Does *A Priori* Phonological Knowledge Improve Cross-Lingual Robustness of Phonemic Contrasts?

Lucy Skidmore[1][0000−0001−7422−446X] and Alexander Gutkin[2][0000−0001−6327−4824]

[1] Speech and Hearing Research Group, University of Sheffield, United Kingdom
lskidmore1@sheffield.ac.uk
[2] Google Research, London, United Kingdom
agutkin@google.com

Abstract. For speech models that depend on sharing between phonological representations an often overlooked issue is that phonological contrasts that are succinctly described language-internally by the phonemes and their respective featurizations are not necessarily robust across languages. This paper extends a recently proposed method for assessing the cross-linguistic consistency of phonological features in phoneme inventories. The original method employs binary neural classifiers for individual phonological contrasts trained solely on audio. This method cannot resolve some important phonological contrasts, such as retroflex consonants, cross-linguistically. We extend this approach by leveraging prior phonological knowledge during classifier training. We observe that since phonemic descriptions are articulatory rather than acoustic the model input space needs to be grounded in phonology to better capture phonemic correlations between the training samples. The cross-linguistic consistency of the proposed method is evaluated in a multilingual setting on held-out low-resource languages and classification quality is reported. We observe modest gains over the baseline for difficult cases, such as cross-lingual detection of aspiration, and discuss multiple confounding factors that explain the dimensions of the difficulty for this task.

Keywords: Phonology · Cross-lingual models · Low-resource languages.

1 Introduction

As the smallest constituents of phonological structure, distinctive features (DFs) can be used to provide a linguistically rich and language-independent representation schema for speech [7]. In contrast to the abstract and language-dependent phonemic representations commonly adopted in speech applications, such as automatic speech recognition (ASR) and text-to-speech (TTS), a unit of speech is instead represented by a set of phonologically derived characteristics. In a monolingual setting, one would typically use the phonemes of a language as the basic sound units and derive their feature encodings from the corresponding DFs.

Accepted for 22nd International Conference on Speech and Computer (SPECOM), October 7–9, 2020, St. Petersburg, Russia. This is a preprint.
The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-60276-5_51.

Encoding speech in this way not only allows comparison of the structure of various phonemes, but also provides a flexible framework for modeling inter- and intra-speaker pronunciation variability.

Various DF representations have been integrated successfully into the ASR and TTS pipelines over the years. In ASR, adding DF detection to the recognition pipelines has been shown to increase recognition performance in monolingual [18,22,23,31,33,34,41], multilingual [35,38,39] and low-resource settings [5,40]. However, detection accuracy of individual DFs can vary widely, with the difference between lowest and highest detector accuracy reported to be as high as 60% [35]. In recent years, more accurate DF detection has been achieved using state-of-the-art deep learning methods [12,17,21,29]. In TTS, where one typically starts from monolingual phonemic pronunciation dictionaries and phoneme inventories, DFs were also shown to be beneficial, especially in multilingual scenario [3,30,42].

As noted in [16], it is not clear a priori whether all DFs will be useful or valid in a multilingual setting. If feature descriptions were *phonetic* rather than phonemic, and *acoustic* rather than articulatory, one would expect a close correspondence between phonetic features and the acoustic signal. Similar observations motivated other recent research aiming for more phonetic realism [25]. The reality, however, is different. In practice, one often starts with phoneme inventories, the pronunciation dictionaries based on these inventories and the DF representations based on the word-level dictionary-based phonemic transcriptions. This procedure potentially introduces multiple sources of problems such as suboptimal design of the original phoneme inventories and under-specified phonemic pronunciations. Additional complications arise due to the choice of DF system for featurization as many competing feature systems are in use today. These issues are further exacerbated in multilingual scenarios due to linguistic diversity among languages.

One possible way of framing the question of practical utility and empirical validity of the chosen features in a multilingual resource sharing setting was proposed in [16], where the consistency of DF descriptions was evaluated on a cross-lingual task in terms of classification quality on phoneme-size spans of connected speech. One of the main empirical findings of that work is that the postulated contrasts that generally hold within a language are not necessarily robust across languages. This method is useful in several application scenarios: design of multilingual phoneme inventories with optimal DF sharing, derivation of phoneme inventories from speech in low- and zero-resource language documentation and evaluation of DF detector errors in ASR.

In this paper we continue the line of research in [16] by examining some of the cases where phonemic contrasts do not hold cross-linguistically. We investigate whether extending the original method by integrating prior linguistic knowledge into the model can improve its performance across the board. Although there are several open-source linguistic ontologies providing useful types of typological information, such as aerial and phylogenetic features of Glottolog [10] and World Atlas of Language Structures (WALS) [11], in this work we limit the scope

of prior knowledge sources to two phonological typologies: PHOIBLE [24] and PanPhon [26]. We hypothesize that the use of phonological grounding alone is sufficient for cross-lingual phonemic contrast resolution.

2 Distinctive Features and Their Typologies

Distinctive features were first established in phonological analysis in the 1950s [15], after which various approaches to their representation have been proposed. For this investigation, distinctive feature values are considered as binary — each speech sound is represented by a set of DFs that are either present or absent (see [7,9] for an overview of alternative feature systems). A sample representation for the phoneme /n/, taken from [9], which follows the binary representation scheme introduced in Chomsky and Halle's *The Sound Pattern of English* (SPE) [2] is [+consonantal, +sonorant, −continuant, +nasal, +coronal].

PHOIBLE is a free database of cross-linguistic phonological data compiled from many linguistic sources. The online 2014 edition [24] includes 2155 phoneme inventories with 2160 segment types found in 1672 distinct languages. The feature system in PHOIBLE aims to be descriptively adequate cross-linguistically and is likely to change as new languages are added. Overall the feature system consists of 37 "binary" features (such as labiodental and spreadGlottis that for the simple phonemic segments take the ternary values: present (+), absent (−) and not applicable (∅). For complex segments, such as diphthongs, tuples of the above values are used. For example, the value of a vowel feature syllabic for diphthong /Ew/ is a pair (+, −).

PanPhon is a resource consisting of a database that relates over 5,000 IPA segments (simple and complex) to their definitions in terms of about 23 articulatory features and a Python package to manipulate the segments and their feature representations [26]. Unlike PHOIBLE, which documents the actual snapshot of contemporary phonological knowledge of the world's languages from the standpoint of linguistic theory, PanPhon's mission is to develop a methodologically solid resource to facilitate research in NLP. One of the nice features of PanPhon is its great flexibility, which is achieved as follows: The resource contains a core set of approximately 146 segments represented in IPA and their corresponding features. The non-trivial segments are derived from this set using formal rules that describe the application of diacritics and modifiers, the feature specifications that provide the necessary context for the modification and articulatory feature changes required if the diacritic or modifier is applied. Similar to PHOIBLE, a ternary system is used to represent each of the articulatory features loosely based on well-established phonological classes.

3 Method and Corpora

We follow and extend the methodology proposed in [16]: to consider a phonemic contrast to be consistent or robust across languages, it needs to be easily predicted on heldout languages. This is operationalized as follows: a particular

Table 1. The six languages used in the experiments.

Name	Code	Family	Documentation	URL
Bengali	bn	Indo-Aryan	Kjartansson et al. [19]	http://www.openslr.org/37/
Gujarati	gu	Indo-Aryan	He et al. [13]	http://www.openslr.org/78/
Marathi	mr	Indo-Aryan	He et al. [13]	http://www.openslr.org/64/
Kannada	kn	Dravidian	He et al. [13]	http://www.openslr.org/79/
Telugu	te	Dravidian	He et al. [13]	http://www.openslr.org/66/
Sundanese	su	Malayo-Polynesian	Kjartansson et al. [19]	http://www.openslr.org/44/

phonemic contrast is presented as a binary classification problem. An instance of this problem consists of a span of a speech signal (e.g., a vowel in surrounding context) and a positive or negative label (e.g., front vowel vs. back vowel). A classifier is trained on a multi-speaker, multi-language dataset withholding one or more languages. We then evaluate the trained classifier on the held-out data and report its quality in terms of Area Under (resp. Over) the receiver operating characteristic Curve (AUC, resp. AOC). If the binary contrast in question is cross-linguistically consistent, we expect it to be readily predictable on held-out languages.

For cases where cross-linguistic consistency does not hold, we propose to extend this method by grounding the task on the contextual phonological knowledge provided by PHOIBLE and PanPhon. This is realized by augmenting the acoustic input features with dense categorical DF encodings. We hypothesize that a certain contrast that cannot be resolved cross-lingually from the speech signal alone may correlate with other contrasts that are robust. Such correlations may in theory be captured by including the full DF context in classifier training. At evaluation time, since the phonological context is unavailable, these categorical input features are set to 'not applicable' (∅).

Languages and Phoneme Inventories We use a smaller subset of the languages previously used for the experiments in [16]. Six languages from South and Southeast Asia were chosen for the experiments: three languages from the Indo-Aryan family (Bengali, Gujarati and Marathi), two languages from the Dravidian family (Kannada and Telugu) and Sundanese, a Malayo-Polynesian language. Open-source speech corpora for these languages are available, as shown in Table 1, which for each language shows its BCP-47 language code [27], corpus documentation reference and the corresponding location in the Open Speech and Language Resources (OpenSLR) repository [28]. All datasets consist of multi-speaker 48 kHz audio and the corresponding transcriptions. In this work we restrict the experiments to female speakers only to constrain the spectral variability due to gender-specific pitch differences. The Indo-Aryan and Dravidian languages are interesting to investigate because, on the one hand, they exhibit considerable phonological variation within each group, and on the other, share several cross-group similarities [4]. The inclusion of Malayo-Polynesian language is justified on the grounds of close historic contacts between the languages from this family, such as Javanese and Sundanese, with the Dravidian languages [14].

We reuse the phoneme inventories from prior work [16], which borrowed the South Asian phoneme inventories from [3] and Malayo-Polynesian phoneme in-

Table 2. Phoneme inventories grouped by language families.

		Phonemes (in IPA notation)
Shared		a b dʒ e f g h i k l m ŋ o p r s tʃ u
Indo-Aryan	bn	bʰ dʒʰ ɖ ɖʰ j kʰ n̪ tʃʰ ʈ ʈʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ ʈʰ ʊ̯ ɐ̯ ɔ̯
	gu	bʰ dʒʰ ɖ ɖʰ j kʰ n̪ tʃʰ ʈ ʈʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ ʈʰ ʊ l ɳ ə
	mr	bʰ dʒʰ ɖ ɖʰ j kʰ n̪ tʃʰ ʈ ʈʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ ʈʰ ʊ l ɳ ə dz dzʰ lʰ mʰ n̪ʰ ts ʊʰ
Dravidian	te	ɖ j n̪ ʈ ɖ l ɲ ɳ ʂ ʦ ʊ bʰ ɖʰ kʰ tʃʰ ʈʰ ɖʰ gʰ ʦʰ ʃ æ
	kn	ɖ j n̪ ʈ ɖ l ɲ ɳ ʂ ʦ ʊ bʰ ɖʰ kʰ tʃʰ ʈʰ ɖʰ gʰ ʦʰ ʃ dʒʰ
Malayo-Polynesian	su	d j n t w x z ɑɪ ɑʊ ə ɲ ʃ ʔ ɔɪ ɤ

Table 3. Details of the corpora used in the experiments.

Code	Speakers	Utterances	Words		Segments	Duration (seconds)
			Total	Unique		
bn	23	7,499	42,177	7,684	51,945	45 353.60
gu	18	2,853	23,065	8,172	27,481	15 464.40
mr	10	1,719	17,103	2,889	20,131	10 864.30
kn	24	2,897	14,780	8,050	18,882	15 533.70
te	24	3,351	11,220	4,186	15,784	9 819.65
su	20	2,401	21,848	3,169	26,742	11 541.40
Total	119	20,720	130,193	–	160,965	108 577.05

ventories from [43]. These phoneme inventories were designed with multilingual speech applications in mind, where languages use a unified underlying phonological representation, which is leveraged to make the most of the available data and eliminate phonemic scarcity by conflating similar phonemes into a single representative phoneme. The phoneme inventories for all languages use International Phonetic Alphabet (IPA) and are shown in Table 2 grouped by their language families. A subset of phonemes that is common to all the inventories is shown as "Shared" in the first row of the table. While the inventories do not map one-to-one to the inventories provided by existing typological resources, such as PHOIBLE, there is nevertheless a significant correlation between them.

Basic overview of the corpora is provided in Table 3. There are 119 female speakers in the combined dataset of 20,720 utterances corresponding to just over 30 hours of speech and 130,193 words. Word-level phonemic transcriptions containing 160,965 segments in total were provided by proprietary lexicons using phoneme inventories from Table 2. In order to determine segment boundaries, transcriptions where force-aligned with the acoustic parametrization of the audio using standard Hidden Markov Model (HMM)-based recipe [44]. The acoustic parametrization was obtained by downsampling the audio to 16 kHz and parametrizing it into HTK-style Mel Frequency Cepstral Coefficients (MFCC) [6] using 10 msec frame shift. The dimension of the MFCC parameters is 39 (13 static + Δ + $\Delta\Delta$ coefficients).

Phonemic Contrasts Each of the DF contrasts can be represented by two sets of phonemes, one for which the feature is present, and one where it is absent. Table 4 shows a list of phoneme groups, together with the corresponding phonemes selected from our corpora, to study such contrasts. For the binary

Table 4. Distinctive features and corresponding phonemes.

Feature		Corresponding Phonemes
FRONT	(+)	e eː æ i iː
	(−)	a aː o oː ɔ ɤ u uː ə
HIGH	(+)	i iː u uː
	(−)	e eː æ o oː ɔ ɤ ə a aː
SG	(+)	b^h dz^h $ɖ^h$ h k^h l^h m^h $n̥^h$ $ʈʃ^h$ $ʂ^h$ $ɖ^h$ g^h $ʈ^h$ $ʊ^h$
	(−)	a aː b d dz dʒ ɖ̪ e eː f i iː j k l m n n̥ o o p r s t ts tʃ ʂ u uː w x z æ ŋ ɔ ɖ ə g ɤ ɭ ɲ ɳ ʂ ʃ ʈ ʊ ʔ
CONT	(+)	a aː e eː f h i iː j l l^h o oː r s u uː w x z æ ɔ ə ɤ ɭ ʂ ʃ ʊ $ʊ^h$
	(−)	b b^h d dz dz^h dʒ ɖ̪ $ɖ^h$ k k^h m m^h n n̥ $n̥^h$ p t ts tʃ ʂ $ʈʃ^h$ $ʂ^h$ ŋ ɖ $ɖ^h$ g g^h ɲ ɳ ʈ $ʈ^h$ ʔ

classification task, the former set of phonemes, provides the positive examples, while the later one provides the negative examples. PHOIBLE and PanPhon assign compatible feature values for the phonemes and contrasts shown.

We investigate four contrasts. The 'front–back' contrast, denoted FRONT in the table, is defined as a combination of features: front vowel (+) is taken to mean [+FRONT, −BACK] in both PHOIBLE and PanPhon, and back vowel (−) is based on [−FRONT, +BACK]. We reproduce this experiment from [16] as a sanity check because we use different training data. In the original work this contrast was found to be consistent cross-linguistically. We extend the vowel experiments to 'high-low' contrast (HIGH), which for high vowels (+) is defined as [+HIGH, −LOW] and for low vowels (−) as [−HIGH, +LOW] in both typologies. The class of low vowels contains the close-mid back unrounded vowel /ɤ/, which is unique to Sundanese in our language set. The next two contrasts are particularly interesting to predict. In both cases the positive class (+) is formed by the set of spectrally diverse phonemes. The SPREADGLOTTIS laryngeal feature (SG) includes all the aspirated consonants in its positive class. The CONTINUANT manner of articulation feature (denoted CONT) specify the openness (+) or complete closure (−) of the vocal tract during the phonation. We don't restrict the +CONT class to consonants (fricatives and liquids) by also including all the vowels.

4 Experiments, Results and Discussion

Experiment Setup We use MFCCs prepared during the phoneme alignment stage (described in Section 3) as acoustic parameters. Admittedly, the use of MFCCs may be too restrictive: other representations, such as F0 or auditory-derived features, may be better suited to model the acoustic cues that signal the contrasts in each scenario [32]. Although we previously demonstrated moderate gains of other acoustic features types over the MFCCs on a similar task [8], in this work we limit the scope of investigation to MFCCs to keep the number of experiments manageable.

For each phonemic contrast three experiment configurations are constructed. For the baseline configuration, a single training example consists of 40 acoustic frames. It is constructed by stacking the frames corresponding to the particular phoneme plus its right and left context frames, possibly padding with zeros if the

context is too short. Phonemes longer than 40 frames are ignored. The PHOIBLE and PanPhon configurations are constructed by extending the baseline input features with 37 and 23 categorical features describing the phonemic segment, respectively. For each DF, the input features corresponding to the classification labels are masked out (set to 'unspecified' value ∅) in the training data. At evaluation time, since no phonological information is available to PHOIBLE and PanPhon configurations, all the input categorical features are set to ∅. The training sets for PHOIBLE and PanPhon are doubled by the simple data augmentation technique: each training example is cloned once and the categorical portion of its input features is masked out, so that the model can also learn to generalize in the absence of phonological context.

The training and evaluation sets in our experiments always consist of disjoint sets of languages and speakers. For each dataset we also limit the number of training examples to 50,000 and evaluation examples to 10,000. In order to keep the overall set of training labels balanced, with equal number of positive and negative examples, we employ a simple under-sampling approach [20]. If enough examples are available, we sample an equal number of them from every language in the training set. Conversely, an imbalance in a language is preferred over the lack of training examples. It is important to note that we do not guarantee that the number of training examples is the same across speakers of a language.

We use mean and standard deviation computed over the training set input features to scale the training as well as evaluation sets. We employ vanilla feed-forward Deep Neural Network (DNN) binary classifier from TensorFlow [1]. A simple two-layer architecture with 200 Softplus [46] units in each layer, dropout probability of 0.2 [37], Adadelta optimizer [45] and the learning rate of 0.6 with a large batch size of 6000 [36] were determined by tuning on the development set. A single classifier is trained on five languages and evaluated on the sixth held-out language. Overall we construct 72 classifiers: one for each of the six held-out languages, three input feature configurations (baseline, PHOIBLE and PanPhon) and four phonemic contrasts (high, front, cont and sg). Each training/evaluation experiment is repeated three times resulting in 216 experiments overall and statistics for the Area Over the Curve (AOC) metric are accumulated. We use AOC for better readability, since the Area Under the ROC Curve (AUC) values are generally high.

Results and Discussion Table 5 shows the AOC values for the detection high, front, cont and sg DFs across multiple training configurations. Each row in the table represents the held-out language on which the classifier trained on five languages is evaluated. Each AOC value is the mean over three runs. Confidence interval (95%) range computed using t-test over sample size $n = 3$ is shown alongside each AOC mean. As can be seen from the table, the front vs. back vowel contrast front is very robust across languages having the lowest AOC values among all the contrasts begin tested. This result confirms the result for front reported in [16, Table 6] on different data. The second best contrast which is very consistent cross-lingually is the cont manner of articulation contrast. This result is somewhat contrary to our expectations as the positive class +cont is

Table 5. AOC for HIGH, FRONT, CONT and SG DFs on held-out languages.

L.	HIGH			FRONT		
	Baseline	PanPhon	PHOIBLE	Baseline	PanPhon	PHOIBLE
bn	6.68 (±0.41)	6.68 (±2.95)	5.64 (±0.27)	1.97 (±0.11)	2.14 (±1.86)	2.78 (±0.21)
gu	7.06 (±0.12)	7.96 (±1.41)	7.64 (±0.30)	0.69 (±0.12)	0.82 (±0.21)	0.72 (±0.05)
mr	7.03 (±0.66)	7.67 (±0.75)	7.93 (±0.52)	2.11 (±0.11)	2.58 (±0.79)	2.03 (±0.16)
kn	7.65 (±0.61)	9.30 (±1.78)	8.63 (±0.25)	0.60 (±0.01)	0.56 (±0.19)	**0.50** (±0.06)
te	6.00 (±0.09)	8.36 (±2.45)	7.50 (±0.44)	1.61 (±0.05)	1.44 (±0.20)	**1.32** (±0.07)
su	3.09 (±0.16)	5.32 (±1.02)	4.83 (±1.09)	1.90 (±0.13)	3.69 (±0.36)	3.57 (±0.23)
L.	CONT			SG		
	Baseline	PanPhon	PHOIBLE	Baseline	PanPhon	PHOIBLE
bn	2.00 (±0.16)	3.18 (±0.05)	3.00 (±0.88)	8.07 (±1.89)	6.61 (±5.00)	5.69 (±5.45)
gu	3.17 (±0.69)	3.16 (±0.09)	3.09 (±0.37)	10.79 (±3.62)	10.58 (±0.62)	10.54 (±0.21)
mr	2.94 (±0.33)	3.45 (±0.18)	3.57 (±0.57)	14.77 (±3.13)	14.36 (±0.71)	14.65 (±1.57)
kn	2.16 (±0.50)	2.29 (±0.24)	2.13 (±0.74)	13.45 (±3.32)	12.14 (±0.38)	12.34 (±0.25)
te	1.82 (±0.05)	**1.53** (±0.20)	**1.46** (±0.30)	13.39 (±3.20)	13.30 (±1.83)	13.42 (±0.64)
su	1.16 (±0.50)	1.83 (±0.20)	1.66 (±0.42)	18.68 (±2.86)	16.63 (±2.03)	18.15 (±0.59)

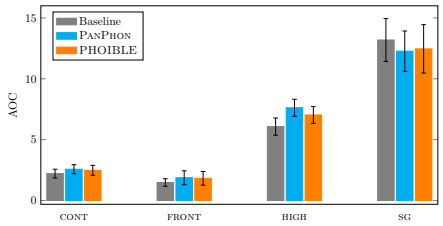
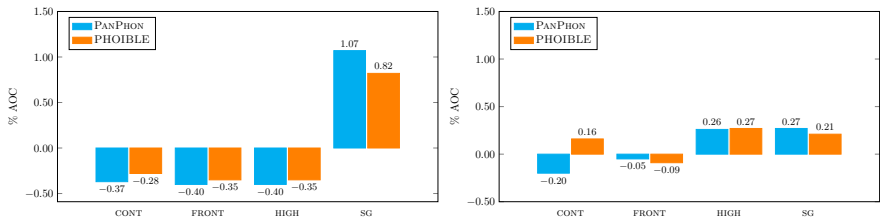


Fig. 1. Average AOC per DF classification across all held-out languages.

very heterogeneous including sounds like fricatives and vowels. The high vs. low vowel contrast HIGH is not as robust across languages as the FRONT contrast, but is also reasonably consistent cross-lingually, with the best predictions among the held-out configurations obtained for Sundanese. The worst performing configurations are found for the contrast SG that separates the aspirated sounds from the rest. With the exception of Bengali, this contrast is not robust across languages. We hypothesize that this contrast is hard to detect cross-lingually because the negative class -SG is very heterogeneous (including all the unaspirated consonants and vowels) and aspiration is acoustically more ambiguous compared to other contrasts we considered.

As can be seen from Table 5, the inclusion of phonological context in the classifier's input feature space leads only to the minor occasional statistically significant improvements over the baseline (shown in bold). In most of the other cases when the mean AOC values for the PHOIBLE and PanPhon configurations are lower than the baseline, the improvements are not statistically significant because of the overlap in confidence intervals. In order to summarize the performance of the classifiers across all the held-out languages for each phonemic contrast, we recomputed the statistics per each contrast, with Figure 1 showing the AOC means and confidence intervals computed for the sample size $n = 18$ (three runs for six languages). As can be seen from the figure, the phonologically



(a) Cross-lingual setup. (b) Language-internal setup.

Fig. 2. Relative improvements (%) in AOC over the baselines.

grounded classifiers do not improve (on average) over the corresponding baselines for CONT, FRONT and HIGH contrasts. This is likely due to the task being already unambiguous enough for the baseline classifier, where the introduction of additional context actually increases the confusion. Note that the inclusion of phonological context from both the PHOIBLE and PANPHON sources improves the detection of aspiration contrast SG. This improvement, however, is not statistically significant because of the confidence intervals overlap.

In order to further evaluate the influence of phonemic grounding on the detection of each phonemic contrast, we compared the relative improvements in AOC over the baselines for our current cross-lingual setup, shown in Figure 2 (a), with the language-internal setup shown in Figure 2 (b). The language-internal configurations are constructed similarly to the cross-lingual ones (also 216 configurations overall), with the only difference that the training and evaluation data is confined to one language. As can be seen from the figures, phonemic grounding has very little influence on the classifier performance in language-internal case. We hypothesize that this is due to the fact that the phonological context that DFs provide already exists implicitly in the structure of the acoustic training data (i.e. the phonemes in the positive and negative classes) and therefore representing it explicitly does not show an effect. In the cross-lingual case, however, phonological context helps detection of SG, the most "difficult" contrast under investigation.

In order to assess the discriminatory power of the phonological typologies alone for our task, we also constructed classifiers without relying on acoustics. For each phonemic contrast we constructed three types of classifiers for PHOIBLE and PANPHON typologies: Naive Bayes (NB), linear regression (LR) and support vector machine with linear kernel function (SVM). Stratified k-fold cross-validation with contrast task-dependent value of k constrained by the minimal negative or positive class size (as shown in Table 4) was employed. Crucially, the input features that directly correlate with the labels were masked out during training and evaluation. For example, when constructing and evaluating the classifier for contrast HIGH, both HIGH and LOW input features were set to $\varnothing$. The average AOC values over k runs are shown in Table 6 for each contrast, classifier and typology type. Comparing these classifiers' performance with the classifiers trained on the full acoustic and phonological data (Table 5) it is evident that the classifiers trained on phonology alone are significantly less accurate. Apart

Table 6. Average AOC for phonological input features alone.

Source	HIGH ($k = 4$)			FRONT ($k = 5$)			CONT ($k = 25$)			SG ($k = 14$)		
	NB	LR	SVM	NB	LR	SVM	NB	LR	SVM	NB	LR	SVM
PHOIBLE	22.92	25.00	31.25	30.00	30.00	25.00	10.00	6.00	16.00	33.93	50.29	57.44
PANPHON	29.17	37.50	37.50	25.00	40.00	20.00	12.00	14.00	16.00	46.13	54.46	54.46

from CONT, the best performing contrast in Table 6, which also correlates reasonably with the results for the full training in Table 5, the rest of the classifiers struggle to detect the contrasts in question based on the phonological context alone. The classifier for SG performs the worst. It is interesting to note that, while these classifiers are generally useful on their own, apart from the very unreliable detector for SG, their accuracy increases significantly once we combine the phonological (articulatory) input space with the acoustics. One confounding factor that may explain detection inaccuracies in this scenario are the typological features themselves – in the case of PHOIBLE and PANPHON the remaining phonological features in the context of the contrast itself may not be enough (e.g., due to their potential ambiguity or wrong definitions) to signal the contrast. This merits further research into the design of feature inventories that are highly consistent in multilingual settings.

5 Conclusion

The results from this investigation provide a starting point for further research on the impact of a priori phonological knowledge on cross-lingual DF classification. The modest gains on the baseline recorded for SG classification require further contextualisation through experimentation on a wider group of DFs. In addition, it would be of value to explore the impact of phonological processes such as assimilation and co-articulation on DF detection accuracy. Exploring an alternative network configuration may also be beneficial, such as training a sequence model over feature detectors or embeddings. In summary, it is clear that it is not only the relationship between acoustic representations of speech and phonological feature inventories that is complex — the internal relationship between individual DFs within feature inventories is also impactful and should be taken into consideration when designing feature inventories for use in multilingual settings.

Acknowledgement. The authors would like to thank Cibu Johny for his help with the experiments, and Işın Demirşahin and Rob Clark for fruitful discussions.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: A System for Large-Scale Machine Learning. In: Proc. 12th Symposium on Operating Systems Design and Implementation (OSDI). pp. 265–283. USENIX Association (2016)

2. Chomsky, N., Halle, M.: The Sound Pattern of English. Harper & Row, New York (1968)
3. Demirsahin, I., Jansche, M., Gutkin, A.: A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech. In: Proc. of 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU). pp. 80–84. ISCA, Gurugram, India (2018). https://doi.org/10.21437/SLTU.2018-17
4. Emeneau, M.: India as a Linguistic Area. Language **32**(1), 3–16 (1956). https://doi.org/10.2307/410649
5. Fu, T., Gao, S., Wu, X.: Improving Minority Language Speech Recognition Based on Distinctive Features. In: Proc. of International Conference on Intelligent Science and Big Data Engineering. pp. 411–420. Springer, Lanzhou, China (August 2018). https://doi.org/10.1007/978-3-030-02698-1_36
6. Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. In: Proc. of 10th International Conference on Speech and Computer (SPECOM). vol. 1, pp. 191–194. Patras, Greece (2005)
7. Gussenhoven, C.: Understanding Phonology. Routledge, London, 4th edn. (2017). https://doi.org/10.4324/9781315267982
8. Gutkin, A.: Eidos: An Open-Source Auditory Periphery Modeling Toolkit and Evaluation of Cross-Lingual Phonemic Contrasts. In: Proc. of 1st Joint Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop (SLTU-CCURL 2020). pp. 9–20. European Language Resources Association (ELRA), Marseille, France (May 2020)
9. Hall, T.A.: Distinctive Feature Theory. Mouton de Grutyer, Berlin, Germany (2001). https://doi.org/10.1515/9783110886672
10. Hammarström, H., Forkel, R., Haspelmath, M., Bank, S.: Glottolog 4.2.1. Max Planck Institute for the Science of Human History, Jena, Germany (2020). https://doi.org/10.5281/zenodo.3754591
11. Haspelmath, M., Dryer, M.S., Gil, D., Comrie, B.: The World Atlas of Language Structures. Oxford University Press, Oxford (2005). https://doi.org/10.5281/zenodo.3731125
12. He, D., Yang, X., Lim, B.P., Liang, Y., Hasegawa-Johnson, M., Chen, D.: When CTC training meets acoustic landmarks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5996–6000. IEEE, Brighton, UK (2019). https://doi.org/10.1109/ICASSP.2019.8683607
13. He, F., Chu, S.H.C., Kjartansson, O., Rivera, C., Katanova, A., Gutkin, A., Demirsahin, I., Johny, C., Jansche, M., Sarin, S., Pipatsrisawat, K.: Open-source Multispeaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In: Proc. of 12th Language Resources and Evaluation Conference (LREC). pp. 6494–6503. European Language Resources Association (ELRA), Marseille, France (May 2020)
14. Hoogervorst, T.: Detecting pre-modern lexical influence from South India in Maritime Southeast Asia. Archipel: Études interdisciplinaires sur le monde insulindien (89), 63–93 (2015). https://doi.org/10.4000/archipel.490
15. Jakobson, R., Fant, G., Halle, M.: Preliminaries to Speech Analysis: the Distinctive Features and their Correlates. MIT Press, Cambridge, MA (1952)
16. Johny, C., Gutkin, A., Jansche, M.: Cross-Lingual Consistency of Phonological Features: An Empirical Study. In: Proc. of Interspeech 2019. pp. 1741–1745. ISCA, Graz, Austria (September 2019). https://doi.org/10.21437/Interspeech.2019-2184

17. Karaulov, I., Tkanov, D.: Attention Model for Articulatory Features Detection. In: Proc. of Interspeech 2019. pp. 1571–1575. ISCA, Graz, Austria (2019). https://doi.org/10.21437/Interspeech.2019-3020
18. Kirchhoff, K., Fink, G.A., Sagerer, G.: Conversational speech recognition using acoustic and articulatory input. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 3, pp. 1435–1438. IEEE, Istanbul, Turkey (June 2000). https://doi.org/10.1109/ICASSP.2000.861883
19. Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M., Ha, L.: Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In: Proc. of 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU). pp. 52–55. ISCA, Gurugram, India (2018). https://doi.org/10.21437/SLTU.2018-11
20. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence **5**(4), 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0
21. Merkx, D., Scharenborg, O.: Articulatory Feature Classification Using Convolutional Neural Networks. In: Proc. of Interspeech. pp. 2142–2146. Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-2275
22. Metze, F., Waibel, A.: A Flexible Stream Architecture for ASR Using Articulatory Features. In: Proc. of 7th International Conference on Spoken Language Processing (ICSLP). pp. 2133–2136. ISCA, Denver, CO (September 2002)
23. Momayyez, P., Waterhouse, J., Rose, R.: Exploiting complementary aspects of phonological features in automatic speech recognition. In: Proc. of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). pp. 47–52. IEEE, Kyoto, Japan (December 2007). https://doi.org/10.1109/ASRU.2007.4430082
24. Moran, S., McCloy, D.: PHOIBLE 2.0. Max Planck Institute for Evolutionary Anthropology, Jena, Germany (2019), http://phoible.org/
25. Mortensen, D.R., Li, X., Littell, P., Michaud, A., Rijhwani, S., Anastasopoulos, A., Black, A.W., Metze, F., Neubig, G.: AlloVera: A Multilingual Allophone Database. arXiv preprint arXiv:2004.08031 (2020)
26. Mortensen, D.R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., Levin, L.: PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In: Proc. of COLING. pp. 3475–3484. Osaka, Japan (December 2016)
27. Phillips, A., Davis, M.: BCP 47 – Tags for Identifying Languages. IETF Trust (2009)
28. Povey, D.: Open SLR. John Hopkins University, US (2020), http://www.openslr.org/resources.php
29. Qu, L., Weber, C., Lakomkin, E., Twiefel, J., Wermter, S.: Combining Articulatory Features with End-to-End Learning in Speech Recognition. In: Proc. of International Conference on Artificial Neural Networks (ICANN). pp. 500–510. Springer, Rhodes, Greece (2018). https://doi.org/10.1007/978-3-030-01424-7_49
30. Rallabandi, S., Black, A.: Variational Attention using Articulatory Priors for Generating Code Mixed Speech using Monolingual Corpora. Proc. of Interspeech pp. 3735–3739 (2019). https://doi.org/10.21437/Interspeech.2019-1103
31. Rasipurama, R., Magimai-Doss, M.: Articulatory Feature Based Continuous Speech Recognition Using Probabilistic Lexical Modeling. Computer Speech and Language **36**, 233–259 (2016). https://doi.org/10.1016/j.csl.2015.04.003
32. Repp, B.H.: Categorical Perception: Issues, Methods, Findings. In: Speech and Language: Advances in Basic Research and Practice, vol. 10, pp. 243–335. Elsevier (1984)

33. Rose, R., Momayyez, P.: Integration of Multiple Feature Sets for Reducing Ambiguity in ASR. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. IV–325–IV–328. IEEE, Honolulu, HI (April 2007). https://doi.org/10.1109/ICASSP.2007.366915
34. Siniscalchi, S.M., Lee, C.H.: A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Communication **51**(11), 1139–1153 (2009). https://doi.org/10.1016/j.specom.2009.05.004
35. Siniscalchi, S.M., Svendsen, T., Lee, C.H.: Toward a Detector-Based Universal Phone Recognizer. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4261–4264. IEEE, Las Vegas, NV (April 2008). https://doi.org/10.1109/ICASSP.2008.4518596
36. Smith, S.L., Kindermans, P.J., Ying, C., Le, Q.V.: Don't Decay the Learning Rate, Increase the Batch Size. arXiv preprint arXiv:1711.00489 (2017)
37. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research (JMLR) **15**(56), 1929–1958 (2014)
38. Stüker, S., Schultz, T., Metze, F., Waibel, A.: Integrating Multilingual Articulatory Features Into Speech Recognition. In: Proc. of EuroSpeech. pp. 1033–1036. ISCA, Geneva, Switzerland (September 2003)
39. Stüker, S., Schultz, T., Metze, F., Waibel, A.: Multilingual Articulatory Features. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. I144–I147. IEEE, Hong Kong (April 2003). https://doi.org/10.1109/ICASSP.2003.1198737
40. Stüker, S., Waibel, A.: Porting Speech Recognition Systems to New Languages Supported by Articulatory Feature Models. In: Proc. of 13th International Conference on Speech and Computer (SPECOM). St. Petersburg, Russia (May 2009)
41. Tolba, H., Selouani, S., O'Shaughnessy, D.: Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. I–837–I–840. IEEE, Orlando, FL (May 2002). https://doi.org/10.1109/ICASSP.2002.5743869
42. Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D.R., Black, A.W., Levin, L., Dyer, C.: Polyglot Neural Language Models: A Case Study in Cross-Lingual Phonetic Representation Learning. In: Proc. of 2016 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 1357–1366. ACL, San Diego, CA (Jun 2016). https://doi.org/10.18653/v1/N16-1161
43. Wibawa, J.A.E., Sarin, S., Li, C.F., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., Ha, L.: Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In: Proc. of 11th Conference on Language Resources and Evaluation (LREC). pp. 1610–1614. European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)
44. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department (2006)
45. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. arXiv preprint arXiv:1212.5701 (2012)
46. Zheng, H., Yang, Z., Liu, W., Liang, J., Li, Y.: Improving Deep Neural Networks Using Softplus Units. In: Proc. of International Joint Conference on Neural Networks (IJCNN). pp. 1–4. IEEE (2015). https://doi.org/10.1109/IJCNN.2015.7280459