11
	multi	**ɔ: 26.82**	ʉ: 19.94	ɵ: 15.22	**ɔ: 51.01**	del: 9.04	ʊ: 6.93	**ɔ: 61.47**	o: 14.55	ɵ: 8.70
	cardinal	**ɔ: 44.87**	ɵ: 30.37	del: 8.13	**ɔ: 63.59**	ɵ: 7.14	o: 5.98	**ɔ: 82.24**	o: 3.49	ɑ: 2.94
o	*nst*	u: 31.49	ʊ: 21.08	**o: 18.38**	**o: 26.56**	ɔ: 25.73	spn: 12.38	**o: 62.97**	u: 14.61	del: 7.16
	mono	u: 30.03	**o: 27.29**	ʊ: 13.74	**o: 49.58**	ɔ: 16.43	spn: 6.83	**o: 65.12**	ɔ: 9.95	u: 7.34
	multi	**o: 31.18**	ɔ: 19.75	del: 12.80	**o: 36.84**	ɔ: 26.46	del: 7.56	**o: 58.92**	ɔ: 18.03	del: 6.46
	cardinal	**o: 51.18**	ɵ: 25.97	ɔ: 8.07	**o: 59.98**	ɔ: 17.05	del: 5.32	**o: 51.17**	ɔ: 30.33	del: 7.04
u	*nst*	**u: 48.10**	ʊ: 22.65	ɵ: 14.71	**u: 69.48**	o: 15.51	del: 5.52	**u: 71.72**	spn: 7.12	ʊ: 6.70
	mono	ʊ: 43.06	**u: 22.88**	del: 12.89	**u: 75.67**	o: 11.13	del: 3.53	**u: 69.26**	o: 11.48	del: 5.58
	multi	**u: 39.39**	ʊ: 31.06	del: 11.31	**u: 79.20**	ʊ: 8.62	o: 3.08	**u: 60.95**	o: 16.30	del: 6.06
	cardinal	o: 34.25	ɵ: 27.37	del: 19.23	o: 50.45	ɔ: 14.66	del: 12.47	o: 27.56	spn: 24.84	ɔ: 12.45
y	*nst*	ʉ: 25.71	del: 23.94	**y: 12.22**	**y: 70.68**	ʏ: 10.44	i: 4.46	**y: 55.06**	i: 13.21	ʉ: 10.20
	mono	ʏ: 25.36	**y: 23.73**	del: 22.22	**y: 64.64**	ʏ: 12.06	ø: 8.08	**y: 47.86**	del: 13.47	ʏ: 10.95
	multi	**y: 42.58**	del: 17.64	ʏ: 15.48	**y: 63.03**	ʏ: 15.44	i: 6.69	**y: 47.39**	ø: 11.40	del: 10.18
	cardinal	del: 23.53	ø: 23.10	**y: 18.05**	ø: 34.74	ʏ: 19.48	e: 14.46	ʏ: 27.90	ø: 15.45	ɵ: 10.44
ø	*nst*	**ø: 23.91**	ɵ: 19.99	ʉ: 18.54	**ø: 84.93**	œ: 6.14	del: 2.57	**ø: 59.34**	ɔ: 9.68	del: 9.47
	mono	**ø: 28.47**	del: 16.41	ʉ: 13.66	**ø: 50.58**	ʉ: 9.77	ɵ: 8.55	**ø: 52.82**	del: 8.90	œ: 6.77
	multi	**ø: 31.39**	del: 16.48	ʏ: 13.75	**ø: 48.59**	ʏ: 20.86	ʉ: 5.22	**ø: 48.69**	œ: 10.95	ɵ: 10.91
	cardinal	**ø: 48.94**	del: 12.84	ɵ: 12.64	**ø: 77.60**	ʏ: 6.60	e: 4.71	**ø: 56.62**	œ: 11.49	del: 8.40
œ	*nst*	del: 41.26	ø: 16.81	**œ: 6.50**	**œ: 79.91**	ø: 7.32	ɛ: 2.73	**œ: 35.79**	ɔ: 22.98	del: 13.20
	mono	**œ: 26.91**	ɵ: 18.33	del: 16.75	ɵ: 24.70	**œ: 18.77**	ɔ: 11.65	**œ: 46.23**	del: 12.39	ɔ: 6.77
	multi	**œ: 26.12**	del: 18.05	ø: 9.94	**œ: 19.21**	ø: 16.98	ɵ: 16.92	**œ: 41.97**	del: 12.05	ɵ: 10.26
	cardinal	**œ: 39.94**	ɵ: 17.67	del: 12.13	**œ: 44.53**	ɵ: 12.54	ø: 12.02	**œ: 64.69**	del: 6.83	ɵ: 6.25

the baseline on 8 out of the 10 shared reference vowels. On average, the correct recognition rates for Danish increase by 9% points over the baseline with the *multi* model, and by 13.8% points with the *cardinal* model. However, looking at recognition rates of the *cardinal* model, we see that it performs exceptionally below average on two reference vowels only: [u, y]. After analyzing the *cardinal* decision boundary plots in Figure 10.6 and measuring vowel distribution in the relabeled transcripts, we discover that these two have become minority vowels in the *cardinal* transcripts. This is a result of their cardinal vowels being too far out in the vowel space. Therefore, it is likely that the *cardinal* models recognize these vowels less because they constitute less than 1% of all vowel tokens in the *cardinal* training data. If we exclude them from the average, the correct recognition rates of the remaining vowels increase to 22.3% points above the baseline.

Among the models evaluated on Norwegian and Swedish: the models fine-tuned on the relabeled transcripts perform better than the baseline on half of the shared reference vowels, but the average recognition rates on Norwegian decrease by 6.6% (*mono*), 7.7% (*multi*), and 7.9% (*cardinal*) points, and on Swedish by 0.7% (*mono*), 5.2% (*multi*), and 9.3% (*cardinal*) points below the baseline. However, here we see again that the *cardinal* model's recognition rates on the minority vowels [u, y] are outliers. Excluding them from the average increases the mean recognition rates of the remaining vowels to 6.7% points above the baseline for Norwegian and 3.9% points above the baseline for Swedish.

We also investigate the models' predictions for vowels encountered in only one of the training languages and vowels not found in the training languages. The recognition rates of the vowels found in only one of the training languages vary greatly depending on both the transcript type and the training and evaluation languages. In particular, the recognition rates for [ɑ], which appears in Danish and Swedish, and [ə], which appears in Danish and Norwegian, are very low: <3% in most cases. This indicates that the presence of a language without these two vowel labels in the training data interferes with cross-lingual transfer of [ɑ, ə] to and from Danish. On the other hand, Danish does not seem to interfere to such an extent with vowel transfer from Norwegian to Swedish and vice versa. Namely, the recognition rates for [ɪ, ʏ, ɵ, ʉ,

Table 11.7: Top 3 predictions and their prediction rates in % for unseen vowels (*del* indicates a deletion). The results are the average over the three experiment runs for each model.

ref	top hyps ttype	1	2	3
ɒ	*nst*	ɾ: 35.06	del: 15.77	ɔ: 15.51
	mono	o: 24.43	del: 20.23	ɾ: 14.33
	multi	del: 21.14	ɵ: 20.10	ɾ: 18.28
	cardinal	ɔ: 25.49	del: 20.66	ɾ: 18.77
ʌ	*nst*	del: 25.18	ə: 13.32	ɑ: 13.21
	mono	del: 22.38	ɑ: 17.51	ə: 12.14
	multi	del: 30.56	ə: 11.51	ɪ: 8.32
	cardinal	del: 26.42	ə: 23.05	ɾ: 9.76
æ	*nst*	ɛ: 50.46	del: 17.51	a: 8.95
	mono	a: 47.89	ɛ: 15.88	del: 14.90
	multi	ɛ: 40.16	a: 20.52	del: 15.62
	cardinal	ɛ: 37.52	a: 29.53	del: 15.20

ʊ] range from 60% to 89%. Regardless of the evaluation language, the baselines outperform the other models on all vowels except [a], where *multi* and *cardinal* perform better but still under 3%.

There are three language-unique vowels in our corpus that we refer to as *unseen* when encountered in the evaluation language: [ɒ, ʌ], found in Danish, and [æ], found in Norwegian. Since a cross-lingual model will never predict nor be able to recognize phone labels that were not seen in training, the recognition rates on unseen vowels are always 0%. The top 3 predictions for the 3 unseen vowels are shown in Table 11.7. As a consonant, [ɾ] is not a plausible prediction for the Danish vowel [ɒ].[2] The top predictions of the other models are closer to [ɒ] in the vowel space, with the *cardinal* model's [ɔ] being the closest both in the cardinal vowel space and the Danish mean vowel space

[2] It likely stems from the lexical similarity between Danish and Norwegian words containing the character sequences ⟨or⟩ or ⟨år⟩, which is a typical spelling of the Danish [ɒ]. In these sequences, the Danish ⟨r⟩ is almost always silent, whereas the Norwegian is either pronounced as [ɾ] or fused with the following consonant when followed by an alveolar Grønnum (1998); Kristoffersen (2000).

(Figure 10.2). The top vowel predictions for the other two unseen vowels are all plausible. As seen in Figure 10.6, they correspond to the vowel categories from the training languages which have the most overlap with a given reference vowel in the vowel space.

Conclusions

12

We have presented a formant-based vowel categorization approach aimed at improving vowel recognition in cross-lingual ASR by reducing confusions stemming from possible notational inconsistencies and phonetic variation of vowels in speech. Specifically, we have performed three types of categorizations: monolingual language-dependent (*mono*), multilingual language-dependent (*multi*), and language-independent (*cardinal*), and investigated their effects on cross-lingual phone recognition using a trilingual corpus comprising Danish, Norwegian, and Swedish.

Our analyses show that the models fine-tuned on the new vowel categories reduce cross-lingual phone error rates on all three languages, as well as phone feature edit distances on Danish and Swedish. The best-performing models are consistent within languages and across variations of sample size and experiment reruns, but different across languages. Namely, the *cardinal* models outperform the baselines in terms of PER on all three languages. They achieve the best performance among the models evaluated on Danish, whereas on Norwegian and Swedish, the best performers are the *mono* models. Moreover, the *cardinal* models result in the highest margins of improvement over the baseline on Danish compared to the best performing models on Norwegian and Swedish.

When it comes to the performance on dialect regions, only weak and statistically non-significant correlations were observed between the models' performance gain on a dialect region and the region's mean vowel distance from the capital. Though still non-significant, these correlations were strongest for Danish dialect regions. Finally, an analysis of individual phone predictions reveals that most shared non-minority vowels benefit significantly from *cardinal* categorization (especially Danish), while all categorization types reduce the recognition rates of vowels absent from one or more training languages. At the same time, a visual comparison of top phone predictions and re-

categorized vowel plots indicates that having the same vowel category overlap in the vowel space across languages increases the vowel recognition rates, whereas a cross-lingual mismatch in vowel categories leads to vowel confusions.

Based on these findings, we can see that cross-lingual vowel recognition remains a challenge, even in the case of a trilingual corpus with three geographically and typologically close languages with similar vowel systems. Nevertheless, we also see that converting vowels into a shared set of formant-based vowel categories can lead to higher recognition rates. Therefore, we propose that future research efforts on formant-based cross-lingual vowel recognition include a larger and more diverse set of languages, use a single shared set of cardinal vowel categories for all languages, and evaluate the resulting transcripts and models in downstream applications, such as ASR or speech synthesis. To deal with the added linguistic diversity, future studies could address additional segmental and suprasegmental features, such as diphthongization, tone, and prosodic prominence, for example, by including measurements of the third formant, vocal intensity, and fundamental frequency. In particular, including the third and possibly higher formants in the analysis could potentially eliminate the need to perform separate categorizations of rounded and unrounded vowels.

Part V

Extrinsic Evaluation of Formant-Based Vowel Representations

Experimental Setup

13

13.1 Introduction

In the previous set of experiments, we saw that formant-based vowel categorization with language-specific vowel sets and language-independent cardinal vowels can significantly improve cross-lingual vowel recognition on vowels that are shared between the evaluation language and across all training languages. This was especially true for the vowel categories that overlapped in the normalized vowel space across all of the investigated languages. On the other hand, a cross-lingual mismatch in vowel categories or absence of a vowel category from one or more training languages was found to increase vowel confusions.

This has prompted a follow-up study of a new set of formant-based vowel categorization methods with a language-universal vowel set that would investigate whether cross-lingual phone recognition models trained on language-universal formant-based vowel representations could generalize to a wider and more diverse set of unseen languages, real-world speech data, and downstream speech recognition. Namely, this study will categorize monophthong vowels from the NST subcorpora, based on their normalized $F_1 - F_2$ values, in a new way derived from the *cardinal* categorization technique analyzed in the previous experiments. We term this vowel categorization method *language-universal* because it converts the language-specific vowel sets of the different training languages into a single unified set of vowel categories shared universally by all training languages. The previous language-independent categorization (*cardinal*) involved vowel categorization with respect to a set of cardinal vowels with hypothetical language-independent formant values, which proved too extreme for the point vowels in Danish, Norwegian, and Swedish as measured from the NST corpus. The difference between the previous language-independent

and the new language-universal categorization is that the new categorization does not rely on hypothetical language-independent formant values for universal cardinal vowel categories, but rather on cardinal vowels whose formant values are determined from the NST corpus.

The crucial question, then, is which cardinal vowel categories to choose for this kind of categorization given that vowel systems can differ vastly across languages. When deciding which vowel categories to include in the cardinal vowel sets, we follow three principles: 1) the vowel set should consist only of cardinal vowels, 2) the selected vowel categories should be symmetrical and evenly spread out across the abstract vowel quadrilateral, and 3) the vowel set should contain top n most common such vowels across studied languages of the world, as documented by PHOIBLE, an online repository of cross-linguistic phonological inventory data (UCLA Phonological Segment Inventory Database, 2019). This way we ensure the selected vowel categories are the most likely to occur in a given language, as well as that they cover the entire vowel space.

Since vowel inventory sizes of the world's languages can range from 2 to 17 cardinal vowels, according to PHOIBLE, are not always evenly distributed and symmetrical, and might feature a number of less frequent vowels, there is no one set of vowels that can cover vowel distinctions in every language. For this reason, we investigate three levels of language-universal vowel categorization, with three sets of cardinal vowels of different sizes: *uni-5*, *uni-10*, and *uni-16*, where the number indicates the size of the cardinal vowel set. *uni-5* and *uni-10* levels consist of 5 and 10 primary vowels respectively. Unrounded front vowels, unrounded open central vowels, and rounded back vowels are often referred to as primary as they are more common across languages than their rounded/unrounded counterparts. Still, a number of languages distinguish one or more of the less common, secondary vowels, which have opposite roundedness to their primary counterparts. For this reason, the third level of vowel categorization *uni-16* includes both primary and secondary vowel categories: 8 unrounded-rounded pairs, or 16 vowels in total. The placement of vowel categories for each categorization level is shown in Table 13.1. We expect these three configurations of vowel categories would respectively cover most vowel distinctions in languages with small, medium, and large vowel

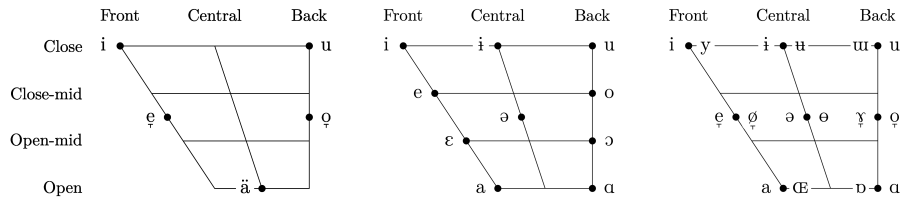


Figure 13.1: The placement of vowel categories on the abstract vowel quadrilateral for each categorization level from left to right: *uni-5*, *uni-10*, and *uni-16*.

inventories.

Although these sets of cardinal vowels are still unlikely to match vowel distinctions in most languages, they will allow us to investigate four questions: 1) whether language-universal vowel categorization can improve cross-lingual vowel recognition, 2) how the different levels of language-universal vowel categorization relate to the actual vowel inventories of the target languages, 3) whether improvement in cross-lingual vowel recognition with formant-based vowel categories can translate to improvement in downstream ASR, and 4) how changes in individual vowel recognition rates affect downstream word recognition.

These questions are addressed in a two-stage evaluation pipeline: intrinsic evaluation in terms of phone error rate and phone feature Hamming edit distance and extrinsic evaluation in terms of word error rate. In addition to intrinsic evaluation on the NST corpus, we also perform both intrinsic and extrinsic evaluation on five parliamentary speech corpora: Danish, Icelandic, Catalan, Serbian, and Finnish, and five low-resource noisy telephone speech data sets from the Babel program: Lao, Zulu, Amharic, Mongolian, and Javanese. The evaluation languages are chosen for their typological and phonological diversity and distance from the training languages in the NST corpus. At the same time, parliamentary and telephone speech data allow us to expand our evaluation of formant-based vowel representations from clean read speech in the previous set of experiments to more challenging speech domains, including real-world spontaneous and conversational speech, noisy and lossy audio quality, and low-resource languages.

13.2 Data Preparation

For this set of experiments, we will once again use the NST corpus for formant-based vowel categorization and intrinsic evaluation of the phone recognition models trained on the resulting relabeled transcripts. The cross-lingual phone recognition models will then be evaluated on parliamentary and noisy telephone speech corpora both intrinsically in terms of phone and phone feature error rates and extrinsically in terms of word error rate. The data preparation methods for the NST corpus are already described in Section 10.2, so this section will describe the data preparation methods for only the parliamentary and noisy telephone speech data.

13.2.1 Parliamentary Data

Of all the parliamentary corpora, Serbian parliament speech corpus is the only one that does not come with standard training-development-testing splits. Therefore, we create a standard split for this corpus by manually selecting 20 testing speakers (10 male and 10 female) and 10 development speakers (5 male and 5 female). The selected male and female speakers are matched for speech duration so that the evaluation partitions have a balanced gender distribution. More detailed corpus statistics and gender distribution by partition are shown in Table 13.1.

Table 13.1: Serbian parliament speech corpus partitions and their size in hours, total number of utterances, tokens, types, and speakers. Speaker counts by gender are given in parentheses as (**F**emale + **M**ale).

Part.	Hours	Utterances	Tokens	Types	Speakers (F+M)
train	874.48	284,546	7,215,124	122,014	598 (215+383)
dev	10.22	2,679	82,852	13,660	10 (5+5)
test	11.21	3,465	92,293	14,079	20 (10+10)
total	895.91	290,690	7,390,269	123,628	628 (230+398)

For the rest of the parliamentary corpora, we use the standard corpus splits for model training and evaluation. Generally speaking,

we use only the development partitions for the intrinsic evaluation of the cross-lingual phone recognition models, creation of cross-lingual pronunciation lexicons, and choosing the best n-gram language model for the extrinsic evaluation. The monolingual ASR models with cross-lingual pronunciation lexicons are evaluated extrinsically on both the development and test partitions.

Both inference with the cross-lingual phone recognition models and extrinsic evaluation with the monolingual hybrid HMM/DNN ASR systems require the input audio data to be in the same format: single-channel WAV with a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 16 kHz. Danish, Catalan, and Finnish parliament speech data are originally available in this format, so they need not undergo any conversion, while Icelandic and Serbian speech data are encoded in a compressed format: MP3 and FLAC respectively. Therefore, we use the sound processing tool SoX to convert the audio data from these two corpora into the WAV audio format (Bagwell et al., 2015).

When it comes to text preprocessing, all corpora except the Serbian parliament corpus are released with normalized utterance transcripts. Text normalization for ASR involves expanding common abbreviations, numbers, dates, and symbols, as well as removing punctuation, capitalization, and unspoken parenthetical remarks and references, in order to make the text as close to the actual speech in the utterances. We normalize Serbian utterance transcripts in accordance with these common practices. First, we remove punctuation, capitalization, and unspoken remarks, and then manually identify and expand abbreviations, numbers, dates, and symbols using a conversion table. Since Serbian is an inflected language where nouns, pronouns, adjectives, and numerals are marked for case, number, and gender, creating an exhaustive conversion table for every token requiring expansion is not feasible. For this reason, we perform only a simplified token expansion in the nominative case unless the token is directly preceded by a preposition commonly denoting a different grammatical case. In the case of number token expansion, we look at the case endings of the nouns commonly found directly after the given number token when determining its inflected forms.

Certain abbreviations such as acronyms can be pronounced in multiple ways depending on the speaker and context, and there is

no way of knowing which form is used without listening to the audio. These and other inconsistent token expansions, both in this corpus and the other parliamentary corpora, will affect the phonetic alignments for both the cross-lingual phone recognition and monolingual ASR model evaluation, but are not expected to have a substantial effect on the overall error rates, because their frequency counts are relatively low.

13.2.2 Low-Resource and Noisy Telephone Data

All Babel data sets come with standard training-development-testing splits. However, the test data is not publicly available. Therefore, we use the development data, which has no speaker overlap with the training data, as the test partition, and set aside randomly selected 10% of the training samples as the development partition. As with the parliamentary corpora, the development partitions are used for the intrinsic evaluation of the cross-lingual phone recognition models, creation of cross-lingual pronunciation lexicons, and selecting the best n-gram language model for the extrinsic evaluation, while both the development and test partitions are used for the extrinsic evaluation with monolingual hybrid ASR systems. More detailed Babel statistics by language are shown in Table 13.2.

Most of the audio files in the Babel data sets are encoded in the A-law audio format, which is a single-channel, lossy audio format typically used to compress telephone signals. They are sampled at 8 kHz and stored in the NIST Sphere file format. A few audio file are in the single-channel WAV format with a 24-bit linear PCM sample encoding (PCM_S24LE) sampled at 48 kHz. Since our models expect inputs in the WAV format, we use the sound processing tool sph2pipe to convert the NIST Sphere audio files to WAV (, LDC), and the tool SoX to resample the WAV inputs to 16 kHz (Bagwell et al., 2015).

Since Babel data sets contain noisy conversational speech, the utterances feature numerous instances of verbal and non-verbal noise, such as cough, laugh, overlapping speech, disfluencies, clicking, ringing, etc. These are marked with various placeholder tokens in the utterance transcripts. We do not need so many different noise tokens since we are only interested in intelligible speech, so we merge the

Table 13.2: Babel data set partitions per language and their size in hours, total number of utterances, tokens, types, and speakers. The train and dev speakers are the same.

Lao					
	Hours	**Utterances**	**Tokens**	**Types**	**Speakers**
train	58.99	59,578	540,302	6,112	733
dev	6.58	6,620	59,919	2,804	
test	10.56	11,342	96,576	2,910	119
total	76.13	77,540	696,797	6,679	852
Zulu					
	Hours	**Utterances**	**Tokens**	**Types**	**Speakers**
train	55.89	54,746	365,267	53,272	718
dev	6.23	6,083	41,165	11,035	
test	10.42	10,505	66,887	14,649	119
total	72.54	71,334	473,319	64,126	837
Amharic					
	Hours	**Utterances**	**Tokens**	**Types**	**Speakers**
train	39.37	37,299	252,840	31,024	478
dev	4.31	4,145	27,908	6,710	
test	11.64	10,315	71,669	12,326	121
total	55.32	51,759	352,417	38,944	599
Mongolian					
	Hours	**Utterances**	**Tokens**	**Types**	**Speakers**
train	41.87	40,398	362,696	20,595	492
dev	4.58	4,489	39,923	5,340	
test	11.31	11,145	98,449	8,945	120
total	57.76	56,032	501,068	25,153	612
Javanese					
	Hours	**Utterances**	**Tokens**	**Types**	**Speakers**
train	41.09	41,873	278,055	13,770	480
dev	4.45	4,653	30,509	4,033	
test	11.36	11,269	77,937	6,328	120
total	56.90	57,795	386,501	16,634	600

numerous placeholder tokens into four broader categories: unknown or unintelligible speech, verbal noise, non-verbal noise, and silence. For this step, we follow the noise token conversion rules found in a preprocessing script in Kaldi's `babel` recipe.[1] This is the only text preprocessing step required as the utterance transcripts are already normalized. We use the transcripts in the standard orthography for languages and experiments. The romanized transcripts in the Lao, Amharic, and Mongolian data sets are not used.

13.3 Formant-Based Vowel Categorization with a Language-Universal Vowel Set

For this set of experiments, we use the same phonetic corpus alignments and vowel formant estimates and normalization conducted on the NST corpus in the previous set of experiments described in Section 10.3. Therefore, this section will only describe the methods for vowel categorization with a language-universal vowel set.

All three levels follow the same categorization procedure. They only differ in the number of vowel categories they distinguish. As mentioned before, this categorization procedure is based on the *cardinal* categorization from Chapter 10, but instead of categorizing vowel points with respect to hypothetical formant values of cardinal vowels, we determine the positions of cardinal vowels in the vowel space based on $F_1 - F_2$ measurements of the point vowels: [i, u, a, ɑ] in the NST corpus. This is achieved by extracting all point vowels from all three NST subcorpora together, taking the mean of each point vowel for each speaker in the corpus, and then calculating the grand mean over all speaker mean point vowels. The obtained four grand means constitute the four point cardinal vowels of our language-universal vowel systems. Connecting the four points with line segments results in a quadrilateral that resembles the abstract vowel quadrilateral. We use the four point cardinal vowels to determine points along the quadrilateral that will serve as the inner cardinal vowels. Since the Norwegian vowel

[1] The script is part of the original Kaldi recipe for Babel data sets and can be found here: `https://github.com/kaldi-asr/kaldi/blob/master/egs/babel/s5d/local/prepare_acoustic_training_data.pl`.

system used in the Norwegian NST lexicon does not feature the vowel [a], we use the measurements of the Norwegian [æ]-vowels in the calculation of the cardinal vowel [a]. As seen from Figure 10.2 in Section 10.3, the Norwegian [æ] is relatively close to the Swedish [a] based on their realizations in the NST subcorpora. Figure 13.2 illustrates the creation of the point vowel quadrilateral from the normalized formant measurements of the point vowels in the NST corpus.

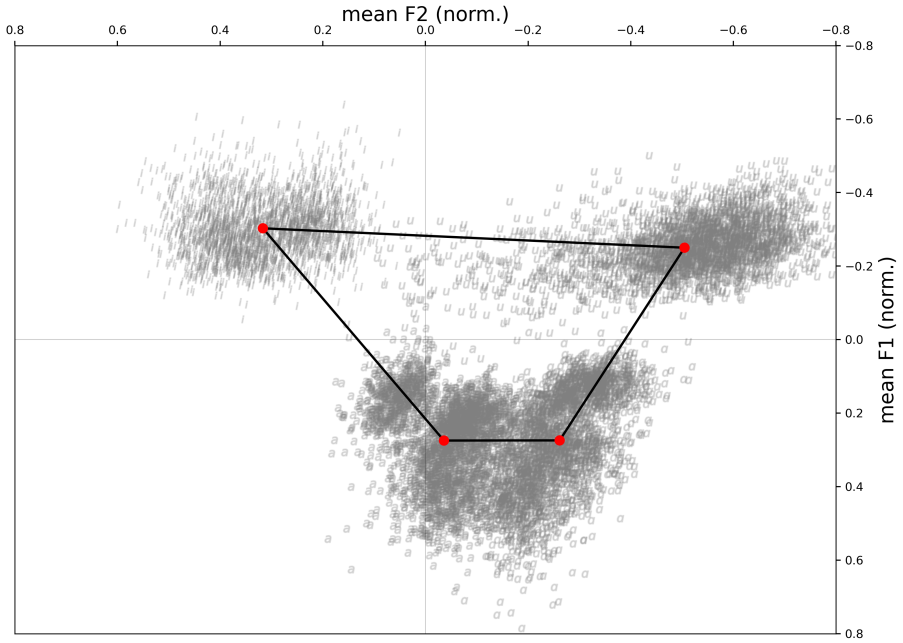


Figure 13.2: The point vowel quadrilateral used for determining the central values of the cardinal vowel categories for each level of language-universal vowel categorization. The red markers are the central values of the point vowels: [i, u, a, ɑ]. They are the grand means of the speakers' means for each of the point vowel categories as measured from the NST corpus, which are plotted in gray.

As introduced in Section 13.1, the *uni-5* categorization level distinguishes 5 vowel categories: [i, e̝, ä, o̝, u]. The diacritical mark for lowering (̝) below [e] and [o] is a phonetic symbol indicating that these vowels are slightly lower (more open) than their usual positions, while the two dots over [a] indicate a more central vowel than plain [a]. The positions of the three inner cardinal vowels [e̝, ä, o̝] are determined simply as the midpoints of the left, right, and bottom side of the point

vowel quadrilateral. Figure 13.3 compares the abstract and data-driven vowel categories for this categorization level.

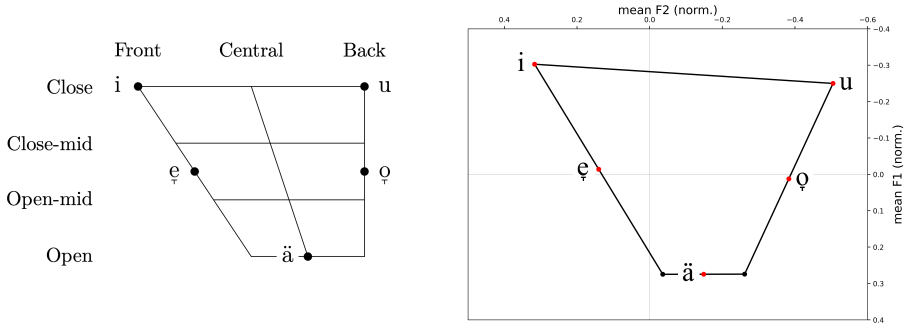


Figure 13.3: The *uni-5* vowel categories on the abstract vowel quadrilateral (left) and NST point vowel quadrilateral (right). The vowel labels left and right of the position marker indicate unrounded and rounded lip shape respectively.

The *uni-10* categorization level distinguishes 10 vowel categories: [i, e, ɛ, a, ɑ, ɔ, o, u, ɨ, ə]. The positions of the front [e, ɛ] and back inner cardinal vowels [ɔ, o] are determined as the points dividing the left and right sides of the point vowel quadrilateral into three equal parts. The position of the close central unrounded vowel [ɨ] is determined as the midpoint of the top side of the point vowel quadrilateral. The position of the mid central vowel [ə] is obtained as the cross-section of the mid and central line segments $\overline{\text{e̞o̞}}$ and $\overline{\text{ɨä}}$. Figure 13.4 compares the abstract and data-driven vowel categories for this categorization level.

Finally, The *uni-16* categorization level distinguishes 16 vowel categories, 8 unrounded: [i, e̞, a, ɑ, ɣ̞, ɯ, ɨ, ə] and 8 rounded: [y, ø̞, œ, ɒ, o̞, u, ʉ, ɵ]. As previously explained, the four points of the NST point vowel quadrilateral are obtained from the primary point vowels, which are unrounded when front and open: [i, a, ɑ], and rounded when back and close: [u]. Their rounded/unrounded counterparts, secondary point vowels, are given the same four points of the quadrilateral even though their average $F_1 - F_2$ values are not exactly matched with those of the primary point vowels. This is done to make the data-driven quadrilaterals closer to the abstract quadrilateral, as well as to avoid a skewed unrounded vowel quadrilateral due to the absence of the close

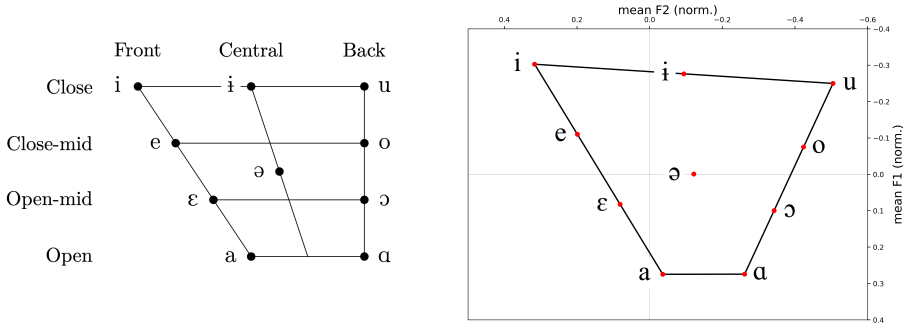


Figure 13.4: The *uni-10* vowel categories on the abstract vowel quadrilateral (left) and NST point vowel quadrilateral (right). The vowel labels left and right of the position marker indicate unrounded and rounded lip shape respectively.

back unrounded vowel [ɯ] from the vowel systems of the three Scandinavian languages. The positions of the pairs of inner unrounded and rounded cardinal vowels are determined in the same way as described in the paragraphs on the *uni-5* and *uni-10* categorization levels. A comparison on abstract and data-driven vowel categories for *uni-16* categorization level is shown in Figure 13.4.

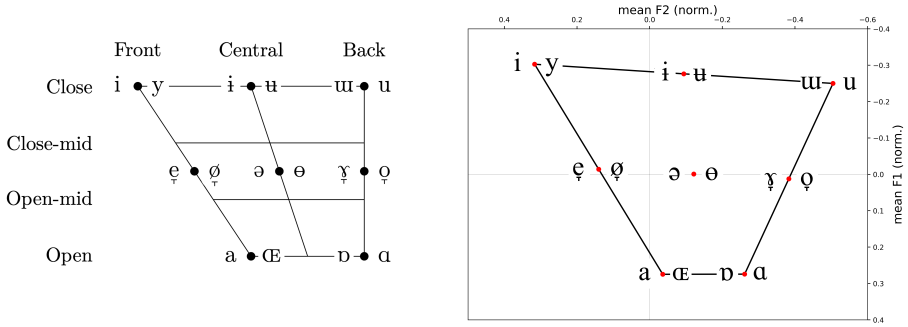


Figure 13.5: The *uni-16* vowel categories on the abstract vowel quadrilateral (left) and NST point vowel quadrilateral (right). The vowel labels left and right of the position marker indicate unrounded and rounded lip shape respectively.

All three vowel categorizations are performed in the same way: by selecting the appropriate cardinal vowel set, computing the central values of its cardinal vowel categories, and, then, recategorizing the source vowels based on their Euclidean distance to the cardinal vowels. This means that each source monophthong is classified as its first

nearest cardinal vowel in the normalized $F_1 - F_2$ space. With *uni-5* and *uni-10* categorization, all monophthongs are recategorized with respect to the same single language-universal vowel set, whereas, with *uni-16*, unrounded and rounded monophthongs are recategorized separately: unrounded monophthongs are recategorized with respect to the unrounded, and rounded with respect to the rounded cardinal vowel set.

As in the previous set of experiments presented in Chapter 4 10, vowel tokens whose normalized formant values are more than 2 standard deviations (std) from the mean are excluded from any data-driven categorization. These tokens are considered outliers which might result from errors in phonetic alignment or formant estimation. However, the outlier vowels also have to be relabeled in order to effectively replace the original vowel sets with the language-universal cardinal vowel sets. This manual recategorization of outliers is performed based on the abstract vowel quadrilaterals. If the outlier vowel representation is in the target cardinal vowel set, it is left unchanged. If it is not in the target cardinal vowel set, it is converted to its nearest cardinal vowel on the abstract vowel quadrilateral. For example, for the *uni-5* and *uni-10* cardinal vowel sets, all outliers with one of the following representations [i, ɪ, y, ʏ] will be relabeled as [i]. Table 13.3 shows the percentage of outlier and recategorized vowel tokens for each categorization level and NST subcorpus out of both all monophthong tokens and total phone tokens. The recategorized tokens include both data-driven and abstract recategorization. Figures 13.6-13.8 show the clustering decision boundaries for each of the categorization methods in relation to the original vowel distributions of Danish, Norwegian, and Swedish respectively, while Figure 13.9 contains all three previous figures for easier cross-lingual comparison. Tables 13.4 and 13.5 show how each of the categorization methods affects the distribution of unrounded and rounded monophthongs respectively in each subcorpus.

Figure 13.6: The decision boundaries of each vowel category and vowel categorization level for Danish. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

Figure 13.7: The decision boundaries of each vowel category and vowel categorization level for Norwegian. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

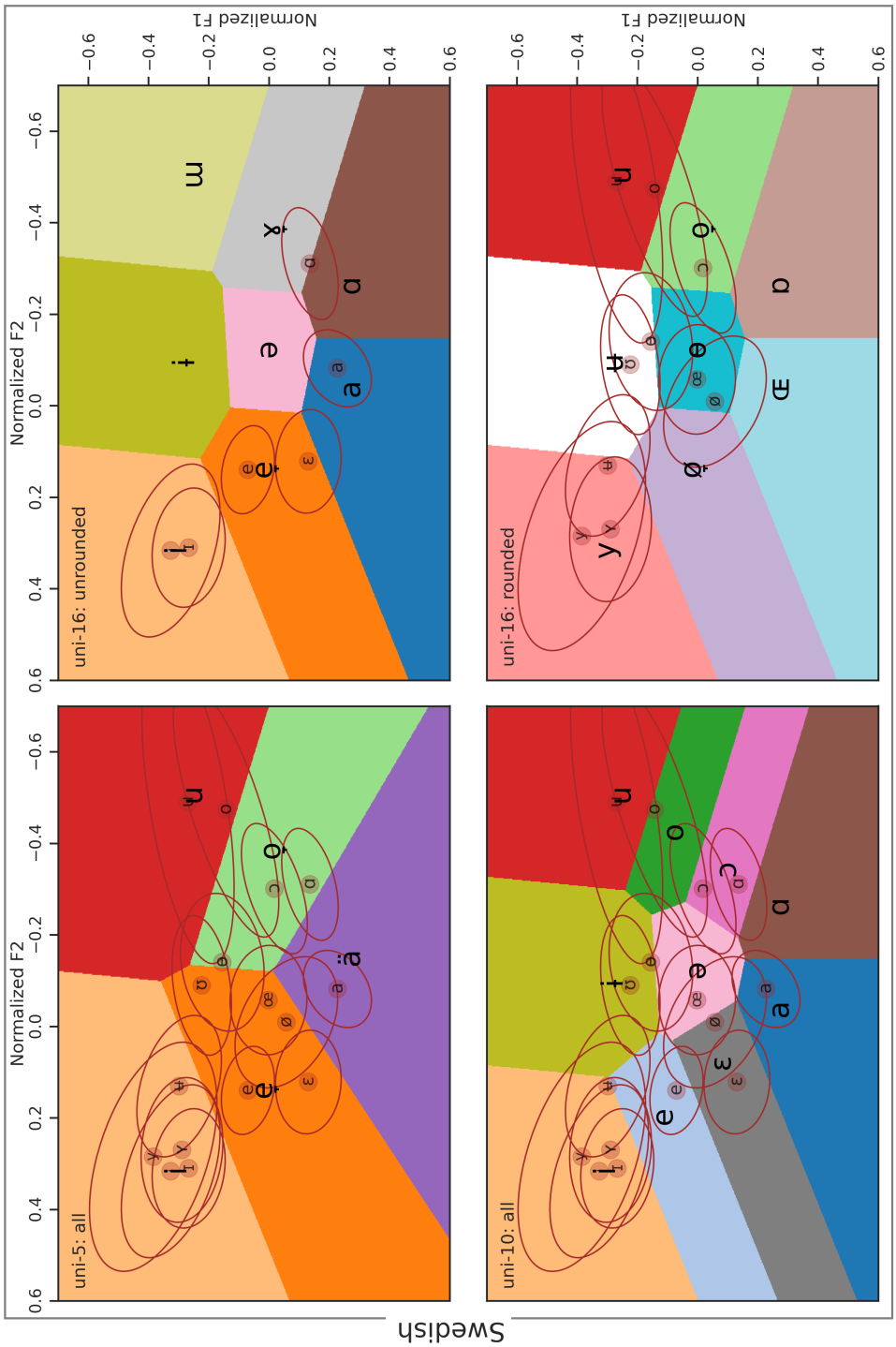


Figure 13.8: The decision boundaries of each vowel category and vowel categorization level for Swedish. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

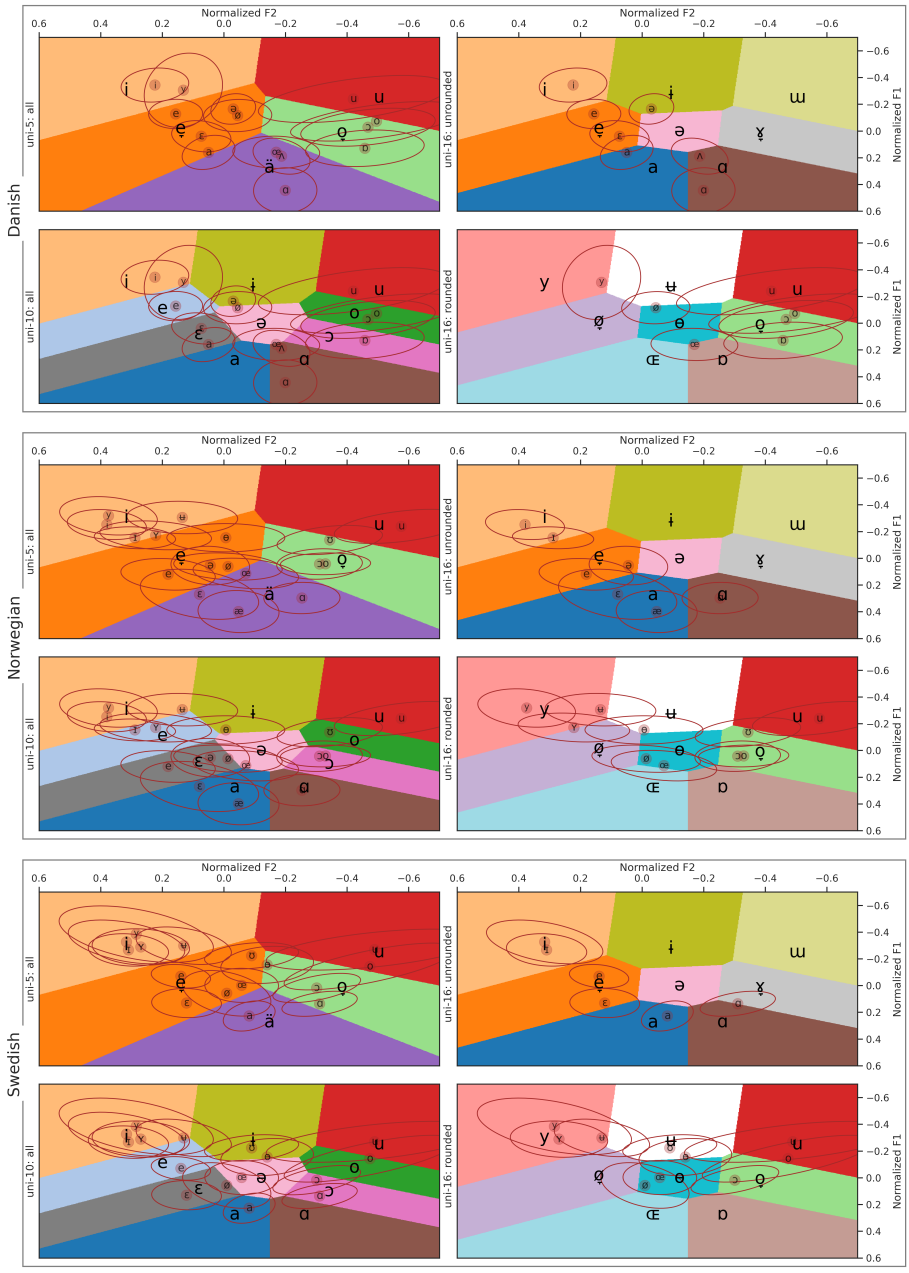


Figure 13.9: The decision boundaries of each vowel category for each of the three categorization levels per language. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses plotted over the decision plots represent the mean of the original vowels and mean vowel spread 2 std from the mean.

Table 13.3: Categorization statistics for each categorization level and language in the NST corpus. For each language and each categorization level, the numbers in the left column indicate % of all monophthong tokens, and the ones in the right % of all phone tokens.

	Danish		**Norwegian**		**Swedish**	
	% monoph.	% total	% monoph.	% total	% monoph.	% total
outliers	4.23	1.76	4.78	1.80	4.01	1.45
uni-5	89.17	37.15	92.67	34.83	94.25	34.17
uni-10	67.97	28.32	69.01	25.94	67.32	24.41
uni-16	77.01	32.08	75.66	28.44	78.15	28.34

13.4 Intrinsic Evaluation: Multilingual and Cross-Lingual Phone Recognition

The three language-universal vowel categorization approaches are assessed intrinsically in a set of phone recognition experiments on 13 different languages spread across two types of speech domains, in-domain: clean read speech (NST corpus), and out-of-domain: parliamentary and low-resource noisy telephone speech (Babel). All phone recognition models are created by fine-tuning the 2-billion parameter pretrained multilingual wav2vec 2.0 model, XLSR-53 (Baevski et al., 2020b; Conneau et al., 2020), on a small subset of the NST corpus. The fine-tuning setup is the same as described in the Section 10.4. As before, we fine-tune and evaluate each model three times, using the same data and hyperparameters, and report the mean error rates and their standard deviation (std) over the three experiment runs to ensure the observed results are not coincidental.

In-domain intrinsic evaluation involves both training and evaluating phone recognition models on the NST corpus transcribed with different types of vowel categorization methods, original dictionary-based transcripts (*nst*), LanguageNet pretrained g2p-based, (*lnet*), and formant-based (*uni-5*, *uni-10*, and *uni-16*). We perform two sets of in-domain experiments: multilingual, where we train on samples from all three NST subcorpora together and evaluate individually on each subcorpus, and cross-lingual, where we train on samples from two of the NST subcorpora and evaluate on the third unseen subcor-

Table 13.4: Unrounded monophthong vowel distribution for each of the categorization levels for each NST subcorpus. The numbers indicate a percentage of total monophthongs.

vowel	Danish				Norwegian				Swedish			
	nst	*uni-5*	*uni-10*	*uni-16*	*nst*	*uni-5*	*uni-10*	*uni-16*	*nst*	*uni-5*	*uni-10*	*uni-16*
i	9.48	19.68	13.19	14.82	5.74	19.91	15.24	15.19	4.41	23.79	18.72	17.18
ɪ	/	/	/	/	11.13	/	/	/	9.73	/	/	/
ɨ	/	/	9.73	8.50	/	/	4.01	2.06	/	/	6.72	3.11
e	12.70	/	14.40	/	8.68	/	15.48	/	21.30	/	14.69	/
e̞	/	31.36	/	19.96	/	31.08	/	19.87		31.51	/	19.04
ɛ	11.37	/	12.73	/	7.02	/	14.31	/	11.02	/	14.03	/
æ	/	/	/	/	2.73	/	/	/	/	/	/	/
a	8.89	/	12.21	14.14	/	/	12.84	14.60	19.70	/	14.39	16.19
ä	/	26.91	/	/	/	27.96	/	/	/	22.92	/	/
ɑ	5.53	/	11.98	12.14	13.20	/	11.68	11.78	5.44	/	6.01	6.54
ʌ	19.79	/	/	/	/	/	/	/	/	/	/	/
	15.29	/	8.97	8.51	24.31	/	9.30	6.64	/	/	8.59	5.88
ɣ̞	/	/	/	3.77	/	/	/	2.40	/	/	/	3.46
ɯ	/	/	/	1.22	/	/	/	0.26	/	/	/	0.18

Table 13.5: Rounded monophthong vowel distribution for each of the categorization levels for each NST subcorpus. The numbers indicate a percentage of total monophthongs.

vowel	Danish				Norwegian				Swedish			
	nst	*uni-5*	*uni-10*	*uni-16*	*nst*	*uni-5*	*uni-10*	*uni-16*	*nst*	*uni-5*	*uni-10*	*uni-16*
y	1.60	/	/	1.19	0.95	/	/	3.63	1.15	/	/	4.62
ʏ	/	/	/	/	0.95	/	/	/	0.86	/	/	/
ø	1.36	/	/	/	1.12	/	/	/	2.36	/	/	/
ø̞	/	/	/	0.95	/	/	/	2.75	/	/	/	2.67
œ	0.71	/	/	/	0.83	/	/	/	1.67	/	/	/
ɶ	/	/	/	0.46	/	/	/	1.03	/	/	/	1.04
ɵ	/	/	/	1.24	3.60	/	/	3.46	3.36	/	/	3.70
ɤ	/	/	/	1.73	2.68	/	/	2.49	3.44	/	/	4.35
ɒ	2.56	/	/	1.21	/	/	/	1.55	/	/	/	0.80
ɔ	3.52	/	5.82	/	6.45	/	7.48	/	6.25	/	6.50	/
o̞	/	13.40	/	4.72	/	14.95	/	7.74	/	13.84	/	5.32
o	3.41	/	5.42	/	6.14	/	5.54	/	3.44	/	4.97	/
ʊ	/	/	/	/	1.75	/	/	/	3.73	/	/	/
u	3.78	8.64	5.55	5.45	2.72	6.1	4.11	4.54	2.16	7.98	5.40	5.92

pus. All models are trained on 3000 random samples from the training set of each NST subcorpus, and evaluated on the development set of each NST subcorpus. We choose 3000 for the number of fine-tuning samples per training language as as our previous phone recognition experiments with different numbers of fine-tuning samples from Chapter 10 showed that the performance of cross-lingual models plateaus when increasing the number of samples beyond 3000. With in-domain evaluation, we evaluate against references transcribed with the same vowel categorization method on which the model was trained, as we have all five types of transcripts for the entire NST corpus.

Out-of-domain intrinsic evaluation involves evaluating the phone recognition models fine-tuned on all three NST subcorpora (same models as used for in-domain multilingual evaluation) cross-lingually and cross-domain on five parliamentary speech corpora in different languages: Danish, Icelandic, Catalan, Serbian, and Finnish, as well as five language packs from the low-resource noisy telephone speech corpus Babel: Lao, Zulu, Amharic, Mongolian, and Javanese. With out-of-domain evaluation, we evaluate all models, including the ones trained on formant-based vowel categories, only against references transcribed using canonical vowel representations (dictionary- or g2p-based), as we do not have formant-based or phonetically annotated transcripts for these corpora. Regardless, analyzing and comparing the performance results of all models across the different languages, domains, and metrics should allow us to measure how closely the models trained on different vowel categorization methods can approach the expected canonical references.

As in the previous set of experiments, all fine-tuned phone recognition models are evaluated in terms of both phone error rate (PER) and phone feature Hamming edit distance (PFHED) (Mortensen et al., 2016). PER is a standard metric that shows the ratio of errors (number of deleted, inserted, and substituted phones) in the hypothesis to the total number of phones in the reference transcript, averaged over all utterances in the evaluation set. With PER, each phone error, insertion, deletion, or substitution, carries the same weight of 1. On the other hand, PFHED gives the same weight to insertions and deletions as PER, 1, but less weight to substitution errors. Namely, it converts all phone tokens in the reference and hypothesis transcripts

into 24-dimensional articulatory/acoustic feature vectors, and then computes the Hamming edit distance between the substituted phones, giving a weight of 1/24 to each feature edit between the reference and hypothesis feature vectors. While PER is useful for downstream ASR tasks where exact transcripts are preferred, PFHED shows us how close the references and hypotheses are in pronunciation.

13.5 Cross-Lingual Pronunciation Lexicons

The next step in the evaluation pipeline involves creating pronunciation lexicons for the extrinsic evaluation of formant-based vowel categorization on out-of-domain speech data: parliamentary and noisy telephone speech. As introduced earlier, the extrinsic evaluation entails training and evaluating modular HMM-DNN ASR systems using monolingual acoustic and language models and cross-lingual pronunciation lexicons derived from the different proposed vowel categorization methods. This will allow us to investigate whether cross-lingual formant-based vowel representations can also be used to differentiate words in word-based speech recognition tasks, as well as to relate changes in individual vowel recognition rates to downstream word recognition rates. To this end, we evaluate and compare ASR systems trained with three types of pronunciation lexicons: monolingual gold standard lexicons, cross-lingual baselines, and cross-lingual formant-based lexicons.

Monolingual gold standard lexicons are standard lexicons used in modular ASR systems. They provide canonical pronunciations for all words in the vocabulary, taken either from human-curated dictionaries or produced by grapheme-to-phoneme conversion software. Modern HMM/DNN ASR systems trained with pronunciation models based on gold standard lexicons can achieve SOTA or near-SOTA results. We train gold standard ASR systems and evaluate them on each evaluation data set from the selected out-of-domain speech corpora. This will give us an estimate of a lower bound on word error rates that can be achieved on these corpora with current modular ASR systems.

For each parliamentary and Babel language, we create one or two gold standard lexicons, depending on the available resources for a given language. Specifically, we use manually transcribed lexicons for

Danish and Icelandic. For Danish, we use the Danish NST lexicon that accompanies the Danish NST subcorpus. For Icelandic, we use the Althingi lexicon accompanying the Althingi, Icelandic parliamentary speech corpus. Since these two lexicons do not cover the entire vocabularies of their corresponding parliamentary corpora, we use them to train grapheme-to-phoneme transducers, which we, then, use to expand the original lexicons to include the whole vocabularies. We use orthographic lexicons for Serbian and Finnish, where the pronunciation transcripts are the same as their corresponding orthographic transcripts. These two languages have near phonemic orthography, where there is an almost one-to-one grapheme-to-phoneme correspondence. For Danish, Serbian, and Finnish, we also use pretrained grapheme-to-phoneme models to transcribe the gold standard lexicons. These models are part of the LanguageNet project (Hasegawa-Johnson et al., 2020), a number of G2P transducers for various languages, trained with Phonetisaurus, an open-source tool for training, compiling, and evaluating grapheme-to-phoneme models for speech recognition (Novak et al., 2012, 2016). Finally, for Catalan, we create a gold standard lexicon using eSpeak-NG, a open-source formant synthesizer and rule-based grapheme-to-phoneme converter (eSpeak NG, 2016), since a LanguageNet G2P model for Catalan was unavailable.

As opposed to the monolingual lexicons, the cross-lingual lexicons are created using grapheme-to-phoneme models trained on the reference-hypothesis pairs produced by multilingual phone recognition models fine-tuned on Danish, Norwegian, and Swedish utterances from the NST corpus when applied cross-lingually on the parliamentary and telephone speech data. Namely, the phone recognition models, which were fine-tuned on a small trilingual subset of the NST training data transcribed using the five different vowel categorization methods: *nst*, *lnet*, *uni-5*, *uni-10*, and *uni-16*, are, first, used to transcribe the whole development sets of each parliamentary and telephone speech corpus. Subsequently, these cross-lingual transcripts are used as training data for different G2P models trained with Phonetisaurus (Novak et al., 2012, 2016). Finally, each resulting G2P model is used to transcribe the whole vocabulary of its corresponding parliamentary or telephone speech corpus. We call the outputs of these G2P models cross-lingual lexicons. In particular, they are referred to as:

xl-nst, *xl-lnet*, *xl-uni-5*, *xl-uni-10*, and *xl-uni-16*, depending on which type of vowel categories they contain. *xl-nst* and *xl-lnet* are used as baseline cross-lingual lexicons, while *xl-uni-5*, *xl-uni-10*, and *xl-uni-16* are formant-based cross-lingual lexicons, which will be compared with the baselines in our experiments and analysis of results. In total, we create 150 cross-lingual lexicons: 3 phone recognition models (experiment runs) for each of the 5 types of vowel categorization applied to each of the 10 out-of-domain corpora (and languages), and evaluate them as part of 150 corresponding ASR systems.

Table 13.6: Absolute vocabulary sizes of the parliamentary corpora covered by the monolingual and cross-lingual lexicons. For cross-lingual lexicons, the numbers show the mean absolute vocabulary size and standard deviation over the three phone recognition experiment runs.

	Danish	**Icelandic**	**Catalan**	**Serbian**	**Finnish**
manual	296,400	195,113	/	/	/
orth	/	/	/	123,631	757,165
g2p	283,520	/	48,650	123,602	757,165
xl-nst	294,547 ± 208	180,815 ± 461	48,475 ± 186	123,391 ± 257	756,746 ± 373
xl-lnet	294,692 ± 55	181,248 ± 4	48,639 ± 2	123,614 ± 2	757,142 ± 8
xl-uni-5	294,719 ± 32	181,143 ± 22	48,621 ± 4	123,585 ± 15	757,057 ± 44
xl-uni-10	294,580 ± 167	180,968 ± 212	48,559 ± 87	123,545 ± 72	756,911 ± 192
xl-uni-16	294,605 ± 34	181,063 ± 15	48,589 ± 21	123,528 ± 21	756,965 ± 44

Table 13.7: Percentage of the parliamentary vocabularies covered by the monolingual and cross-lingual lexicon. For cross-lingual lexicons, the numbers show the mean percentage and standard deviation over the three phone recognition experiment runs.

	Danish	**Icelandic**	**Catalan**	**Serbian**	**Finnish**
monoling.	100.0	100.0	100.0	100.0	100.0
xl-nst	99.4 ± 0.1	92.7 ± 0.2	99.6 ± 0.4	99.8 ± 0.2	99.9 ± 0.0
xl-lnet	99.4 ± 0.0	92.9 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
xl-uni-5	99.4 ± 0.0	92.8 ± 0.0	99.9 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
xl-uni-10	99.4 ± 0.1	92.8 ± 0.1	99.8 ± 0.2	99.9 ± 0.1	100.0 ± 0.0
xl-uni-16	99.4 ± 0.0	92.8 ± 0.0	99.9 ± 0.0	99.9 ± 0.0	100.0 ± 0.0

Tables 13.6 and 13.8 show the absolute vocabulary size covered by each monolingual and cross-lingual lexicon for each of the extrinsic

Table 13.8: Absolute vocabulary sizes of the Babel data sets covered by the monolingual and cross-lingual lexicons. For cross-lingual lexicons, the numbers show the mean absolute vocabulary size and standard deviation over the three phone recognition experiment runs.

	Lao	**Zulu**	**Amharic**	**Mongolian**	**Javanese**
g2p	6,680	64,127	38,945	25,154	16,636
xl-nst	6,119 ± 223	62,844 ± 692	38,488 ± 158	23,812 ± 523	15,725 ± 469
xl-lnet	6,619 ± 13	63,929 ± 38	38,934 ± 2	24,924 ± 32	16,483 ± 16
xl-uni-5	6,328 ± 40	63,616 ± 42	38,603 ± 21	24,323 ± 15	16,210 ± 23
xl-uni-10	6,261 ± 119	63,277 ± 304	38,548 ± 104	23,997 ± 446	15,971 ± 301
xl-uni-16	6,259 ± 60	63,284 ± 190	38,584 ± 14	24,052 ± 105	15,970 ± 67

Table 13.9: Percentage of the Babel vocabularies covered by the monolingual and cross-lingual lexicon. For cross-lingual lexicons, the numbers show the mean percentage and standard deviation over the three phone recognition experiment runs.

	Lao	**Zulu**	**Amharic**	**Mongolian**	**Javanese**
g2p	100.0	100.0	100.0	100.0	100.0
xl-nst	91.6 ± 3.3	98.0 ± 1.1	98.8 ± 0.4	94.7 ± 2.1	94.5 ± 2.8
xl-lnet	99.1 ± 0.2	99.7 ± 0.1	100.0 ± 0.0	99.1 ± 0.1	99.1 ± 0.1
xl-uni-5	94.7 ± 0.6	99.2 ± 0.1	99.1 ± 0.1	96.7 ± 0.1	97.4 ± 0.1
xl-uni-10	93.7 ± 1.8	98.7 ± 0.5	99.0 ± 0.3	95.4 ± 1.8	96.0 ± 1.8
xl-uni-16	93.7 ± 0.9	98.7 ± 0.3	99.1 ± 0.0	95.6 ± 0.4	96.0 ± 0.4

evaluation corpora, while Tables 13.9 and 13.9 the percentage of total vocabulary covered by the same lexicons. Since we have three cross-lingual lexicons for each type of vowel categorization (from the three phone recognition experiment runs), for these lexicons, we provide the mean vocabulary size and standard deviation both rounded to the closest whole number. We use only the currently official scripts in tokenization and vocabulary construction for each language. As introduced in Part III, eight of the ten investigated languages use alphabets as official writing scripts, of which Mongolian is the only language using a version of the Cyrillic alphabet rather than Latin, while two, namely, Lao and Amharic, use abugidas. The vocabulary is word-based and measured simply by tokenizing the utterance transcripts on white space, since all investigated languages use white space to denote word separation. As a result, the vocabulary size reflects not only the corpus

size and lexical diversity, but also certain characteristics of the language in question, such as morphological structure, word formation processes, and writing conventions. As can be seen from the tables, most of the cross-lingual lexicons do not cover the whole vocabulary. The missing words are the ones that could not be transcribed by their respective cross-lingual G2P models.

13.6 Extrinsic Evaluation: Monolingual Speech Recognition

The three language-universal vowel categorization approaches are assessed extrinsically in a set of word-based speech recognition experiments on 10 different languages divided into two types of speech domains, parliamentary speech: Danish, Icelandic, Catalan, Serbian, and Finnish, and low-resource noisy telephone speech: Lao, Zulu, Amharic, Mongolian, and Javanese. The purpose of the extrinsic evaluation of formant-based vowel categorization is to investigate whether improvement in cross-lingual vowel recognition with formant-based vowel representations can translate to improvement in word-based speech recognition, i.e. whether cross-lingual formant-based vowel representations can be used to differentiate words in addition to providing more acoustically grounded descriptions of vowel realizations.

The setup for each extrinsic experiment is the same: we train and evaluate modular HMM-DNN ASR systems, each consisting of a monolingual acoustic and n-gram language model coupled with a cross-lingual pronunciation model derived from the cross-lingual lexicons described in the previous section. For each ASR system, we use the training set in its entirety of a given language to train the acoustic and language model, the development set to train the cross-lingual G2P model that will transcribe the cross-lingual lexicon, and, then, evaluate the resulting system on both the development and test sets. The vocabulary in the lexicons comprises all words from the entire corpus, including training, development, and test partitions. This is common practice in the development of modular ASR systems, as the separate pronunciation model can easily be expanded to include target-domain words without the need to also retrain the acoustic

and language models. As described in the previous section, the total number of trained and extrinsically evaluated ASR systems with cross-lingual pronunciation models is 150: 3 experiment runs per each of the 5 vowel categorization types, two baselines (*xl-nst* and *xl-lnet*) and three formant-based (*xl-uni-5*, *xl-uni-10*, and *xl-uni-16*) for each of the 10 evaluation corpora and languages. Additionally, for each evaluation corpus, we also train and evaluate a fully monolingual ASR system with canonical pronunciation lexicons to have a perspective of what is currently regarded as competitive performance.

13.6.1 Evaluation on Parliamentary Speech

For each parliamentary corpus, the acoustic model is trained on the full training set, and evaluated in terms of word error rate on the development and test sets. The model training and evaluation procedure follows Kaldi's `sprakbanken` recipe that trains from scratch a HMM-DNN hybrid acoustic model with a Time-Delay Neural Network (TDNN) (Peddinti et al., 2015). The model is based on monophone and triphone segmentation GMM acoustic models (Bing-Hwang Juang et al., 1986) and an i-vector speaker adaptation model (Dehak et al., 2011).[2] This is the same model used to validate the Danish parliament speech corpus, *FT Speech*.

We train TDNN acoustic models consisting of 6 layers with an affine transform, ReLU activation, and a *renorm* component. The models are trained with the lattice-free maximum mutual information (LF-MMI) objective, which maximizes the log probability of the correct phone sequence (Povey et al., 2016). They are trained for 5 epochs on mini-batches of 128 chunks, sequences processed in parallel, where each chunk contains 150 feature frames. The input feature frames consist of 40-dimension high-resolution MFCCs and 100-dimension i-vectors. We use a learning rate that decays from 0.001 to 0.0001 during training and clip parameters at a Frobenius norm of 2.0. The same acoustic model architecture and hyperparameters are used for all experiments on the parliamentary corpora.

[2]The original recipe was created for the Danish NST subcorpus and can be found at: `https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/run.sh`.

When it comes to language modeling, we use n-gram language models (LMs) trained exclusively on in-domain parliamentary text data. We use pretrained LMs when available and estimate our own for the corpora without pretrained models. More specifically, for the Catalan, Serbian, and Finnish parliament corpora, we estimate multiple 3-gram, 4-gram, and 5-gram language models with modified Kneser-Ney smoothing on parliamentary training text data, and evaluate them on development text data in terms of perplexity. We select the language model with the lowest perplexity to use for the evaluation of the acoustic and pronunciation models, and report the word error rate individually for each development and test set. All models are estimated and evaluated using the SRI Language Modeling Toolkit (Stolcke, 2002). As training text data for the Catalan and Serbian LMs, we use only the transcripts of the training utterances from the two parliamentary speech corpora, which contain roughly 6 and 7 million word tokens respectively. The Finnish corpus comes with a 35-million-word-token in-domain text corpus created from the full-length transcripts of parliamentary sessions. We use this text corpus to train Finnish LMs. For all three corpora, a 4-gram LM is chosen for the final evaluation.

On the other hand, the Danish and Icelandic parliament corpora come with pretrained and evaluated in-domain LMs, which can be used without modification. The Danish LM is a 4-gram language model with Witten-Bell smoothing trained on a 43-million-word-token in-domain text corpus (Kirkedal et al., 2020), whereas the Icelandic LM is a pruned 3-gram language model with Kneser-Ney smoothing trained on a 30-million-word-token in-domain corpus (Helgadóttir et al., 2017). In all five cases, there is no overlap between any of the evaluation subsets (development or test utterances) and the text data used for language modeling.

13.6.2 Evaluation on Low-Resource Noisy Telephone Speech

For each selected Babel language pack, the acoustic model is trained on the full training set, and evaluated in terms of word error rate individually on the development and test sets. The model architecture and

training procedure follow the `discophone` Kaldi recipe (Feng et al., 2021) that trains a HMM-DNN hybrid acoustic model with a factorized Time-Delay Neural Network (TDNNF) (Povey et al., 2018). The TDNNF model training relies on frame-level phone alignments obtained by forced alignment with HMM-GMM acoustic models trained beforehand. The recipe was originally created for a phone recognition task and applied on the GlobalPhone and Babel corpora.[3]

We train TDNNF acoustic models consisting of 12 layers, with a hidden dimension of 1024, bottleneck dimension of 128, and skip connections. The models are trained with the LF-MMI objective for 12 epochs on mini-batches of 128 chunks. The input features are the same as for the acoustic models trained on parliamentary data and consist of 40-dimension high-resolution MFCCs and 100-dimension i-vectors. The training hyperparameters are taken from the Kaldi's Wall Street Journal recipe.[4] The same model architecture and hyperparameters are used for all experiments on the Babel corpora.

When it comes to language modeling, we do not have any additional in-domain text data for any of the Babel languages, so we rely only on the training utterance transcripts for LM estimation. Compared with the text corpora used in the parliamentary ASR systems, these are much smaller in size, ranging from 300-600 thousand word tokens. As with parliamentary LMs, for all Babel corpora, we estimate multiple n-gram language models with different smoothing methods and parameters, and evaluate them on the development text data in terms of perplexity. Then, the language model with the lowest perplexity is selected for the evaluation of the acoustic and pronunciation models. For each Babel corpus, this ends up being a 3-gram language model with modified Kneser-Ney smoothing.

[3]The recipe is available at: `https://github.com/pzelasko/kaldi/tree/discophone/egs/discophone`.

[4]The recipe can be found at: `https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain/tuning/run_tdnn_1g.sh`

Results

14

14.1 Introduction

In this chapter, we present and interpret the performance results of the phone and speech recognition models. We start with general corpus-level metrics: phone error rate (PER), phone feature Hamming edit distance (PFHED), and word error rate (WER), which are calculated on the standard evaluation partitions (development and test subsets) of the investigated corpora: NST corpus, five different parliamentary corpora, and five language packs from Babel, a noisy telephone speech corpus. Subsequently, we perform a deeper analysis by looking specifically at the vowel predictions of both the phone and speech recognition models on different languages. Finally, for each language and categorization level, we interpret the phone prediction results by comparing the prediction rates for each reference vowel with the amount of overlap between the position of the reference vowel in the abstract vowel space and the position of the hypothesis vowels in the normalized F_1-F_2 space.

14.2 Multilingual and Cross-Lingual Phone Recognition

We begin by evaluating the language-universal formant-based vowel representations intrinsically, as we did in the previous set of experiments in Chapter 11. To investigate the effect of the different levels of vowel categorization on the overall performance on the phone recognition task, we first take a look at the PER and PFHED results of each multilingual and cross-lingual model on the NST corpus. Subsequently, we evaluate the multilingual phone recognition models cross-lingually on the parliamentary and Babel data sets.

14.2.1 Performance on the NST Corpus

Table 14.1 shows the mean PERs and Table 14.2 the mean PFHEDs of both the multilingual and cross-lingual phone recognition models fine-tuned and evaluated on the NST corpus. The multilingual models, which are fine-tuned on 3000 samples from each of the three subcorpora (9000 in total), are applied cross-lingually on the parliamentary and telephone data in the further experiments. On the other hand, the cross-lingual phone recognition models, which are fine-tuned on 3000 samples from two of the NST languages and evaluated on the third, heldout, language, are created and evaluated only for the purpose of comparing their performance with the performance of the cross-lingual experiments from Chapter 11. They are not used in any further experiments.

Firstly, we see that all multilingual models outperform the cross-lingual models on all evaluation languages both in terms of PER and PFHED. This is not surprising as the multilingual models are fine-tuned on the evaluation languages, whereas the cross-lingual ones are not. Next, we observe that the *nst* multilingual models, which are fine-tuned on the *nst* transcriptions whose monophthong vowel set consists of 20 unique vowel categories, outperform all other multilingual models on all three languages, including the other baseline, *lnet*, with a monophthong vowel set of 16 vowel categories. This is also not surprising since the *nst* transcriptions are dictionary-based and exhibit less phonetic variability than the *uni* transcriptions. Additionally, they were designed to be comparable across the three NST subcorpora and allow the model to better leverage the cross-lingual lexical similarities among the three Scandinavian languages.

What is surprising, however, is that the multilingual *lnet* models have higher PERs than the *uni* models in all cases, except the *uni-10* and *uni-16* models on Danish. Like the *nst* ones, the *lnet* transcriptions are dictionary-based and were expected to be more predictable in the multilingual evaluation scenario. This suggests that the lack of cross-lingual consistency among the *lnet* transcription systems is interfering with the models' ability to capture general phonetic patterns. We suspect that most interference is caused by the Danish *lnet* transcriptions. The current *lnet* G2P model produces pronunciation

transcripts that are very similar to their corresponding orthographic transcripts, which could be problematic for languages with an opaque orthography, such as Danish. Namely, it has a smaller phone inventory than the *nst* transcription system, with 10 vowel categories as opposed to 14 in the *nst* system. This is likely the reason why the multilingual *lnet* model performs better on Danish than on Norwegian and Swedish, opposite to the trend exhibited by the multilingual *nst* model.

Table 14.1: Mean PERs and std of all multilingual and cross-lingual models averaged over three experiment runs. The best results for each evaluation language are shown in bold.

exp. type	ttype	Danish	Norwegian	Swedish
multiling.	*nst*	**11.82±1.03**	**7.66±0.82**	**9.26±0.92**
	lnet	22.36±0.61	25.28±0.57	24.34±1.41
	uni-5	20.21±2.47	15.20±1.04	16.08±0.40
	uni-10	24.70±0.99	19.30±0.98	22.33±3.01
	uni-16	24.37±0.54	17.00±0.67	21.01±1.08
cross-ling.	*nst*	53.29±0.44	39.69±0.40	42.09±0.72
	lnet	55.60±0.36	50.88±0.40	50.38±0.64
	uni-5	**40.27±0.70**	**31.46±0.68**	**32.93±0.69**
	uni-10	46.95±1.38	36.70±0.66	38.30±0.13
	uni-16	46.56±0.51	34.03±0.50	36.98±0.26

When it comes to the cross-lingual models, we can see that the *uni* models consistently outperform both baselines on all three languages, in terms of both PER and PFHED. As was the case with the multilingual models, the *lnet* baseline performs the worst in terms of PER, especially on Norwegian and Swedish, where we can see a difference of over 11 and 8 percentage points respectively. Again, this is likely caused by the discrepancy between the Danish *lnet* transcriptions system, on one hand, and the Norwegian and Swedish ones, on the other. At the same time, we can see that having fewer vowel categories in the transcription system does not guarantee better cross-lingual performance.

Table 14.2: Mean PFHEDs and std of all multilingual and cross-lingual models averaged over three experiment runs. The best results for each evaluation language are shown in bold.

exp. type	ttype	Danish	Norwegian	Swedish
multiling.	*nst*	**2.15±0.20**	**1.43±0.14**	**1.70±0.11**
	lnet	2.90±0.12	2.12±0.02	1.80±0.20
	uni-5	2.93±0.56	1.99±0.35	2.19±0.16
	uni-10	3.03±0.32	2.20±0.31	2.65±0.63
	uni-16	3.25±0.14	2.05±0.16	2.53±0.24
cross-ling.	*nst*	7.88±0.25	6.21±0.10	6.74±0.20
	lnet	8.13±0.34	8.02±0.10	6.59±0.10
	uni-5	**6.03±0.11**	5.71±0.18	**6.21±0.13**
	uni-10	6.49±0.07	5.87±0.08	6.26±0.12
	uni-16	6.42±0.10	**5.72±0.04**	6.27±0.09

Returning to the *uni* models, we can see that shifting to a unified language-universal vowel set, and effectively ruling out unseen vowel categories at inference time, greatly reduces both the PER and PFHED scores on all languages. In almost all cases, the *uni-5* models are the best performing models. The only exception is the Danish-Swedish *uni-16* model, which performs very slightly better on Norwegian in terms of PFHED. Since the *uni-5* transcription system collapses the large vowel inventories of the Scandinavian languages into only 5 broad vowel categories, it is not surprising that the models dealing with fewer vowel categories will exhibit fewer vowel confusions. However, by comparing the *uni-10* and *uni-16* models, we see that reducing the number of phone categories does not necessarily lead to higher phone recognition rates. In fact, the *uni-16* cross-lingual models, which are fine-tuned to distinguish among 16 vowel categories, perform better on average than the *uni-10* models on all three languages both in terms of PER and PFHED. This tells us that the *uni-16* vowel categorization level seems more suitable for cross-lingual phone prediction on the Scandinavian languages. This most likely stems from the fact that the *uni-16* level distinguishes vowels based on roundedness, which is an

important distinctive feature in all three Scandinavian languages.

Since the *uni* cross-lingual models are fine-tuned and evaluated on the same subsets of the NST subcorpora, we can compare their performance to that of the cross-lingual models fine-tuned and evaluated on formant-based vowel representations with language-specific vowel sets, which we examined in the previous set of experiments, in Part IV. Here, we see that the language-universal cross-lingual models outperform all language-specific models both in terms of PER and PFHED. The difference in performance is likely caused by the main difference between these two types of models. Specifically, this is the degree of cross-lingual overlap between the corresponding vowel categories in the F_1-F_2 vowel space and the lack of unseen vowels. Even though the size of the vowel inventory could also contribute to a difference in performance, its effects are likely not as pronounced. As we have seen when comparing the *uni-10* and *uni-16* models, we can see that the *uni-16* models, which are fine-tuned to distinguish 16 vowel categories, also outperform the language-specific models, which are fine-tuned to distinguish 18-19 vowel categories. Nevertheless, this does not guarantee that these models would generalize to other languages or that the language-universal formant-based representations could be used in other tasks, such as to provide semantic information. This is why we will also examine their applications to other languages, domains, and speech recognition tasks.

14.2.2 Performance on Parliamentary Speech

When it comes to the parliamentary corpora, we do not have their formant-based transcriptions. Therefore, in this case, we compare the predictions of all our models to the same dictionary-based reference transcripts. It should be noted that for these and all subsequent experiments, we use only the multilingual models from the previous subsection, i.e. the ones trained on all three Scandinavian languages. We refer to them as cross-lingual from now on, because we apply them cross-lingually to the parliamentary and telephone speech data. To distinguish them from the cross-lingual models in the previous subsection, we give them the prefix *xl-* when referring to them in the current and following sections.

Table 14.3: Mean PERs and std of all cross-lingual models evaluated on the parliamentary speech data averaged over three experiment runs. The best results for each evaluation language are shown in bold.

		dev-balanced	dev-other
Danish	*xl-nst*	**17.06 ± 0.57**	**15.53 ± 0.69**
	xl-lnet	63.99 ± 0.87	63.55 ± 0.84
	xl-uni-5	45.29 ± 0.57	44.65 ± 0.63
	xl-uni-10	42.37 ± 0.19	41.59 ± 0.27
	xl-uni-16	42.98 ± 0.40	42.19 ± 0.38
			dev
Icelandic	*xl-nst*		67.91 ± 0.75
	xl-lnet		73.07 ± 0.68
	xl-uni-5		**64.99 ± 0.74**
	xl-uni-10		65.61 ± 0.21
	xl-uni-16		67.43 ± 1.17
		clean-dev	other-dev
Catalan	*xl-nst*	54.30 ± 1.18	55.64 ± 1.08
	xl-lnet	58.29 ± 1.86	59.96 ± 1.88
	xl-uni-5	52.12 ± 1.42	54.06 ± 1.43
	xl-uni-10	52.02 ± 1.11	53.70 ± 1.07
	xl-uni-16	**51.41 ± 0.69**	**53.05 ± 0.71**
			dev
Serbian	*xl-nst*		56.91 ± 1.38
	xl-lnet		63.49 ± 2.39
	xl-uni-5		**44.09 ± 1.08**
	xl-uni-10		53.29 ± 0.45
	xl-uni-16		49.23 ± 0.87
		2016-dev-seen	2016-dev-unseen
Finnish	*xl-nst*	59.70 ± 1.22	59.03 ± 0.99
	xl-lnet	56.46 ± 1.55	56.43 ± 1.54
	xl-uni-5	**44.17 ± 1.42**	**42.18 ± 1.37**
	xl-uni-10	55.74 ± 0.44	55.21 ± 0.40
	xl-uni-16	50.24 ± 1.33	49.36 ± 1.28

Table 14.4: Mean PFHEDs and std of all cross-lingual models evaluated on the parliamentary speech data averaged over three experiment runs. The best results for each evaluation language are shown in bold.

		dev-balanced	dev-other
Danish	*xl-nst*	**8.36 ± 0.23**	**7.31 ± 0.21**
	xl-lnet	20.17 ± 0.28	19.50 ± 0.34
	xl-uni-5	10.49 ± 0.31	9.54 ± 0.31
	xl-uni-10	10.60 ± 0.18	9.53 ± 0.11
	xl-uni-16	10.93 ± 0.07	9.85 ± 0.14
			dev
Icelandic	*xl-nst*		28.70 ± 3.34
	xl-lnet		**23.56 ± 0.08**
	xl-uni-5		25.68 ± 0.50
	xl-uni-10		27.33 ± 1.76
	xl-uni-16		28.96 ± 0.56
		clean-dev	other-dev
Catalan	*xl-nst*	19.19 ± 2.43	22.60 ± 2.72
	xl-lnet	18.38 ± 0.52	21.43 ± 0.61
	xl-uni-5	**16.55 ± 0.46**	**19.76 ± 0.53**
	xl-uni-10	17.40 ± 1.59	20.67 ± 1.78
	xl-uni-16	18.78 ± 0.73	22.22 ± 0.89
			dev
Serbian	*xl-nst*		27.78 ± 1.20
	xl-lnet		38.82 ± 1.53
	xl-uni-5		25.65 ± 0.71
	xl-uni-10		26.30 ± 0.29
	xl-uni-16		**25.10 ± 0.20**
		2016-dev-seen	2016-dev-unseen
Finnish	*xl-nst*	37.55 ± 5.52	30.93 ± 4.36
	xl-lnet	**28.95 ± 0.24**	**24.26 ± 0.23**
	xl-uni-5	30.22 ± 0.77	24.75 ± 0.57
	xl-uni-10	33.89 ± 3.65	27.82 ± 2.79
	xl-uni-16	36.33 ± 1.54	29.52 ± 1.02

The mean PERs and PFHEDs of all the cross-lingual models on the parliamentary speech corpora are shown in Tables 14.3 and 14.4 respectively. The first parliamentary corpus we examine is the Danish *FT Speech*. Since the Danish NST subcorpus was part of the fine-tuning data for all of the cross-lingual models, the models are, technically, not applied cross-lingually in this case. Rather, they are applied across different domains, since parliamentary speech is different from read speech in a quiet office environment. Nonetheless, it is the same language, and the PER and PFHED scores reflect it. Namely, most of the models, with the exception of *xl-lnet*, perform better on the Danish parliament corpus than on the other parliament corpora, especially in terms of PFHED. Additionally, the *xl-nst* models outperform the other models on both metrics and evaluation data sets by a large margin. The likely reason for such a large margin between the *xl-nst* scores and the rest is the fact that we compare the models' predictions to the references transcribed using the NST lexicons. This is because both the *xl-nst* predictions and the reference transcripts are based on the same *nst* transcription system.

When it comes to the performance of the cross-lingual models on the other four languages, we see that the *xl-uni* models consistently outperform the baselines on all languages in terms of PER. In the case of Serbian and Finnish, the difference in performance is considerable. On Serbian, the *xl-uni-5* model outperforms the *xl-nst* by 12.82 and the *xl-lnet* by 19.4 percentage points. On Finnish, the *xl-uni-5* model outperforms the *xl-nst* by 15.53-16.85 and the *xl-lnet* by 12.28-14.25 percentage points. The *xl-uni-5* model is also the best performing model on Icelandic, while the best performing model on Catalan is the *xl-uni-16*. However, these results are not as convincing as in the case of Serbian and Finnish.

Looking at the PFHED results for the same corpora, we can see that the best performing models here do not correspond to the best performing models in terms of PERs, except in the case of Danish. The most surprising results are the *xl-lnet* results on the Icelandic and Finnish corpora. The *xl-lnet* models are the best performing models in terms of PFHED on these two languages, whereas they were among the worst in terms of PER. The lowest PFHED results on the Catalan corpus are achieved by the *xl-uni-5* model, and on the Serbian by

the *xl-uni-16* model. However, both of these results are close to the baselines, so it is difficult to determine which model performs best.

A possible reason for the discrepancies between the PER and PFHED results is the way Panphon's implementation of PFHED treats different types of errors. Namely, PHFED has a higher penalty for insertion and deletion errors, and, therefore, favors models that make more substitutions. This is in line with our observations. For example, the *xl-lnet* models made more substitutions than the other models on the Icelandic corpus (7.5 percentage points more on average). Phone error rate, on the other hand, gives the same weight to all three types of errors. As a result, the distribution of error types has no effect on it.

Since the PER and PFHED results are very general and, at times, conflicting, they do not tell us how the models perform on each vowel specifically. For this reason, we are going to look at each model's performance in more detail in Section 14.4, by investigating how the performance on individual vowels relates to the position of the vowel category in the abstract and formant-based vowel spaces.

14.2.3 Performance on Low-Resource and Noisy Telephone Speech

Moving on to the low-resource noisy telephone speech data, we should note that these are very challenging data sets even for models trained or fine-tuned them. Previous research involving zero-shot cross-lingual experiments on Lao, Zulu, and Amharic have reported phone error rates ranging 70-78% (Żelasko et al., 2020; Gao et al., 2021; Xu et al., 2022).

The mean PERs and PFHEDs of all the cross-lingual models on the selected Babel data sets are shown in Table 14.5. As we can see, the PER results are much higher than on the previous two corpora, which is in line with our expectation. The *xl-uni-5* models achieve the best PER on Zulu, Amharic, Mongolian, and Javanese. The lowest PER result on Lao is achieved by the *xl-uni-10* models. It is markedly lower than the *xl-lnet* baseline, but not significantly lower than the *xl-nst*.

When it comes to the PFHED results, the smallest edit distances on Lao, Zulu, and Javanese are achieved by the *xl-uni-10* models, while the *xl-lnet* models have the lowest PFHEDs on Amharic and Mongo-

lian. Unlike the phone error rates, the PFHED results of the different models are quite close, and most of them do not seem significantly better than the rest. Since these results only provide a general overview, we are going to look at each model's performance in more detail in Section 14.4, by looking at the performance on individual reference vowels and its relation to the position of the vowel category in the abstract and formant-based vowel spaces.

14.3 Monolingual Speech Recognition with Cross-Lingual Pronunciation Lexicons

In this section, we provide a general extrinsic evaluation of the language-universal formant-based vowel representations in terms of word error rate (WER) on the five parliamentary corpora and five languages from the Babel noisy telephone speech corpus. For this purpose, we train and evaluate monolingual hybrid HMM/DNN ASR systems with different cross-lingual pronunciation lexicons obtained from the multilingual phone recognition models that we evaluated in the previous section. For each evaluation language, we train and evaluate systems with 5 different cross-lingual pronunciation lexicons, 2 baselines: *xl-nst* and *xl-lnet*, and 3 formant-based ones: *xl-uni-5*, *xl-uni-10*, and *xl-uni-16*. Additionally, for each evaluation language, we train and evaluate systems with monolingual pronunciation lexicons to be able to compare our results to what is currently considered competitive performance. For each type of pronunciation lexicon, we run 3 experiments and measure mean word error rate and standard deviation (std) to obtain more reliable results.

14.3.1 Performance on Parliamentary Speech

The mean WERs on the five parliamentary corpora are presented in the five Tables 14.6-14.10. On each corpus, the lowest WERs are consistently achieved by the models with the *xl-lnet* lexicons. The results of the remaining models are quite close to one another. Moreover, they exhibit high variance, so we cannot choose the best performing *xl-uni* model with certainty.

Table 14.5: Mean PERs and PFHEDs and std of all cross-lingual models evaluated on Babel data averaged over three experiment runs. The best results for each evaluation language are shown in bold.

		PER	PFHED
Lao	*xl-nst*	81.19 ± 2.49	8.73 ± 0.25
	xl-lnet	89.21 ± 1.63	8.70 ± 0.44
	xl-uni-5	86.53 ± 1.05	**8.25 ± 0.03**
	xl-uni-10	**80.50 ± 0.66**	8.28 ± 0.40
	xl-uni-16	82.00 ± 1.40	8.51 ± 0.16
Zulu	*xl-nst*	83.93 ± 2.12	12.61 ± 0.92
	xl-lnet	84.64 ± 2.49	11.33 ± 0.23
	xl-uni-5	**72.34 ± 0.62**	**11.42 ± 0.05**
	xl-uni-10	77.31 ± 1.01	11.95 ± 0.69
	xl-uni-16	76.52 ± 1.58	12.30 ± 0.31
Amharic	*xl-nst*	80.76 ± 1.19	15.94 ± 1.34
	xl-lnet	82.86 ± 1.94	**13.59 ± 0.09**
	xl-uni-5	**76.58 ± 0.34**	14.72 ± 0.00
	xl-uni-10	76.83 ± 0.38	15.46 ± 0.97
	xl-uni-16	77.28 ± 1.55	16.04 ± 0.39
Mongolian	*xl-nst*	91.57 ± 1.61	14.68 ± 1.24
	xl-lnet	91.48 ± 1.43	**12.63 ± 0.09**
	xl-uni-5	**86.03 ± 0.09**	13.86 ± 0.15
	xl-uni-10	86.73 ± 0.51	14.50 ± 1.07
	xl-uni-16	87.98 ± 0.81	15.18 ± 0.39
Javanese	*xl-nst*	82.65 ± 2.28	8.73 ± 0.57
	xl-lnet	83.90 ± 2.70	8.06 ± 0.29
	xl-uni-5	**74.73 ± 0.76**	**7.92 ± 0.04**
	xl-uni-10	76.43 ± 0.83	8.28 ± 0.65
	xl-uni-16	77.54 ± 2.35	8.55 ± 0.28

Table 14.6: Mean WERs and std of all ASR models trained and evaluated on Danish *FT Speech* averaged over three experiment runs. The best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the ASR models with monolingual lexicons are shown on top and provide current approximate SOTA results.

	dev-balanced	dev-other	test-balanced	test-other
nst	13.06 ± 0.09	13.03 ± 0.07	13.84 ± 0.01	13.58 ± 0.11
lnet	13.37 ± 0.09	13.27 ± 0.09	14.13 ± 0.04	13.80 ± 0.09
xl-nst	33.38 ± 5.72	33.57 ± 5.72	34.15 ± 5.84	34.03 ± 5.79
xl-lnet	**26.01 ± 0.24**	**26.06 ± 0.18**	**26.48 ± 0.19**	**26.34 ± 0.16**
xl-uni-5	32.14 ± 2.74	32.22 ± 2.70	32.97 ± 2.65	32.79 ± 2.67
xl-uni-10	35.83 ± 1.74	36.17 ± 1.63	36.74 ± 1.82	36.62 ± 1.69
xl-uni-16	34.98 ± 1.83	35.17 ± 1.75	35.62 ± 1.72	35.61 ± 1.85

However, the consistently lower error rates of the models with the *xl-lnet* lexicons seem unusual, since the *xl-lnet* models were not convincingly better on the phone recognition task. For this reason, we try to find explanation for this discrepancy in performance. We look at the vocabulary size comparison of the different lexicons in Section 13.5, and, more specifically, the portion of vocabulary covered by each lexicon displayed in Table 13.7. We perform correlation analysis to investigate whether vocabulary coverage could significantly impact the models' performance. The correlation plots for each parliamentary corpus are presented in Figure 14.1. For each corpus, they show the WER of each model as a function of the vocabulary coverage of that model's *xl*-lexicon. The plots contain the regression line, 95% confidence interval, correlation coefficient, and p-value. These measures indicate that there is, indeed, a statistically significant correlation between the lexicon's vocabulary coverage and the model's performance.

Since out-of-vocabulary words cannot be recognized by hybrid ASR systems and the *xl-lnet* lexicons have the largest vocabulary of all *xl*-lexicons, it is likely that the larger vocabulary size contributes to the models performance. The most plausible reason for the *xl-lnet* lexicons having the widest vocabulary coverage is due to the *xl-lnet* phone recognition models having the lowest deletion error rates compared with the other *xl*-models. This is because G2P converters, which were used to transcribe the *xl*-lexicons, are more likely to output deletions

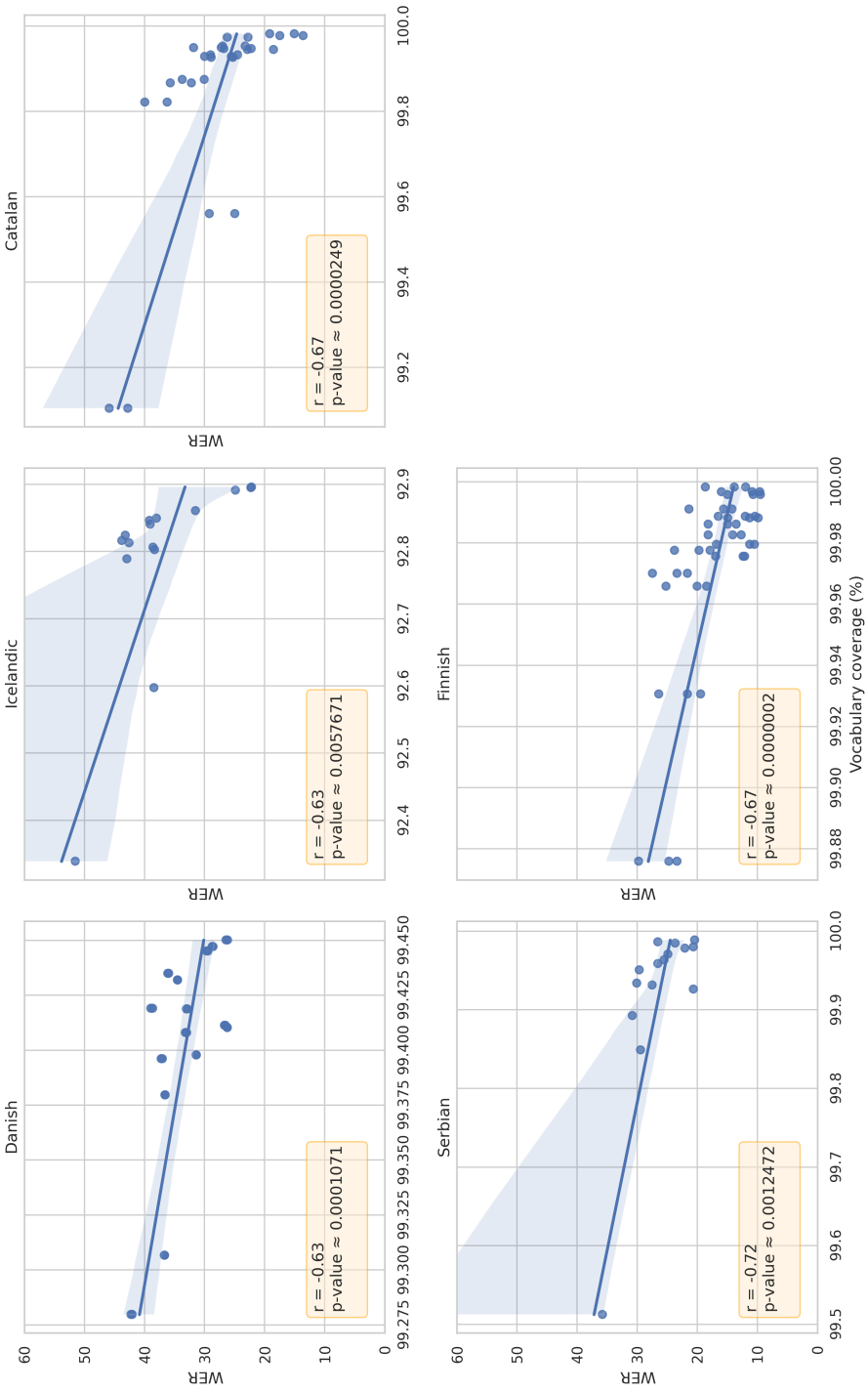


Figure 14.1: Correlation analyses of the parliamentary ASR models' WERs as a function of the vocabulary coverage of their respective cross-lingual lexicons. The plots show the data points fitted with a regression line and a 95% confidence interval.

Table 14.7: Mean WERs and std of all ASR models trained and evaluated on the Icelandic *Althingi* corpus averaged over three experiment runs. The best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the ASR models with monolingual lexicons are shown on top and provide current approximate SOTA results.

	dev	test
althingi	12.63 ± 0.05	12.37 ± 0.05
xl-nst	42.14 ± 8.31	42.29 ± 8.23
xl-lnet	**22.98 ± 1.23**	**23.14 ± 1.23**
xl-uni-5	39.83 ± 2.20	40.15 ± 2.22
xl-uni-10	39.65 ± 1.88	40.06 ± 1.83
xl-uni-16	39.64 ± 2.15	39.97 ± 2.12

and empty strings when they are trained on transcripts containing more deletions. However, it remains unclear to what extent the number of out-of-vocabulary words can affect a model's performance. It seems likely that it should depend on the number of out-of-vocabulary tokens in the evaluation set.

Still, the overall WERs do not provide information on how individual vowel predictions in the pronunciation lexicons relate to the ASR model's word predictions. This is investigated in Section 14.5, where we look at phone prediction rates for each reference vowel in correctly and incorrectly recognized words and relate them the vowel prediction rates of the phone recognition models, as well as the positions of the vowel categories in the abstract and formant-based vowel spaces.

14.3.2 Performance on Low-Resource and Noisy Telephone Speech

The mean WERs on all five Babel data sets are presented in the Table 14.11. We observe the same situation as in the case of the parliamentary corpora. Namely, on each data set, the lowest WERs are consistently achieved by the models with the *xl-lnet* lexicons. The results of the remaining models are quite close and exhibit high variance, so we cannot choose the best performing *xl-uni* model with certainty. Once

Table 14.8: Mean WERs and std of all ASR models trained and evaluated on the Catalan parliament corpus averaged over three experiment runs. The best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the ASR models with monolingual lexicons are shown on top and provide current approximate SOTA results.

	clean-dev	other-dev	clean-test	other-test
espeak	7.11 ± 0.03	6.91 ± 0.04	10.11 ± 0.05	10.53 ± 0.04
xl-nst	32.33 ± 7.79	32.76 ± 7.98	35.52 ± 7.71	36.23 ± 7.72
xl-lnet	**17.33 ± 3.94**	**17.18 ± 4.03**	**20.10 ± 3.87**	**20.98 ± 3.77**
xl-uni-5	24.97 ± 2.07	25.03 ± 2.06	28.10 ± 2.22	29.58 ± 2.06
xl-uni-10	22.93 ± 2.84	22.69 ± 2.94	25.87 ± 2.77	27.07 ± 2.94
xl-uni-16	30.34 ± 4.24	30.54 ± 4.47	33.21 ± 4.17	34.22 ± 4.52

again, the word error rates of the models with the *xl-lnet* lexicons are unusually lower than the rest. The explanation for this discrepancy in performance is likely the same as before.

The vocabulary size comparison of the different lexicons in Section 13.5, and, more specifically, the portion of vocabulary covered by each lexicon displayed in Table 13.9 reveal again that the *xl-lnet* lexicons have the widest lexicon coverage compared to the other *xl*-lexicons. We perform correlation analysis to investigate whether the impact of vocabulary coverage on the models' performance could be significant. The correlation plots for the Babel data sets are presented in Figure 14.2. For each language, they show the WER of each model as a function of the vocabulary coverage of that model's *xl*-lexicon. The regression lines, 95% confidence intervals, correlation coefficients, and p-values indicate that there is a statistically significant correlation between the lexicon's vocabulary coverage and the model's performance. The reason for this again seems to stem from the fact that the *xl-lnet* phone recognition models make fewer deletion errors compared to the other *xl*-models. However, it is not clear why this happens.

However, the overall WERs on Babel data do not explain how individual vowel predictions in the pronunciation lexicons relate to the ASR model's word predictions. For this reason, in Section 14.5, we will analyze phone prediction rates for each reference vowel in correctly and incorrectly recognized words and how they relate to the vowel

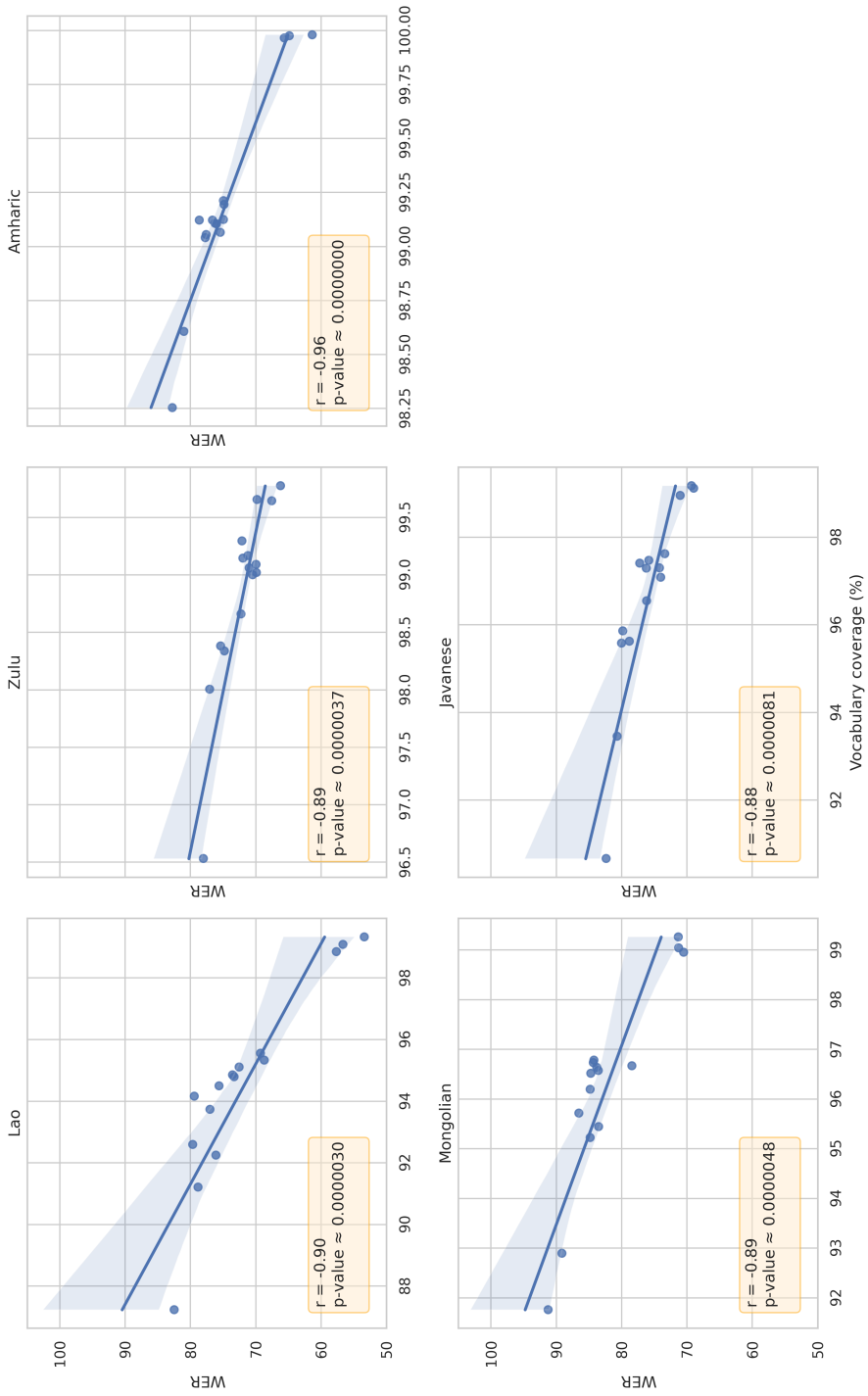


Figure 14.2: Correlation analyses of the Babel ASR models' WERs as a function of the vocabulary coverage of their respective cross-lingual lexicons. The plots show the data points fitted with a regression line and a 95% confidence interval.

Table 14.9: Mean WERs and std of all ASR models trained and evaluated on the Serbian parliament corpus averaged over three experiment runs. The best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the ASR models with monolingual lexicons are shown on top and provide current approximate SOTA results.

	dev	test
orth	10.19 ± 0.03	11.30 ± 0.06
lnet	10.22 ± 0.04	11.21 ± 0.04
xl-nst	29.39 ± 4.45	30.27 ± 4.45
xl-lnet	**22.88 ± 2.15**	**23.59 ± 2.51**
xl-uni-5	24.57 ± 3.58	25.65 ± 3.74
xl-uni-10	24.63 ± 3.05	25.72 ± 3.01
xl-uni-16	25.21 ± 3.99	26.35 ± 4.24

prediction rates of the phone recognition models, as well as the positions of the vowel categories in the abstract and formant-based vowel spaces.

14.4 Phone Prediction Analysis of Cross-Lingual Phone Recognition Results

In this section, we break down the performance of each cross-lingual phone recognition model by prediction rates on individual reference vowels. We do this for each corpus and cross-lingual model separately. Namely, we present the results of one cross-lingual model on all evaluation languages from a corpus collection side by side in the same table. For each evaluation language, we show the top phone predictions and their prediction rates for all reference vowels in the language. This allows us to calculate the mean vowel recognition rate across all reference vowels in the evaluation language and compare the prediction and recognition rates of each and all reference vowels across languages within a corpus collection. To put it more concretely, this means that we have five tables: *(xl-)nst*, *(xl-)lnet*, *(xl-)uni-5*, *(xl-)uni-10*, and *(xl-)uni-16*, per corpus collection: NST corpus, parliamentary corpus collection, and Babel corpus collection, which adds up to 15

Table 14.10: Mean WERs and std of all ASR models trained and evaluated on the Finnish parliament corpus averaged over three experiment runs. The best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the ASR models with monolingual lexicons are shown on top and provide current approximate SOTA results.

	dev-seen	dev-unseen	test-seen	test-unseen	2020-test
orth	9.74 ± 0.04	9.95 ± 0.07	7.18 ± 0.05	6.21 ± 0.06	7.41 ± 0.05
lnet	9.57 ± 0.10	9.85 ± 0.09	7.16 ± 0.04	6.13 ± 0.04	7.40 ± 0.07
xl-nst	23.02 ± 6.43	22.52 ± 6.67	19.81 ± 6.03	18.31 ± 6.00	24.09 ± 6.51
xl-lnet	**15.04 ± 1.23**	**14.44 ± 1.38**	**11.84 ± 1.45**	**10.38 ± 1.14**	**16.56 ± 1.54**
xl-uni-5	18.28 ± 3.26	17.64 ± 3.09	15.80 ± 3.13	14.17 ± 3.07	20.57 ± 3.02
xl-uni-10	19.61 ± 3.31	18.84 ± 3.11	16.92 ± 3.39	15.24 ± 3.01	20.96 ± 3.88
xl-uni-16	17.17 ± 4.09	16.98 ± 4.00	14.58 ± 3.88	13.72 ± 3.43	19.66 ± 3.92

tables in total.

Before we analyze performance on individual vowels, we will look at the overall vowel recognition rates of all phone recognition models on the NST corpus, parliamentary speech corpora, and Babel languages. Table 14.12 shows the mean and standard deviation of the vowel recognition rates of all phone recognition models on the three corpus collections.

The results on the NST corpus reveal that all *uni*-models on average outperform both baselines on all three evaluation languages. Overall, the *uni-5* models yield the best results on all evaluation languages. On Danish, the *uni-5* models improve 30.05 percentage points over *nst* and 35.51 percentage points over the *lnet* baseline. On Norwegian, the *uni-5* models improve 22.94 percentage points over *nst* and 38.68 percentage points over the *lnet* baseline. On Swedish, they improve 24.47 percentage points over *nst* and 39.74 percentage points over the *lnet* baseline. This aligns with our expectations based on the previous set of cross-lingual experiments in Chapter 11, where we found that vowel recognition rates increase for vowel categories shared by all fine-tuning languages when their positions in the normalized $F_1 - F_2$ space overlap.

The results on the parliamentary corpora reveal that the *xl-uni* models on average outperform both baselines on all evaluation languages, with the exception of all *xl-uni* models on Danish and *xl-uni-10* on Serbian and Finnish. On the Danish parliament corpus, the

Table 14.11: Mean WERs and std of all monolingual ASR models trained and evaluated on the Babel data sets and averaged over three experiment runs. For each language, the best results among the models trained with cross-lingual lexicons (xl-models) are shown in bold. The WERs of the models with monolingual lexicons are shown as current approximate SOTA results.

		dev	test
Lao	*lnet*	39.18 ± 0.03	42.57 ± 0.12
	xl-nst	74.40 ± 6.16	75.79 ± 5.63
	xl-lnet	**53.13 ± 1.83**	**55.94 ± 1.83**
	xl-uni-5	73.18 ± 4.25	74.81 ± 4.18
	xl-uni-10	73.33 ± 3.08	75.01 ± 2.74
	xl-uni-16	75.03 ± 2.79	76.69 ± 2.59
Zulu	*lnet*	51.91 ± 0.04	53.79 ± 0.06
	xl-nst	73.21 ± 3.42	74.49 ± 3.35
	xl-lnet	**66.01 ± 1.19**	**67.88 ± 1.47**
	xl-uni-5	69.87 ± 0.18	71.79 ± 0.39
	xl-uni-10	70.39 ± 3.39	72.51 ± 3.23
	xl-uni-16	70.75 ± 1.96	72.72 ± 1.59
Amharic	*lnet*	39.11 ± 0.10	42.25 ± 0.08
	xl-nst	77.84 ± 3.03	79.27 ± 2.70
	xl-lnet	**61.67 ± 2.21**	**63.98 ± 1.86**
	xl-uni-5	73.33 ± 0.65	75.46 ± 0.48
	xl-uni-10	74.76 ± 3.37	77.00 ± 2.86
	xl-uni-16	75.44 ± 0.88	77.36 ± 0.48
Mongolian	*lnet*	46.30 ± 0.10	49.60 ± 0.17
	xl-nst	86.24 ± 2.94	87.53 ± 2.74
	xl-lnet	**69.15 ± 0.84**	**71.07 ± 0.38**
	xl-uni-5	80.50 ± 2.62	82.18 ± 2.63
	xl-uni-10	84.13 ± 2.70	85.72 ± 2.45
	xl-uni-16	82.92 ± 0.61	84.41 ± 0.60
Javanese	*lnet*	51.66 ± 0.03	56.34 ± 0.10
	xl-nst	75.28 ± 3.36	78.50 ± 3.35
	xl-lnet	**65.39 ± 1.05**	**69.80 ± 0.91**
	xl-uni-5	72.18 ± 1.80	75.63 ± 1.63
	xl-uni-10	73.06 ± 3.13	76.86 ± 2.82
	xl-uni-16	74.89 ± 1.85	78.70 ± 1.78

Table 14.12: Mean and std of vowel recognition rates of all phone recognition models on the NST corpus, parliamentary speech corpora, and Babel languages. Higher is better. Bolded results are the best recognition rates for a language. The asterisk in the superscript (*) denotes that the result is significantly above the *xl-nst* baseline and dagger (†) that it is significantly above the *xl-lnet*.

	(xl-)nst	*(xl-)lnet*	*(xl-)uni-5*	*(xl-)uni-10*	*(xl-)uni-16*
NST					
Danish	26.00 ± 0.79	20.54 ± 0.86	**56.05 ± 0.27***†	42.58 ± 1.74*†	42.53 ± 1.51*†
Norwegian	44.73 ± 0.61	28.99 ± 0.61	**67.67 ± 0.68***†	54.24 ± 0.77*†	58.98 ± 1.40*†
Swedish	45.44 ± 0.37	30.17 ± 2.08	**69.91 ± 0.52***†	54.25 ± 0.64*†	57.85 ± 0.65*†
Parliament					
Danish	**85.80 ± 0.96**	38.67 ± 0.94	24.69 ± 0.48	30.53 ± 0.33	29.56 ± 0.54
Icelandic	20.69 ± 3.54	17.88 ± 0.37	**32.02 ± 1.08**†	26.46 ± 0.96†	19.60 ± 3.43
Catalan	35.15 ± 0.94	29.96 ± 3.04	**39.61 ± 0.15**	38.20 ± 0.49	38.81 ± 2.49
Serbian	33.27 ± 3.08	45.06 ± 5.79	**73.82 ± 0.41***	41.80 ± 1.84	56.18 ± 2.45*
Finnish	20.53 ± 2.77	29.46 ± 2.87	**52.96 ± 0.33**	25.74 ± 2.02	38.55 ± 2.83*
Babel					
Lao	**30.68 ± 3.94**	25.27 ± 2.74	15.97 ± 1.16	29.75 ± 2.47	25.03 ± 1.74
Zulu	21.09 ± 2.74	31.60 ± 4.16	**52.71 ± 3.94***	35.97 ± 1.63	38.36 ± 2.79
Amharic	17.59 ± 0.17	13.88 ± 0.66	**30.82 ± 0.53***	26.49 ± 0.99	27.60 ± 0.46
Mongolian	12.92 ± 1.98	18.82 ± 1.74	**32.32 ± 0.75**†	26.35 ± 1.55	22.28 ± 0.85
Javanese	20.63 ± 3.83	27.44 ± 2.85	**41.32 ± 3.16***	32.47 ± 1.59	31.31 ± 3.56

best and considerably higher than all other vowel recognition rates are achieved by the *xl-nst* models. As explained before, this is not surprising since these models are not actually applied cross-lingually in this case. Overall, the *uni-5* models yield the best results on all evaluation languages excluding Danish. On Icelandic, they improve 11.33 percentage points over the *nst* and 14.14 percentage points over the *lnet* baseline. On Catalan, they improve 4.46 percentage points over the *nst* and 9.65 percentage points over the *lnet* baseline. On Serbian, they improve 40.55 percentage points over the *nst* and 28.76 percentage points over the *lnet* baseline. Finally, on Finnish, they improve 32.43 percentage points over *nst* and 23.5 percentage points over the *lnet* baseline.

The results on the Babel corpora reveal that the *xl-uni* models on average outperform both baselines on all evaluation languages except Lao. On the Lao data set, the highest vowel recognition rates are achieved by *xl-nst* models, but this result is only slightly better than that of the *xl-uni-10* model. On all four other Babel languages, the *xl-uni* models score higher than both baselines, with the *uni-5* model outperforming all other models. As was the case with the NST and parliamentary corpora, on average, the *uni-5* models yield the best results on all evaluation languages, except Lao. On Zulu, they improve 31.62 percentage points over *nst* and 21.11 percentage points over the *lnet* baseline. On Amharic, they improve 13.59 percentage points over *nst* and 16.94 percentage points over the *lnet* baseline. On Mongolian, they improve 19.4 percentage points over *nst* and 13.5 percentage points over the *lnet* baseline. Finally, on Javanese, they improve 20.69 percentage points over *nst* and 13.88 percentage points over the *lnet* baseline.

To test the significance of our results, we perform two-sample t-tests for independent samples with unequal variances (also known as Welch's t-test). Namely, we test whether each of the *(xl-)uni* vowel recognition rates is significantly higher than either of the baselines for each of the evaluation languages, which adds up to 78 comparisons. We use the Holm-Bonferroni method (Holm, 1979) to correct the significance threshold for multiple comparisons, which gives the threshold of 0.0015. The statistically significant results, the ones below 0.0015, are marked in Table 14.12. We can see that the results of all *uni* models

on the NST corpus are statistically significant. On the parliamentary corpora, the *xl-uni-5* results are statistically significant on Icelandic compared with the *xl-lnet* baseline, and on Serbian, compared with the *xl-nst*. The *xl-uni-10* results are significant on Icelandic compared with the *xl-lnet* baseline, while the *xl-uni-16* results are significant on Serbian and Finnish compared with the *xl-nst*. On the Babel corpora, only the *xl-uni-5* are statistically significant compared with the *xl-nst* for Zulu, Amharic, and Javanese, and compared with the *xl-lnet* for Mongolian.

These results suggest that fine-tuning models on language-universal vowel categories can indeed improve cross-lingual vowel recognition, especially when dealing with only a small number of broad categories, such as *uni-5*. However, they do not tell us how these models perform on individual vowels and how their performance relates to the arrangement of the reference and predicted vowel categories in the vowel space. This will be investigated throughout the remainder of this section.

14.4.1 Phone Prediction on the NST Corpus

Tables 14.13-14.17 provide phone prediction rates for individual reference vowels of each of the five cross-lingual models respectively on the NST corpus. Looking at the *nst* baseline prediction rates, the predictions on Norwegian and Swedish are more accurate on average, and more closely aligned, especially on the vowels: [i, ɛ, ɔ, u, y, ø, œ, ɪ, ʏ, ʉ, ʊ], and [ɵ]. This is presumably due to Norwegian and Swedish having more similar vowel systems. When it comes to the other baseline, *lnet*, it is less accurate than the *nst* on most of the vowels. This is most likely a result of the fact that LanguageNet G2P models for different languages are not cross-linguistically compatible. This is especially evident for the Danish LanguageNet model. Its transcriptions are almost completely orthographic rather than phonemic and vowel inventory smaller than the Danish phonological vowel inventory. This is likely why the *lnet* is the worst performing model and why its performance on the Danish NST subcorpus is the least accurate of all models and NST subcorpora. On the other hand, the recognition rates of the *uni*-models seem more balanced across the three languages.

Their recognition rates are still higher on Norwegian and Swedish than on Danish, but the difference in performance is not as great as for the *nst* baseline.

We will now examine the vowels that occur in the vowel inventories of both the evaluation language and all of the fine-tuning languages. For the NST languages, these include the following vowels: [i, e / e̞, ɛ, ɑ, ɔ, o, u, y, ø/ ø̞].[1] The performance results indicate that the *uni* models generally achieve the largest improvement over the baseline on these vowels. However, this does not mean that all of these vowels see the same amount of improvement on every evaluation language. In fact, some *uni*-models perform below the baseline on certain vowels. For example, the most improvement is seen on the vowels [i, e, o], and [u]. We can see from the prediction tables that recognition improves considerably for the Danish and Norwegian [i] with all three *uni* models, but decreases for the Swedish [i]. We can also see that recognition of [e] and [o] improves for almost all *uni* models on all three languages. Recognition of the Danish vowel [u] improves the most with all three *uni* models, especially the *uni-16*. However, the performance on the Swedish [u] decreases with all three *uni* models. Relating these changes in performance to the organization of vowel categories for the different categorization methods (Figures 13.6-13.9), we can see that, in the original data, these vowels have different means and spreads across the three languages. Converting them to the language-universal vowel categories seems to help recognition in general, but not all vowels benefit from the conversion equally. For instance, separating the broad [e̞] and [o̞] categories into the narrower [e] and [ɛ], and [o] and [ɔ] does not seem to help improve the recognition rates for [ɛ] and [ɔ].

When it comes to the unrounded vowels [y] and [ø] / [ø̞], the performance results seem mixed. For example, the recognition rates for the Danish [y] increase considerably from the baselines to the *uni-16* models (from 23.92% and 8.9% to 73.61%). They improve somewhat for the Norwegian [y] (from 79.51% and 62.04% to 80.28%), but they decrease for the Swedish [y] (57.01% and 39.67% to 40.82%). Further-

[1] Although the mid front vowels [e̞] and [ø̞] are technically not the same categories as the close-mid vowels [e] and [ø], we group them together because they overlap in the vowel space and the broader [e̞] and [ø̞] encompass most of the vowel tokens from their corresponding narrower categories e and [ø].

Table 14.13: Phone prediction rates of the *nst* cross-lingual models on the NST corpus. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel in the three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, *spn* stands for spoken noise, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish					Norwegian					Swedish				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
i	i: **50.88**	ɪ: 26.44	del: 8.44	y: 2.56	k: 1.14	i: **71.71**	ɛ: 8.16	e: 5.51	del: 4.79	ɪ: 3.78	i: **85.38**	ɪ: 7.37	del: 1.84	ə: 1.40	e: 1.16
e	e: **30.41**	ɪ: 25.95	del: 12.10	i: 11.64	ə: 5.48	e: **55.37**	ɛ: 19.49	del: 13.59	spn: 5.54	a: 1.08	ə: 41.86	e: **28.04**	del: 12.58	ɛ: 4.76	i: 2.54
ɛ	ɛ: **34.01**	del: 20.48	e: 15.08	ɪ: 9.99	ə: 8.48	ɛ: **70.41**	e: 9.18	a: 8.08	del: 7.98	spn: 1.34	ɛ: **52.86**	æ: 10.44	e: 10.05	ə: 7.41	del: 7.27
ɑ	ɑ: **46.75**	del: 22.69	a: 11.22	ɾ: 6.22	æ: 3.50	a: 38.96	ɑ: **37.51**	ɔ: 7.47	del: 6.97	ʌ: 1.69	ɑ: **84.09**	del: 6.79	o: 3.39	ɔ: 0.93	æ: 0.84
ɔ	ɵ: 37.66	o: 22.50	del: 12.12	ɔ: **9.65**	ʊ: 7.42	ɔ: **58.02**	del: 10.52	œ: 5.73	ø: 4.45	r: 3.80	ɔ: **61.11**	o: 20.30	del: 4.98	ʊ: 3.09	ɑ: 2.47
o	u: 30.82	ʊ: 24.01	del: 14.96	o: **14.89**	spn: 3.64	ɔ: 28.21	o: **27.32**	spn: 10.20	r: 7.85	del: 7.70	o: **59.41**	u: 18.77	del: 7.41	ɔ: 5.57	ɾ: 1.60
u	u: **39.86**	ɵ: 21.68	ʊ: 17.39	del: 9.41	ʉ: 4.85	u: **74.68**	o: 14.46	del: 3.64	ʊ: 2.21	ɔ: 1.34	u: **79.81**	spn: 6.99	ʊ: 3.61	del: 1.99	o: 1.95
y	y: **23.92**	ʉ: 19.57	del: 19.29	ʃ: 11.68	ʏ: 9.02	y: **79.51**	i: 5.67	ʏ: 3.54	del: 3.26	e: 2.73	y: **57.01**	i: 29.00	ʉ: 4.34	ʏ: 3.29	ɾ: 1.31
ø	ø: **25.12**	ɵ: 18.39	del: 16.42	ʉ: 13.97	y: 6.15	ø: **82.80**	œ: 8.57	del: 2.71	e: 1.73	ɛ: 1.08	ø: **56.28**	ɔ: 13.96	del: 6.96	œ: 6.43	ɾ: 3.52
œ	del: 30.33	œ: **22.63**	ø: 16.92	ɾ: 5.93	ɵ: 4.44	œ: **81.87**	ø: 6.48	ɛ: 4.04	del: 2.28	e: 2.20	œ: **42.57**	ɔ: 27.98	del: 13.14	ø: 3.68	ɑ: 3.02
ɪ	-	-	-	-	-	ɪ: **80.04**	e: 6.47	i: 5.00	del: 3.41	ɛ: 2.46	ɪ: **85.37**	i: 5.60	del: 3.00	ə: 2.68	e: 0.88
a	ɛ: 31.84	del: 15.89	e: 11.95	ə: 9.01	ɑ: 8.40	-	-	-	-	-	ɑ: 44.41	ə: 31.39	ɾ: 6.83	del: 4.37	g: 3.51
ə	ə: **46.15**	del: 25.00	ɪ: 3.23	e: 3.09	t: 2.26	e: 47.96	del: 15.38	a: 15.22	ɛ: 5.25	r: 4.88	-	-	-	-	-
ʏ	-	-	-	-	-	ʏ: **56.39**	ɪ: 11.59	œ: 6.53	ɵ: 5.53	y: 4.09	ʏ: **76.63**	y: 11.97	ɪ: 4.52	ɵ: 2.11	ʉ: 1.60
ɵ	-	-	-	-	-	ɵ: **53.88**	ɔ: 16.55	ʉ: 16.34	ʊ: 2.97	ʏ: 2.96	ɵ: **57.44**	ʊ: 17.55	ɔ: 12.28	ʏ: 2.79	œ: 2.23
ʉ	-	-	-	-	-	ʉ: **83.45**	y: 6.33	del: 1.76	ø: 1.36	ɵ: 1.14	ʉ: **77.51**	ɵ: 12.15	o: 1.92	del: 1.76	ə: 0.94
ʊ	-	-	-	-	-	ʊ: **59.12**	ɔ: 16.10	o: 8.31	u: 5.69	del: 4.06	del: 37.06	ʊ: **35.33**	ə: 10.99	u: 6.06	ɔ: 2.90
ɒ	ɾ: 38.78	ɔ: 17.87	del: 14.59	o: 8.54	ɑ: 7.15	-	-	-	-	-	-	-	-	-	-
ʌ	del: 23.51	ɾ: 16.51	ə: 12.14	ɔ: 11.08	ɑ: 7.19	-	-	-	-	-	-	-	-	-	-
æ	-	-	-	-	-	ɛ: 53.83	del: 15.55	e: 9.38	a: 8.19	r: 4.05	-	-	-	-	-
mean	-	-	-	-	**25.54**	-	-	-	-	**43.98**	-	-	-	-	**45.35**

Table 14.14: Phone prediction rates of the *lnet* cross-lingual models on the NST corpus. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel in the three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, *spn* stands for spoken noise, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish					Norwegian					Swedish				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
i	ɪː 38.57	del: 20.47	iː **17.23**	ɛː 3.45	j: 2.72	iː **70.88**	ɪː 13.67	del: 4.60	ɛː 2.33	e: 2.04	iː **71.54**	ɪː 9.00	del: 4.22	ɛː 2.55	e: 1.48
e	ɛː 24.13	del: 23.53	e: **17.81**	æː 4.92	ɪː 4.81	ɛː 36.80	e: **34.63**	a: 7.19	del: 6.79	ɑː 4.40	e: **26.49**	ɛː 25.89	əː 15.92	æː 13.57	del: 6.96
ɛ	ɛː **37.53**	del: 19.19	e: 12.81	ɪː 5.15	æː 4.42	ɛː **46.99**	e: 28.12	a: 6.93	del: 5.73	ɑː 3.51	ɛː **35.51**	e: 19.88	əː 14.20	æː 14.00	del: 6.31
ɔ	ɔː **76.39**	del: 6.81	ɑː 5.27	u: 2.77	o: 1.64	ɔː **40.80**	u: 20.60	o: 17.82	del: 4.04	r: 3.44	ɔː **51.19**	o: 23.85	u: 7.99	del: 5.22	iː 2.47
o	ɔː 28.48	del: 21.41	u: 15.69	o: **11.78**	ɑː 10.04	u: 37.13	ɔː 23.60	o: **14.05**	del: 6.72	r: 3.98	ɔː 70.65	o: **8.09**	del: 7.73	ɑː 3.14	ɾː 2.08
u	uː **65.93**	ʉː 10.86	del: 7.03	ɔː 6.58	o: 2.94	uː **31.46**	ɵː 27.62	ɔː 11.57	ʉː 7.44	o: 5.49	o: 34.98	ɔː 27.46	uː **11.43**	del: 9.49	iː 3.73
y	del: 32.64	ʉː 18.40	j: 10.78	yː **8.90**	ʏː 7.37	yː **62.04**	ʏː 10.90	iː 8.01	del: 4.34	ɪː 3.78	yː **39.67**	ʏː 17.75	iː 14.41	øː 8.21	kː 6.24
ø	øː **44.31**	del: 22.83	ʉː 10.68	uː 3.98	ɾː 2.64	øː **39.66**	œː 38.09	y: 6.04	del: 5.37	r: 2.83	øː **58.74**	o: 14.18	del: 9.82	ɔː 6.10	ɑː 3.02
ɪ	-	-	-	-	-	iː 48.09	ɪː **20.59**	del: 8.00	l: 6.19	uː 3.34	iː 49.51	del: 15.05	ɪː **13.28**	ɛː 2.97	eː 1.68
a	del: 26.24	ɑː 25.33	ɛː 18.07	eː 7.17	æː 5.15	-	-	-	-	-	ɑː 41.03	ɛː 11.92	del: 8.89	eː 7.63	ɾː 7.12
ɑ	-	-	-	-	-	aː 51.12	ɑː **22.47**	del: 5.54	oː 3.76	ɔː 3.23	ɑː **53.52**	ɛː 11.38	eː 7.74	ɾː 4.57	əː 4.37
ə	ɛː 25.79	del: 25.39	eː 11.39	ɪː 5.46	əː **5.27**	ɛː 36.07	eː 30.96	del: 8.99	aː 6.68	r: 4.25	-	-	-	-	-
ʏ	-	-	-	-	-	yː 51.71	ʏː **13.31**	del: 9.49	iː 6.83	r: 4.30	yː 27.28	ʏː **24.24**	iː 14.36	øː 8.27	kː 6.92
ʉ	-	-	-	-	-	ɵː 54.35	ʉː **21.37**	uː 9.05	del: 3.69	yː 2.88	uː 64.70	ʉː **14.49**	del: 4.34	øː 2.13	ɔː 2.05
æ	-	-	-	-	-	ɛː 36.62	eː 29.61	del: 9.34	aː 6.66	r: 4.51	-	-	-	-	-
ɵ	-	-	-	-	-	-	-	-	-	-	uː 62.21	ʉː 13.99	del: 4.92	øː 2.52	ɔː 2.29
œ	-	-	-	-	-	-	-	-	-	-	øː 45.81	del: 16.22	oː 13.69	ɔː 7.66	ɾː 3.96
mean	-	-	-	-	**20.54**	-	-	-	-	**28.99**	-	-	-	-	**29.78**

Table 14.15: Phone prediction rates of the *uni-5* cross-lingual models on the NST corpus. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel in the three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, *spn* stands for spoken noise, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish					Norwegian					Swedish				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
i	i: **80.62**	del: 10.97	e̞: 3.25	u: 0.76	t: 0.69	i: **83.67**	e̞: 8.36	del: 5.12	spn: 0.67	r: 0.60	i: **71.84**	e̞: 17.03	del: 6.59	ä: 0.92	spn: 0.87
e̞	e̞: **56.61**	i: 19.68	del: 14.46	ä: 1.74	r: 1.29	e̞: **70.19**	i: 10.24	del: 9.31	ä: 3.40	spn: 2.13	e̞: **60.77**	ä: 17.64	del: 8.72	i: 5.16	o̞: 3.29
ä	ä: **42.28**	del: 22.74	e̞: 16.93	r: 4.92	o̞: 4.87	ä: **60.27**	e̞: 14.63	del: 12.06	o̞: 7.06	r: 1.76	ä: **88.77**	o̞: 3.33	del: 3.28	e̞: 2.76	r: 0.45
o̞	o̞: **44.85**	del: 19.01	e̞: 10.18	u: 8.75	ä: 6.10	o̞: **52.12**	e̞: 11.05	del: 10.52	ä: 9.84	u: 7.92	o̞: **61.19**	ä: 21.68	del: 4.96	e̞: 4.64	u: 4.28
u	u: **59.37**	del: 17.50	o̞: 6.60	e̞: 4.47	i: 2.91	u: **67.09**	o̞: 8.03	del: 7.03	i: 6.94	e̞: 6.25	u: **60.95**	o̞: 18.27	del: 6.80	i: 4.17	e̞: 3.54
mean	-	-	-	-	**56.05**	-	-	-	-	**67.67**	-	-	-	-	**69.91**

Table 14.16: Phone prediction rates of the *uni-10* cross-lingual models on the NST corpus. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel in the three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, *spn* stands for spoken noise, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish					Norwegian					Swedish				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
i	i: **83.77**	del: 7.63	e: 2.25	ɨ: 0.74	ɛ: 0.66	i: **82.98**	e: 6.14	del: 4.32	ɨ: 2.98	ɛ: 0.73	i: **62.98**	e: 19.56	del: 7.40	ɨ: 4.50	ə: 0.94
e	e: **44.62**	i: 27.81	del: 13.93	ɛ: 2.12	ɾ: 1.41	e: **51.90**	i: 18.21	del: 10.72	ɨ: 5.85	ɛ: 4.28	e: **46.36**	ɛ: 21.11	del: 9.37	i: 5.59	ə: 5.53
ɛ	ɛ: **34.95**	e: 25.36	del: 18.66	i: 3.95	ɾ: 3.29	ɛ: **44.05**	e: 17.30	del: 13.78	ə: 7.66	a: 5.45	ɛ: **49.14**	a: 24.72	del: 9.68	e: 3.75	ə: 3.20
a	a: **35.78**	ɛ: 19.98	del: 19.55	ɾ: 6.82	e: 3.03	a: **50.87**	ɛ: 14.06	del: 13.67	ɑ: 7.50	ə: 3.79	a: **67.74**	ɑ: 17.50	del: 3.95	ə: 2.80	ɛ: 2.19
ɑ	ɑ: **35.47**	del: 17.40	a: 14.25	ɾ: 10.27	ɔ: 6.49	ɑ: **59.03**	ɔ: 13.41	del: 9.27	a: 8.18	ɛ: 2.73	ɑ: **72.53**	a: 11.43	ɔ: 4.88	del: 4.83	ə: 1.14
ɔ	ɔ: **27.76**	del: 17.83	ə: 14.08	o: 10.04	ɾ: 9.34	ɔ: **46.65**	o: 11.88	del: 11.42	ɑ: 8.48	ə: 5.92	ɔ: **43.89**	ɑ: 30.81	o: 7.03	del: 5.29	a: 2.63
o	o: **40.41**	del: 13.54	u: 12.06	ə: 7.83	ɨ: 6.07	o: **38.81**	u: 15.04	ɔ: 12.36	del: 10.43	spn: 5.37	o: **48.04**	ɔ: 24.21	u: 6.57	del: 5.34	ə: 4.77
u	u: **64.01**	del: 9.85	o: 6.16	ɨ: 5.33	ə: 4.45	u: **73.77**	o: 6.04	del: 5.21	ɔ: 3.01	i: 2.98	u: **61.78**	o: 18.39	del: 6.14	ɨ: 2.54	i: 1.84
ɨ	del: 29.51	ɨ: **18.89**	i: 14.81	e: 12.76	ə: 3.94	ɨ: **46.27**	i: 19.36	del: 10.04	e: 6.05	ə: 3.69	ɨ: **32.91**	ə: 17.79	del: 11.73	e: 10.97	i: 6.47
ə	ə: **23.19**	del: 22.61	e: 13.22	ɛ: 10.34	i: 5.91	ə: **31.44**	del: 14.07	e: 10.19	i: 7.63	a: 6.58	ə: **40.90**	a: 11.51	ɛ: 10.96	del: 9.39	ɑ: 6.78
mean	-	-	-	-	**42.58**	-	-	-	-	**54.24**	-	-	-	-	**54.25**

Table 14.17: Phone prediction rates of the *uni-16* cross-lingual models on the NST corpus. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel in the three evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, *spn* stands for spoken noise, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish					Norwegian					Swedish				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
i	i: **76.74**	del: 9.02	y: 5.93	e̞: 2.69	j: 0.64	i: **85.91**	e̞: 4.92	del: 3.87	y: 1.54	i: 1.11	i: **73.08**	e̞: 14.08	del: 4.83	i: 2.37	spn: 1.12
e̞	e̞: **50.27**	i: 21.96	del: 13.16	a: 1.77	r: 1.75	e̞: **60.62**	i: 14.48	del: 9.40	a: 3.71	i: 2.38	e̞: **53.07**	a: 19.59	del: 9.42	i: 5.98	ə: 4.41
a	a: **37.67**	e̞: 20.30	del: 18.92	r: 5.78	ə: 2.27	a: **59.20**	e̞: 15.64	del: 10.70	ə: 4.22	ɑ: 3.95	a: **71.75**	ɑ: 16.06	del: 4.12	ə: 2.16	e̞: 1.75
ɑ	ɑ: **24.91**	del: 18.51	a: 12.29	r: 9.22	o̞: 7.71	ɑ: **58.68**	a: 12.47	del: 8.02	ɤ̞: 5.87	e̞: 2.93	ɑ: **77.92**	a: 11.00	del: 4.59	ɤ̞: 0.91	ə: 0.91
ɤ̞	del: 24.96	ɤ̞: **15.54**	o̞: 12.71	ə: 7.87	r: 7.64	ɤ̞: **25.64**	ɑ: 16.05	del: 13.87	ə: 11.39	e̞: 11.31	ɑ: 51.32	ɤ̞: **11.88**	del: 9.68	o̞: 7.31	ə: 3.95
ɯ	del: 39.26	u: 6.84	ə: 5.94	r: 5.59	o̞: 4.96	del: 28.84	ɨ: 15.59	e̞: 14.49	i: 8.40	ɯ: **5.55**	ɑ: 17.98	del: 15.12	ɤ̞: 13.27	ə: 12.35	i: 6.64
ɨ	del: 34.14	i: 16.49	e̞: 11.39	i: **9.38**	ʉ: 3.43	ɨ: **34.73**	i: 25.75	del: 13.19	e̞: 7.49	a: 4.61	ɨ: **24.54**	del: 18.63	e̞: 16.76	i: 14.79	ə: 11.84
ə	del: 25.10	e̞: 21.55	ə: **15.41**	r: 5.03	i: 4.09	ə: **25.38**	e̞: 20.64	del: 15.72	a: 12.03	i: 7.17	ə: **26.84**	a: 23.47	del: 13.89	ɑ: 13.12	e̞: 9.43
y	y: **73.61**	del: 6.89	ʃ: 6.02	u: 5.49	ʉ: 2.96	y: **80.28**	ʉ: 7.96	i: 6.23	del: 2.05	ø̞: 1.22	y: **40.82**	ʉ: 17.02	del: 14.92	ø̞: 10.64	i: 8.45
ø̞	ø̞: **31.16**	y: 17.36	del: 14.20	u: 14.09	ʉ: 5.23	ø̞: **31.83**	ʉ: 23.04	y: 16.37	ɵ: 7.94	del: 5.02	ø̞: **43.77**	del: 13.04	ɵ: 11.10	ʉ: 7.08	œ: 5.42
œ	del: 20.08	œ: **16.44**	u: 13.26	ɵ: 8.94	o̞: 7.81	ɵ: 22.01	o̞: 16.12	del: 12.26	œ: **12.12**	ø̞: 9.26	œ: **32.84**	del: 13.50	o̞: 13.27	a: 8.20	ɒ: 7.21
ɒ	del: 19.51	o̞: 17.12	r: 16.78	ɵ: 9.59	ɒ: **8.54**	o̞: 47.93	ɒ: **11.63**	del: 10.40	ɵ: 6.98	r: 3.66	ɒ: **40.77**	ɑ: 20.72	o̞: 19.46	del: 5.25	u: 3.02
o̞	o̞: **47.81**	u: 17.49	ɵ: 10.08	del: 9.71	ʉ: 3.65	o̞: **56.50**	u: 15.38	del: 7.15	spn: 5.54	ɵ: 2.47	o̞: **68.62**	ɒ: 8.84	u: 6.06	ɵ: 3.50	ɑ: 3.36
u	u: **79.35**	o̞: 4.41	ʉ: 4.21	del: 3.90	ɵ: 3.40	u: **80.17**	o̞: 8.51	del: 3.09	ʉ: 2.31	spn: 1.92	u: **67.30**	o̞: 18.57	del: 5.43	spn: 1.75	ʉ: 0.89
ʉ	ʉ: **40.08**	y: 19.33	del: 11.91	u: 8.82	ø̞: 4.47	ʉ: **55.17**	y: 19.36	del: 6.14	u: 4.98	o̞: 2.99	ʉ: **38.77**	ɵ: 20.99	o̞: 7.75	ø̞: 7.07	u: 6.37
ɵ	ɵ: **36.21**	del: 15.04	ø̞: 13.46	ʉ: 12.43	u: 5.55	ɵ: **31.99**	ʉ: 14.56	o̞: 11.16	u: 9.25	del: 8.89	ɵ: **53.78**	o̞: 14.04	ɒ: 5.72	œ: 5.61	del: 4.96
mean	-	-	-	-	**42.53**	-	-	-	-	**58.98**	-	-	-	-	**57.85**

more, the recognition rates for the Danish [ø] improve somewhat, but they decrease for both the Norwegian and Swedish [ø]. The reason for this might be due to Danish having quite different distributions of [y] and [ø] from their distributions in Norwegian and Swedish. Therefore, converting to language-universal categories benefits Danish more the other two languages. At the same time, Norwegian and Swedish have more than one rounded vowel category high in the front. They have [y], [ʏ], and [ʉ], while Danish has only [y]. Merging and splitting these three categories on formant values might be the source of confusion and interference for the transformer's language model.

Next, we look at the vowels that are missing from the phonological systems of one or both fine-tuning languages, but which are part of at least one of the language-universal vowel sets. These are the following vowels: [a / ä, ə, ʉ, ɵ, ɒ]. For Danish, the recognition rate for a improves from 5.23% for the *nst* and ≈0% for the *lnet* model to 42.28% for the *uni-5*, 35.78% for the *uni-10*, and 37.67% for the *uni-16* model. For Swedish, it improves from 0.45% for the *nst* and 3.07% for the *lnet* model to 88.77% for the *uni-5*, 67.74% for the *uni-10*, and 71.75% for the *uni-16* model. The low baseline recognition rates probably stem from the fact that this vowel category has quite different realizations across the three Scandinavian languages, which can be seen in Figure 10.2.

For Danish, the recognition rate for [ə] degrades from 46.15% for the *nst* but improves over the *lnet* baseline to 23.19% for the *uni-5* and 15.41% for the *uni-10* model. For Norwegian, the recognition rate for [ə] improves from 3.21% for the *nst* and ≈0% for the *lnet* baseline to 31.44% for the *uni-10* and 25.38% for the *uni-16* model. The recognition rate for Danish ɒ improves from ≈0% for the *nst* to 8.54% for the *uni-16* model. The recognition rates for Norwegian and Swedish [ʉ] deteriorate from 83.45% and 77.51% for the *nst* to 55.17% and 38.77% for the *uni-16* model. However, these rates are still an improvement over the *lnet* baseline, which scores 21.37% on Norwegian and 14.49% on Swedish. The situation is similar for Norwegian and Swedish [ɵ], whose *nst* baseline recognition rates degrade from 53.88% and 57.44% to 31.99% and 53.78% for the *uni-16* model. The relatively high and cross-lingually aligned performances of the *nst* model on Norwegian and Swedish ʉ and ʉ seem to be a result of these two vowels having

very similar cross-lingual distributions in the normalized $F_1 - F_2$ space, especially the vowel ʉ, which has almost the same mean and spread in both languages. Another likely reason is that they are due to the lexical and phonotactic similarity between Norwegian and Swedish, which can be captured by the transformer model. Converting the phonological vowel categories to the formant-based categories affects the transformer's language model and likely leads to reduced performance.

Next, we will look at the vowels that we have introduced as part of the three language-universal categorization levels but which otherwise do not occur in the phonological inventories of the NST languages: [ɨ, ʏ̞, ɯ̟, œ]. The performance on these vowels is always markedly below the mean vowel recognition rate. In most cases, it is less than half the mean recognition rate and among the worst results compared with the rest of the vowels. For example, the mean recognition rates of the *xl-uni-16* models on the vowel [œ] are 16.44% for Danish, 12.12% for Norwegian, and 32.84% for Swedish. On the vowel [ɨ], they are 9.38%, 34.73%, and 24.54% for the three languages respectively. On the vowel [ʏ̞], they are 15.54%, 25.64%, and 11.88%, while, on the vowel [ɯ̟], they are ≈0%, 5.55%, and 1.85% for the three languages respectively.

Finally, looking at all of the *uni* models' predictions, can we somehow infer the most likely vowel inventory of an unseen language? Without knowing anything about the target language, it would be difficult to determine if it has additional vowels that do not occur in our language-universal vowel sets. For example, there are some Scandinavian vowels that are not in any of the *uni* vowel sets, such as: [ɪ, ʏ, œ, ʊ, æ, ʌ], so our models can never predict them.

Table 14.18 shows the distribution of vowels predicted by the *uni* models compared with the distribution of the same vowels in the *nst* (dictionary-based) reference transcripts. We merge the vowel categories [e] and [e̞], [ø] and [ø̞], [o] and [o̞], and [a] and [ä] from the different models, so we can show their frequencies side by side. The vowels [ɨ], [ʏ̞], [ɯ̟], and [œ], which do not occur in the *nst* vowel systems of any of the evaluation languages, are highlighted in orange. For all three languages, these four vowels are among the least frequently predicted by the *uni-10* and *uni-16* models. As we discussed earlier, they are also among the most challenging for our models to recognize.

Table 14.18: Vowel prediction distribution of the *uni* models compared with the distribution of the same vowels in the *nst* reference transcripts. All distributions are measured on the development sets of the three NST subcorpora. The vowels highlighted in pale orange are not found in the *nst* vowel systems of any of the evaluation languages. The hyphen (-) means that the vowel category does not occur in a model's vowel set.

	Danish				Norwegian				Swedish			
	reference	*uni-5*	*uni-10*	*uni-16*	reference	*uni-5*	*uni-10*	*uni-16*	reference	*uni-5*	*uni-10*	*uni-16*
ə	14.06	-	7.61	4.42	23.23	-	6.90	4.29	-	-	8.34	4.51
e	13.34	32.69	17.91	22.25	10.29	36.07	16.92	21.22	22.31	27.67	13.95	16.70
ɛ	12.58	-	12.15	-	7.02	-	11.88	-	11.33	-	13.96	-
i	9.86	30.13	23.85	24.17	5.82	25.51	21.12	20.58	4.44	19.88	14.35	15.91
a	8.76	16.55	9.74	10.21	-	20.75	10.44	13.14	19.05	32.99	17.96	20.77
ɑ	5.36	-	7.25	5.10	13.07	-	9.78	8.57	5.93	-	11.22	12.25
ɔ	4.10	-	4.50	-	6.12	-	7.68	-	5.97	-	6.14	-
o	3.73	11.25	5.40	7.00	5.62	11.42	4.26	7.04	3.59	13.36	5.51	7.32
u	3.52	9.38	6.82	8.58	2.73	6.26	5.04	5.80	2.47	6.10	4.27	5.01
ɒ	2.16	-	-	1.49	-	-	-	0.71	-	-	-	1.53
y	1.49	-	-	4.15	0.93	-	-	5.57	1.05	-	-	2.18
ø	1.35	-	-	2.34	1.19	-	-	2.03	2.45	-	-	2.33
ʉ	-	-	-	2.97	2.59	-	-	3.46	2.94	-	-	3.00
ɵ	-	-	-	3.37	4.19	-	-	2.20	3.62	-	-	4.34
ɨ	-	-	4.77	1.72	-	-	5.98	2.94	-	-	4.30	1.86
ʏ	-	-	-	1.82	-	-	-	1.95	-	-	-	1.04
ɯ	-	-	-	0.03	-	-	-	0.11	-	-	-	0.07
œ	-	-	-	0.39	-	-	-	0.37	-	-	-	1.18

These two are related and in part stem from the distribution of these vowels in the fine-tuning data, which can be found in Tables 13.4 and 13.5, in Chapter 13. However, comparing these two tables with the predicted vowel frequencies from Table 14.18, we can see that vowel prediction distribution is not always the same as vowel distribution in the fine-tuning data. Therefore, it is possible that the low prediction frequency and recognition rate of these vowels could tell us that they are likely not part of the Scandinavian vowel systems.

At the same time, vowel prediction distributions show that certain vowels have disproportionately high frequencies in the outputs of the *uni* models. These are the vowel [i] for all three *uni* models on Danish and Norwegian, vowel [e] for the *uni-5* and *uni-16* models on Danish and Norwegian, and the vowels [e] and [a] for the *uni-5* model on Swedish. This could indicate that these categories are too

broad for a single vowel and that they could potentially be split into more categories. In the case of these three languages, this would be a correct assumption, as all three vowel systems have a four-level organization of front unrounded vowels: [i], [e], [ɛ], and [a], as well as a distinction between front and back open unrounded vowels: [a] and [ɑ]. Nevertheless, more research is needed to investigate how the outputs of the phone recognition models could be modified to relate the predicted vowels to the phonological vowel categories of the target language.

14.4.2 Phone Prediction on Parliamentary Speech

As previously explained, we use a different approach to evaluate vowel prediction performance on the parliamentary corpora. Namely, since we do not have formant-based transcriptions for these data sets, we evaluate the phone recognition models against their dictionary-based reference transcriptions. Tables 14.19-14.23 provide phone prediction rates for individual reference vowels of each of the five cross-lingual models respectively on the parliamentary corpus. We will first take a look at the models' performance on the reference vowels that exist in a given model's vowel set. Subsequently, we will analyze the performance on the reference vowels that do not exist in the vowel set of a given model. Finally, we will discuss the possibility of inferring the vowel inventory of an unseen language based on the models' predictions.

Starting with the *xl-nst* model's performance on Danish, we can see from Table 14.19 that this model has all Danish reference vowels in its vowel set and that it achieves consistently high vowel recognition rates: 73.29%-93.64% (85.80% on average across all reference vowels). As discussed earlier, this model has "seen" Danish during fine-tuning and its transcriptions are based on the original NST lexicons, so its high vowel recognition rates are not surprising. Moving on to the *xl-lnet* baseline, this model has also "seen" Danish but its transcriptions are based on the LanguageNet G2P model for Danish. It achieves somewhat worse recognition rates on the vowels in its vowel system than the the *xl-nst* baseline, but still relatively high. However, since three of the Danish reference vowels are missing from its vowel set, this

brings its average vowel recognition rate to 38.67%. Compared with the baselines, the *xl-uni* models all achieve lower vowel recognition rates. The main reason for this is that they also do not have the full set of Danish reference vowels in their vowel sets. Namely, the *xl-uni-5* is missing 9 vowels, *xl-uni-10* is missing 5, and the *xl-uni-16* is missing 4. Still, their performance on just the known reference vowels is overall worse than the baseline performances on those same vowels. We believe that this could be caused, at least in part, by the differences in the underlying language models between the *xl-uni* models and the baselines. In other words, the formant-based vowel representations in the utterance transcripts interfere with the transformer's ability to capture the underlying language model, because the same word will have different transcripts with each realization, many of which might not be dependent only on the phonetic context, but also on the broader situational context and the speaker.

Now when we consider actual cross-lingual performance, for example on Icelandic, we can see that both baselines perform considerably worse: the *xl-nst* 20.69% and *xl-lnet* 17.88%, despite the fact that none of the Icelandic reference vowels are missing from the *xl-nst* vowel set and only one is missing from the *xl-lnet*. This suggests that the baseline Scandinavian vowel systems are not compatible cross-lingually with the Icelandic vowel system. Comparing these results with the performance of the *xl-uni* models, we can see that the *xl-uni-5* model, which has only 3 out of 8 of the Icelandic reference vowels in its vowel set, outperforms both baselines because it achieves a relatively high recognition rate on the three known vowels: [i, a, u]. The *xl-uni-10* model, which knows 5 out of 8 Icelandic vowels, also outperforms both baselines, while the *xl-uni-16*, which like the *xl-uni-5* knows only 3 of the Icelandic vowels, outperforms only the *xl-lnet* baseline and performs slightly below the *xl-nst* baseline. However, if we look at only the performance on the known reference vowels, all three *xl-uni* models achieve higher than baseline recognition rates on almost all of the vowels (the exception is the performance of the *xl-uni-10* and *xl-uni-16* models on the vowel u).

Analyzing the cross-lingual vowel recognition on Catalan, we can see that the *xl-nst* baseline, which has all Catalan vowels in its vowel set, achieves 35.15%, the *xl-lnet* baseline, which is missing one of

the vowels, achieves 29.96%. The three *xl-uni* models have similar average vowel recognition rates, which are somewhat higher than both baselines, but their performance on individual vowels differs across the models. For example, the *xl-uni-5* model, which does not know 4 out 9 Catalan vowels, outperforms both baselines on the 5 reference vowels that it knows. Likewise, the *xl-uni-16*, which knows 6 Catalan vowels, outperforms both baselines on 3 known vowels, outperforms only one of the baselines on one known vowel, and has approximately the same performance as baseline on two known vowels. However, using the *xl-uni-10* model, which distinguishes four degrees of vowel height, helps improve the recognition of open-mid vowels [ɛ] and [ɔ], but it does so at the expense of the recognition rates on the close-mid vowels [e] and [o]. Therefore, its overall recognition rate remains close to the rates of the other two *xl-uni* models, despite having a 10-vowel system that most resembles the Catalan phonological vowel system.

When it comes to cross-lingual vowel recognition on Serbian, we can see that both baselines know all 5 Serbian reference vowels, but their overall recognition rates of 33.27% and 45.06% suggest that their large vowel sets are not well aligned with the Serbian vowel inventory. The *xl-uni* models also know all Serbian vowels, but they do not perform equally well on all vowels. Namely, the best performance is achieved by the *xl-uni-5* model. At 73.82%, it is substantially higher than both baselines. This result is not surprising as we expected that the *xl-uni-5* model's 5-vowel partition of the vowel space would match the Serbian phonological 5-vowel system. In comparison, the *xl-uni-16* model also scores above both baselines, but we can see a reduction in the recognition rates on almost all vowels caused by more frequent vowel confusions. What is interesting to note here is that confusions between rounded and unrounded vowels are rare. For instance, fewer than 1% of [i] vowel tokens are falsely recognized as [y], and fewer than 1% of [e] vowel tokens are falsely recognized as [ø].[2] Therefore, most of the confusions happen between vowels adjacent in the vowel space. Since we do not have narrow phonetic transcriptions of the parliamentary corpora, we cannot know whether some of the incorrectly predicted vowels correspond to different allophonic realizations of the

[2]This cannot be seen from Table 14.23 because it only shows top 3 predictions for each reference vowel.

reference vowels. On the other hand, using using the *xl-uni-10* model, which distinguishes four degrees of vowel height, on Serbian vowels introduces many confusions between the open-mid and close-mid vowels, i.e. [e] and [ɛ], and [o] and [ɔ], thus bringing its overall vowel recognition rate below the *xl-lnet* baseline.

Finally, moving on to cross-lingual vowel recognition on Finnish, the baseline vowel recognition rates are 20.53% and 29.46%, and both baselines know all Finnish reference vowels. We should remember that Finnish is the first non-Indo-European language we have considered so far, and the only from the whole parliamentary corpus collection. This means that lexical and phonotactic similarities between the evaluation language and the fine-tuning languages are minimal. For this reason, the baseline phone recognition models will likely not be able to "guess" the correct vowel using their knowledge of the Scandinavian language models.

The best performance on Finnish is achieved by the *xl-uni-5* model, which has 5 out of 8 Finnish vowels in its vowel system. This model consistently outperforms both baselines on all known reference vowels individually, except the vowel [i], which it can recognize correctly only 44.78% of the time. This recognition rate disagrees with the rates achieved on the other four languages. Namely, the *xl-uni-5* model has a consistent recognition rate of over 80% on all four previously analyzed languages. This could suggest that the formant-based vowel space delimited by the mean values of the four point vowels derived from the Scandinavian NST corpus might not provide the optimal position for the Finnish vowel [i]. The *xl-uni-16*, which knows 7 out of 8 Finnish reference vowels, also scores above the baseline. Although the introduction of front rounded vowels helps to improve the performance on the vowels [y] and [ø], the larger number of vowel categories also leads to more confusions and reduced performance on the primary vowels. Finally, the *xl-uni-10* model, which like the *xl-uni-5* recognizes only 5 out 8 Finnish vowels, suffers even more confusions than the other two *xl-uni* models. As we have seen with the previous evaluation languages, additional levels of vowel height lead to confusions particularly between the open-mid and close-mid vowels which are not distinguished in Finnish ([e] vs. [ɛ], and [o] vs. [ɔ]), bringing the overall vowel recognition rate below the *xl-lnet* base-

line. Another peculiarity that we have observed with Finnish is very high vowel deletion rate across all models, both baseline and *xl-uni*. At this point, it is not clear why this happens, but could be a result of vowel and consonant lengthening, which is characteristic of Finnish phonology and was indicated by duplicating the phone tokens in the reference transcriptions.

Regarding the performance on the reference vowels that do not occur in the *xl-uni* models' vowel sets, we will analyze each model separately. With five vowels, the *xl-uni-5* has the most limited vowel system of all the *xl-uni* models. This has proved useful for languages with smaller vowels sets of primary vowels, but it also means that its performance drops with the number of vowels outside its vowel set. Its predictions on the unseen reference vowels of the different languages tell us that the model usually predicts them as their closest vowel that is in its vowel set. For example, [ɑ] is frequently predicted as [a], [ɛ] as [e], [ɔ] as [o], and [ʊ] as [u]. When it comes to the rounded front vowels, they are often predicted as their unrounded counterparts, or closes unrounded counterparts. For instance, [y] as [i], [ø] as [e], and [œ] as [a]. Therefore, the model is generally able to recognize the broadest vowel categories.

With ten vowels, the *xl-uni-10* model's predictions of the unseen vowels look more diverse. Namely, it also generally predicts them as their closest vowels in the vowel space that it can find in its vowel set. However, its predictions are more evenly distributed across several closely related vowels. For example, [y] is variably predicted as [i] or [ɨ], [ø] as [ɨ], [ɛ], [e], and [ə], and [œ] as [ɑ], [a], or [ə].

With 16 vowels, the *xl-uni-16* model has the largest vowel set of all investigated models. It is also the only model that can distinguish corresponding rounded and unrounded vowels. This model has generally performed better than the *xl-uni-10* on the languages of the parliamentary corpora despite having more vowels in its vowel sets and potential for vowel confusions. We believe that this is not only because it was able to recognize the front rounded vowels in the languages that feature them, but also because its 8-primary-vowel set seems more in line with the languages of the parliamentary corpora. As mentioned previously, this model had no problem distinguishing rounded and unrounded vowels, so its predictions looked more closely related to

Table 14.19: Phone prediction rates of the *xl-nst* cross-lingual models on the parliamentary corpora. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language.

	Danish			Icelandic			Catalan			Serbian			Finnish		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	a: **84.97**	del: 7.48	ɛ: 3.02	ɑ: 33.15	del: 21.44	ə: 14.48	ɑ: 35.29	a: **16.34**	del: 13.42	ɑ: 32.98	ə: 16.11	a: **14.98**	ɑ: 37.70	del: 30.04	a: **10.67**
e	e: **90.19**	del: 3.43	i: 1.89	-	-	-	e: **38.67**	ɛ: 35.77	del: 7.36	ə: 29.19	e: **25.93**	ɛ: 18.69	del: 27.18	e: **26.81**	ɛ: 16.99
i	i: **92.60**	del: 3.41	e: 1.34	i: **58.31**	ɪ: 16.93	del: 9.65	i: **52.43**	ɪ: 26.98	del: 8.52	i: **51.35**	ɪ: 17.61	del: 10.64	del: 30.37	i: **23.92**	ɪ: 15.96
o	o: **87.36**	del: 5.19	u: 3.39	-	-	-	o: **35.39**	ɔ: 23.19	del: 13.06	o: **35.20**	ɔ: 23.44	del: 13.49	o: **34.13**	del: 21.52	ɔ: 19.12
u	u: **93.64**	del: 2.42	o: 1.86	u: **65.99**	del: 7.49	ʊ: 6.08	u: **69.28**	ʊ: 6.55	ɔ: 5.51	u: **57.76**	ʊ: 9.62	del: 9.05	u: **33.39**	del: 29.55	ʊ: 13.45
y	y: **84.48**	del: 9.26	ø: 1.86	-	-	-	-	-	-	-	-	-	del: 32.60	ʉ: 23.09	ɵ: 10.35
ø	ø: **89.22**	œ: 2.98	del: 2.91	-	-	-	-	-	-	-	-	-	ø: **37.41**	del: 18.61	ə: 8.29
œ	œ: **84.91**	del: 5.99	ø: 4.75	del: 25.87	œ: **25.07**	ø: 15.36	-	-	-	-	-	-	-	-	-
ɑ	ɑ: **90.16**	del: 5.92	ʌ: 0.84	-	-	-	-	-	-	-	-	-	-	-	-
ɒ	ɒ: **91.69**	del: 2.03	ʌ: 1.68	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	ɔ: **86.01**	o: 4.10	del: 3.42	ɔ: **30.06**	o: 28.15	del: 18.35	o: 26.51	ɔ: **20.83**	del: 17.22	-	-	-	-	-	-
ə	ə: **82.05**	del: 13.21	e: 1.13	-	-	-	ə: **35.35**	ɑ: 14.85	del: 13.44	-	-	-	-	-	-
ɛ	ɛ: **73.29**	del: 12.22	e: 10.47	ɛ: **36.63**	del: 18.98	e: 11.15	ɛ: **30.31**	del: 18.65	e: 15.44	-	-	-	-	-	-
ʌ	ʌ: **86.45**	del: 6.53	ə: 2.34	-	-	-	-	-	-	-	-	-	-	-	-
ɪ	-	-	-	ɪ: **19.49**	del: 19.22	ə: 18.80	-	-	-	-	-	-	-	-	-
ʏ	-	-	-	del: 21.97	ə: 21.62	ɵ: 12.72	-	-	-	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	ʊ: **26.09**	u: 20.72	ɔ: 18.42	-	-	-	-	-	-
æ	-	-	-	-	-	-	-	-	-	-	-	-	del: 32.68	ə: 14.20	a: 12.03
mean	-	-	**85.80**	-	-	**20.69**	-	-	**35.15**	-	-	**33.27**	-	-	**20.53**

Table 14.20: Phone prediction rates of the *xl-lnet* cross-lingual models on the parliamentary corpora. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray do **not** occur in the model's vowel set.

	Danish			Icelandic			Catalan			Serbian			Finnish		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	a: **85.57**	del: 3.68	ɛ: 2.62	ɑ: 35.21	del: 12.01	a: **11.13**	a: **34.90**	ɑ: 22.92	e: 9.82	a: **33.66**	ɑ: 20.74	e: 8.72	ɑ: 35.44	a: **18.63**	del: 17.38
e	e: **30.96**	i: 24.44	ɛ: 8.15	-	-	-	e: **56.44**	i: 14.58	ɛ: 9.66	e: **47.82**	ɛ: 22.08	del: 4.85	e: **46.30**	ɛ: 20.45	del: 14.36
i	i: **89.04**	ə: 2.74	e: 2.37	i: **78.82**	ɪ: 4.52	del: 3.93	i: **80.52**	ɪ: 4.71	del: 3.51	i: **74.51**	del: 4.92	ɪ: 4.00	i: **46.80**	del: 17.97	e: 5.09
o	o: **89.80**	ɔ: 2.32	u: 2.06	-	-	-	o: **35.32**	ɔ: 27.02	del: 6.88	ɔ: 33.04	o: **24.66**	del: 8.91	ɔ: 27.28	o: **24.06**	del: 13.32
u	u: **92.35**	o: 2.45	del: 1.61	u: **48.27**	o: 16.19	ɔ: 11.79	u: **65.97**	o: 7.89	ɔ: 6.60	u: **54.42**	ɔ: 12.99	o: 8.72	u: **34.26**	del: 23.17	ɔ: 16.26
y	y: **80.37**	del: 6.96	u: 4.16	-	-	-	-	-	-	-	-	-	ʉ: 28.08	del: 20.92	u: 13.19
ø	ø: **78.16**	y: 8.89	t: 2.89	-	-	-	-	-	-	-	-	-	ø: **47.84**	e: 10.00	del: 8.78
œ	ø: 41.13	r: 35.28	del: 4.47	ø: 32.97	del: 19.56	ɔ: 5.56	-	-	-	-	-	-	-	-	-
ɑ	a: 31.75	r: 25.35	e: 8.33	-	-	-	-	-	-	-	-	-	-	-	-
ɒ	r: 79.20	o: 3.72	del: 2.76	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	ɔ: **60.38**	u: 18.02	o: 4.44	ɔ: **38.92**	o: 30.09	del: 8.74	ɔ: **38.36**	o: 22.27	ɑ: 11.37	-	-	-	-	-	-
ə	ɛ: 25.46	e: 24.57	ə: **22.69**	-	-	-	e: 30.45	ɑ: 14.69	ɛ: 8.67	-	-	-	-	-	-
ɛ	ɛ: **44.37**	e: 22.09	ə: 11.50	ɛ: **28.19**	e: 25.93	del: 10.11	e: 26.61	ɛ: **18.51**	del: 14.36	-	-	-	-	-	-
ʌ	r: 38.80	o: 22.01	e: 6.41	-	-	-	-	-	-	-	-	-	-	-	-
ɪ	-	-	-	i: 26.18	e: 17.17	del: 11.74	-	-	-	-	-	-	-	-	-
ʏ	-	-	-	u: 16.26	del: 14.22	e: 13.71	-	-	-	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	o: 33.71	u: 33.01	ɔ: 11.69	-	-	-	-	-	-
æ	-	-	-	-	-	-	-	-	-	-	-	-	del: 21.08	ɛ: 17.46	e: 13.36
mean	-	-	**38.67**	-	-	**17.88**	-	-	**29.96**	-	-	**45.06**	-	-	**29.46**

Table 14.21: Phone prediction rates of the *xl-uni-5* cross-lingual models on the parliamentary corpora. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Danish			Icelandic			Catalan			Serbian			Finnish		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	a: **47.37**	e: 37.48	del: 11.21	a: **83.65**	e: 5.62	del: 5.02	a: **91.33**	e: 3.70	del: 2.07	a: **72.02**	e: 14.44	del: 5.07	a: **75.62**	del: 11.75	e: 5.15
e	e: **64.07**	i: 25.17	del: 4.71	-	-	-	e: **89.31**	i: 4.74	del: 2.45	e: **80.51**	a: 6.28	i: 5.38	e: **77.27**	del: 11.50	a: 6.55
i	i: **82.95**	e: 12.20	del: 2.99	i: **80.66**	e: 12.78	del: 3.41	i: **87.07**	e: 5.35	del: 3.18	i: **80.71**	e: 10.27	del: 4.58	i: **44.78**	del: 28.16	e: 14.79
o	u: 45.59	o: **38.88**	del: 6.78	-	-	-	o: **55.92**	a: 17.19	e: 7.36	o: **65.80**	a: 10.61	del: 7.90	o: **61.88**	del: 15.55	a: 10.56
u	u: **85.69**	o: 6.05	e: 3.52	u: **67.94**	o: 15.02	del: 6.22	u: **76.85**	o: 8.50	del: 3.62	u: **68.17**	o: 9.08	del: 6.56	u: **58.12**	del: 22.32	o: 9.03
y	i: 68.17	del: 16.14	e: 10.35	-	-	-	-	-	-	-	-	-	del: 31.48	i: 25.11	e: 24.45
ø	e: 65.30	del: 11.69	o: 8.16	-	-	-	-	-	-	-	-	-	e: 51.07	del: 20.96	a: 9.36
œ	a: 67.51	del: 14.54	e: 10.09	del: 25.89	e: 24.64	a: 23.48	-	-	-	-	-	-	-	-	-
ɑ	a: 78.07	del: 11.29	e: 4.58	-	-	-	-	-	-	-	-	-	-	-	-
ɒ	o: 64.41	a: 18.58	del: 7.37	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	o: 60.78	u: 24.61	del: 7.61	o: 41.87	del: 26.23	a: 8.57	a: 46.93	o: 19.41	del: 16.90	-	-	-	-	-	-
ə	e: 34.97	del: 20.04	i: 16.98	-	-	-	a: 40.96	e: 31.99	del: 13.09	-	-	-	-	-	-
ɛ	e: 53.87	a: 17.89	del: 16.21	e: 36.57	del: 22.94	a: 20.76	e: 48.95	del: 18.80	a: 12.03	-	-	-	-	-	-
ʌ	a: 57.00	e: 12.60	del: 12.16	-	-	-	-	-	-	-	-	-	-	-	-
ɪ	-	-	-	e: 41.06	i: 22.02	del: 18.72	-	-	-	-	-	-	-	-	-
ʏ	-	-	-	e: 43.92	del: 17.87	o: 10.27	-	-	-	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	u: 55.18	o: 17.89	del: 10.80	-	-	-	-	-	-
æ	-	-	-	-	-	-	-	-	-	-	-	-	a: 45.80	del: 26.09	e: 14.38
mean	-	-	**24.69**	-	-	**32.02**	-	-	**39.61**	-	-	**73.82**	-	-	**52.96**

Table 14.22: Phone prediction rates of the *xl-uni-10* cross-lingual models on the parliamentary corpora. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Danish			Icelandic			Catalan			Serbian			Finnish		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	a: **36.15**	ɛ: 33.69	del: 14.34	a: **36.92**	ɑ: 25.15	del: 14.60	a: **69.70**	ɑ: 15.27	del: 5.14	a: **37.90**	ɑ: 20.01	ɛ: 11.31	ɑ: 30.13	a: **26.51**	del: 22.99
e	e: **51.76**	i: 17.14	ɛ: 11.86	-	-	-	e: **46.30**	ɛ: 36.25	del: 6.13	ɛ: 40.32	e: **32.88**	del: 7.10	ɛ: 30.40	e: **28.22**	del: 24.83
i	i: **71.09**	e: 12.51	i: 7.68	i: **69.81**	e: 17.13	del: 5.75	i: **79.89**	e: 8.60	del: 4.82	i: **69.72**	e: 15.68	del: 6.24	i: **35.29**	del: 26.72	e: 14.14
o	o: **34.99**	u: 34.37	i: 8.82	-	-	-	ɔ: 21.94	o: **15.42**	del: 13.49	ɔ: 28.32	o: **25.87**	del: 10.75	o: **26.90**	del: 21.69	ɔ: 19.71
u	u: **63.63**	i: 19.78	o: 6.05	u: **49.06**	o: 14.98	del: 9.12	u: **57.19**	i: 12.52	o: 9.78	u: **47.46**	i: 13.04	o: 11.50	u: **41.79**	del: 25.11	o: 10.19
ʏ	i: 36.15	i: 36.05	del: 17.13	-	-	-	-	-	-	-	-	-	del: 33.99	i: 20.08	e: 15.78
ø	i: 29.17	ə: 26.97	e: 17.48	-	-	-	-	-	-	-	-	-	del: 25.00	ɛ: 23.91	ə: 19.66
œ	ɑ: 31.90	a: 23.11	ə: 15.56	del: 28.57	ə: 21.66	ɛ: 11.40	-	-	-	-	-	-	-	-	-
ɑ	ɑ: **57.80**	a: 24.08	del: 7.10	-	-	-	-	-	-	-	-	-	-	-	-
ɒ	ɔ: 37.92	ɑ: 22.09	o: 15.91	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	o: 42.04	ɔ: **24.09**	u: 13.40	ɔ: **34.16**	o: 19.69	del: 18.83	ɑ: 34.30	ɔ: **23.06**	del: 14.49	-	-	-	-	-	-
ə	i: 24.23	del: 20.63	e: 13.92	-	-	-	a: 21.61	ɛ: 20.96	del: 13.87	-	-	-	-	-	-
ɛ	ɛ: **37.31**	e: 23.96	a: 11.61	ɛ: **42.97**	del: 15.92	a: 14.59	ɛ: **43.84**	e: 19.92	del: 13.48	-	-	-	-	-	-
ʌ	ɑ: 28.13	a: 24.47	del: 14.26	-	-	-	-	-	-	-	-	-	-	-	-
ɪ	-	-	-	e: 38.07	del: 20.71	i: 12.27	-	-	-	-	-	-	-	-	-
ʏ	-	-	-	del: 20.79	ə: 19.41	i: 17.09	-	-	-	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	u: 37.84	o: 16.01	i: 14.77	-	-	-	-	-	-
æ	-	-	-	-	-	-	-	-	-	-	-	-	a: 35.74	del: 30.27	ɛ: 13.57
mean	-	-	**30.53**	-	-	**26.46**	-	-	**38.20**	-	-	**41.80**	-	-	**25.74**

Table 14.23: Phone prediction rates of the *xl-uni-16* cross-lingual models on the parliamentary corpora. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Danish			Icelandic			Catalan			Serbian			Finnish		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	a: **51.60**	e: 24.36	del: 13.68	a: **42.56**	ɑː 23.07	del: 14.72	a: **71.14**	ɑː 15.48	del: 5.52	a: **42.74**	ɑː 19.99	ə: 12.42	a: **30.44**	ɑː 28.24	del: 25.36
e	e: **48.57**	i: 28.28	del: 7.85	-	-	-	e: **77.31**	i: 7.55	a: 5.12	e: **62.45**	a: 9.78	del: 7.41	e: **59.50**	del: 20.04	a: 8.04
i	i: **78.25**	e: 7.51	ɨ: 6.97	i: **81.54**	e: 8.44	del: 4.62	i: **86.65**	del: 4.40	e: 3.33	i: **80.27**	e: 6.57	del: 5.69	i: **46.86**	del: 27.73	e: 7.60
o	u: 48.24	o: **25.22**	del: 10.17	-	-	-	o: **34.36**	del: 12.73	ɒː 9.64	o: **43.87**	del: 12.67	ɑː 9.81	o: **43.20**	del: 22.42	ɵ: 6.37
u	u: **74.72**	ʉ: 13.89	del: 3.45	u: **58.45**	o: 9.42	del: 9.30	u: **65.16**	ʉ: 11.70	del: 6.02	u: **56.94**	ʉ: 10.23	del: 8.80	u: **51.17**	del: 25.86	o: 5.43
y	y: **46.16**	ʉ: 32.91	del: 11.99	-	-	-	-	-	-	-	-	-	del: 32.10	y: **18.16**	ʉ: 17.18
ø	ø: **28.99**	ɵ: 26.72	ʉ: 24.86	-	-	-	-	-	-	-	-	-	ø: **37.05**	del: 18.93	ɵ: 13.38
œ	œ: 36.20	ɒː 18.19	del: 17.63	del: 32.49	ɵ: 17.84	ø: 9.58	-	-	-	-	-	-	-	-	-
ɑ	ɑː **52.87**	a: 28.71	del: 9.83	-	-	-	-	-	-	-	-	-	-	-	-
ɒ	o: 36.35	ɒː **33.98**	del: 6.76	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	o: 51.03	u: 22.67	del: 8.94	o: 36.18	del: 29.12	u: 5.52	del: 23.17	ɑː 21.01	ɒː 17.09	-	-	-	-	-	-
ə	del: 23.81	ɨ: 18.96	ə: **17.16**	-	-	-	a: 26.52	e: 18.34	ə: **17.73**	-	-	-	-	-	-
ɛ	e: 42.35	del: 19.13	a: 16.57	del: 28.65	e: 22.04	a: 19.72	e: 37.63	del: 23.04	a: 15.30	-	-	-	-	-	-
ʌ	a: 30.16	ɑː 29.38	del: 13.48	-	-	-	-	-	-	-	-	-	-	-	-
ɪ	-	-	-	e: 30.07	i: 23.61	del: 22.70	-	-	-	-	-	-	-	-	-
ʏ	-	-	-	del: 23.58	ø: 11.54	ə: 10.58	-	-	-	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	u: 43.60	del: 14.59	o: 13.79	-	-	-	-	-	-
æ	-	-	-	-	-	-	-	-	-	-	-	-	a: 39.31	del: 35.52	e: 5.76
mean	-	-	**29.56**	-	-	**19.60**	-	-	**38.81**	-	-	**56.18**	-	-	**38.55**

the reference vowels, even for the unseen vowels. For instance, the unseen vowel [ɛ] was often predicted as [e] or [a], the vowel [ɪ] as [e] or [i], vowel [ɔ] as [o], [u], or [ɒ], vowel [ʊ] as [u] or [o], vowel [ʏ] as [ø], and vowel [œ] as [ɶ], [ɒ], or [ɵ].

14.4.3 Phone Prediction on Noisy Telephone Speech

We use the same approach here as when evaluating phone predictions on parliamentary speech in the previous section. Namely, since we do not have formant-based transcriptions for Babel data sets, we evaluate the phone recognition models against dictionary-based reference transcriptions of these data sets. Tables 14.24-14.28 provide phone prediction rates for individual reference vowels of each of the five cross-lingual models respectively on the parliamentary corpus. It should be pointed out that none of these languages are Indo-European and are both typologically and geographically distant to the three Scandinavian languages in our fine-tuning data. Moreover, noisy telephone conversations are generally a more challenging type of speech, which was also evident from the overall phone and word error rates on these languages, which we reported in the previous sections.

We begin by looking at the models' performance on Lao on the reference vowels that exist in a given model's vowel set. The baseline models know 7 out 9 of Lao's reference vowels, and achieve mean vowel recognition rates of 30.68% and 25.27%. The uni models generally outperform both baselines on each individual known reference vowel except on the vowel [ɑ]. While the individual recognition rates of the *xl-uni-5* model on the known reference vowels are all above 58%, this model has only four of Lao's 9 vowels in its vowel set, so the average recognition rate ends up being below both baselines. However, we see that Lao has only one open unrounded vowel, labeled as [ɑ] in its reference transcription system, and our *xl-uni-5* model uses also only one open (central) unrounded vowel, originally labeled as [ä].[3]

[3]We strip the diacritics for centralization ([̈]) and lowering ([̞]) from the vowels in the *xl-uni-5* and *xl-uni-16* transcription systems when evaluating on the parliamentary and Babel data. These features are rarely marked in phonological systems as there are few language where the distinctions between a and [ä], or [e] and [e̞] are phonemic.

Therefore, the difference between the representations [a] and [ɑ] in this case is a matter of convention.[4] As we can see from Table 14.26, the *xl-uni-5* model predicts Lao's [ɑ] as [a] 36.18% of the time. This means that if we had used [ɑ] to represent the open central unrounded vowel in the *xl-uni-5* vowel system, the vowel recognition rate of this model on Lao would have been 33.25%, which is above both baselines.

The *xl-uni-10* and *xl-uni-16* models have wider coverage of Laos vowels, 7 out of 9, but their lower performance on individual vowels bring down their average close to the baselines. Of all our evaluation languages, Lao is the only one that has the two back unrounded vowels [ɤ] and [ɯ]. However, the *xl-uni-16* model, which has these two vowels in its vowel set, could not recognize most of them. As discussed previously, these two vowels were very rare in the fine-tuning data as neither occurs as a phoneme in the phonological systems of the Scandinavian languages. This is likely the reason why the model could not recognize them. Instead, it often predicted both of them as [e], or [ɤ] as [a], and [ɯ] as [u].

Looking at the performance results on Zulu, the baseline models which have seen all 6 of its reference vowels, achieve mean vowel recognition rates of 21.09% and 31.60%. The *xl-uni-5* and *xl-uni-16* models outperform both baselines on each seen vowel individually and on average. The *xl-uni-10* model also outperforms both baselines, but not on all vowels individually and by a smaller margin. As was the case with the previously analyzed languages, this one also shows that increasing the number of vowel categories and, especially, levels of vowel height, decreases the recognition rates of individual vowels.

On Amharic, the baseline models which have both seen 6 of its 7 reference vowels, achieve mean vowel recognition rates of 17.59% and 13.88%. The*xl-uni-5* model outperforms both baselines on each seen vowel individually and on average. While the *xl-uni-10* and *xl-uni-16* models also outperform both baselines on average, performance on individual seen vowels decreases compared with the *xl-uni-5* model. More specifically, the recognition rates of the primary vowels decrease when an additional level of vowel height is introduced in the *xl-uni-10* vowel space. Amharic has one rare vowel which does not occur in the

[4] Of course, [a] and [ɑ] differ phonetically, but we can expect that their realizations would overlap to a certain extent in the $F_1 - F_2$ space.

Scandinavian languages: [ɨ], but is seen by the *xl-uni-10* and *xl-uni-16* models. The *xl-uni-10* model was able to recognize 10.62% of those vowels, and *xl-uni-16* recognized only 5.51% of them. This seems much lower than these models achieve on the other seen Amharic vowels, especially the primary vowels whose recognition rates range 21.43-58-68%.

On Mongolian, the *xl-nst* baseline which has seen all 7 Mongolian reference vowels achieves a recognition rate of 12.92%, while the *xl-lnet* baseline, which has seen 6 of them, achieves 18.82%. The *xl-uni-5* model achieves the highest performance on this language because it has the highest recognition rates on all individual seen vowels. It is worth noting that the reference phonological system used to transcribe the Mongolian data does not have the mid front unrounded vowel [e], or the close-mid front unrounded vowel [e̞], only the open-mid [ɛ]. The models were able to detect this and no model produced too many confusions with [e], even the *xl-uni-5* and *xl-uni-16* which do not distinguish [e] and [ɛ]. The only *xl-uni* model that distinguishes these two vowels, *xl-uni-10*, outperformed both baselines on this particular vowel. However, it was less successful at differentiating the mid back vowels: [o] and [ɔ], on which it outperformed only one of the baselines.

On Javanese, the *xl-nst* baseline which has seen all 10 Javanese reference vowels achieves a recognition rate of 20.86%, while the *xl-lnet* baseline, which has seen 9 of them, achieves 27.44%. As before, the *xl-uni-5* model achieves the highest performance on this language because it has the highest recognition rates on all individual seen vowels compared to the other models, despite having seen only 5 of the 10 Javanese vowels. Increasing the number of vowel categories from 5 to 10 and 16 leads to higher recognition rates on the added vowels but at the expense of the initial vowel. Having more but narrower categories in the vowel set leads to more frequent vowel confusions among adjacent categories, e.g. [e] and [ɛ], or [a] and [ɑ].

Regarding the performance on the reference vowels that do not occur in the *xl-uni* models' vowel sets, we will analyze the performance on four pairs of vowels: [ɛ] and [ɔ], [ə] and [ɨ], [ɪ] and [ʊ], and [ɤ] and [ɯ]. As previously mentioned, vowel confusions frequently happen among vowels that are adjacent in the vowel space. For example, when [ɛ] and [ɔ] are not in the model's vowel set, it commonly substitutes [ɛ]

Table 14.24: Phone prediction rates of the *xl-nst* cross-lingual models on the Babel data. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Lao			Zulu			Amharic			Mongolian			Javanese		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
e	e: **41.21**	del: 12.85	ɛ: 11.90	e: **25.16**	del: 23.13	ə: 8.20	e: **30.75**	del: 21.78	i: 10.19	-	-	-	e: **29.64**	del: 19.91	ə: 8.78
i	i: **47.54**	del: 13.02	ɪ: 9.11	del: 29.35	i: **19.49**	ə: 5.52	i: **31.59**	del: 28.59	ɪ: 5.66	del: 24.96	i: **21.17**	ə: 5.13	i: **26.56**	del: 25.47	ɪ: 6.80
o	o: **30.84**	ɔ: 23.58	del: 15.15	o: **29.46**	del: 20.45	ɔ: 8.12	del: 28.09	o: **24.75**	ɑ: 9.31	del: 27.36	o: **21.63**	ə: 6.85	o: **25.94**	del: 19.21	ɑ: 10.36
u	u: **36.94**	del: 12.61	o: 8.35	u: **25.37**	del: 24.60	ɔ: 6.21	del: 28.98	u: **22.50**	ɔ: 6.38	u: **22.42**	del: 19.44	ə: 5.68	del: 24.55	u: **23.55**	o: 7.77
ɑ	ɑ: **35.65**	del: 19.96	a: 7.89	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	del: 20.97	ɔ: **19.13**	ɑ: 13.51	ɔ: **31.65**	del: 19.37	o: 9.93	-	-	-	del: 26.25	ɑ: 14.11	ɔ: **12.62**	del: 24.77	ɑ: 16.74	ɔ: **14.22**
ɛ	ɛ: **30.05**	del: 24.56	e: 12.88	-	-	-	-	-	-	del: 29.41	ɛ: **10.84**	ə: 7.79	del: 31.50	ɛ: **25.78**	ə: 5.18
ɤ	del: 25.30	ø: 9.71	ə: 9.35	-	-	-	-	-	-	-	-	-	-	-	-
ɯ	del: 26.57	ə: 8.16	u: 6.59	-	-	-	-	-	-	-	-	-	-	-	-
a	-	-	-	del: 25.08	ɑ: 19.56	a: **15.05**	del: 31.71	ɑ: 20.59	a: **14.40**	del: 29.82	ɑ: 10.83	a: **8.89**	del: 22.96	ɑ: 22.14	a: **16.61**
ə	-	-	-	-	-	-	del: 32.50	ə: **27.86**	ɑ: 6.01	-	-	-	del: 26.14	ə: **24.60**	ɑ: 9.71
ɨ	-	-	-	-	-	-	del: 40.23	ə: 7.08	ɑ: 5.65	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	-	-	-	del: 29.70	o: 9.36	ə: 6.34	del: 26.81	ɔ: 12.78	ʊ: **9.37**
ɪ	-	-	-	-	-	-	-	-	-	-	-	-	del: 26.15	ɪ: **20.86**	e: 7.83
mean			**30.68**			**21.09**			**17.59**			**12.92**			**20.63**

Table 14.25: Phone prediction rates of the *xl-lnet* cross-lingual models on the Babel data. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Lao			Zulu			Amharic			Mongolian			Javanese		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
e	e: **47.31**	i: 13.57	ɛ: 8.76	e: **36.09**	del: 16.25	ɛ: 6.58	e: **36.57**	del: 15.02	i: 14.70	-	-	-	e: **36.75**	del: 15.37	ɛ: 6.99
i	i: **65.91**	del: 6.06	e: 5.39	i: **30.44**	del: 22.31	e: 4.84	i: **42.08**	del: 19.14	e: 5.07	i: **36.28**	del: 14.38	e: 6.29	i: **39.40**	del: 19.31	e: 5.73
o	ɔ: 25.48	o: **20.20**	del: 11.02	o: **23.41**	del: 17.82	ɔ: 11.87	del: 23.11	ɔ: 17.06	o: **14.15**	del: 23.59	ɔ: 12.27	o: **10.11**	ɔ: 19.91	o: **18.18**	del: 15.24
u	u: **28.76**	ɔ: 15.03	del: 10.19	u: **25.16**	del: 19.06	ɔ: 8.84	u: **30.50**	del: 19.06	ɔ: 10.96	u: **24.27**	del: 15.58	ɔ: 7.06	u: **21.07**	del: 20.05	ɔ: 13.87
ɑ	ɑ: **17.31**	a: 17.18	del: 14.45	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	ɔ: **35.51**	del: 10.59	ɑ: 9.11	ɔ: **47.50**	del: 10.67	o: 7.72	-	-	-	ɔ: **20.91**	del: 17.02	ɑ: 12.34	ɔ: **25.96**	del: 15.98	ɑ: 11.21
ɛ	ɛ: **35.79**	del: 14.78	e: 13.83	-	-	-	-	-	-	del: 22.02	ɛ: **16.60**	e: 9.32	del: 25.24	ɛ: **24.19**	e: 9.23
ɤ	ø: 15.21	del: 14.88	e: 11.71	-	-	-	-	-	-	-	-	-	-	-	-
ɯ	del: 17.76	e: 7.82	u: 7.37	-	-	-	-	-	-	-	-	-	-	-	-
a	-	-	-	a: **32.79**	del: 15.74	ɑ: 10.55	a: **29.33**	del: 19.86	ɑ: 17.09	del: 20.19	a: **17.12**	ɑ: 8.78	a: **36.75**	del: 13.17	ɑ: 13.06
ə	-	-	-	-	-	-	del: 33.29	e: 9.91	ɑ: 7.07	-	-	-	del: 23.64	e: 11.36	ɑ: 8.25
ɨ	-	-	-	-	-	-	del: 29.54	e: 11.80	i: 5.50	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	-	-	-	del: 22.12	ɔ: 15.55	o: 5.03	del: 21.08	ɔ: 15.40	u: 9.44
ɪ	-	-	-	-	-	-	-	-	-	-	-	-	del: 23.13	i: 17.29	e: 9.94
mean	-	-	**25.27**	-	-	**31.60**	-	-	**13.88**	-	-	**18.82**	-	-	**27.44**

Table 14.26: Phone prediction rates of the *xl-uni-5* cross-lingual models on the Babel data. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Lao			Zulu			Amharic			Mongolian			Javanese		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
e	e: **68.02**	i: 16.02	del: 6.10	e: **61.75**	del: 14.99	i: 4.94	e: **59.23**	i: 16.96	del: 11.55	-	-	-	e: **62.10**	del: 13.48	a: 7.97
i	i: **75.98**	e: 7.18	del: 5.84	i: **37.90**	del: 27.37	e: 9.18	i: **57.24**	del: 17.04	e: 9.41	i: **48.83**	del: 16.76	e: 11.12	i: **48.41**	del: 20.37	e: 10.72
o	o: **59.48**	u: 13.48	del: 7.59	o: **43.49**	del: 17.04	a: 8.06	o: **37.01**	del: 27.27	a: 10.27	o: **31.47**	del: 23.10	u: 9.79	o: **47.22**	a: 16.37	del: 12.77
u	u: **58.56**	i: 7.72	del: 7.36	u: **42.77**	del: 24.37	o: 5.59	u: **39.92**	del: 22.12	o: 8.53	u: **36.95**	del: 17.68	e: 10.35	u: **42.38**	del: 20.84	o: 10.31
ɑ	a: 36.18	del: 22.37	e: 9.65	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	a: 24.96	o: 19.18	del: 18.93	o: 27.59	del: 25.93	a: 11.64	-	-	-	del: 31.30	a: 20.10	e: 8.17	a: 28.03	del: 26.81	o: 13.95
ɛ	e: 32.18	del: 26.41	a: 15.95	-	-	-	-	-	-	del: 33.22	e: 16.70	i: 12.31	del: 36.85	a: 18.32	e: 16.21
ɤ	e: 27.28	a: 19.61	del: 17.76	-	-	-	-	-	-	-	-	-	-	-	-
ɯ	e: 21.24	del: 20.40	u: 11.27	-	-	-	-	-	-	-	-	-	-	-	-
a	-	-	-	a: **75.12**	del: 10.15	e: 4.38	a: **86.89**	del: 5.66	e: 2.74	a: **63.99**	del: 11.83	e: 7.07	a: **81.91**	del: 6.11	e: 4.18
ə	-	-	-	-	-	-	del: 42.15	e: 16.05	a: 15.50	-	-	-	del: 29.85	a: 20.56	e: 17.81
ɨ	-	-	-	-	-	-	del: 35.85	e: 20.69	a: 9.77	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	-	-	-	del: 28.80	o: 16.65	a: 9.94	del: 24.97	o: 18.78	u: 12.61
ɪ	-	-	-	-	-	-	-	-	-	-	-	-	del: 28.20	e: 23.50	i: 13.47
mean	-	-	**15.97**	-	-	**52.71**	-	-	**30.82**	-	-	**32.32**	-	-	**41.32**

Table 14.27: Phone prediction rates of the *xl-uni-10* cross-lingual models on the Babel data. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Lao			Zulu			Amharic			Mongolian			Javanese		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
e	e: **58.52**	i: 11.08	del: 10.59	e: **36.21**	del: 21.26	ɛ: 12.26	e: **44.17**	del: 17.67	i: 11.77	-	-	-	e: **28.48**	del: 22.21	ɛ: 16.50
i	i: **67.33**	del: 9.00	e: 8.23	del: 30.27	i: **27.17**	e: 8.08	i: **47.24**	del: 22.93	e: 6.87	i: **36.87**	del: 22.11	e: 8.07	i: **39.56**	del: 24.05	e: 8.86
o	o: **37.61**	del: 14.72	ɔ: 11.89	o: **26.14**	del: 22.21	ɔ: 6.66	del: 30.23	o: **21.37**	ɑ: 9.90	del: 29.93	o: **18.59**	ə: 5.65	o: **20.29**	del: 19.78	ɑ: 12.88
u	u: **41.99**	del: 11.22	o: 7.89	u: **29.78**	del: 25.39	o: 5.70	del: 31.02	u: **21.43**	o: 6.69	u: **25.31**	del: 20.76	e: 7.26	u: **29.86**	del: 23.93	o: 8.94
ɑ	ɑ: **22.31**	a: 21.71	del: 21.13	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	ɔ: **27.52**	del: 18.86	ɑ: 13.88	ɔ: **33.90**	del: 17.63	o: 10.77	-	-	-	del: 28.15	ɔ: **14.56**	ɑ: 13.04	del: 23.70	ɔ: **19.09**	ɑ: 16.58
ɛ	ɛ: **52.04**	del: 16.64	e: 8.80	-	-	-	-	-	-	del: 27.69	ɛ: **20.63**	e: 11.13	ɛ: **36.84**	del: 27.45	a: 9.75
ɤ	del: 21.49	ɛ: 18.47	ə: 13.15	-	-	-	-	-	-	-	-	-	-	-	-
ɯ	del: 24.68	e: 10.96	ɨ: 8.62	-	-	-	-	-	-	-	-	-	-	-	-
a	-	-	-	a: **47.57**	del: 16.73	ɑ: 11.23	a: **55.50**	del: 17.00	ɑ: 11.84	a: **37.73**	del: 22.15	ɑ: 6.74	a: **52.58**	del: 13.73	ɑ: 13.56
ə	-	-	-	-	-	-	del: 40.23	ə: **11.93**	ɛ: 11.40	-	-	-	del: 30.92	ə: **12.71**	ɛ: 9.99
ɨ	-	-	-	-	-	-	del: 36.90	ɨ: **10.62**	ɛ: 9.69	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	-	-	-	del: 32.14	o: 8.43	ɔ: 7.13	del: 27.68	o: 10.59	u: 7.73
ɪ	-	-	-	-	-	-	-	-	-	-	-	-	del: 32.15	e: 13.27	ɛ: 10.71
mean	-	-	**29.75**	-	-	**35.97**	-	-	**26.49**	-	-	**26.35**	-	-	**32.47**

Table 14.28: Phone prediction rates of the *xl-uni-16* cross-lingual models on the Babel data. The table shows top 3 phone predictions and their prediction rates in % for each reference vowel in the five evaluation languages. The prediction rates are the average over the three experiment runs for each model. *del* signifies a deletion error, while the hyphen (-) indicates that the reference vowel does not occur in the given evaluation language. Bolded results are correct predictions. The bottom row provides the mean recognition rate across all reference vowels in a given language. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	Lao			Zulu			Amharic			Mongolian			Javanese		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
e	e: **55.82**	i: 18.71	del: 9.58	e: **44.96**	del: 20.75	i: 5.00	e: **46.02**	i: 18.41	del: 17.35	-	-	-	e: **47.15**	del: 18.65	i: 6.13
i	i: **73.25**	del: 7.61	e: 4.92	i: **32.54**	del: 29.73	e: 5.27	i: **50.98**	del: 22.04	e: 5.25	i: **42.87**	del: 21.55	e: 6.17	i: **45.18**	del: 22.32	e: 6.69
o	o: **49.52**	del: 13.54	u: 10.91	o: **32.58**	del: 21.59	ɑ: 5.54	del: 29.29	o: **28.89**	ɑ: 7.98	del: 30.63	o: **20.86**	u: 5.59	o: **33.00**	del: 18.58	ɑ: 10.45
u	u: **46.36**	del: 10.94	o: 5.98	u: **33.53**	del: 26.09	o: 5.14	del: 29.09	u: **25.32**	o: 5.92	u: **29.06**	del: 21.05	e: 5.70	u: **33.76**	del: 23.48	o: 8.35
ɑ	ɑ: **22.53**	a: 22.01	del: 21.91	-	-	-	-	-	-	-	-	-	-	-	-
ɔ	del: 25.53	ɑ: 15.18	o: 13.50	del: 27.28	o: 17.48	ɑ: 8.11	-	-	-	del: 34.45	ɑ: 13.44	a: 4.78	del: 30.62	ɑ: 15.66	a: 8.91
ɛ	del: 31.87	e: 25.26	a: 13.25	-	-	-	-	-	-	del: 38.24	e: 10.00	i: 9.70	del: 40.98	a: 12.57	e: 10.05
ɤ	del: 22.87	ɵ: 10.81	a: 9.64	-	-	-	-	-	-	-	-	-	-	-	-
ɯ	del: 24.18	e: 7.59	ʉ: 6.27	-	-	-	-	-	-	-	-	-	-	-	-
a	-	-	-	a: **49.71**	del: 17.17	ɑ: 10.73	a: **58.68**	del: 16.80	ɑ: 10.92	a: **39.41**	del: 21.79	ɑ: 6.31	a: **54.13**	del: 13.84	ɑ: 12.70
ə	-	-	-	-	-	-	del: 41.20	ə: **14.14**	a: 9.24	-	-	-	del: 30.99	ə: **13.41**	a: 9.72
ɨ	-	-	-	-	-	-	del: 40.75	e: 8.96	i: 5.64	-	-	-	-	-	-
ʊ	-	-	-	-	-	-	-	-	-	del: 33.61	o: 12.52	ɑ: 6.33	del: 28.52	o: 13.29	u: 9.90
ɪ	-	-	-	-	-	-	-	-	-	-	-	-	del: 32.43	e: 16.80	i: 11.99
mean	-	-	**25.03**	-	-	**38.36**	-	-	**27.60**	-	-	**22.28**	-	-	**31.31**

with [e] and [a], and [ɔ] with [o] and [ɑ]. When [ə] and [ɨ] are not in the model's vowel set, it often replaces them with [e] or [a]. When [ɪ] and [ʊ] are not in the model's vowel set, it tends to replace [ɪ] with [i], [e], or [ɛ], and [ʊ] with [o] or [u]. Finally, when [ɤ] and [ɯ] are excluded from the model's vowel set, [ɤ] is typically replaced with [e], [ɛ], [a], or [ə], and [ɯ] with [e], [u], or [ɨ].

Finally, a general trend toward high vowel deletion rates was observed on all Babel languages with all models. The deletion rates of all vowels in the Babel languages are generally higher than they were on the parliamentary corpora in Indo-European languages, and are on par with the vowel deletion rates on Finnish. We have not carried out a systematic analysis of deletions, so it is unclear what is causing them. We expect that finding ways to curb the deletion rates would help improve cross-lingual phone recognition on these data sets.

14.4.4 Inferring the Vowel Inventory of an Unseen Language

In the previous sections, we saw that the *xl-uni-5* model had the best performance on most of the evaluation languages in terms of vowel recognition rate. However, we also know that this model has the smallest vowel set and broadest vowel categories. We have also seen that, for many languages, in particular, those with larger vowel sets, these five vowel categories do not allow us to distinguish all their contrastive vowel qualities. Therefore, it is also worth asking which cross-lingual models are capable of detecting the relevant vowel contrasts in a given target language. Regrettably, we cannot answer this question conclusively yet, as we have investigated only a very small number of languages with differing vowel systems. We will, however, describe and visualize some qualitative observations on several of the investigated languages. They will exemplify how a high correspondence between the model's vowel system and the vowel system of the target language help the model to infer the vowel inventory of the language.

Namely, we will look at diagrams called *dendrograms*, which result from the hierarchical clustering of the cross-lingual phone predictions on individual parliamentary and Babel corpora. We analyze the phone prediction dendrograms of the *xl-uni* models with vowel sets

most similar to the vowel system of a target language, and compare them to the dendrograms of the baselines. They are shown in Figures 14.3-14.5. The clusters in the diagrams can be interpreted as groups of phones that tend to have similar patterns of confusions. The representations clustered lower on the y-axis have more similarity, while the representations that branch off higher are more distant. This visualization technique for phone prediction analysis is based on the work by Żelasko et al. (2022).

We first look at the dendrograms of the baselines and the *xl-uni-5* model on Serbian, shown in Figure 14.3. We see that the *xl-uni-5* model whose 5-category vowel system closely matches the vowel system of Serbian is able to detect all 5 of the Serbian reference vowels, as well as make a greater distinction between the predicted vowel categories. In contrast, the categories predicted by the baselines, which have more vowels and higher vowel error rates, do not seem as distinct from each other.

Next, we first look at the dendrograms of the baselines and the *xl-uni-10* model on Catalan, depicted in Figure 14.4. Now, we see that the *xl-uni-10* model with a 10-category vowel set is a closer match for the vowel system of Catalan, which has 9 reference vowels. It is able to detect and distinguish 8 of the 9 Catalan vowels (all except [ʊ], which is not in its vowel set). On the other hand, the *xl-nst* cannot differentiate between [e] and [ɛ], or [o] and [ɔ] at the same clustering threshold. It also detects an additional vowel, vowel [ɪ], which is outside of Catalan's vowel inventory. Unlike the *xl-nst*, the *xl-lnet* can distinguish between the open-mid and close-mid vowels, but fails to distinguish the vowel [ə].

We will now look at the dendrograms of the baselines and the *xl-uni-16* model on Finnish, displayed in Figure 14.5. Here, we see that the *xl-uni-16* model with a 16-category vowel set is better at distinguishing Finnish vowels than the baselines. It is able to detect and distinguish all 8 of the Finnish vowels (if [a] is considered close enough to the Finnish [æ]). On the other hand, the *xl-nst* cannot differentiate between [i] and [y], at the same clustering threshold, and generally finds more similarity between the predicted phones. It also detects an additional vowel, vowel [ʉ], which is not in Finnish's vowel inventory. The *xl-lnet* baseline does not distinguish the vowel [æ],

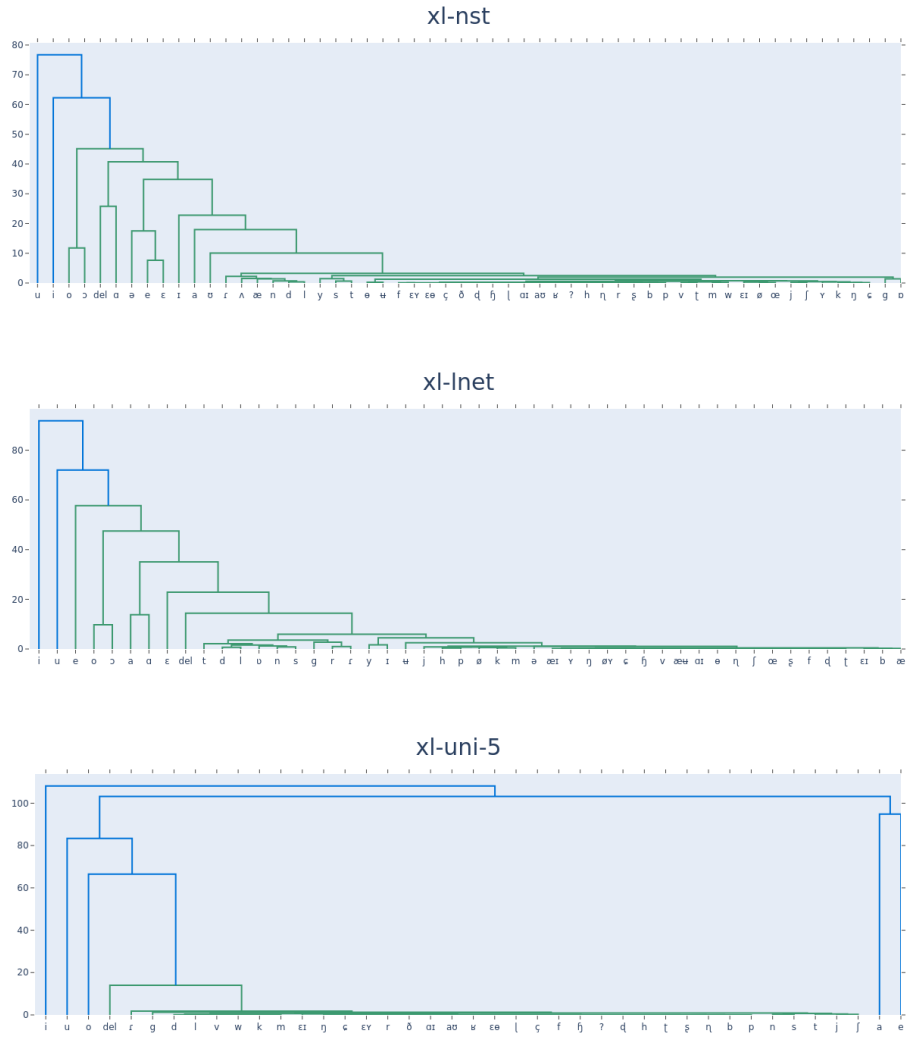


Figure 14.3: Hierarchical clustering of the cross-lingual phone predictions on the Serbian parliament corpus, visualized as dendrograms. The clustering threshold for all subfigures is set to 60.

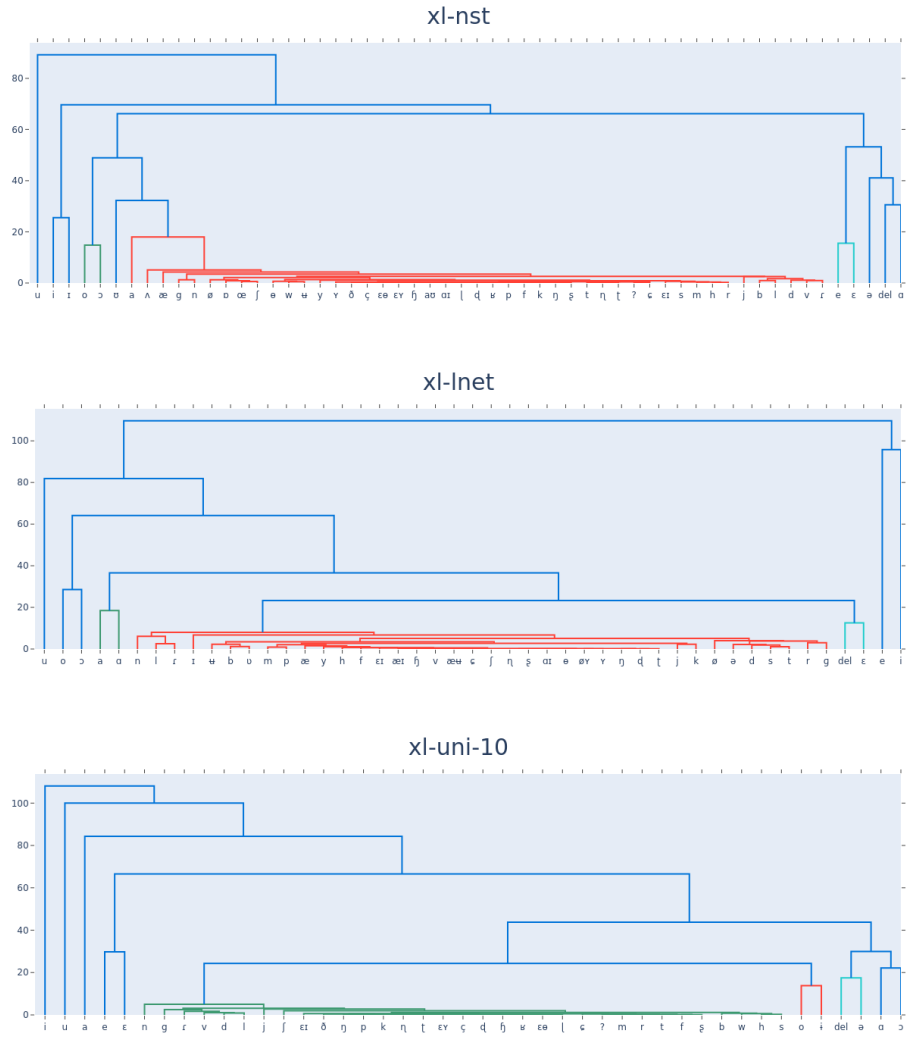


Figure 14.4: Hierarchical clustering of the cross-lingual phone predictions on the Catalan parliament corpus, visualized as dendrograms. The clustering threshold for all subfigures is set to 20.

and is generally worse at separating the relevant from irrelevant vowel categories.

On the other hand, if we look at the dendrograms of phone predictions on Lao, shown in Figure 14.6, we can see that none of our models is able to distinguish all of its 9 vowels. As we remember, Lao has two rare vowels, mid back unrounded [ɤ] and close back unrounded [ɯ]. The only model that saw these two representations was the *xl-uni-16*, which means that it was the only model that could predict them. However, these vowels are not distinctive in the languages of our fine-tuning corpus and were thus quite rare in the *xl-uni-16* transcriptions. As a result, this model was not good at recognizing them nor distinguishing them as relevant. Additionally, this model could also not predict the open-mid vowels [ɛ] and [ɔ] as they were not in its vowel set. From the remaining three models shown in the figure, the *xl-uni-10* and *xl-lnet* could distinguish both of them, while the *xl-nst* had a harder time distinguishing [o] and [ɔ]. Still, none of them could predict the unrounded back vowels [ɤ] and [ɯ]. Hence, we see that none of the investigated vowel categorization methods was appropriate for Lao, which is a likely reason that none of the models was particularly good at recognizing its vowels.

As we can see from these examples, when the model's vowel system and training data align with the vowel system of the target language, the model is better at inferring the vowel inventory of this language. This does not prove that these models are better at recognizing reference vowels, but it shows that they can make more intuitive and phonologically relevant vowel predictions.

14.5 Phone Prediction Analysis of Monolingual Speech Recognition Results

In the previous section, we looked at individual vowel prediction rates of the phone recognition models. Here, we are going to examine individual phone prediction rates of the word-based ASR systems. However, since these systems predict whole words and not individual phones, we cannot directly measure their phone error rates. Therefore, we compute the top phone predictions and their prediction rates for

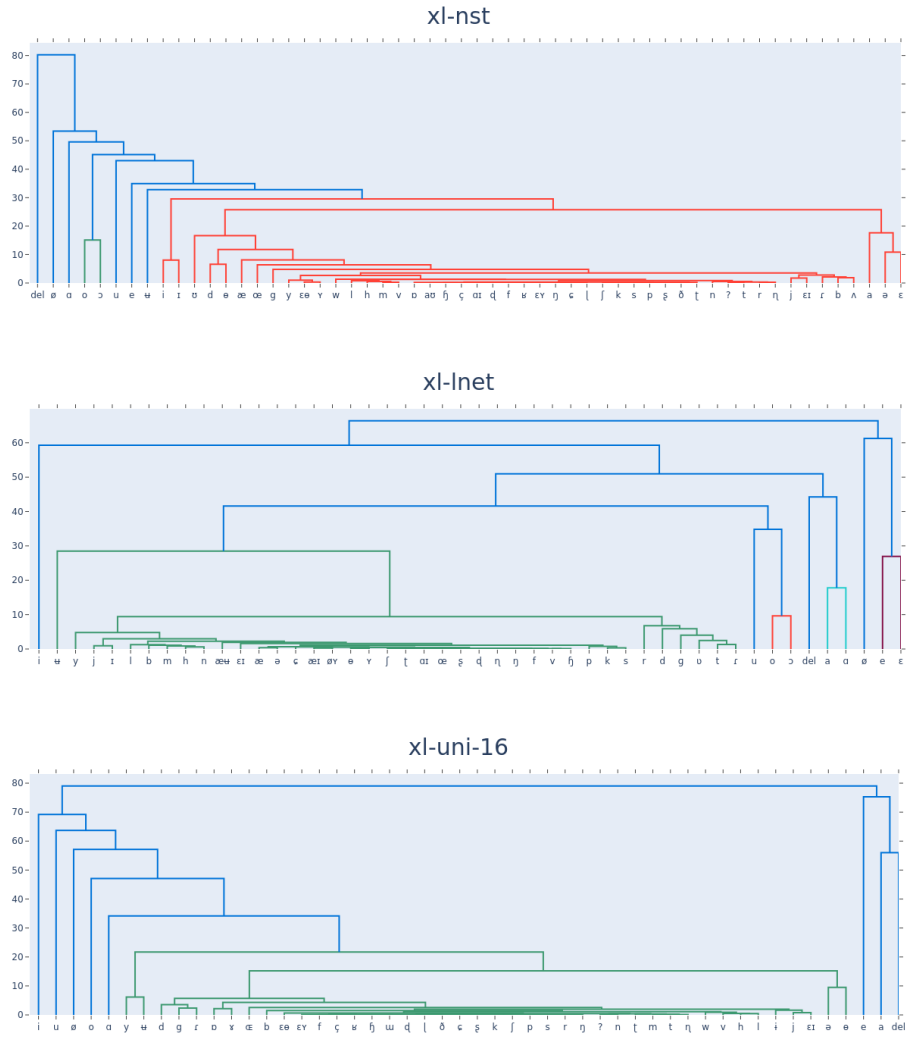


Figure 14.5: Hierarchical clustering of the cross-lingual phone predictions on the Finnish parliament corpus, visualized as dendrograms. The clustering threshold for all subfigures is set to 30.

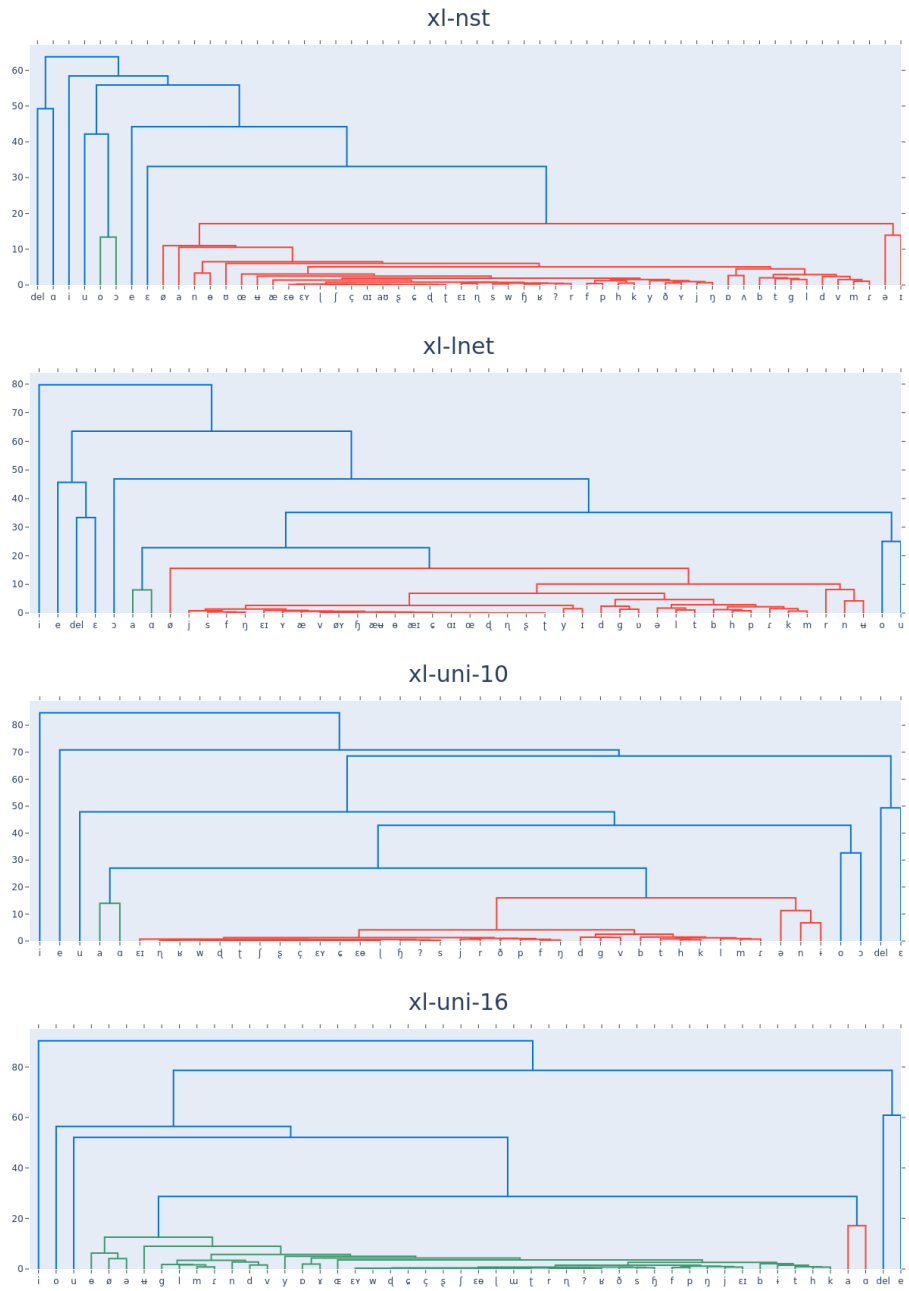


Figure 14.6: Hierarchical clustering of the cross-lingual phone predictions on the Lao data set from the telephone speech corpus Babel, visualized as dendrograms. The clustering threshold for all subfigures is set to 20.

each reference vowel in the test sets of the evaluation corpora based on their G2P-predicted pronunciation from the cross-lingual lexicon. We then compare the phone prediction rates for each reference vowel to its word recognition rate.

The prediction rates of a particular vowel are obtained by counting the number of its tokens in the test set(s) according to the original (canonical) pronunciations, and calculating what percentage of those vowels is predicted as which phone in the given cross-lingual pronunciation lexicon. The word recognition rates of a particular vowel are measured by counting the number of word tokens in the test set(s) that contain this vowel in their original pronunciation and calculating the percentage of them that was correctly recognized by the ASR system. Both prediction rates and word recognition rates are averaged across all experiment runs and evaluation languages. The evaluation languages include both parliamentary and Babel data sets, but exclude Danish as this language is seen by the *xl* phone recognition models and would bias the results. These results are shown in Table 14.29a-c. Although these results do not show a direct relationship between the predicted phones and the word recognition rate, they should tell us how phone predictions for an individual reference vowel correlate with the system's performance on words containing the same vowel.

As we see from the table, the phone predictions for different vowels vary by cross-lingual lexicon and are roughly in line with the predictions of the phone recognition models. However, the word recognition rates look very similar across the different ASR systems despite them using different pronunciation lexicons. Looking at the five most common vowels: [a, e, i, o, u], which are found in all five cross-lingual lexicons and whose recognition rate improves the most over the baselines, we see that their corresponding word recognition rates are relatively stable across the different ASR systems. Moreover, larger improvements in vowel recognition, such as those seen with the *xl-uni-5* models on the vowels [a] and [o], are even found to correlate with worse performance on word recognition. When it comes to the front rounded vowels: [y, ø, œ], they were also associated with similar word recognition rates across the ASR systems, with [y] and [ø] having two of the three highest recognition rates, even when they were outside the system's lexicon. Finally, the results for the rarest vowels:

Table 14.29a: Phone prediction rates for each cross-lingual lexicon for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	top vowel predictions					% correct
	1	2	3	4	5	words
xl-nst						
a	del: 48.79	ɑ: 20.88	ə: 9.14	a: **3.07**	ɛ: 1.99	33.64
e	del: 37.87	e: **21.90**	ə: 15.31	ɛ: 8.63	ɑ: 2.02	45.36
i	del: 44.43	i: **20.06**	ɪ: 12.33	ə: 3.30	ɑ: 3.10	41.17
o	del: 35.11	ɔ: 23.99	o: **18.04**	ɑ: 5.39	ə: 2.28	45.58
u	del: 47.78	u: **20.16**	ʊ: 6.30	ɔ: 4.54	ɑ: 2.80	35.47
y	del: 37.24	ʉ: 21.12	ɵ: 15.29	d: 6.41	ə: 4.35	77.17
æ	del: 37.87	ə: 26.23	æ: **7.22**	ɛ: 6.75	a: 4.13	77.83
ø	ø: **37.48**	del: 23.27	ə: 11.14	ɔ: 8.25	œ: 5.67	80.82
œ	del: 38.34	œ: **21.30**	ø: 14.90	ɔ: 5.48	k: 2.66	47.76
ɑ	del: 59.44	ɑ: **21.05**	a: 3.90	ɛ: 1.87	ə: 1.42	5.95
ɔ	del: 55.83	ɑ: 16.06	ɔ: **11.13**	o: 2.59	ə: 1.37	10.53
ə	del: 52.72	ə: **21.93**	ɑ: 8.48	e: 3.11	ɛ: 3.04	20.55
ɛ	del: 53.01	ɛ: **17.88**	e: 8.90	ə: 4.55	æ: 1.70	23.10
ɤ	del: 42.28	ø: 27.86	ɔ: 9.44	a: 4.34	ə: 3.66	5.45
ɨ	del: 79.63	ə: 3.25	ɑ: 2.19	ɛ: 1.90	e: 1.15	5.79
ɪ	del: 31.69	ə: 19.83	ɪ: **18.43**	e: 9.97	i: 5.33	38.84
ɯ	del: 66.01	u: 4.43	ə: 4.24	ɔ: 4.13	ɪ: 2.99	6.00
ʊ	del: 41.55	ʊ: **15.44**	ɔ: 15.07	u: 11.22	o: 4.03	38.94
ʏ	del: 33.93	ə: 22.47	ɵ: 12.73	ø: 4.39	ɾ: 2.91	44.94
xl-lnet						
a	del: 34.59	ɑ: 25.78	a: **13.06**	e: 4.99	ɛ: 2.89	35.39
e	e: **33.75**	del: 23.42	ɛ: 14.86	ə: 3.76	i: 3.70	45.25
i	i: **41.31**	del: 28.73	ɪ: 5.00	e: 2.80	t: 1.98	42.81
o	ɔ: 29.27	del: 26.36	o: **14.21**	ɑ: 7.78	u: 2.99	47.71
u	del: 37.30	u: **20.24**	ɔ: 16.25	o: 4.66	ɑ: 2.20	37.27
y	ʉ: 43.94	del: 27.75	d: 5.37	u: 3.31	y: **1.65**	80.06
æ	del: 29.73	ɛ: 18.27	e: 16.38	ɑ: 10.43	a: 7.33	80.92
ø	ø: **54.55**	del: 10.52	e: 7.23	ʉ: 5.61	ɛ: 2.75	83.08
œ	ø: 38.88	del: 26.62	ɔ: 10.37	ɑ: 3.04	o: 2.61	55.84
ɑ	del: 36.37	a: 18.82	ɑ: **10.63**	e: 10.16	g: 3.91	7.04
ɔ	del: 40.88	ɑ: 16.23	ɔ: **14.72**	a: 6.82	o: 5.39	11.59
ə	del: 47.53	e: 17.06	ɑ: 10.10	ɛ: 3.76	a: 2.41	22.70
ɛ	del: 38.16	e: 27.76	ɛ: **8.53**	i: 2.22	æ: 2.21	24.60
ɤ	ø: 44.82	del: 20.66	e: 10.36	p: 5.56	ʉ: 4.18	5.27
ɨ	del: 69.43	e: 10.09	ø: 2.22	ɑ: 2.08	i: 1.31	8.25
ɪ	i: 31.78	del: 19.87	e: 18.98	ɪ: **3.71**	ɾ: 3.05	44.72
ɯ	del: 41.90	ɔ: 10.72	ʉ: 9.02	e: 8.89	ø: 6.04	6.61
ʊ	del: 29.26	o: 20.40	u: 16.73	ɔ: 15.27	ʉ: 1.96	40.12
ʏ	ʉ: 23.78	del: 21.61	e: 17.03	u: 10.94	ø: 3.81	53.49

Continued on next page.

Table 14.29b: Phone prediction rates for each cross-lingual lexicon for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

Continued from previous page.

	top vowel predictions					% correct
	1	2	3	4	5	words
xl-uni-5						
a	a: **55.73**	del: 31.09	e: 4.67	o: 1.61	s: 0.75	34.35
e	e: **59.74**	del: 24.63	a: 4.77	i: 3.99	o: 1.20	45.65
i	i: **42.59**	del: 37.50	e: 7.00	a: 2.27	ɛɪ: 1.64	41.54
o	o: **43.82**	del: 28.88	a: 9.27	u: 5.88	e: 3.85	46.58
u	del: 40.95	u: **39.04**	o: 4.44	e: 3.54	a: 2.87	36.59
y	del: 35.93	i: 30.00	e: 17.74	d: 3.96	g: 3.53	77.21
æ	a: 40.30	del: 29.04	e: 17.28	d: 5.67	b: 1.74	78.62
ø	e: 57.63	del: 21.32	d: 5.67	a: 5.02	i: 3.37	81.15
œ	del: 40.92	e: 19.98	a: 16.59	o: 12.14	k: 1.67	47.03
ɑ	del: 59.34	a: 22.13	e: 3.05	v: 2.21	g: 1.76	5.63
ɔ	del: 54.90	a: 19.43	o: 9.73	e: 2.68	u: 2.11	10.75
ə	del: 49.66	e: 20.75	a: 17.49	i: 1.60	d: 1.34	22.33
ɛ	del: 53.89	e: 22.48	a: 9.12	d: 1.87	i: 1.47	22.66
ɤ	e: 52.49	del: 27.78	a: 7.64	o: 5.11	n: 1.47	4.90
ɨ	del: 73.27	e: 8.31	a: 5.78	i: 2.70	d: 0.98	7.31
ɪ	del: 30.41	e: 28.51	i: 24.16	r: 4.61	a: 2.25	39.81
ɯ	del: 57.87	e: 15.51	u: 9.89	a: 4.56	i: 4.35	5.81
ʊ	del: 37.65	u: 31.78	o: 11.42	a: 4.01	e: 3.79	39.47
ʏ	e: 43.68	del: 30.50	u: 5.14	o: 4.80	i: 3.38	45.64
xl-uni-10						
a	del: 39.16	a: **29.05**	ɑ: 14.43	ɛ: 3.38	ə: 2.38	34.42
e	del: 32.17	ɛ: 25.73	e: **22.29**	a: 3.90	i: 3.24	45.75
i	del: 38.29	i: **35.55**	e: 8.53	ɛ: 2.17	ɑ: 1.52	42.08
o	del: 32.93	ɔ: 19.60	o: **16.69**	ə: 7.17	ɑ: 6.53	46.71
u	del: 42.34	u: **27.93**	o: 6.92	ɨ: 3.66	ə: 2.37	36.67
y	del: 36.16	ɨ: 24.59	e: 15.61	i: 4.18	g: 3.57	77.87
æ	a: 37.48	del: 36.93	ɛ: 9.21	d: 6.05	b: 1.76	78.81
ø	del: 36.71	ɛ: 21.41	ɨ: 12.05	ə: 11.82	e: 4.63	81.10
œ	del: 43.59	ə: 20.84	ɛ: 8.07	ɑ: 6.57	a: 3.70	48.65
ɑ	del: 59.62	a: 16.91	ɑ: **7.90**	ɛ: 1.63	d: 1.48	5.70
ɔ	del: 51.36	ɔ: **11.24**	ɑ: 10.38	a: 8.68	o: 4.29	11.02
ə	del: 51.54	ɛ: 13.72	a: 11.24	ɑ: 4.17	e: 4.11	21.89
ɛ	del: 48.54	ɛ: **25.31**	a: 7.34	e: 6.15	ɑ: 0.99	23.05
ɤ	ɛ: 44.60	del: 34.63	ə: 7.09	i: 2.28	a: 2.25	4.65
ɨ	del: 74.50	ɛ: 5.84	a: 2.95	ɑ: 2.62	ɨ: **1.95**	6.94
ɪ	e: 37.34	del: 32.06	i: 9.87	ɛ: 3.89	r: 2.63	40.29
ɯ	del: 63.44	o: 5.43	e: 5.17	ɛ: 4.66	ɨ: 3.90	5.97
ʊ	del: 40.40	u: 21.99	o: 9.79	ɨ: 7.50	ɔ: 4.13	39.07
ʏ	del: 33.04	ə: 20.69	ɨ: 14.37	e: 10.35	ɛ: 5.79	46.65

Continued on next page.

Table 14.29c: Phone prediction rates for each cross-lingual lexicon for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

Continued from previous page.

	top vowel predictions					% correct
	1	2	3	4	5	words
xl-uni-16						
a	a: **38.94**	del: 38.16	ɑ: 8.93	ə: 2.98	e: 2.05	34.52
e	e: **46.56**	del: 32.83	i: 4.39	a: 4.24	ə: 3.10	46.18
i	i: **41.08**	del: 39.69	e: 4.07	ɛɪ: 2.04	ɑ: 1.49	41.93
o	del: 37.20	o: **28.99**	u: 4.51	ɑ: 4.26	ɒ: 4.11	46.76
u	del: 42.54	u: **34.95**	o: 3.79	ʉ: 2.22	ɑ: 1.87	36.82
y	del: 33.33	y: **22.97**	ʉ: 11.29	ø: 11.15	d: 4.86	78.96
æ	a: 43.31	del: 38.28	e: 4.34	d: 3.91	ə: 2.16	79.97
ø	ø: **58.12**	del: 15.54	ə: 8.88	ɵ: 3.93	e: 3.87	81.85
œ	del: 51.41	ø: 11.05	ɵ: 8.80	œ: 5.92	ə: 3.88	48.89
ɑ	del: 61.32	a: 17.03	ɑ: **7.71**	n: 2.69	e: 1.42	5.67
ɔ	del: 59.64	a: 10.16	ɑ: 8.68	o: 5.38	n: 1.54	10.87
ə	del: 54.18	a: 17.95	e: 8.18	ə: **6.64**	ɑ: 1.58	21.45
ɛ	del: 58.54	e: 14.82	a: 10.38	i: 2.03	ə: 1.53	23.01
ɤ	del: 47.13	œ: 25.74	a: 6.37	ɵ: 5.59	n: 4.89	4.99
ɨ	del: 80.02	a: 3.55	e: 1.91	ɑ: 1.32	d: 1.25	6.96
ɪ	del: 33.96	i: 28.46	e: 19.26	ɾ: 3.87	ə: 1.43	41.19
ɯ	del: 69.81	u: 3.68	a: 3.14	o: 3.05	e: 2.58	6.00
ʊ	del: 44.09	u: 25.44	o: 8.65	ʉ: 4.73	a: 1.83	39.16
ʏ	del: 37.52	ø: 18.17	ə: 8.55	ɵ: 6.95	e: 4.84	47.12

[ɤ, ɯ], which were present in only the *xl-uni-16* lexicon, and [ɨ] in the *xl-uni-10* and *xl-uni-16*, were found to correlate with some of the poorest word recognition rates (4.65%-8.25%). To an extent, this stems from the fact that each of these vowels were part of only one of the evaluation languages: [ɤ, ɯ] of Lao, and [ɨ] of Amharic, which generally had worse word error rates than the parliamentary corpora.

To see how phone predictions are distributed within correctly and incorrectly predicted words, we can separate the overall prediction rates into prediction rates for the correctly recognized words and the incorrectly predicted words only, as shown in Table 14.30a-c. The left side of the table shows top 5 predictions and their prediction rates for the reference vowels in the correctly recognized words, whereas the right side shows top 5 predictions and their prediction rates for the

reference vowels in the incorrectly predicted words. Concretely, for each ASR system with a cross-lingual lexicon derived from an *xl* phone recognition model and each reference vowel, we first find all correctly recognized and all incorrectly predicted words that feature the reference vowel. Then, we use the cross-lingual lexicon used by the ASR system to obtain all predicted vowels for that reference vowel and measure their prediction rates. In other words, for each reference vowel, the prediction rates over the correct words tell us what percentage of correctly recognized words which feature that vowel in the reference lexicon was predicted as which phone in the given cross-lingual lexicon. Conversely, the prediction rates over the incorrect words tell us what percentage of incorrectly predicted words which feature a given vowel in the reference lexicon was predicted as which phone in its corresponding cross-lingual lexicon. Granted, these results do not show a direct relationship between the predicted phones and the predicted words. Nevertheless, intuitively and on average, they should tell us which phone predictions help the ASR systems the most and which ones are likely to make them worse.

We first look at the prediction rates for the incorrectly predicted words to see which prediction patterns are most likely to aggravate the ASR systems' performance. As we can see from the table, the top result for all ASR systems and across all reference vowels is a deletion error with high deletion rates (over 50%). This indicates that vowel deletion errors adversely affect the ASR systems' performance making them more likely to incorrectly predict the whole word. However, we can also see that correctly recognizing a reference vowel does not guarantee that the whole word will be recognized. The systems using the *xl-lnet* lexicon have a noticeably lower vowel deletion rate on average compared to the other ones. Namely, the average deletion rate with the *xl-lnet* lexicons is 54.2%, while with the other lexicons they are 67.23% with *xl-nst*, 61.7% with *xl-uni-5*, 63.94% with *xl-uni-10*, and 65.27% with *xl-uni-16*. As we noted before, the *xl-lnet* phone recognition model had lower phone deletion rates, which likely resulted in the wider vocabulary coverage of the *xl-lnet* lexicons and, ultimately, lower word error rates of their corresponding ASR systems. Here, we see further evidence that phone deletion errors lead to higher word error rates.

Table 14.30a: Phone prediction rates for each cross-lingual lexicon in correctly recognized and incorrectly predicted words for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

	correctly recognized words					incorrectly predicted words				
	1	2	3	4	5	1	2	3	4	5
xl-nst										
a	ɑː 33.17	del: 30.38	əː 15.07	a: **3.04**	ɛː 2.79	del: 68.20	ɑː 7.93	a: **3.10**	əː 2.88	nː 1.52
e	eː **30.91**	əː 22.03	del: 21.64	ɛː 12.09	ɪː 1.48	del: 63.02	eː **7.95**	əː 4.90	ɛː 3.26	ɑː 3.04
i	iː **29.88**	del: 26.29	ɪː 19.76	əː 4.39	ɑː 3.38	del: 67.51	iː **7.57**	ɪː 2.87	ɑː 2.75	əː 1.92
o	ɔː 33.91	oː **23.07**	del: 20.21	ɑː 6.15	əː 2.46	del: 62.66	oː **8.74**	ɔː 5.64	ɑː 3.97	əː 1.97
u	uː **34.73**	del: 24.77	ʊː 12.02	ɔː 7.31	oː 3.90	del: 68.14	uː **7.27**	ɑː 3.06	ɔː 2.10	oː 1.64
y	del: 34.59	ʉː 22.66	ɵː 16.55	d: 6.91	əː 4.44	del: 53.73	ʉː 11.48	ɵː 7.41	əː 3.73	d: 3.31
æ	del: 32.55	əː 29.78	æː **8.11**	ɛː 7.47	d: 4.43	del: 62.78	əː 9.61	ɛː 3.41	a: 3.09	æː **3.02**
ø	øː **42.60**	del: 15.85	əː 12.59	ɔː 9.58	œː 6.39	del: 60.34	øː **11.93**	əː 3.85	ɛː 2.45	œː 2.08
œ	del: 29.82	œː **26.14**	øː 18.97	ɔː 6.96	k: 3.42	del: 61.07	œː **8.38**	øː 4.04	ɑː 2.85	əː 2.41
ɑ	ɑː **49.20**	del: 19.33	a: 10.68	æː 3.54	əː 3.26	del: 71.63	ɑː **12.49**	a: 1.84	nː 1.56	ɛː 1.46
ɔ	ɑː 37.91	ɔː **25.05**	del: 21.77	oː 4.89	ʌː 2.27	del: 72.55	ɑː 5.33	ɔː **4.30**	oː 1.45	nː 1.41
ə	əː **40.46**	del: 25.32	ɑː 14.98	eː 5.80	ɛː 5.02	del: 73.71	əː **7.73**	ɑː 3.51	ɛː 1.52	eː 1.05
ɛ	ɛː **36.92**	del: 20.19	eː 19.20	əː 8.21	æː 3.67	del: 75.76	ɛː **4.69**	ɑː 2.04	əː 2.01	eː 1.76
ɤ	øː 52.82	ɔː 19.34	a: 7.72	əː 7.16	del: 6.09	del: 73.44	øː 6.37	ɑː 3.39	nː 2.05	a: 1.43
ɨ	del: 65.38	əː 11.20	ɛː 4.59	eː 3.34	œː 2.77	del: 83.42	ɑː 2.52	ɛː 1.18	əː 1.13	nː 0.84
ɪ	əː 26.27	ɪː **24.50**	del: 18.27	eː 13.31	iː 6.87	del: 58.49	əː 6.98	ɪː **6.31**	eː 3.30	ɾː 2.90
ɯ	del: 39.18	ɔː 17.73	əː 17.37	ɪː 12.64	nː 2.79	del: 72.34	uː 4.89	ɑː 3.30	øː 2.94	nː 2.25
ʊ	ʊː **26.17**	ɔː 25.09	uː 18.91	del: 16.74	oː 6.11	del: 71.49	ɔː 2.99	ɑː 2.71	ʊː **2.49**	əː 2.14
ʏ	əː 28.42	del: 22.91	ɵː 16.47	øː 5.96	ɔː 3.33	del: 57.02	əː 9.98	ɵː 4.89	ɾː 3.83	dː 1.84
xl-lnet										
a	ɑː 36.92	del: 21.22	a: **14.85**	eː 5.85	ɛː 3.82	del: 55.28	a: **10.29**	ɑː 8.52	eː 3.66	nː 2.02
e	eː **41.75**	ɛː 19.04	del: 11.86	əː 5.00	iː 3.77	del: 50.56	eː **14.98**	ɛː 5.07	iː 3.53	nː 2.62
i	iː **53.63**	del: 15.91	ɪː 7.11	eː 2.58	ɛː 1.91	del: 53.22	iː **17.77**	eː 3.22	nː 2.28	tː 2.23
o	ɔː 35.80	del: 17.40	oː **16.60**	ɑː 9.29	uː 3.42	del: 50.09	ɔː 11.96	oː **7.87**	ɑː 3.79	nː 2.37
u	uː **29.52**	ɔː 24.43	del: 20.39	oː 6.63	ʉː 2.74	del: 58.31	uː **8.71**	ɔː 6.08	ɑː 2.33	oː 2.21
y	ʉː 46.84	del: 26.20	dː 5.45	uː 2.88	ʉː 1.73	del: 40.47	ʉː 20.25	uː 6.84	dː 4.77	eː 4.20
æ	del: 28.32	ɛː 19.30	eː 17.04	ɑː 10.73	a: 7.64	del: 41.31	eː 10.95	ɛː 9.77	ɑː 8.00	a: 4.73
ø	øː **57.59**	del: 7.49	eː 7.36	ʉː 5.95	ɛː 2.79	del: 35.55	øː **29.47**	eː 6.08	dː 3.78	ʉː 2.75
œ	øː 43.88	del: 22.06	ɔː 11.69	oː 2.79	ɑː 2.78	del: 48.70	øː 14.63	ɑː 4.31	ɔː 3.99	eː 3.55
ɑ	a: 29.89	eː 15.80	del: 15.69	ɑː **13.65**	gː 5.82	del: 54.09	a: 9.33	ɑː **8.04**	eː 5.33	nː 3.04
ɔ	ɑː 29.25	ɔː **21.28**	del: 16.27	a: 11.96	oː 9.81	del: 59.60	ɔː **9.74**	ɑː 6.33	a: 2.91	eː 2.40
ə	del: 30.91	eː 27.31	ɑː 15.15	ɛː 5.65	lː 3.04	del: 66.00	eː 5.67	ɑː 4.48	a: 2.07	ɛː 1.67
ɛ	eː 44.95	del: 15.66	ɛː **12.18**	æː 3.79	ʊː 2.71	del: 65.69	eː 6.70	ɛː **4.06**	iː 2.16	tː 1.99
ɤ	øː 61.96	eː 13.37	pː 8.91	ʉː 6.33	del: 4.76	del: 45.24	øː 18.30	eː 5.72	a: 3.97	nː 3.57
ɨ	del: 59.36	eː 22.08	øː 5.04	ɑː 1.37	ɔː 1.07	del: 74.84	eː 3.64	ɑː 2.46	iː 1.47	lː 1.43
ɪ	iː 39.05	eː 22.20	del: 10.16	ɪː **4.64**	ɛː 3.41	del: 48.84	iː 10.07	eː 9.37	ɾː 3.25	nː 2.64
ɯ	ɔː 25.28	ʉː 21.81	eː 19.08	del: 13.89	iː 9.51	del: 57.99	øː 9.27	a: 3.90	nː 3.63	ɛː 3.17
ʊ	oː 30.60	uː 24.75	ɔː 21.63	del: 9.61	ʉː 2.94	del: 60.69	ɔː 5.12	oː 4.09	uː 3.90	tː 2.69
ʏ	ʉː 27.87	eː 18.44	del: 15.94	uː 12.13	øː 4.32	del: 44.25	eː 11.38	ʉː 7.42	uː 6.22	ɾː 4.23

Continued on next page.

Table 14.30b: Phone prediction rates for each cross-lingual lexicon in correctly recognized and incorrectly predicted words for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

Continued from previous page.

	correctly recognized words					incorrectly recognized words				
	1	2	3	4	5	1	2	3	4	5
xl-uni-5										
a	a: **77.22**	del: 10.77	e: 5.65	o: 1.87	s: 0.87	del: 55.76	a: **29.66**	e: 3.49	o: 1.29	n: 1.23
e	e: **78.44**	del: 9.28	a: 4.27	i: 4.03	o: 1.10	del: 53.87	e: **24.10**	a: 5.70	i: 3.91	n: 2.14
i	i: **60.95**	del: 20.22	e: 8.45	ɛɪ: 2.71	t: 1.62	del: 62.06	i: **16.52**	e: 4.93	a: 3.98	n: 1.81
o	o: **56.50**	del: 15.09	a: 10.49	u: 7.27	e: 4.00	del: 57.22	o: **17.76**	a: 6.77	e: 3.55	u: 3.02
u	u: **62.21**	del: 21.00	o: 5.72	e: 3.65	i: 1.40	del: 62.55	u: **13.95**	a: 4.76	e: 3.42	o: 3.06
y	del: 33.90	i: 31.56	e: 18.43	d: 4.12	g: 3.77	del: 50.68	i: 18.63	e: 12.67	a: 3.94	d: 2.83
æ	a: 42.40	del: 26.17	e: 18.64	d: 6.05	b: 1.84	del: 48.93	a: 25.72	e: 7.89	d: 3.02	m: 2.68
ø	e: 62.10	del: 17.36	d: 5.66	a: 4.97	i: 3.56	del: 50.05	e: 25.23	d: 5.72	a: 5.41	n: 2.25
œ	del: 34.49	e: 22.70	a: 18.40	o: 14.87	k: 1.95	del: 60.78	e: 11.57	a: 11.01	o: 3.69	ɾ: 1.48
ɑ	a: 53.62	del: 27.42	g: 3.76	e: 3.26	j: 2.54	del: 70.94	a: 10.69	e: 2.97	n: 2.11	v: 2.10
ɔ	a: 40.14	del: 26.03	o: 20.60	u: 3.74	e: 2.27	del: 69.66	a: 8.85	o: 4.16	e: 2.90	n: 2.02
ə	e: 36.39	a: 27.49	del: 26.05	i: 1.82	d: 1.30	del: 72.91	a: 7.63	e: 5.35	d: 1.38	i: 1.37
ɛ	e: 42.21	del: 31.88	a: 13.10	d: 2.73	i: 1.53	del: 72.01	e: 6.25	a: 5.84	n: 1.63	t: 1.53
ʏ	e: 83.89	o: 8.06	a: 6.10	u: 0.68	i: 0.57	del: 59.56	e: 15.95	a: 9.43	n: 3.17	i: 1.70
ɨ	del: 54.07	e: 20.71	a: 10.91	i: 6.81	s: 1.29	del: 79.79	e: 4.10	a: 4.04	i: 1.31	n: 1.10
ɪ	e: 35.54	i: 32.61	del: 18.19	ɾ: 5.36	k: 1.50	del: 56.12	e: 13.71	i: 6.37	a: 4.51	ɾ: 3.03
ɯ	u: 42.63	e: 18.42	del: 16.91	i: 13.70	m: 5.51	del: 69.23	e: 14.71	a: 5.80	n: 2.09	i: 1.75
ʊ	u: 52.79	o: 17.50	del: 13.69	e: 4.46	i: 2.58	del: 67.13	a: 6.54	u: 5.94	o: 3.94	e: 2.96
ʏ	e: 54.81	del: 19.99	u: 6.66	o: 5.83	i: 3.76	del: 54.11	e: 18.64	a: 5.31	ɾ: 2.89	i: 2.52
xl-uni-10										
a	a: **39.36**	ɑ: 22.23	del: 20.70	ɛ: 4.37	ə: 3.71	del: 60.66	a: **17.03**	ɑ: 5.34	ɛ: 2.22	n: 1.43
e	ɛ: 35.30	e: **29.30**	del: 17.70	a: 4.04	i: 3.60	del: 58.73	e: **9.43**	ɛ: 8.19	a: 3.66	i: 2.58
i	i: **51.17**	del: 20.42	e: 12.25	ɛ: 2.30	ɑ: 1.49	del: 63.97	i: **13.10**	e: 3.18	a: 2.38	ɛ: 1.99
o	ɔ: 26.77	o: **21.61**	del: 18.15	ə: 9.90	ɑ: 8.13	del: 61.34	o: **7.24**	ɔ: 5.82	ɑ: 3.45	a: 2.80
u	u: **44.78**	del: 21.48	o: 11.09	ɨ: 6.28	ə: 3.82	del: 64.02	u: **10.41**	a: 2.62	o: 2.58	ɑ: 2.19
y	del: 34.16	ɨ: 26.13	e: 16.45	i: 3.99	g: 3.91	del: 51.24	ɨ: 13.02	e: 9.37	i: 5.60	a: 2.91
æ	a: 40.77	del: 33.42	ɛ: 9.74	d: 6.64	b: 1.91	del: 57.93	a: 17.81	ɛ: 6.03	d: 2.54	m: 2.29
ø	del: 34.90	ɛ: 22.41	i: 13.24	ə: 12.58	d: 4.63	del: 49.95	ɛ: 14.08	e: 7.61	ə: 6.27	a: 4.11
œ	del: 37.37	ə: 25.17	ɛ: 8.97	ɑ: 7.44	a: 3.79	del: 62.55	ə: 7.67	ɛ: 5.32	ɑ: 3.93	a: 3.42
ɑ	a: 44.43	del: 26.31	ɑ: **12.68**	d: 3.44	j: 2.44	del: 70.91	a: 7.58	ɑ: **6.27**	ɛ: 1.63	n: 1.55
ɔ	ɔ: **24.72**	ɑ: 22.66	a: 18.55	del: 13.10	o: 9.35	del: 71.06	ɔ: **4.29**	ɑ: 4.05	a: 3.61	o: 1.69
ə	del: 28.07	ɛ: 24.90	a: 18.68	e: 6.97	ɑ: 5.84	del: 73.21	a: 4.36	ɛ: 3.41	ɑ: 2.64	ə: **2.07**
ɛ	ɛ: **48.79**	del: 16.95	a: 12.45	e: 11.36	j: 1.33	del: 72.45	ɛ: **7.52**	a: 3.48	e: 2.21	ɑ: 1.28
ʏ	ɛ: 73.79	ə: 12.85	del: 8.19	i: 2.81	a: 0.42	del: 63.77	ɛ: 12.44	a: 4.27	v: 2.37	ɑ: 2.29
ɨ	del: 56.91	ɛ: 15.95	ɑ: 6.74	a: 4.31	ɨ: **3.96**	del: 80.20	ɛ: 2.57	a: 2.50	ɨ: **1.30**	ɑ: 1.29
ɪ	e: 50.22	del: 19.37	i: 13.08	ɛ: 4.47	ɾ: 2.57	del: 58.51	e: 10.50	i: 3.19	ɾ: 2.75	ɛ: 2.67
ɯ	del: 33.45	o: 25.12	e: 19.63	ɑ: 6.40	i: 5.84	del: 70.72	ɛ: 5.45	ɨ: 4.08	a: 3.67	n: 2.56
ʊ	u: 36.33	del: 17.70	o: 15.74	ɨ: 12.56	ɔ: 6.34	del: 68.46	u: 4.26	a: 3.86	o: 2.44	ɑ: 1.94
ʏ	ə: 26.53	del: 23.59	i: 18.26	e: 12.06	ɛ: 6.64	del: 55.30	ə: 6.91	e: 6.31	ɨ: 5.21	ɛ: 3.80

Continued on next page.

Table 14.30c: Phone prediction rates for each cross-lingual lexicon in correctly recognized and incorrectly predicted words for all ASR systems evaluated cross-lingually. The table shows top 5 phone predictions and their prediction rates in % for each reference vowel averaged across the nine evaluation languages. del signifies a deletion error. Bolded results are correct predictions. The reference vowels highlighted in gray are **not** found in the model's vowel set.

Continued from previous page.

	correctly recognized words					incorrectly recognized words				
	1	2	3	4	5	1	2	3	4	5
xl-uni-16										
a	a: **55.96**	del: 18.61	ɑ: 12.85	ə: 4.77	e: 2.02	del: 61.08	a: **18.98**	ɑ: 4.33	e: 2.07	n: 1.24
e	e: **65.35**	del: 15.73	i: 4.93	a: 4.28	ə: 4.27	del: 61.26	e: **15.32**	a: 4.17	i: 3.48	n: 1.85
i	i: **59.81**	del: 21.93	e: 4.81	ɛɪ: 3.41	ɑ: 1.44	del: 64.37	i: **15.06**	e: 3.03	a: 2.60	n: 1.60
o	o: **37.96**	del: 24.12	u: 5.66	ɒ: 5.64	ɑ: 4.94	del: 62.50	o: **11.63**	a: 3.03	ɑ: 2.94	u: 2.30
u	u: **57.65**	del: 20.65	o: 5.21	ʉ: 3.80	ɵ: 2.21	del: 64.97	u: **11.68**	a: 2.79	o: 2.33	ɑ: 1.99
y	del: 31.95	y: **23.79**	ø: 11.79	ʉ: 11.60	d: 5.01	del: 44.48	y: **16.33**	ʉ: 8.74	ø: 5.90	d: 3.58
æ	a: 46.63	del: 35.34	e: 4.52	d: 4.37	ə: 2.30	del: 56.37	a: 22.88	e: 3.25	m: 2.65	ɑ: 1.40
ø	ø: **62.30**	del: 11.89	ə: 9.22	e: 3.98	ɵ: 3.94	del: 43.93	ø: **25.60**	ə: 6.29	ɵ: 3.90	a: 3.15
œ	del: 44.88	ø: 14.66	ɵ: 10.31	œ: 7.50	ə: 4.59	del: 67.07	ɵ: 5.19	ɑ: 3.27	ø: 2.40	a: 2.25
ɑ	a: 44.85	del: 27.83	ɑ: **14.17**	n: 3.43	e: 1.59	del: 72.17	a: 8.03	ɑ: **5.62**	n: 2.45	e: 1.36
ɔ	del: 32.29	a: 21.82	ɑ: 19.15	o: 11.31	ɒ: 2.60	del: 72.81	a: 4.55	ɑ: 3.64	o: 2.53	n: 1.83
ə	a: 33.36	del: 28.42	e: 14.98	ə: **11.70**	s: 1.76	del: 74.78	a: 5.62	e: 2.74	ə: **2.59**	ɑ: 1.59
ɛ	del: 37.07	e: 29.39	a: 17.99	ə: 2.80	i: 2.71	del: 75.32	a: 4.43	e: 3.43	i: 1.49	n: 1.37
ɤ	œ: 50.19	del: 16.68	ɵ: 11.46	a: 8.43	n: 7.74	del: 72.71	œ: 5.21	a: 4.64	n: 2.50	e: 1.92
ɨ	del: 71.87	a: 6.15	e: 2.78	ɨ: **2.75**	d: 2.57	del: 82.63	a: 2.71	e: 1.63	ɑ: 1.07	i: 0.99
ɪ	i: 38.99	e: 24.51	del: 21.56	ɾ: 4.46	ə: 1.62	del: 59.95	e: 8.27	i: 6.38	a: 2.70	ɾ: 2.64
ɯ	del: 48.20	u: 14.94	o: 12.07	ʉ: 6.57	e: 5.46	del: 75.52	a: 3.97	ɵ: 2.86	n: 2.74	e: 1.81
ʊ	u: 41.72	del: 24.25	o: 13.31	ʉ: 7.84	ɵ: 2.65	del: 69.35	u: 4.71	a: 3.99	o: 2.71	ɑ: 1.84
ʏ	del: 28.35	ø: 24.22	ə: 11.01	ɵ: 8.38	ʉ: 5.62	del: 58.84	e: 4.50	ø: 4.11	ɵ: 3.63	ɾ: 3.06

However, inspecting the prediction rates for the correctly recognized words, we can see that neither lowering the phone deletion rate or increasing the vowel recognition rate is guaranteed to improve the word recognition rates. For example, the phone prediction rates for the vowel [a] within correctly recognized words show that all three *xl-uni* lexicons have increased the vowel recognition rate compared with both baselines. This is especially evident with the *xl-uni-5* lexicon whose vowel recognition rate increases 74.18% points over the *xl-nst* and 62.37% points over the *xl-lnet* baseline. At the same time, its vowel deletion rate for [a] is almost three times lower compared to the *xl-nst* baseline (10.77% vs. 30.38%) and twice as low compared to the *xl-lnet* (10.77% vs. 21.22%). Yet, the word recognition rates for words containing this vowel are relatively stable across all lexicons. This suggests that there is no clear correlation between cross-lingual vowel recognition and downstream word recognition, and that hybrid

ASR systems are relatively resistant to deviation in pronunciation from the phonological norm.

Conclusions

15

We have presented another formant-based vowel categorization method aimed at improving vowel recognition in cross-lingual ASR. The method involves computing the mean $F_1 - F_2$ values of the cardinal vowels from the NST corpus of Scandinavian languages and using them to create three levels of language-universal vowel systems, which can be used to recategorize the vowels of any spoken language. The three categorization levels differ in the number of vowel categories they use to partition the $F_1 - F_2$ vowel space. The first level, *uni-5*, distinguishes five very broad vowel categories. The second level, *uni-10*, distinguishes ten narrower vowel categories. Finally, the third level, *uni-16*, which adds the rounding dimension, distinguishes eight pairs of vowel categories that differ only in terms of lip rounding, or 16 vowels altogether. We applied each of the categorization levels to the NST corpus to obtain language-universal formant-based vowel representations of Danish, Norwegian, and Swedish. Then we evaluated the resulting vowel representations using cross-lingual phone recognition models in-domain, on the NST corpus, and out-of-domain, on five parliamentary speech corpora and on five low-resource languages from the noisy telephone speech corpus, Babel.

As we have seen in the previous chapter, formant-based vowel categorization with language-universal vowel systems can improve cross-lingual vowel recognition, both on in-domain NST data and on unseen languages and speech domains. However, the performance results are highly dependent on the target language and its vowel system, as well as the number of vowel categories in the categorization system. They are also likely dependent on the languages and domains used for determining the language-universal vowel categories, as well as on the fine-tuning data.

In particular, the models trained on the *uni-5* representations achieved the best vowel recognition rates on most of the evaluated languages, even the ones that contrast more than five vowels. Increasing

the number of vowel categories almost always led to increased vowel confusions and lower recognition rates. This was especially evident when increasing the number of categories along the same dimension. For this reason, the models trained on the *uni-10* representations tended to have worse recognition rates, because they had to distinguish four levels of vowel height, i.e. one height level more than the *uni-5* and *uni-16* models.

However, the model with the best vowel recognition rate was not necessarily able to recognize all reference vowels of the target language. For example, despite having the highest vowel recognition rates, the *uni-5* model could not predict vowels outside of its limited vowel set. This makes it unsuitable for languages with larger vowel systems. On the other hand, for languages with smaller vowel systems, the *uni-10* and *uni-16* models often predicted vowel categories not found in their vowel inventories. Nevertheless, we found that the models with the larger vowel sets (*uni-10* and *uni-16*) were sometimes able to infer the vowel inventory of a language by simply suppressing predictions for the vowels outside the inventory.

To assess their wider applicability in ASR, we also evaluated the investigated vowel representations as part of a downstream speech recognition task. Namely, we used the phone recognition models to create cross-lingual pronunciation lexicons for monolingual hybrid ASR systems. The hybrid ASR systems were trained and evaluated on the same parliamentary and noisy telephone speech corpora. These experiments, however, did not reveal many conclusive patterns as the hybrid systems tended to have similar word error rates regardless of their cross-lingual lexicon, both overall and on individual reference vowels. High phone deletion rates of the phone recognition model was the only issue we observed to be harmful to the hybrid systems' performance. This resulted in both a high number of deleted vowels in the cross-lingual lexicon and lower vocabulary coverage, which were both found to be detrimental to performance. Apart from phone deletion, the hybrid systems seem robust to deviation in pronunciation from the phonological norm.

Part VI

Conclusion

Discussion 16

16.1 Discussion of Research Questions

Over the course of the thesis, we have presented four versions of a formant-based vowel categorization method aimed at improving vowel recognition in cross-lingual speech recognition. The main goal of this method was to uncover the phonetic quality of spoken vowels from their formant frequencies. This goal was rooted in the assumption that phonetic quality is more consistent across languages than language-specific phonological qualities, and is, thus, more likely to transfer to unseen languages, including the still numerous low-resource languages throughout the world. Therefore, it has the potential to improve speech recognition for languages with little to no speech data available.

Specifically, we have investigated whether and to what extent formant-based vowel categories obtained from a trilingual speech corpus of Danish, Norwegian, and Swedish could transfer to unseen languages, both in-domain, within the same corpus, and out-of-domain, on real-world speech data. Although they are no longer considered low-resource, the three Scandinavian languages are still substantially less resourced than the highest-resource languages. Moreover, their rich and diverse vowel systems cover most of the cardinal vowels which offers portability to a larger and more varied set of languages. Finally, they also comprise a large trilingual corpus suitable for experiments in multilingual and cross-lingual phonetic transfer due to its phonetic diversity and high signal-to-noise ratio.

Specifically, we have performed four types of vowel categorizations: monolingual language-dependent (*mono*), multilingual language-dependent (*multi*), language-independent (*cardinal*), and language-universal (*uni*) with three different vowel set sizes. We have, then, investigated their effects on cross-lingual phone recognition, first,

using the trilingual Scandinavian corpus, and, then, on additional more challenging speech domains, including parliamentary and noisy telephone speech data.

We now return to the research questions posed in Section 1.2, summarize our key findings for each question, and discuss the answers they provide. Our first question was:

RQ1 Can we derive phonetic vowel representations, which are consistent across languages, from the measurements of vowel formant frequencies using *language-specific* vowel categories?

This question was addressed in Part IV, where we investigated how the first three types of vowel categorizations: monolingual language-dependent (*mono*), multilingual language-dependent (*multi*), and language-independent (*cardinal*) affected the performance of cross-lingual phone recognition models on the three Scandinavian languages: Danish, Norwegian, and Swedish.

Our analyses of the experiment results showed that the models fine-tuned on the new vowel categories reduced cross-lingual phone error rates on all three languages, as well as phone feature edit distances on Danish and Swedish. The best-performing models were consistent within languages and across variations of sample size and experiment reruns, but different across languages. Namely, the *cardinal* models outperformed the baselines in terms of phone error rate on all three languages. They achieved the best performance among the models evaluated on Danish, whereas on Norwegian and Swedish, the best performers were the *mono* models. Moreover, the *cardinal* models resulted in the highest margins of improvement over the baseline on Danish compared to the best performing models on Norwegian and Swedish. We speculate that the performance improvement was higher for Danish because its phonological system is more distant to those of Norwegian and Swedish than they are to each other. For this reason, its vowel system is also less compatible with the vowel systems of

the other two languages, and, could, thus, benefit the most from the recategorization.

When it comes to the performance on dialect regions, only weak and statistically non-significant correlations were observed between the models' performance gain on a dialect region and the region's mean vowel distance from the capital. Therefore, the answer to the question whether formant-based vowel categorization could improve cross-lingual phone recognition on under-represented language varieties, such as regional dialects, remains inconclusive.

Finally, the analysis of individual phone predictions revealed that most non-minority vowels that were shared by all three languages benefited from the *cardinal* categorization (especially Danish). On the other hand, absence of a vowel from one or both training languages led to reduced vowel recognition rates for all categorization types. At the same time, a visual comparison of top phone predictions and recategorized vowel plots indicates that having the same vowel category overlap in the vowel space across languages increases the vowel recognition rates, whereas a cross-lingual mismatch in vowel categories leads to vowel confusions.

Based on these findings, we can see that cross-lingual alignment of vowel representations in the formant space can indeed lead to better cross-lingual vowel transfer. As a result, converting vowels into a shared set of formant-based vowel categories can lead to higher recognition rates. However, cross-lingual vowel recognition remains a challenge, even in the case of a trilingual corpus with three geographically and typologically close languages with similar vowel systems. Finally, to answer our main question, obtaining phonetic vowel representations from the formant frequencies of vowels using these categorization techniques is possible to an extent, but more work is needed to further reduce cross-lingual interference and ensure that the obtained vowel representations are transferable to a wide variety of languages and speech domains.

Our second research question was:

RQ2 Can we derive phonetic vowel representations, which are consistent across languages, from the measurements of vowel formant frequencies using *language-universal* vowel categories?

This question was addressed in Part V, where we investigated how the fourth type of formant-based vowel categorizations, performed with a language-universal vowel set (*uni*), affected the performance of cross-lingual phone recognition models on the three Scandinavian languages, as well as on additional speech data from the parliamentary and conversational telephone speech domains. This study involved the creation of three levels of language-universal vowel systems, which could be used to recategorize the vowels of any spoken language. The three categorization levels differed in the number of vowel categories used to partition the $F_1 - F_2$ vowel space. The first level (*uni-5*) distinguished five very broad vowel categories, the second level (*uni-10*) distinguished ten, and the third level (*uni-16*) distinguished 16, eight unrounded and eight rounded.

The most consequential difference between this categorization method and the previous one, with language-specific vowel sets, was that this method replaced the entire phonological vowel system of each training language with a single vowel system to be shared by all training languages. As we have seen, this significantly increased the cross-lingual overlap of corresponding vowel categories from the training languages in the formant space, and, in turn, led to improvements in cross-lingual vowel and overall phone recognition, on many of the evaluated languages, including the low-resource ones. However, not all evaluation languages saw the same amount of improvement. The performance was highly dependent on the target language and its vowel system, as well as the number of vowel categories in the categorization system. They were also likely dependent on the languages and domains used for determining the vowel categories, as well as on the fine-tuning data, which, in our case, was the NST corpus of Scandinavian languages.

In particular, the models trained on the *uni-5* representations

achieved the best vowel recognition rates on most of the evaluated languages, even the ones that contrast more than five vowels. Increasing the number of vowel categories almost always led to increased vowel confusions and lower recognition rates. This was especially evident when increasing the number of categories along the same dimension, such as having to distinguish four levels (as with the *uni-10* models) of vowel height instead of three (*uni-5* and *uni-16* models).

However, the model with the best vowel recognition rate was not necessarily able to recognize all reference vowels of the target language. For example, despite having the highest vowel recognition rates, the *uni-5* model could not predict vowels outside of its limited vowel set. This made it less suitable for languages with larger vowel systems. On the other hand, for languages with smaller vowel systems, the *uni-10* and *uni-16* models often predicted vowel categories not found in their vowel inventories. Nevertheless, a qualitative analysis revealed that the models with the larger vowel sets were sometimes able to infer the vowel inventory of a language by simply suppressing predictions for the vowels outside the inventory.

Finally, to answer the main question, deriving phonetic vowel representations from the formant frequencies of vowels using the language-universal vowel categorization techniques is again possible to an extent. However, future research should investigate how to best adapt these methods to individual languages, as well as whether the derived representations agree with human judgments of vowel quality.

Our third research question was:

RQ3 Can we show that formant-based vowel representations are useful in word-based speech recognition?

This was a secondary research objective for the study presented in Part V, which was addressed in Sections 13.6, 14.3, and 14.5. There, we evaluated the vowel representations obtained using language-universal formant-based vowel categorization techniques as part of a downstream speech recognition task. With this set of experiments we

wanted to show whether formant-based vowel representations could be phonologically relevant, i.e. used to recognize lexical items.

For this purpose, we used the phone recognition models to create cross-lingual pronunciation lexicons for monolingual hybrid ASR systems. The hybrid ASR systems were, then, trained and evaluated on the same parliamentary and noisy telephone speech corpora that we used for the evaluation of the phone recognition models.

However, these experiments did not reveal many conclusive patterns, as the hybrid systems tended to have similar word error rates regardless of their cross-lingual lexicon, both overall and on individual reference vowels. High phone deletion rates of the phone recognition model was the only issue we observed to be harmful to the hybrid systems' performance. This resulted in both a high number of deleted vowels in the cross-lingual lexicon and lower vocabulary coverage, which were both found to be detrimental to performance. Apart from phone deletion, the hybrid systems seemed robust to deviation in pronunciation from the phonological norm.

Our fourth and final research question was:

RQ4 Can we use Danish parliamentary data to create a large speech corpus that will significantly expand publicly available ASR resources for Danish?

This was a secondary research objective addressed in Chapter 7. In this chapter, we described the creation and the evaluation of the *FT Speech* corpus from the recorded meetings of the Danish Parliament, known as the Folketing. *FT Speech* was introduced at the Interspeech conference and released publicly in 2020. With over 1800 hours of speech, it remains the largest publicly available ASR corpus for Danish to date.

Prior to its release, there had been very few public speech corpora in Danish, and almost none of them included real-world spontaneous speech. In fact, the only available and suitable corpus for ASR at the

time was the Danish NST subcorpus, which we used in vowel categorization experiments. However, this corpus contains only read speech recorded in a quiet office environment. Moreover, it was recorded in the 1990's and is already around 30 years old. Finally, with 320 hours of speech, it is also about 5.5 times smaller than *FT Speech*. Therefore, we believe that *FT Speech* is a significant contribution to the repertoire of publicly available ASR resources for Danish.

16.2 Limitations

As we noted in the beginning, speech is a very complex signal and can be analyzed at various levels, starting from the most concrete acoustic level to the most abstract phonological and even pragmatic levels. However, these levels are not always easy to separate, because one and the same feature or cluster of features can be informative at multiple levels. Furthermore, the features at one level can interact with and affect the features at a different level. For example, stress, which is manifested through changes in loudness, length, and pitch, and used to signify prosodic prominence in languages such as English, can affect the phonetic quality of vowels, making unstressed vowel appear shorter, weaker, and more centralized.

One of the main limitations of our study is that, while trying to capture and isolate vowel quality, we have largely ignored other phonetic features and the influence they might have on vowel quality. For example, we have ignored stress, pitch accent, Danish stød, vowel length, and vowel reduction. This might have skewed the distribution of certain vowel categories and affected both the vowel normalization and categorization procedures.

Another limitation is that we have also ignored the temporal dimension of speech and treated vowels and their formants as stationary. This simplified view of formant frequencies made it impossible to investigate how vowel quality changes from segment to segment, what happens at segment boundaries, and whether diphthongs can be analyzed phonetically and compared across languages.

Furthermore, we have also ignored additional vowel features, such as nasalization, rhotacization, pharyngealization, and tone, which are distinctive in many languages. Therefore, our phone recognition

models would not be able to predict vowels with such features. When it comes to the feature of lip rounding, we have only partially investigated it. Namely, we treated the rounded vowels as completely separate from the unrounded ones, which meant that a rounded vowel phoneme could never be recategorized as an unrounded one, and vice versa. This was done because our study was limited to only the first two formant frequencies. Expanding the formant analysis to include the third and higher formants might enable us to capture some of these additional features.

Since we did not have manually segmented speech data and formant trackings, we relied on automatic forced alignment and formant estimates, which might have introduced errors in both the segmentation and formant calculations of the speech signals. These errors would have propagated to the vowel categorization step in the pipeline and affected its outcome. Moreover, starting always from the canonical dictionary-based transcriptions at the forced alignment stage means that we do not allow segment deletions and assimilation, which are very common sound changes, especially in spontaneous speech.

Our goal was to obtain general phonetic vowel representations that can be recognized across languages regardless of their phonological vowel inventory. However, we have based our vowel categories on a very small set of languages, namely the three Scandinavian languages. While these languages do have rich and diverse vowel systems, they do not cover all possible vowels and dimensions of vowel variation. The cross-lingual evaluation of the obtained vowel representations was also performed on a rather small sample of languages. Showing that vowel representations are general enough to be apply to any language would require a much larger pool of evaluation languages.

Another important limitation is concerned with our evaluation procedure. Namely, we did not have ground-truth phonetic transcriptions for any of the evaluation languages due to a general scarcity of phonetically annotated corpora. For this reason, we could only evaluate the formant-based vowel representations generated by the phone recognition models against either the formant-based vowel representations or the dictionary-based vowel representations from the evaluation corpora. We also did not conduct any human judgment studies to evaluate how humans would perceive the vowel representations obtain

through formant-based categorization.

Finally, the use of a transformer model architecture, as well as a pre-trained speech model in particular, could have introduced some confounding factors into our study. Specifically, the attention mechanism of the transformer model allows it to learn both short- and long-distance relationships from the input signal. This might bias the model toward the lexical content in the fine-tuning languages, making it more likely to produce the common phonotactic and lexical patterns of the languages it has seen, which are likely less transferable to unseen languages than completely language-independent phonetic patterns. At the same time, the pre-trained speech model we have used has seen a number of different languages during pre-training, and we do not know if and to what extent these language affect its predictions after fine-tuninig.

17 Outlook

While the accuracy, robustness, and multilinguality of ASR systems have increased significantly over the last decade, the immense complexity and variability of speech and language in general have thwarted our efforts to create a universal speech recognizer that works for everyone regardless of what language they use or how they use it. Nevertheless, technological development is advancing fast, owing in large part to increased diversity, accessibility, and inclusivity, which we initially sought to expand. Therefore, maintaining these ideals will inevitably bring us ever closer to our ultimate goals. Here, we outline a number of promising avenues to explore along the way.

Phonetic representations. Addressing the limitations of this study, such as those listed in the previous chapter, would be one way of contributing to improved cross-lingual speech recognition. For example, considering additional vowel and other phonetic features could increase their applicability to more types of speech sounds. Taking the temporal dimension of speech sounds into account would allow us to track sound changes and transitions. Studying prosodic patterns across languages could reveal whether pragmatic or emotional cues have a cross-lingual basis. Finally, expanding the study to not only more training and evaluation languages, but also to a more diverse set of speakers, such as children, non-native speakers, and speakers with speech disorders, could considerably broaden its applications.

Data collection and annotation. Self-supervised learning has made it possible to port large pre-trained models to the target domain using a relatively small amount of target domain data. For this reason, efforts to collect and annotate speech data for under-resourced languages and language varieties should be encouraged. Additionally, phonetic analyses and annotations would make it possible to evaluate

cross-lingual phone recognition models on ground-truth phonetic representations.

Robust transfer learning. Since data collection and annotation efforts cannot keep up with linguistic creativity and variation, we should also prioritize robustness in transfer learning. Making efficient large language and speech models that can learn more from less data would greatly improve their potential for cross-lingual and cross-domain transfer. For example, recent speech models based on the Conformer architecture (Gulati et al., 2020) enable accelerated training and inference without the need for web-scale data (Rekesh et al., 2023; Puvvada et al., 2024). This makes it easier and cheaper to fine-tune and extend such models to additional languages. Parameter-efficient fine-tuning (PEFT) is another promising approach to reducing the computational and storage costs of adapting and deploying large pre-trained models for downstream tasks. Certain PEFT methods have already been shown to be better than fine-tuning the full model on small amounts of target data, as well as better at generalizing to out-of-domain scenarios (Hu et al., 2022; Liu et al., 2022; Lester et al., 2021). These and other breakthroughs in the field of transfer learning are some of the trends that are likely to shape the future of machine learning.

Bibliography

Aalto University, Department of Signal Processing and Acoustics. 2023. Aalto Finnish Parliament ASR Corpus 2008-2020, version 2.

Nikki Adams, Bills, Aric, Conners, Thomas, Dubinski, Eyal, Harper, Mary, Lin, Willa, Melot, Jennifer, Ray, Jessica, Rytting, Anton, Shen, Wade, Silber, Ronnie, Tzoukermann, Evelyne, Wong, Jamie, Fiscus, Jonathan G., and Judith Bishop. 2017. IARPA Babel Zulu Language Pack IARPA-babel206b-v0.1e.

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. 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)*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

Patti Adank, Roel Smits, and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5):3099–3107.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan,

Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

Xavier Anguera, Jordi Luque, and Ciro Gracia. 2014. Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 1405–1409. ISCA.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2020a. Effectiveness of self-supervised pre-training for speech recognition.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Chris Bagwell, Rob Sykes, and Pascal Giard. 2015. SoX: Sound eXchange [Computer program]. Version 14.4.2, retrieved 20 Mar 2019, `https://sourceforge.net/projects/sox/`.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd

International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Hans Basbøll. 2005. *The Phonology of Danish*, 1 edition. Oxford University Press, Oxford, United Kingdom.

Daniel Benowitz, Aric Bills, Thomas Conners, Eyal Dubinski, Jonathan G. Fiscus, Mary Harper, Melanie Heighway, Hanh Le, Jennifer Melot, Akiko Onaka, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Judith Bishop. 2017. IARPA Babel Lao Language Pack IARPA-babel203b-v3.1a.

Aric Bills, Thomas Conners, Anne David, Luanne Cruz, Eyal Dubinski, Jonathan Fiscus, Ketty Gann, Mary Harper, Michael Kazi, Hanh Le, Nicolas Malyska, Jennifer Melot, Jessica Ray, Fred Richardson, Anton Rytting, Jacqui Zwanenburg, and Judith Bishop. 2020a. IARPA Babel Javanese Language Pack IARPA-babel402b-v1.0b.

Aric Bills, Thomas Conners, Anne David, Eyal Dubinski, Jonathan Fiscus, Ketty Gann, Mary Harper, Michael Kazi, Lynn-Li Lim, Nicolas Malyska, Jennifer Melot, Jessica Ray, Anton Rytting, Sinney Shen, Rosanna Smith, and Judith Bishop. 2020b. Iarpa babel mongolian language pack iarpa-babel401b-v2.0b.

Aric Bills, Conners, Thomas, David, Anne, Dubinski, Eyal, Fiscus, Jonathan G., Gann, Ketty, Harper, Mary, Kazi, Michael, Le, Hanh, Malyska, Nicolas, Melot, Jennifer, Phillips, Josh, Ray, Jessica, Roomi, Bergul, Rytting, Anton, Strahan, Tania E., and Judith Bishop. 2019. IARPA Babel Amharic Language Pack IARPA-babel307b-v1.0b.

Bing-Hwang Juang, S. Levinson, and M. Sondhi. 1986. Maximum likelihood estimation for multivariate mixture observations of

markov chains (corresp.). *IEEE Transactions on Information Theory*, 32(2):307–309.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

A. W. Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Paul Boersma and David Weenink. 2018. Praat: Doing phonetics by computer [Computer program]. Version 6.0.49, retrieved 20 Mar 2019, `http://www.praat.org/`.

Gail A. Carpenter and Krishna K. Govindarajan. 1993. Neural Network and Nearest Neighbor Comparison of Speaker Normalization Methods for Vowel Recognition. In Stan Gielen and Bert Kappen, editors, *ICANN '93*, pages 412–415. Springer London, London.

J. C. Catford. 2001. *A Practical Introduction to Phonetics (2nd ed.)*. Oxford University Press.

Dominic S. F. Chan, Adrian Fourcin, Dafydd Gibbon, Björn Granström, Mark A. Huckvale, George Kokkinakis, Knut Kvale, Lori Lamel, Børge Lindberg, Asunción Moreno, Jiannis Mouropoulos, Francesco Senia, Isabel Trancoso, Corin 't Veld, and Jerome Zeiliger. 1995. EUROM – a spoken language resource for the EU – the SAM projects. In *Fourth European Conference on Speech Communication and Technology, EUROSPEECH 1995, Madrid, Spain, September 18-21, 1995*. ISCA.

W. Chan, N. Jaitly, Q. Le, and O. 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)*, pages 4960–4964.

J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori. 2018. Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning,

and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527.

Kyunghyun Cho, 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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 577585, Cambridge, MA, USA. MIT Press.

Hyunju Chung, Eun Jong Kong, Jan Edwards, Gary Weismer, Marios Fourakis, and Youngdeok Hwang. 2012. Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, 131(1):442–454.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady ElSahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen

Chen, Marta Ruiz Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzm'an, Kevin Heffernan, Somya Jain, Justine T. Kao, Ann Lee, Xutai Ma, Alexandre Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Y. Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *ArXiv*, abs/2312.05187.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

David Crystal. 2010. *The Cambridge Encyclopedia of Language*, 3 edition. Cambridge: Cambridge University Press.

S. Dalmia, R. Sanabria, F. Metze, and A. W. Black. 2018. Sequence-based multi-lingual low resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913.

Siddharth Dalmia, Xinjian Li, Alan W Black, and Florian Metze. 2019. Phoneme level language models for sequence based low resource asr. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Sandra Ferrari Disner. 1980. Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, 67(1):253261.

L. Dong, S. Xu, and B. Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Olle Engstrand. 1990. Swedish. *Journal of the International Phonetic Association*, 20(1):4244.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Daris, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Nikola Ljubešić, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2023. Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. Slovenian language resource repository CLARIN.SI.

eSpeak NG. 2016. espeak-ng. Version: 1.51, April 2, 2022, `https://github.com/espeak-ng/espeak-ng`.

Gunnar Fant. 1960. *Acoustic Theory Of Speech Production*. The Hague, The Netherlands, Mouton.

Siyuan Feng, Piotr Żelasko, Laureano Moro-Velázquez, Ali Abavisani, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak.

2021. How phonotactics affect multilingual and zero-shot ASR performance.

Eli Fischer-Jørgensen. 1989. Phonetic analysis of the stød in standard danish. *Phonetica*, 46(13):159.

Folketingstidende. The Office of the Folketing Hansard.

D. B. Fry, Arthur S. Abramson, Peter D. Eimas, and Alvin M. Liberman. 1962. The identification and discrimination of synthetic vowels. *Language and Speech*, 5(4):171–189.

M.J.F. Gales. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech & Language*, 12(2):7598.

Heting Gao, Junrui Ni, Yang Zhang, Kaizhi Qian, Shiyu Chang, and Mark Hasegawa-Johnson. 2021. Zero-Shot Cross-Lingual Phonetic Recognition with External Language Embedding. In *Proc. Interspeech 2021*, pages 1304–1308.

Diana Geneva, Georgi Shopov, and Stoyan Mihov. 2019. *Building an ASR Corpus Based on Bulgarian Parliament Speeches*, pages 188–197. Springer, Cham.

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

Charlotte Gooskens. 2020. The north germanic dialect continuum. In Michael T. Putnam and B. RichardEditors Page, editors, *The Cambridge Handbook of Germanic Linguistics*, Cambridge Handbooks in Language and Linguistics, page 761782. Cambridge University Press.

Charlotte Gooskens and Wilbert Heeringa. 2004. Perceptive evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change*, 16(03).

Alex Graves, 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*, ICML '06, page 369376, New York, NY, USA. Association for Computing Machinery.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II1764II1772. JMLR.org.

Nina Grønnum. 2003. Why are the danes so hard to understand? In Henrik Galberg Jacobsen, Dorthe Bleses, Thomas O. Madsen, and Pia Thomsen, editors, *Take Danish - for instance*, pages 119–130. Syddansk Universitetsforlag. Contact the author if you want an offprint.

Nina Grønnum. 2006. DanPASS – A Danish phonetically annotated spontaneous speech corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

F. Grézl, M. Karafiát, and K. Veselý. 2014. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658.

Nina Grønnum. 1996. Danish vowels scratching the recent surface in a phonological experiment. *Acta Linguistica Hafniensia*, 28(1):5–63.

Nina Grønnum. 1998. Illustrations of the IPA: Danish. *Journal of the International Phonetic Association*, 28(1 & 2):99–105.

Nina Grønnum. 2023. Three quarters of a century of phonetic research on common danish stød. *Nordic Journal of Linguistics*, 46(3):299330.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented

transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Awni Hannun. 2017. Sequence modeling with ctc. *Distill*. Https://distill.pub/2017/ctc.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

Dorte Haltrup Hansen. 2018. The Danish Parliament Corpus 2009–2017, v1. CLARIN-DK-UCPH Centre Repository.

Dorte Haltrup Hansen, Costanza Navarretta, and Lene Offersgaard. 2018. A pilot gender study of the Danish Parliament Corpus. In *Proceedings of the Parlaclarin Workshop at the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Dorte Haltrup Hansen, Costanza Navarretta, Lene Offersgaard, and Jürgen Wedekind. 2019. Towards the automatic classification of speech subjects in the Danish Parliament Corpus. In *DHN*, pages 166–174.

Mark Hasegawa-Johnson, Leanne Rolston, Camille Goudeseune, Gina Anne Levow, and Katrin Kirchhoff. 2020. Grapheme-to-phoneme transduction for cross-language asr. In *Statistical Language and Speech Processing - 8th International Conference, SLSP 2020, Proceedings*, Lecture Notes in Computer Science, pages 3–19, Germany. Springer Science and Business Media Deutschland GmbH. Funding Information: This research was supported by the DARPA LORELEI program. Conclusions and findings are those of the authors, and are not endorsed by DARPA.; 8th International Conference on Statistical Language and Speech Processing, SLSP 2020 ; Conference date: 14-10-2020 Through 16-10-2020.

A. Haubold and J. R. Kender. 2007. Alignment of speech to highly imperfect text transcriptions. In *2007 IEEE International Conference on Multimedia and Expo*, pages 224–227.

Timothy J. Hazen. 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *INTERSPEECH 2006 – ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA.

Wilbert Heeringa, Keith Johnson, and Charlotte Gooskens. 2009. Measuring norwegian dialect distances using acoustic features. *Speech Communication*, 51(2):167183.

G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. 2013. Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8619–8623.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an ASR corpus using Althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Inga Helgadóttir, Róbert Kjaran, Anna Nikulásdóttir, and Jon Gudnason. 2021. Althingi parliamentary speech.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an ASR Corpus Using Althingis Parliamentary Speeches. In *Proc. Interspeech 2017*, pages 2163–2167.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Matthew Honnibal and Ines Montani. spaCy: Industrial-strength natural language processing [Software]. Version: 2.1.8, 8 Aug 2019, `https://spacy.io/`.

Takaaki Hori, Shinji Watanabe, and John Hershey. 2017. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529, Vancouver, Canada. Association for Computational Linguistics.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning. In *Proc. Interspeech 2020*, pages 1037–1041.

Wei-Ning Hsu, 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. *CoRR*, abs/2106.07447.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, 1 edition. Cambridge University Press, Cambridge, United Kingdom.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP2020*.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, USA.

Keith Johnson. 2011. *Acoustic and Auditory Phonetics*, 3 edition. John Wiley and Sons Ltd, Chicester, United Kingdom.

Martin Joos. 1948. Acoustic phonetics. *Language Monographs*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2 edition. Pearson Education International, Upper Saddle River, NJ, United States.

Daniel Jurafsky and James H. Martin. 2020. Automatic speech recognition and text-to-speech. In *Speech and Language Processing*, 3 ed. draft edition, chapter 26. Pearson Education International, Upper Saddle River, NJ, United States.

Martin Karafiát, Murali Karthick Baskar, Shinji Watanabe, Takaaki Hori, Matthew Wiesner, and Jan ernocký. 2019. Analysis of Multilingual Sequence-to-Sequence Speech Recognition Systems. In *Proc. Interspeech 2019*, pages 2220–2224.

S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang. 2019. A comparative study on transformer vs RNN in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456.

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. Interspeech 2019*, pages 1408–1412.

S. Kim, T. Hori, and S. Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839.

S. Kim and M. L. Seltzer. 2018. Towards language-universal end-to-end speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4914–4918.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The lacunae of Danish natural language processing.

In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362, Turku, Finland. Linköping University Electronic Press.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish Parliament Speech Corpus. In *Proc. Interspeech 2020*, pages 442–446.

Andreas Søeborg Kirkedal. 2016. *Danish Stød and Automatic Speech Recognition*. Ph.D. thesis, Copenhagen Business School, Denmark.

Andreas Søeborg Kirkedal. 2018. Acoustic word disambiguation with phonogical features in Danish ASR. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 21–31, Brussels, Belgium. Association for Computational Linguistics.

K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S. . Zhang. 2013. Investigation of multilingual deep neural networks for spoken term detection. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 138–143.

Mary Elizabeth Kohn and Charlie Farrington. 2012. Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. *The Journal of the Acoustical Society of America*, 131(3):2237–2248.

Gjert Kristoffersen. 2000. *The Phonology of Norwegian*, 1 edition. Oxford University Press, Oxford, United Kingdom.

Baybars Külebi. 2021. ParlamentParla - Speech corpus of Catalan Parliamentary sessions.

Baybars Külebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

William Labov, Sharon Ash, and Charles Boberg. 2005. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton.

Peter Ladefoged. 1990. Some reflections on the ipa. *Journal of Phonetics*, 18(3):335–346. Phonetic Representation.

Peter Ladefoged. 2003. *Phonetic Data Analysis*, 1 edition. Blackwell Publishing, Malden, MA, United States.

Peter Ladefoged and D. E. Broadbent. 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1):98104.

Peter Ladefoged and Sandra Ferrari Disner. 2012. *Vowels and Consonants*, 3 edition. Wiley–Blackwell.

Peter Ladefoged and Keith Johnson. 2015. *A Course in Phonetics*, 7 edition. Cengage Learning, Inc., Belmont, CA, United States.

Peter Ladefoged and Ian Maddieson. 1990. Vowels of the worlds languages. *Journal of Phonetics*, 18(2):93–122.

Ernestina Landau, Mijo Lonari, Damir Horga, and Ivo kari. 1995. Croatian. *Journal of the International Phonetic Association*, 25(2):8386.

John Laver. 1994. *Principles of Phonetics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Linguistic Data Consortium (LDC). sph2pipe. Version 2.5, retrieved 20 Mar 2019, `https://www.ldc.upenn.edu/language-resources/tools/sphere-conversion-tools`.

Therese Leinonen. 2011. Aggregate analysis of vowel pronunciation in swedish dialects. *Oslo Studies in Language*, 3(2).

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753.

Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. 2020a. On the Comparison of Popular End-to-End Models for Large Scale Speech Recognition. In *Proc. Interspeech 2020*, pages 1–5.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020b. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Xinjian Li, Siddharth Dalmia, David Mortensen, Juncheng Li, Alan Black, and Florian Metze. 2020c. Towards zero-shot learning for automatic phonemic transcription. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):82618268.

Mona Lindau. 1978. Vowel features. *Language*, 54(3):541–563.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France. European Language Resources Association.

Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. 2024. Parliamentary spoken corpus of serbian ParlaSpeech-RS 1.0. Slovenian language resource repository CLARIN.SI.

B M Lobanov. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.

Hans Lohninger. 2013. Mahalanobis-distanz. In *Grundlagen der Statistik*.

L. Lu, X. Zhang, and S. Renais. 2016. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5060–5064.

André Mansikkaniemi, Peter Smit, and Mikko Kurimo. 2017. Automatic construction of the Finnish parliament speech corpus. In *Proc. Interspeech 2017*, volume 8, pages 3762–3766.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 498–502. ISCA.

James D Miller. 1989. Auditory-perceptual interpretation of the vowel. *The Journal of Acoustical Society of America*, 85(5):22.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. pages 559–584.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

Terrance M Nearey. 1978. *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club. Indiana.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, Donostia–San Sebastián. Association for Computational Linguistics.

Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907938.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur. 2018. Low latency acoustic modeling using temporal convolution and lstms. *IEEE Signal Processing Letters*, 25(3):373–377.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Bolette Pedersen, Anna Braasch, Anders Johannsen, Héctor Martínez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, and Nicolai Hartvig Sørensen. 2016. The SemDaX corpus - sense annotations with scalable sense inventories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 842–847, Portorož, Slovenia. European Language Resources Association (ELRA).

Inge Lise Pedersen. 2003. Traditional dialects of danish and the dedialectalization 1900-2000. *International Journal of the Sociology of Language*, 2003(159).

Anna Persson and T. Florian Jaeger. 2023. Evaluating normalization accounts against the dense vowel space of central swedish. *Frontiers in Psychology*, 14.

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proc. Interspeech 2018*, pages 3743–3747.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Interspeech*, pages 2751–2755.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020a. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Proc. Interspeech 2020*, pages 4751–4755.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020b. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

R. Puggaard-Rode, C. Horslund, and H. Jørgensen. 2022. The rarity of intervocalic voicing of stops in danish spontaneous speech. *Laboratory Phonology*, 13.

Krishna C. Puvvada, Piotr elasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition and translation without web-scale data. In *Interspeech 2024*, pages 3964–3968.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Henning Reetz and Allard Jongman. 2020. *Phonetics: Transcription, production, acoustics, and perception*, 2 edition. Wiley-Blackwell, Chichester.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Thomas Riad. 2014. *The Phonology of Swedish*, 1 edition. Oxford University Press, Oxford, United Kingdom.

Caitlin Richter, Naomi H. Feldman, Harini Salgado, and Aren Jansen. 2017. Evaluating low-level speech features against human perceptual data. *Transactions of the Association for Computational Linguistics*, 5:425–440.

Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition.

O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux. 2020. Speech technology for unwritten languages.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:964–975.

Tanja Schultz. 2002. GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Proceedings of the ICSLP*, pages 345–348.

Tanja Schultz and Tim Schlippe. 2014. GlobalPhone: Pronunciation dictionaries in 20 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 337–341, Reykjavik, Iceland. European Language Resources Association (ELRA).

Språkbanken: The Norwegian Language Bank. 2003a. NST Danish ASR Database (16 khz).

Språkbanken: The Norwegian Language Bank. 2003b. NST Danish dictation (22 khz).

Språkbanken: The Norwegian Language Bank. 2003c. NST Norwegian ASR Database (16 khz).

Språkbanken: The Norwegian Language Bank. 2003d. NST Pronunciation Lexicon for Danish.

Språkbanken: The Norwegian Language Bank. 2003e. NST Pronunciation Lexicon for Norwegian Bokmål.

Språkbanken: The Norwegian Language Bank. 2003f. NST Pronunciation Lexicon for Swedish.

Språkbanken: The Norwegian Language Bank. 2003g. NST Swedish ASR Database (16 khz).

S. S. Stevens and J. Volkmann. 1940. The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*, 53(3):329–353.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904. ISCA.

Helmer Strik and Catia Cucchiarini. 2014. *On Automatic Phonological Transcription of Speech Corpora*. Oxford University Press, Oxford, United Kingdom.

James Tanner, Morgan Sonderegger, and Jane Stuart-Smith. 2022. Multidimensional acoustic variation in vowels across English dialects. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–82, Seattle, Washington. Association for Computational Linguistics.

S. Thomas, S. Ganapathy, and H. Hermansky. 2012. Multilingual mlp features for low-resource lvcsr systems. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4269–4272.

Sibo Tong, Philip N. Garner, and Hervé Bourlard. 2018a. Cross-lingual adaptation of a ctc-based multilingual acoustic model. *Speech Communication*, 104:39 – 46.

Sibo Tong, Philip N. Garner, and Hervé Bourlard. 2018b. Fast language adaptation using phonological information. In *Proc. Interspeech 2018*, pages 2459–2463.

S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.

Hartmut Traunmüller. 1990. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100.

UCLA Phonological Segment Inventory Database. 2019. Lelemi sound inventory (upsid).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 60006010, Red Hook, NY, USA. Curran Associates Inc.

K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova. 2012. The language-independent bottleneck features. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 336–341.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2023. Finnish parliament ASR corpus: Analysis, benchmarks and statistics. *Language Resources and Evaluation*, 57(4):16451670.

S. Watanabe, T. Hori, and J. R. Hershey. 2017a. Language independent end-to-end architecture for joint language identification and speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271.

S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. 2017b. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Allison Wetterlin. 2010. *Tonal Accents in Norwegian*. Linguistische Arbeiten. De Gruyter, Berlin, Germany.

M. Wiesner, O. Adams, D. Yarowsky, J. Trmal, and S. Khudanpur. 2019. Zero-shot pronunciation lexicons for cross-language acoustic model transfer. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1048–1054.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Proc. Interspeech 2022*, pages 2113–2117.

Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, and Najim Dehak. 2022. Discovering phonetic inventories with

crosslingual automatic speech recognition. *Computer Speech & Language*, 74:101358.

Piotr Żelasko, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2020. That Sounds Familiar: An Analysis of Phonetic Representations Transfer Across Languages. In *Proc. Interspeech 2020*, pages 3705–3709.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Franccoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *ArXiv*, abs/2303.01037.

E Zwicker. 1961. Analytical expressions for criticalband rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 33(2):248.

E. Zwicker and E. Terhardt. 1980. Analytical expressions for criticalband rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5):1523–1525.