



Delft University of Technology

Formant-based vowel categorization for cross-lingual phone recognition

Stepanović, Marija; Hardmeier, Christian; Scharenborg, Odette

DOI

[10.1121/10.0036222](https://doi.org/10.1121/10.0036222)

Publication date

2025

Document Version

Final published version

Published in

Journal of the Acoustical Society of America

Citation (APA)

Stepanović, M., Hardmeier, C., & Scharenborg, O. (2025). Formant-based vowel categorization for cross-lingual phone recognition. *Journal of the Acoustical Society of America*, 157(3), 2248-2262.
<https://doi.org/10.1121/10.0036222>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Formant-based vowel categorization for cross-lingual phone recognition

Marija Stepanović,^{1,a)} Christian Hardmeier,¹  and Odette Scharenborg² 

¹Department of Computer Science, IT University of Copenhagen, 2300 Copenhagen, Denmark

²Multimedia Computing Group, Delft University of Technology, 2628 XE Delft, The Netherlands

ABSTRACT:

Multilingual 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 underresourced 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 work introduces formant-based vowel categorization 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 multi-dialect speech corpus of Danish, Norwegian, and Swedish using three categorization techniques. Cross-lingual phone recognition experiments reveal that uniting vowel categories of different languages into a set of shared formant-based categories improves cross-lingual recognition of the shared vowels, but also interferes with recognition of vowels not present in one or more training languages. Cross-lingual evaluation on regional dialects provides inconclusive results. Nevertheless, improved recognition of individual vowels can translate to improvements in overall phone recognition on languages unseen during training. © 2025 Acoustical Society of America.

<https://doi.org/10.1121/10.0036222>

(Received 19 March 2024; revised 23 February 2025; accepted 4 March 2025; published online 27 March 2025)

[Editor: Paavo Alku]

Pages: 2248–2262

I. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) models perform well on a small number of high-resource languages, but often generalize poorly to the large number of under-resourced languages and dialects, especially ones not found in the training data (Scharenborg *et al.*, 2020). 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-lingually in low- and zero-resource scenarios (Feng *et al.*, 2021; Li *et al.*, 2020; Xu *et al.*, 2022; Želasko *et al.*, 2022; Želasko *et al.*, 2020). 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).

One possible explanation of 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 characteristics. As a result, these models have usually relied on combined monolingual

phonological systems (Xu *et al.*, 2022; Želasko *et al.*, 2020), 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).

Vowels are particularly prone to variation and notational inconsistencies (Labov *et al.*, 2005; Tanner *et al.*, 2022). Different languages have different vowel inventories where each vowel category is represented as a discrete phonological symbol (Ladefoged and Maddieson, 1990). Different vowels 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 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.

^{a)}Email: maste@itu.dk

This paper proposes 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 (PERs) on unseen languages, and especially their non-standard regional dialects. More specifically, we investigate the effect of formant-based vowel categorization on cross-lingual phone recognition of three languages: Danish, Norwegian, and Swedish. 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 underresourced speech varieties. Finally, we investigate for which vowels our approach is most effective. Although no longer considered low-resource, these three languages comprise a diverse trilingual corpus suitable for experiments in multilingual and cross-lingual phonetic transfer.

Our method entails the estimation and normalization of vowel formants from a phonologically 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 they occur in most languages, including the three languages studied here, and can be directly associated with vowel formants (Ladefoged and Maddieson, 1990). Minor features of vowel quality, such as nasalization, pharyngealization, rhotacization, phonation, diphthongization, and tone, are not clearly distinctive in all three languages in this study, and are, thus, not included in the analysis. Vowel length is distinctive in the studied languages but cannot be extracted from static formant frequencies alone, and is, hence, also excluded. For example, Danish and Swedish do not have a clear-cut distinction between diphthongs and monophthong vowel-vowel, vowel-consonant, or consonant-vowel sequences (Grønnum, 1998; Riad, 2014).

II. THEORETICAL BACKGROUND

A. Vowel formants

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. Although the vowel space is continuous, the cardinal vowel system allows us to describe the height and backness of 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 cardinal vowels do not accurately reflect their corresponding tongue positions (Ladefoged and Disner, 2012, p. 131). Indeed, 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; Ladefoged and Maddieson, 1990; Lindau, 1978). 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 Phonetic Association, 1999, p. 12).

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.

B. 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 (Traunmüller, 1990; Zwicker, 1961; Zwicker and Terhardt, 1980), 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 speaker's vowel space. The most widely used vowel-extrinsic procedures include Nearey1 and Nearey2 (Adank *et al.*, 2004; Nearey, 1978), 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 (Adank *et al.*, 2004; Carpenter and Govindarajan, 1993; Disner, 1980; Kohn and Farrington, 2012; Lobanov, 1971; Persson and Jaeger, 2023; Richter *et al.*, 2017). 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-linguistic 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, ʌ, 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 ($\bar{F}_{i,s}^L$).

C. Phonological characteristics of Danish, Norwegian, and Swedish

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).

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, 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, pp. 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; Grønnum, 2023). These prosodic differences further reduce the mutual intelligibility of Norwegian and Swedish with Danish (Grønnum, 2003).

The phonological systems of Danish, Norwegian, and Swedish presented here belong to the varieties spoken in and around the capital regions, as they are described in the corpus used in our experiments (see Sec. III B). Although none of the languages have a mandated spoken standard, the capital regions enjoy a relatively high level of cultural influence.³ The consonant sets of Norwegian and Swedish have significant overlap and include 23 consonants each, while Danish has 19 consonants. When it comes to their vowel systems, Danish has 26, Norwegian 19, and Swedish 18 monophthong vowels. Table I shows a side-by-side comparison of the monophthong vowel systems of Danish, Norwegian, and Swedish as described in *Språkbanken: The Norwegian Language Bank* (2003c,d,e). In addition to monophthongs, the Norwegian and Swedish vowel sets also include 5 and 2 diphthongs, respectively, which are excluded from this study as our focus is specifically on monophthongs.

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 [i:, y:, u:] in Norwegian and Swedish). 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. Since the transcription conventions used in the corpus were devised for the purpose of ASR, the presented vowel sets are not identical to the phonological systems described in the established literature (Basbøll, 2005; Engstrand, 1990; Grønnum, 1998; Kristoffersen, 2000; Riad, 2014). The analytical differences between the presented vowel systems and their corresponding systems from phonological studies include small discrepancies in the number of distinguished monophthongs or their IPA labels. This is not surprising for the Scandinavian languages as analytical inconsistencies can also be found across different phonological studies of the same language (Grønnum, 1996; Kristoffersen, 2000, p. 11).

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).⁴ 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).

III. METHODS

A. Motivation

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

TABLE I. Monophthong vowel inventories of Danish, Norwegian, and Swedish.^a

	Danish						Norwegian						Swedish					
	Front		Central		Back		Front		Central		Back		Front		Central		Back	
	−r	+r	−r	+r	−r	+r	−r	+r	−r	+r	−r	+r	−r	+r	−r	+r	−r	+r
Close	i, i:	y, y:	-	-	-	u, u:	ɪ, i:	ʏ, y:	-	ʉ:	-	ʊ, u:	ɪ, i:	ʏ, y:	-	ʉ:	-	ʊ, u:
Close mid	e, e:	ø, ø:	ə	-	-	o, o:	e:	ø:	ə	ø	-	o:	e, e:	ø:	-	ø	-	o:
Open mid	ɛ, ɛ:	œ, œ:	ʌ	-	-	ɔ, ɔ:	ɛ	œ	-	-	-	ɔ	ɛ, ɛ:	œ	-	-	-	ɔ
Open	a, a:	-	-	-	ɑ, ɑ:	ɒ, ɒ:	æ, æ:	-	-	-	ɑ, ɑ:	-	a	-	-	-	ɑ:	-

^aThe triangular colon (:) indicates length. −r and +r denote unrounded and rounded vowels, respectively. Hyphens (-) indicate that a particular vowel type is not part of a given vowel set.

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 with respect to the two primary vowel features height and backness (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, pp. 133–134). Therefore, this kind of clustering might require a large number of diverse languages to achieve better generalization in cross-lingual vowel recognition.

B. Corpus description and preparation

As speech data, we use the ASR databases for Danish, Norwegian, and Swedish created in the 1990s by the now defunct company Nordisk Språkteknologi (NST) (Språkbanken: The Norwegian Language Bank, 2003a,b,f). They 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.⁵ 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 in Danish, Norwegian Bokmål, and Swedish. Since all three 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.

Each subcorpus comprises a large number of utterances and a diverse set of adult speakers: 257 818 utterances by 716 speakers (51.8% female and 48.2% male) in Danish, 381 864 utterances by 959 speakers (55.5% female and 44.5% male) in Norwegian, and 381 865 utterances by 1041 speakers (54.9% female and 45.1% male) in Swedish. It should be noted that each speaker in a subcorpus reads the same list of utterances with their native pronunciation. This means that only the pronunciation varies across the provided dialect regions while the lexical content stays the same. Furthermore, the dialect regions in the subcorpora do not necessarily correspond to the regional varieties recognized in Scandinavian dialectology. For example, some regions are not represented in the subcorpora, such as the Danish island of Bornholm and the Swedish island of Gotland. In some cases, the regions are delimited in such a way as to create a regionally balanced corpus. The exact dialect regions for each subcorpus show phone recognition results by dialect region (see Table V).

All three subcorpora are originally split into a training and test set only. Since there are 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 left-over data from the original test set are 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 are, for Danish: 263.9, 17.9, 18.9, and 20.2 h; for Norwegian: 428.1, 32.5, 33.1, and 33.0 h; and for Swedish: 420.2, 25.4, 28.6, and 27.7 h.

To transcribe the speech data phonologically, we use the accompanying NST pronunciation lexicons of Danish, Norwegian, and Swedish (Språkbanken: The Norwegian Language Bank, 2003c,d,e). 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 to convert all phonological transcripts to the IPA, to be able to compare vowel categories across languages and to the theoretical cardinal vowels.

To be able to represent Danish, Norwegian, and Swedish vowel phonetic qualities cross-linguistically, we strip the dictionary phonological representations of all suprasegmental features, i.e., stress, tone, length, and stød

TABLE II. 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 monophthongs)	5 (2)	9 (1)	4 (0)

(Danish creaky voice) markers. Table II shows the number of phone and vowel types for each of the three languages in the NST corpus.

The three NST vowel systems are shown side by side in terms of their formant values in Fig. 1. The monophthongs shared by all three languages are [i, e, ɛ, a, y, ø, œ, ɔ, o, u]. Danish and Norwegian share only [ə], Danish and Swedish [a], while Norwegian and Swedish have [ɪ, ʏ, ʉ, ɐ, ʊ] in common. Danish has two unique monophthongs [ʌ, ɒ], Norwegian one [æ], and Swedish none.

C. Vowel categorization pipeline

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).⁶ 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 feature-space Maximum Likelihood Linear Regression transforms estimated at the speaker level (Gales, 1998). After this stage, we extract phone alignments for each utterance, which are a by-product of the training procedure, 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 2 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

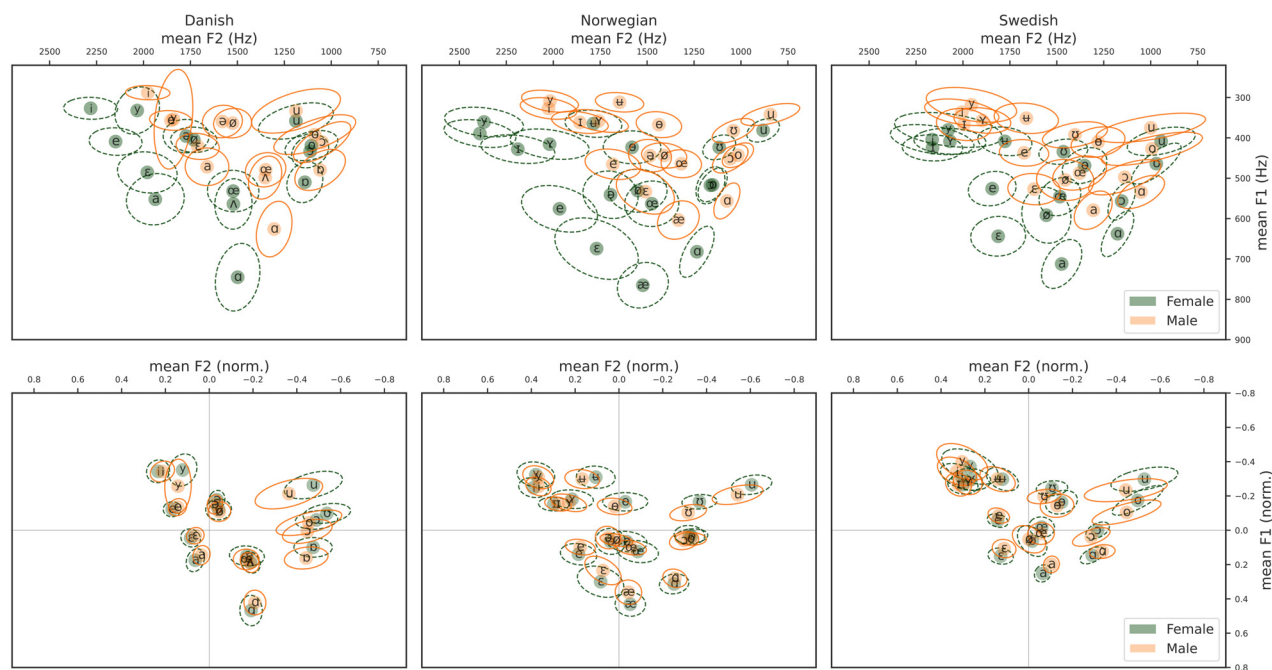


FIG. 1. Vowel spaces for each language in terms of mean raw F_1 and F_2 in Hz (top) and mean normalized F_1 and F_2 (bottom). Female (dashed green circles) and male (solid orange circles) vowel distributions are plotted separately. The locations of the vowel labels represent the grand means for each 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.

F_1-F_2 , are normalized following the procedure explained in Sec. IIB.

The effect of this procedure is visualized in Fig. 1 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 Sec. IIIA), which results in three types 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 from the mean. The outlier vowels over 2 standard deviations 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.

It should be noted that none of the categorization methods changes 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 underwent relabeling in each NST subcorpus. With multi transcripts, about 22% of the Swedish, 25% of the Norwegian, and 26% of the Danish phone tokens underwent a label change. Finally, with cardinal transcripts, about 22% of the Swedish, 25% of the Norwegian, and 25% of the Danish phone tokens underwent a label change. About 4% of all monophthong tokens were considered outliers and excluded from any categorization. Figure 2 shows the clustering decision boundaries for each of the categorization methods in relation to the original vowel distributions.

D. Cross-lingual evaluation

The utility of the three vowel categorization approaches is assessed 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., 2020; Conneau et al., 2021), 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 (original), which is used as the baseline. We further investigate the effect of the number of labeled fine-tuning samples on the cross-lingual (zero-resource) models' performance by fine-tuning on 1000, 2000, 3000, 4000, 5000, and 10 000 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 for wav2vec 2.0 (Baevski and Mohamed 2020; Graves et al., 2006) 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. For optimization, we use AdamW with a learning rate of 3×10^{-5} with 2000 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 in downstream tasks.

All fine-tuned models are evaluated in terms of PER and phone feature Hamming edit distance (PFHD)

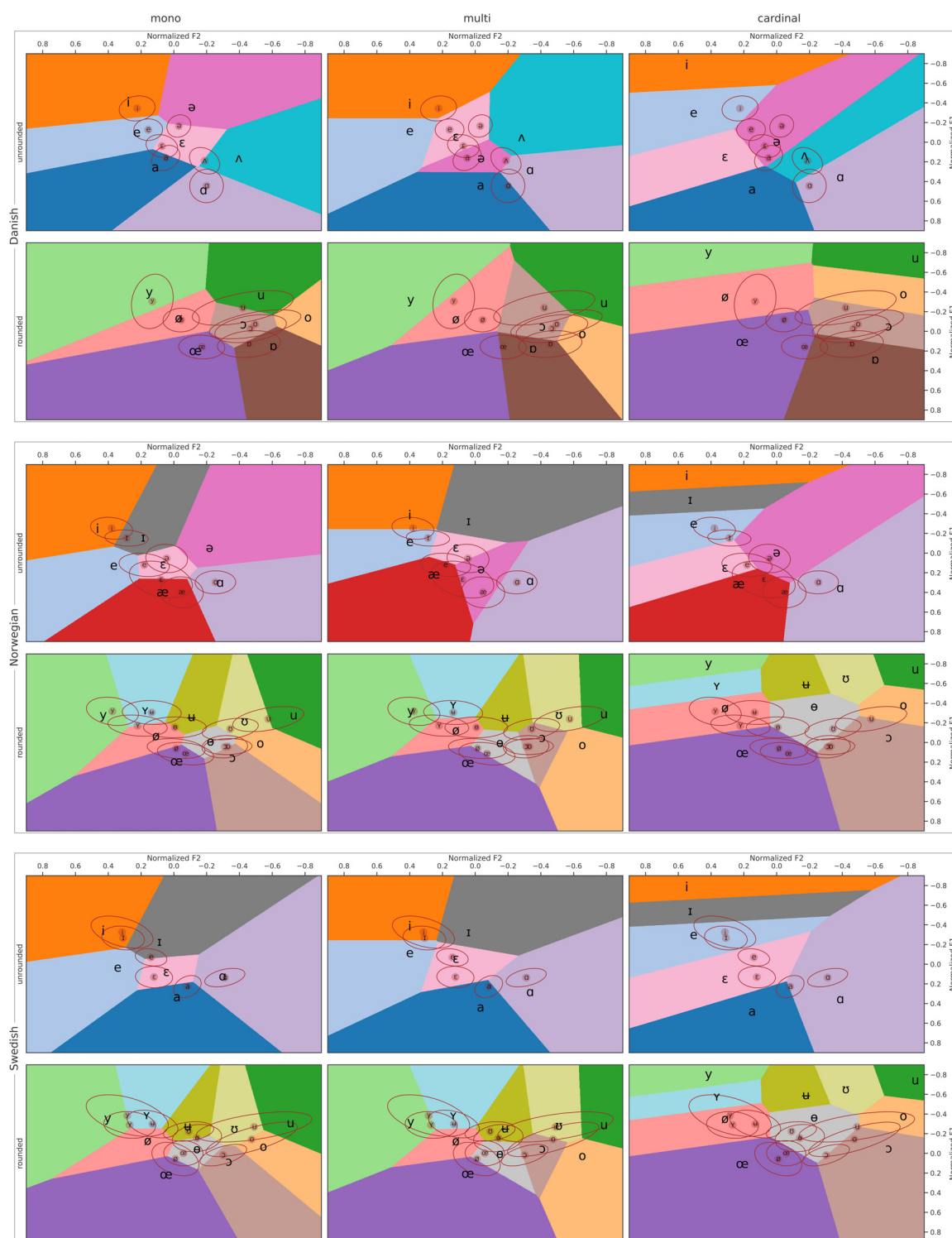


FIG. 2. The decision boundaries of each vowel category for each of the three categorization methods per language. Each vowel cluster has a different color and is labeled with the corresponding IPA symbol, which is located at the cluster centers. The circled vowel symbols and the surrounding ellipses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

(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. With this metric, each phone error (insertion, deletion, or substitution)

carries the same weight of 1. For example, for a reference utterance such as [ð i s i z ə k ʰ æ t] (*This is a cat.*), hypotheses such as [d i s i z a k ɛ t] and [o m s i z r v n t] would have the same PER (55.6%). However, 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.

TABLE III. Mean PERs and standard deviation (%) of all cross-lingual models averaged over three experiment runs. The results in a rectangular frame indicate the best models for each categorization type across different sample sizes.

Samples per training language transcript		1000	2000	3000	4000	5000	10 000
Danish	Original	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 ^a	50.29 ± 0.18 ^a	49.70 ± 0.12 ^a	49.92 ± 0.11 ^a	49.75 ± 0.26 ^a	49.97 ± 0.18 ^a
Norwegian	Original	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 ^a	39.65 ± 0.31 ^a	39.32 ± 0.30 ^a	39.35 ± 0.29 ^a	39.32 ± 0.20 ^a	39.60 ± 0.49 ^a
	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	Original	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 ^a	42.26 ± 0.10 ^a	41.75 ± 0.09 ^a	41.68 ± 0.06 ^a	41.46 ± 0.16 ^a	41.18 ± 0.06 ^a
	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

^aThe best results for each number of fine-tuning samples per training language and each evaluation language.

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. This metric gives the same weight to insertions and deletions as PER, 1, but less weight to substitution errors. Specifically, 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.

The two metrics complement each other and should be analyzed together as both have their strengths and limitations. PER is useful for downstream ASR tasks where exact transcripts are preferred, more likely to correlate with downstream metrics such as character or word error rate, and allows us to compare the performance of our models with the performance of other models in the literature. However, it penalizes substitution errors between phonetically close

representations too strongly. On the other hand, PFHED shows how close the references and hypotheses are in pronunciation, but penalizes deletion and insertion errors too strongly. Since both PER and PFHED measure error rates, lower scores indicate better performance.

Finally, 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 over the three experiment runs.

IV. RESULTS

A. 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 III shows the mean PERs and Table IV the mean PFHEDs of each cross-lingual model fine-tuned on the

TABLE IV. Mean PFHEDs and standard deviation of all cross-lingual models averaged over three experiment runs.

Samples per training language evaluation language transcript		1000	2000	3000	4000	5000	10 000
Danish	Original	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 ^a	6.92 ± 0.08 ^a	6.89 ± 0.04 ^a	7.05 ± 0.02 ^a	6.93 ± 0.01 ^a	7.04 ± 0.04 ^a
	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	Original	6.46 ± 0.02 ^a	6.14 ± 0.07 ^a	6.23 ± 0.06 ^a	6.28 ± 0.06 ^a	6.28 ± 0.02 ^a	6.28 ± 0.11 ^a
	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	Original	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 ^a	6.51 ± 0.08 ^a	6.52 ± 0.04 ^a	6.55 ± 0.01 ^a
	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 ^a	6.53 ± 0.05 ^a	6.54 ± 0.03	6.55 ± 0.05	6.57 ± 0.06	6.60 ± 0.04

^aThe best results for each number of fine-tuning samples per training language and each evaluation language.

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% 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% 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% 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.

B. Phone recognition on regional dialects

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 + standard deviation among the models fine-tuned on different amounts of labeled data and are shown enclosed in a rectangular frame in Table III.

Table V shows the mean PERs and standard deviation of the best cross-lingual models for each evaluation language 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 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 $\geq 4\%$ 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 ($\geq 2\%$ 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 ($\geq 2\%$ 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 region's 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 (original) 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. 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 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.

C. 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

TABLE V. Mean PERs and standard deviation (%) of the best cross-lingual models on each of the evaluation languages 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 heading.

		Copenhagen metropolitan 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)				
Danish											Total		
	Original	53.12	51.56	53.40	53.03	53.89	52.83	52.75	52.94				
		±0.23	±0.13	±0.15	±0.26	±0.19	±0.16	±0.29	±0.19				
	Mono	53.95	52.28	54.39	53.68	55.38	54.78	55.01	54.18				
		±0.08	±0.18	±0.03	±0.20	±0.05	±0.07	±0.06	±0.05				
	Multi	50.53	50.33	52.25	50.55	51.50	53.24	51.07	51.36				
		±0.09	±0.08	±0.10	±0.13	±0.01	±0.13	±0.21	±0.03				
	Cardinal	49.20	49.69	50.12	48.82	49.93	51.06	48.94	49.70				
		±0.24 ^a	±0.12 ^a	±0.15 ^a	±0.02 ^a	±0.11 ^a	±0.13 ^a	±0.17 ^a	±0.12 ^a				
		Oslo metropolitan area	Voss and surroundings		Oslo Outer Fjord	Nordland	Hedmark and Oppland	Sørlandet	Trøndelag	Sunnmøre	Bergen and Outer Vestland	South Vestland	Total
Norwegian		(1.17)	(1.20)	(1.20)	(1.20)	(1.21)	(1.23)	(1.25)	(1.27)	(1.28)	(1.30)		
	Original	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 ^a	±0.17 ^a	±0.22 ^a	±0.18 ^a	±0.33	±0.29 ^a	±0.19	±0.35 ^a	±0.14 ^a	±0.43	±0.29	±0.20 ^a
	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 ^a	±0.24	±0.10 ^a	±0.12	±0.11	±0.32 ^a	±0.03 ^a	±0.08
		Stockholm metropolitan area	Middle Sweden	Östergötland and	West Sweden	Dalarna with surroundings	Norrland	Eastern South Sweden	Gothenburg with surroundings	Västergötland	Western South Sweden	Total	
Swedish		(1.10)	(1.15)	(1.17)	(1.18)	(1.19)	(1.19)	(1.20)	(1.21)	(1.26)			
	Original	42.77	43.66	45.10	42.72	42.15	42.02	43.41	43.07	41.37	45.80	43.21	
		±0.11	±0.24	±0.22	±0.14	±0.07	±0.07	±0.14	±0.07	±0.15	±0.24	±0.12	
	Mono	40.16	40.91	42.47	40.58	41.01	41.57	42.09	40.13	39.61	43.21	41.18	
		±0.15 ^a	±0.08 ^a	±0.04 ^a	±0.07 ^a	±0.03 ^a	±0.09 ^a	±0.01 ^a	±0.11 ^a	±0.04 ^a	±0.18 ^a	±0.05 ^a	
	Multi	45.48	45.86	45.51	44.99	43.92	43.94	44.36	43.66	43.89	48.38	45.00	
		±0.29	±0.31	±0.14	±0.14	±0.27	±0.24	±0.19	±0.25	±0.16	±0.32	±0.22	
	Cardinal	44.15	44.11	42.93	42.69	41.92	42.41	41.96	41.44	42.76	44.29	42.86	
		±0.16	±0.09	±0.28	±0.11	±0.14	±0.07	±0.22	±0.26	±0.21	±0.21	±0.16	

^aThe best results for each dialect region and evaluation language.

the exception of Danish unaspirated stops [b, d, g],⁷ 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 standard deviation of the recognition rates of seen vowels are $45.7 \pm 14.9\%$.

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 VI. 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 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 Fig. 2 and measuring vowel distribution in the relabeled transcripts, we discover that these two have become minority

TABLE VI. 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 indicates deletion error; spn indicates spoken noise.

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

^aCorrect prediction.

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 [a], 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 [a, 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, ø, ʉ, 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 VII. As a consonant, [r] is not a plausible prediction for the Danish vowel [ɒ].⁸ 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 (Fig. 1). The top vowel predictions for the other two unseen vowels are all plausible. As seen in Fig. 2, they correspond to the vowel categories from the training languages which have

the most overlap with a given reference vowel in the vowel space.

V. CONCLUSION

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 PERs 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. We speculate that the performance improvement was higher for Danish because its phonological system is more distant to those of Norwegian and Swedish than the phonological systems of Norwegian and Swedish 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 underrepresented regional dialects, 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. Since this evaluation was performed on a very small sample of languages and dialect regions, the answer to the question whether formant-based vowel categorization could lead to lower PERs on non-standard regional dialects remains inconclusive.

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 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.

TABLE VII. Top 3 predictions and their prediction rates in % for unseen vowels. del indicates deletion. The results are the average over the three experiment runs.

Top hypotheses		1	2	3
ɒ	Original	r: 35.06	del: 15.77	ɔ: 15.51
	Mono	o: 24.43	del: 20.23	r: 14.33
	Multi	del: 21.14	ø: 20.10	r: 18.28
	Cardinal	ɔ: 25.49	del: 20.66	r: 18.77
ʌ	Original	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	r: 8.32
	Cardinal	del: 26.42	ə: 23.05	r: 9.76
æ	Original	e: 50.46	del: 17.51	a: 8.95
	Mono	a: 47.89	e: 15.88	del: 14.90
	Multi	e: 40.16	a: 20.52	del: 15.62
	Cardinal	e: 37.52	a: 29.53	del: 15.20

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, prosodic prominence, and tone, for example, by including temporal formant analysis, 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.

ACKNOWLEDGMENTS

We acknowledge Danish e-infrastructure Consortium (DeiC) for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium, through DeiC National HPC (ga DeiC-ITU-L5-000002).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts of interest to disclose.

DATA AVAILABILITY

Data sharing is not applicable to this article, as no new data were created or analyzed in this study. The study is performed on previously released and publicly available data, which can be obtained at the URL addresses provided in the bibliography.

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

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

³The 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 (Gooskens, 2020).

⁴Local 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).

⁵The metadata actually includes two fields regarding speakers' regional background: *Region of Birth* and *Region of Youth*. The majority of speakers have the same region in both fields. In instances where the two fields differ, we assume *Region of Youth* to be more indicative of a speaker's native regional dialect even though in reality it might not always be the case.

⁶The 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>.

⁷In Danish, [b, d, g] are commonly realized as voiceless unaspirated stops (Grønnum, 1998).

⁸It 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 [r] or fused with the following consonant when followed by an alveolar (Grønnum, 1998; Kristoffersen, 2000).

Adank, P., Smits, R., and van Hout, R. (2004). "A comparison of vowel normalization procedures for language variation research," *J. Acoust. Soc. Am.* **116**(5), 3099–3107.

Baevski, A., and Mohamed, A. (2020). "Effectiveness of self-supervised pre-training for ASR," in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Barcelona, Spain (May 4–8, 2020) (IEEE, New York), pp. 7694–7698.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada (Curran Associates Inc., Red Hook, NY), article 1044.

Basbøll, H. (2005). *The Phonology of Danish*, 1st ed. (Oxford University Press, Oxford, UK).

Boersma, P., and Weenink, D. (2018). "Praat: Doing phonetics by computer (version 6.0.49) [computer program]," available at, <http://www.praat.org/> (Last viewed March 20, 2019).

Carpenter, G. A., and Govindarajan, K. K. (1993). "Neural network and nearest neighbor comparison of speaker normalization methods for vowel recognition," in *ICANN '93*, edited by S. Gielen and B. Kappen (Springer, London, UK), pp. 412–415.

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

Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., and Hwang, Y. (2012). "Cross-linguistic studies of children's and adults' vowel spaces," *J. Acoust. Soc. Am.* **131**(1), 442–454.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech 2021*, pp. 2426–2430.

Disner, S. F. (1980). "Evaluation of vowel normalization procedures," *J. Acoust. Soc. Am.* **67**(1), 253–261.

Engstrand, O. (1990). "Swedish," *J. Int. Phon. Assoc.* **20**(1), 42–44.

Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).

Feng, S., Želasko, P., Moro-Velázquez, L., Abavisani, A., Hasegawa-Johnson, M., Scharenborg, O., and Dehak, N. (2021). "How phonotactics affect multilingual and zero-shot ASR performance," in *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 7238–7242.

Fischer-Jørgensen, E. (1989). "Phonetic analysis of the stød in standard Danish," *Phonetica* **46**(1), 1–59.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). "The identification and discrimination of synthetic vowels," *Lang. Speech* **5**(4), 171–189.

Gales, M. (1998). "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.* **12**(2), 75–98.

Gao, H., Ni, J., Zhang, Y., Qian, K., Chang, S., and Hasegawa-Johnson, M. (2021). "Zero-shot cross-lingual phonetic recognition with external language embedding," in *Proceedings of Interspeech 2021*, pp. 1304–1308.

Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**(1), 103–138.

Gooskens, C. (2020). "The North Germanic dialect continuum," in *The Cambridge Handbook of Germanic Linguistics*, Cambridge Handbooks in Language and Linguistics, edited by M. T. Putnam and B. R. Page (Cambridge University Press, London, UK), pp. 761–782.

Gooskens, C., and Heeringa, W. (2004). "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data," *Lang. Var. Change* **16**(3), 189–207.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). "Connectionist temporal classification: Labelling unsegmented sequence

- data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning* (Association for Computing Machinery, New York), pp. 369–376.
- Grønnum, N. (1996). “Danish vowels: Scratching the recent surface in a phonological experiment,” *Acta Linguist. Haf.* **28**(1), 5–63.
- Grønnum, N. (1998). “Illustrations of the IPA: Danish,” *J. Int. Phon. Assoc.* **28**(1), 99–105.
- Grønnum, N. (2003). “Why are the Danes so hard to understand?” in *Take Danish, for Instance*, edited by H. Galberg Jacobsen, D. Bleses, T. Madsen, and P. Thomsen (Syddansk Universitetsforlag, Odense, Denmark), pp. 119–130.
- Grønnum, N. (2023). “Three quarters of a century of phonetic research on common Danish stød,” *Nord. J. Linguist.* **46**(3), 299–330.
- Heeringa, W., Johnson, K., and Gooskens, C. (2009). “Measuring Norwegian dialect distances using acoustic features,” *Speech Commun.* **51**(2), 167–183.
- International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, 1st ed. (Cambridge University Press, London, UK).
- Jadoul, Y., Thompson, B., and de Boer, B. (2018). “Introducing parselmouth: A python interface to praat,” *J. Phon.* **71**, 1–15.
- Johnson, K. (2011). *Acoustic and Auditory Phonetics*, 3rd ed. (Wiley, Chichester, UK).
- Joos, M. (1948). “Acoustic phonetics,” *Language* **24**(2), 5–136.
- Kohn, M. E., and Farrington, C. (2012). “Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data,” *J. Acoust. Soc. Am.* **131**(3), 2237–2248.
- Kristoffersen, G. (2000). *The Phonology of Norwegian*, 1st ed. (Oxford University Press, Oxford, UK).
- Labov, W., Ash, S., and Boberg, C. (2005). *The Atlas of North American English: Phonetics, Phonology and Sound Change* (De Gruyter Mouton, Berlin).
- Ladefoged, P. (2003). *Phonetic Data Analysis*, 1st ed. (Blackwell, Malden, MA).
- Ladefoged, P., and Broadbent, D. E. (1957). “Information conveyed by vowels,” *J. Acoust. Soc. Am.* **29**(1), 98–104.
- Ladefoged, P., and Disner, S. F. (2012). *Vowels and Consonants*, 3rd ed. (Wiley-Blackwell, Hoboken, NJ).
- Ladefoged, P., and Johnson, K. (2015). *A Course in Phonetics*, 7th ed. (Cengage Learning, Belmont, CA).
- Ladefoged, P., and Maddieson, I. (1990). “Vowels of the world’s languages,” *J. Phon.* **18**(2), 93–122.
- Landau, E., Lončarić, M., Horga, D., and Škarić, I. (1995). “Croatian,” *J. Int. Phonetic Assoc.* **25**(2), 83–86.
- Laver, J. (1994). *Cambridge Textbooks in Linguistics Principles of Phonetics* (Cambridge University Press, London, UK).
- Leinonen, T. (2011). “Aggregate analysis of vowel pronunciation in Swedish dialects,” *OSLA* **3**(2), 75–95.
- Li, X., Dalmia, S., Mortensen, D., Li, J., Black, A., and Metze, F. (2020). “Towards zero-shot learning for automatic phonemic transcription,” *AAAI* **34**(5), 8261–8268.
- Lindau, M. (1978). “Vowel features,” *Language* **54**(3), 541–563.
- Lobanov, B. M. (1971). “Classification of Russian vowels spoken by different speakers,” *J. Acoust. Soc. Am.* **49**(2B), 606–608.
- Lohninger, H. (2013). “Mahalanobis-distanz” (“Mahalanobis distance”), available at http://www.statistics4u.info/fundstat_germ/ee_mahalanobis_distance.html, *Grundlagen der Statistik*.
- Miller, J. D. (1989). “Auditory-perceptual interpretation of the vowel,” *J. Acoust. Soc. Am.* **85**(5), 2114–2134.
- Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. S. (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*, pp. 3475–3484.
- Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*, Indiana University (Bloomington). Linguistics Club. (Bd 224) (Indiana University Linguistics Club, Bloomington, IN).
- Pedersen, I. L. (2003). “Traditional dialects of Danish and the de-dialectalization 1900–2000,” *Int. J. Sociol. Lang.* **2003**(159), 9–28.
- Persson, A., and Jaeger, T. F. (2023). “Evaluating normalization accounts against the dense vowel space of central Swedish,” *Front. Psychol.* **14**, 1165742.
- Reetz, H., and Jongman, A. (2020). *Phonetics: Transcription, Production, Acoustics, and Perception*, 2nd ed. (Wiley-Blackwell, Chichester, UK).
- Riad, T. (2014). *The Phonology of Swedish*, 1st ed. (Oxford University Press, Oxford, UK).
- Richter, C., Feldman, N. H., Salgado, H., and Jansen, A. (2017). “Evaluating low-level speech features against human perceptual data,” *Trans. Assoc. Comput. Linguist.* **5**, 425–440.
- Scharenborg, O., Besacier, L., Black, A., Hasegawa-Johnson, M., Metze, F., Neubig, G., Stüker, S., Godard, P., Müller, M., Ondel, L., Palaskar, S., Arthur, P., Ciannella, F., Du, M., Larsen, E., Merks, D., Riad, R., Wang, L., and Dupoux, E. (2020). “Speech technology for unwritten languages,” *IEEE/ACM Trans. Audio. Speech Lang. Process.* **28**, 964–975.
- Språkbanken: The Norwegian Language Bank (2003a). “NST Danish ASR Database (16 kHz),” available at <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-19>.
- Språkbanken: The Norwegian Language Bank (2003b). “NST Norwegian ASR Database (16 kHz),” available at <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-13>.
- Språkbanken: The Norwegian Language Bank (2003c). “NST pronunciation lexicon for Danish,” available at <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-26>.
- Språkbanken: The Norwegian Language Bank (2003d). “NST pronunciation lexicon for Norwegian Bokmål,” available at <https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-23>.
- Språkbanken: The Norwegian Language Bank (2003e). “NST pronunciation lexicon for Swedish,” available at <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-22>.
- Språkbanken: The Norwegian Language Bank (2003f). “NST Swedish ASR Database (16 kHz),” available at <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-16>.
- Stevens, S. S., and Volkman, J. (1940). “The relation of pitch to frequency: A revised scale,” *Am. J. Psychol.* **53**(3), 329–353.
- Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2022). “Multidimensional acoustic variation in vowels across English dialects,” in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, edited by G. Nicolai and E. Chodroff (Association for Computational Linguistics, Seattle, WA), pp. 72–82.
- Trauttmüller, H. (1990). “Analytical expressions for the tonotopic sensory scale,” *J. Acoust. Soc. Am.* **88**(1), 97–100.
- Wetterlin, A. (2010). *Linguistische Arbeiten Tonal Accents in Norwegian* (De Gruyter, Berlin).
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics), pp. 38–45.
- Xu, Q., Baevski, A., and Auli, M. (2022). “Simple and effective zero-shot cross-lingual phoneme recognition,” in *Proceedings of Interspeech 2022*, pp. 2113–2117.
- Želasko, P., Feng, S., Moro Velázquez, L., Abavisani, A., Bhati, S., Scharenborg, O., Hasegawa-Johnson, M., and Dehak, N. (2022). “Discovering phonetic inventories with crosslingual automatic speech recognition,” *Comput. Speech Lang.* **74**, 101358.
- Želasko, P., Moro-Velázquez, L., Hasegawa-Johnson, M., Scharenborg, O., and Dehak, N. (2020). “That sounds familiar: An analysis of phonetic representations transfer across languages,” in *Proceedings of Interspeech 2020*, pp. 3705–3709.
- Zwicker, E. (1961). “Subdivision of the audible frequency range into critical bands (Frequenzgruppen),” *J. Acoust. Soc. Am.* **33**(2), 248.
- Zwicker, E., and Terhardt, E. (1980). “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” *J. Acoust. Soc. Am.* **68**(5), 1523–1525.