INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Cross-Lingual Consistency of Phonological Features: An Empirical Study

Cibu Johny, Alexander Gutkin, Martin Jansche

Google AI, London, United Kingdom

{cibu,agutkin,mjansche}@google.com

Abstract

The concept of a phoneme arose historically as a theoretical abstraction that applies language-internally. Using phonemes and phonological features in cross-linguistic settings raises an important question of conceptual validity: Are contrasts that are meaningful within a language also empirically robust across languages? This paper develops a method for assessing the cross-linguistic consistency of phonological features in phoneme inventories. The method involves training separate binary neural classifiers for several phonological contrast in audio spans centered on particular segments within continuous speech. To assess cross-linguistic consistency, these classifiers are evaluated on held-out languages and classification quality is reported. We apply this method to several common phonological contrasts, including vowel height, vowel frontness, and retroflex consonants, in the context of multi-speaker corpora for ten languages from three language families (Indo-Aryan, Dravidian, and Malayo-Polynesian). We empirically evaluate and discuss the consistency of phonological contrasts derived from features found in phonological ontologies such as PANPHON and PHOIBLE.

Index Terms: computational phonology, cross-lingual, low-resource languages, neural networks

1. Introduction

Multilingual speech processing tasks often benefit from featural representations of basic sound segments. For example, acoustic models for speech recognition or synthesis can be trained on multilingual data and share parameters across languages. In the simplest case the basic sounds of two or more languages can be pooled and represented with indicator variables ("one hot" encoding). Depending on the amount of overlap among sound inventories, this does not generally lead to optimal parameter sharing and data efficiency. It is therefore standard practice to represent sound units using denser feature encodings which potentially allow for more sharing. This raises the question of the practical utility and empirical validity of the chosen features, which we attempt to address in this paper.

In a monolingual setting, one would typically use the phonemes of a language as the basic sound units and derive their feature encodings from distinctive phonological features. Phonological features can be grounded in acoustic properties (e.g. periodic glottal source), articulatory properties (e.g. nasal), which may overlap (e.g. short/long); or they can be necessitated by phonological processes (e.g. Turkish /a/ as a back vowel for purposes of harmony). Often phonological descriptions are based on articulatory features as recurrent elementary components that form the sound systems of world's languages [1]. Many feature systems, ranging from binary distinctive features [2] to elaborate acoustic landmark-based phonetic-phonological representations [3], have been proposed over the years. Due to their compact combinatorial description, they have been a popular discrete representation pervasive in many areas of pronunciation modeling in speech technology research, such as text-to-speech [4, 5, 6, 7] and automatic speech recognition [8, 9, 10, 11].

Table 1: *The ten languages used in the experiments.*

Language family	Language (Code)
Indo-Aryan (I-A)	Bengali (bn), Gujarati (gu), Marathi (mr), Sinhala (si)
Dravidian (DRAV)	Kannada (kn), Malayalam (ml), Tamil (ta), Telugu (te)
Malayo-Polynesian (M-P)	Javanese (jv), Sundanese (su)

Some have argued that since phonemes are conceptual constructs [12], articulatory features constitute a significantly more viable representation [13]. In addition to being effective in describing the correlations and contrasts between phonemes of an individual language, articulatory features have been shown to be useful in multilingual scenarios [14, 15], where the need to model the phonological sharing succinctly is more acute due to the diversity between languages. It is not clear *a priori* whether all phonological features will be useful or valid in a multilingual setting. If feature descriptions were phonetic rather than phonemic, and acoustic rather than articulatory, we would expect a close correspondence between phonetic features and the acoustic signal. In practice, we typically start from monolingual phonemic pronunciation dictionaries and phoneme inventories. The dictionaries are combined with word-level utterance transcriptions, resulting in phoneme-level transcriptions that often cannot resolve alternative pronunciations. The phoneme inventories are turned into feature encodings derived from cross-linguistic typology resources, such as PHOIBLE [16] and PANPHON [17]. While the recourse to such resources has been shown to be advantageous [18, 19], there remains a question of how accurate a phonological description of a certain language within such resource really is. PHOIBLE, for example, is designed as a repository of multiple phoneme inventories found in the literature for any given language. The sources may not agree with each other and may contain mistakes that lead to unexpected inventories or featurizations.[1]

We propose a method for evaluating the cross-lingual consistency of phonological features in a multilingual setting. We treat cross-linguistic consistency in terms of classification quality on phoneme-size spans of connected speech, evaluated on held-out languages. Section 2 defines the problem and describes our method and data. In Section 3 we apply this method to several phonological features and discuss our empirical findings.

2. Problem, method, and data

To consider a phonemic contrast to be consistent or robust across languages, we require it to be easily predicted on held-out languages. We operationalize this as follows: a particular phonemic contrast is presented as a binary classification prob-

[1]For example, PHOIBLE contains 3 different phoneme inventories for Javanese, consisting of either 29, 30, or 33 segments. The largest of these, inventory GM 1675, contains a spurious vowel /y/ that presumably resulted from the use of non-IPA notation in the underlying article.

http://dx.doi.org/10.21437/Interspeech.2019-2184

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 ɡ̤ʰ i̤ kʰ n tʃʰ ʈ ʈ̤ʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ tʰ ṳ e̤ o̤
	gu	bʰ dʒʰ ɖ ɡ̤ʰ j kʰ n̤ tʃʰ ʈ ʈ̤ʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ tʰ ʊ ɭ ɳ ə
	mr	bʰ dʒʰ ɖ ɡ̤ʰ j kʰ n̤ tʃʰ ʈ ʈ̤ʰ æ ɔ ɖ ɖʰ gʰ ʃ ʈ ʈ̤ʰ ʊ ɭ ɳ ə dz dzʰ lʰ mʰ n̤ʰ ts ʊʰ
	si	ɖ̤ j n̤ ʈ̤ ɖ ɲ ʃ ʈ ʊ ᵐb ᶯg ᶯɖ̤ ᶯɖ æ ə
Dravidian	ta	ɖ̤ j n̤ ʈ̤ ɖ ɭ ɲ ɳ ʂ ʈ ʊ ɻ r n
	ml	ɖ̤ j n̤ ʈ̤ ɖ ɭ ɲ ɳ ʂ ʈ ʊ ɻ r n d t ʃ ə
	te	ɖ̤ j n̤ ʈ̤ ɖ ɭ ɲ ɳ ʂ ʈ ʊ bʰ ɡ̤ʰ kʰ tʃʰ ʈ ʈ̤ʰ ɖʰ gʰ ʈʰ ʃ æ
	kn	ɖ̤ j n̤ ʈ̤ ɖ ɭ ɲ ɳ ʂ ʈ ʊ bʰ ɡ̤ʰ kʰ tʃʰ ʈ ʈ̤ʰ ɖʰ gʰ ʈʰ ʃ dʒʰ
M-P	jv	d j n t w x z ɑɪ ɑʊ ə ɲ ʃ ʔ ɔ ɖ ʈ
	su	d j n t w x z ɑɪ ɑʊ ə ɲ ʃ ʔ ɔ ɪ ɤ

lem. 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). We train a classifier on a multi-speaker, multi-language dataset and hold out 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.

This paradigm directly addresses one potential confounding factor: Suppose a purported contrast is actually not cross-linguistically valid. Then we might be able to spuriously predict it on test data that are very similar to the training data (e.g. a disjoint subset of utterances from the same corpus used for training). One real way in which a contrast might be spurious is due to mislabeling: suppose we label /i/ as a front vowel in one language, but accidentally mislabel it as a back vowel in another language. If our training data and classification model are sufficiently rich, the classifier might in effect learn to perform language identification and correctly reproduce the inconsistent labeling by identifying the language of a test item, despite the fact that the labels are not meaningful. The risk of this is substantially reduced because we always ensure that the set of languages (as well as the set of talkers) is disjoint between training and cross-lingual testing.

This setup cannot control for all potential confounding factors though. It is certainly possible, even likely, that our data resources are suboptimal, that language-specific processing such as automatic time-alignment introduces confounding artifacts, or that the choice of model or training procedure is deficient. The question we are trying to answer is situated within current practice: given existing transcribed corpora and existing phoneme databases, which phonemic contrasts can be predicted consistently across languages? We hold the data resources constant, use standard optimizations to train classifiers, and focus on the choice and nature of the phonological features.

A known, subtle confounding factor is due to well-known mismatches in how different languages group allophones under different phonemes. For example, aspiration is contrastive in Bengali, but not in Spanish. In Bengali the phoneme /p/ (unaspirated) contrasts with an aspirated phoneme, which has [pʰ] and [f] as allophones (our Bengali corpus uses /f/ as the phoneme label). In Spanish, the phoneme /p/ is unmarked for aspiration and could be realized as [pʰ], which contrasts with the phoneme /f/. That means in a given multilingual dataset we may find [f] and [pʰ] sounds labeled differently depending on language, because we are working with phonemic rather than phonetic transcriptions. We will return to this example later.

Table 3: *Phonological contrasts and corresponding phonemes.*

Contrast	Corresponding Phonemes
Close vowel	i iː u uː
Non-close vowel	e eː æ æː o oː ɔ ɤ ə əː a aː
Front vowel	e eː æ æː i iː
Back vowel	o oː ɔ ɤ u uː
Retroflex	ɖ ɖʰ ɭ ɳ ɻ ʂ ʈ ʈʰ ᶯɖ
Non-retroflex	d ɡ̤ ɡ̤ʰ l lʰ n n̤ n̤ʰ r ɾ ʃʷ t ʈ̤ ʈ̤ʰ ᶯɡ̤
High specified	a aː e eː ẹ o oː o̤ æ æː ɔ ə əː ɤ i iː i̤ j k kʰ u uː ṳ w x ŋ g gʰ ɲ
High unspec.	b bʰ d dʒ ɡ̤ ɡ̤ʰ f h l lʰ m mʰ n n̤ n̤ʰ p r s t tʃ tʃʰ ʈ ʈ̤ʰ z ɖ ɖʰ ...
Delayed release	dʒ f s tʃ tʃʰ x z ʂ ʃ
Non-delayed rel.	l lʰ m mʰ n n̤ n̤ʰ r ɾ ŋ ɭ ɲ ɳ ɻ ɾ b bʰ d ɡ̤ ɡ̤ʰ k kʰ p t ʈ̤ ʈ̤ʰ ɖ ɖʰ ...
Anterior	b bʰ d dz dzʰ ɡ̤ ɡ̤ʰ f l m n n̤ p pʰ r s t ts tʰ ʈ ʈ̤ʰ z ɾ ʊ
Non-anterior	dʒ dʒʰ h j k kʰ tɕ tɕʰ tʃ tʃʰ w x ŋ ɖ ɖʰ g gʰ ɭ ɲ ɳ ɻ ʂ ʃ ʈ ʈʰ ʔ

While accurate feature detectors have been built in the past [20, 14, 21], there has been relatively little focus in applying such detectors to evaluation of phonological representation itself. Such an approach not only can be used to improve the quality of the existing cross-lingual linguistic resources but ultimately also for deriving segment inventories for languages where no such resources exist [22].

Corpus details We selected ten languages from three distinct language families for our experiments. The languages and their families are shown in Table 1 along with their corresponding BCP-47 codes [23]. The majority of languages in the mix correspond to South Asia and span two very different language families, namely Indo-Aryan and Dravidian. These 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 [24]. The third, Malayo-Polynesian, group is useful for contrasting the performance of its languages (Javanese and Sundanese) with the languages of South Asia. These multi-speaker multi-gender datasets were collected in the past as part of an ongoing effort to build multi-speaker corpora for low-resource languages in an affordable way and were used for extensive experimentation [15, 25, 26]. The datasets consist of 48 kHz audio and the corresponding hand-curated transcriptions.

Our phoneme inventories have been designed with multilingual speech applications in mind. The South Asian languages use a unified underlying phonological representation. We leverage this unification to make the most of the data we have and eliminate scarcity of data for certain phonemes by conflating similar phonemes into a single representative phoneme [15]. The inventories for Malayo-Polynesian languages also aim for maximum degree of overlap to reflect the close similarity between Javanese, Sundanese and their larger relative Indonesian [26]. The phoneme inventories for all languages use International Phonetic Alphabet (IPA) notation [27] 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 our inventories do not necessarily map to the available typological resources (e.g., PHOIBLE) one-to-one, we made sure that there is a significant correlation between them.

Phonological features Each phoneme segment shown in Table 2 can be represented as a bundle of phonological features. Once the phoneme inventory has been specified in IPA, cross-lingual typological resources can be used to decompose it into a particular feature system. PHOIBLE contains 2160 distinctly notated sounds from 1672 languages, and its feature system consists of 37 ternary features [16]. PANPHON relates over 5,000 distinct sound notations to representations in terms of about 23 binary articulatory features [17]. Decompositions like these can

Table 4: *Spanish–Bengali phoneme asymmetry.*

	P - F		VOICED	
	bn	es	bn	es
bn	**0.4**	2.9	**1.1**	9.2
es	1.6	**0.1**	11.3	**0.3**

be accessed using automated tools [28]. A ternary feature (such as LABIODENTAL) in PHOIBLE can either be present (+), absent (−), or not applicable (∅).

3. Experiments, results, and discussion

Each of the phonological feature contrasts can be represented by two sets of phonemes, one for which the feature is present, and one where it is absent. Table 3 shows a list of phoneme groups, together with the corresponding phonemes selected from our corpus, to study such contrasts. For the binary 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.

Experiment setup The speech data was downsampled to 16 kHz and then parameterized into HTK-style Mel Frequency Cepstral Coefficients (MFCC) [29] using a 10 msec frame shift. The dimension of the MFCC parameters is 39 (13 static + Δ + ΔΔ coefficients).[2] To determine the phoneme time boundaries, the acoustic parameter sequences were force-aligned with the transcriptions [31]. A single training example consists of 40 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 training and evaluation sets in our experiments always consist of disjoint sets of languages and speakers. While the recordings for each language are multi-speaker and often multi-gender, our experiments mostly used female multi-speaker data. 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 [32, 33]. If enough examples are available, we sample 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 [34], further tuning the model hyper-parameters for maximizing the AUC using Vizier [35]. A simple two-layer architecture with 200 Softplus [36] units in each layer, dropout probability of 0.2 [37], Adadelta optimizer [38] and the learning rate of 0.6 with a large batch size of 6000 [39] were found to perform well across our experiments.

Evaluation results and discussion For each classification, we measure the area under the ROC curve (AUC) numbers for every pair of training and evaluation languages, including a language against itself. We have also trained models for each language family and then evaluated that model on languages outside that family. We also have measured every feature defined in PANPHON and PHOIBLE with positive and negative classes indicating the presence and absence of that feature. All these configurations make up around 900+ experiments. Each experiment is repeated multiple times to ensure reliability of the results and median values are reported.

[2] Admittedly, the use of MFCCs is too restrictive: Other representations (e.g., F0) or combinations thereof may be better suited to model the acoustic cues that signal the contrasts [30] in each scenario.

Table 5: *Close vs. non-close vowels.*

	INDO-ARYAN				DRAVIDIAN				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
bn	**3.1**	9.8	9.7	13.0	10.9	15.6	12.4	9.6	13.6	4.9
gu	6.4	**2.3**	6.0	10.4	8.2	15.6	12.4	7.8	11.1	2.3
mr	6.9	6.9	**1.4**	11.3	8.7	15.5	13.1	8.5	10.4	2.7
si	9.3	11.2	11.8	**3.1**	12.4	14.9	15.2	10.2	16.3	3.3
kn	6.8	8.5	6.4	8.9	**1.9**	10.6	7.2	5.0	9.9	1.8
ml	7.3	11.0	8.5	9.4	5.4	**4.9**	7.9	4.9	12.0	2.6
ta	7.2	11.5	8.4	12.8	5.5	10.1	**3.0**	4.5	10.8	3.0
te	7.1	9.5	8.0	9.6	5.3	10.6	7.9	**2.0**	10.4	3.0
jv	8.3	9.2	8.0	11.4	7.7	15.9	13.1	7.5	**4.2**	1.9
su	10.0	8.9	8.5	10.8	7.9	17.0	12.5	8.2	10.3	**0.3**

Table 6: *Front vs. back vowels.*

	INDO-ARYAN				DRAVIDIAN				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
bn	**0.6**	0.5	0.8	2.2	0.8	6.4	8.7	2.2	0.9	2.2
gu	1.9	**0.3**	1.5	1.7	0.7	4.7	5.5	2.3	1.0	1.7
mr	2.9	1.4	**0.3**	5.5	2.6	9.7	11.5	4.8	1.9	4.8
si	4.0	1.1	2.1	**0.4**	1.8	6.7	8.4	3.2	2.9	2.2
kn	1.6	0.4	1.0	1.6	**0.5**	4.6	5.0	1.7	0.7	2.7
ml	1.9	0.4	1.2	1.9	0.7	**1.9**	4.9	1.9	0.9	2.4
ta	2.5	0.6	1.8	1.6	1.0	3.8	**1.6**	2.6	1.0	1.3
te	1.7	0.5	1.0	1.5	0.6	4.4	6.1	**0.9**	0.8	2.6
jv	1.1	0.5	0.8	1.3	0.6	4.4	5.6	2.1	**0.4**	1.7
su	2.7	0.7	1.7	1.4	1.4	6.0	7.0	3.0	1.0	**0.2**

For a brief illustration we return to the contrasts among labial consonants in Bengali and Spanish. Both languages have phonemes that are labeled /f/ and /p/, but as discussed earlier [pʰ] is an allophone of /f/ in Bengali and an allophone of /p/ in Spanish. Table 4 shows the cross-linguistic consistency of two contrasts: the 'p–f' contrast only distinguishes between the phonemic labels /p/ and /f/. We can see that this contrast is robust between Bengali and Spanish, despite the conflicting status of the allophone [pʰ]. On the other hand, voicing is contrastive in both languages, but cannot be predicted as reliably across these two languages.

For the height contrast among vowels, we place vowels marked +HIGH in both PANPHON and PHOIBLE into the class 'close vowel', and those with −HIGH into 'non-close vowel'. As vowel height is strongly associated with F_1 frequency it should be easily predictable within one language. Boundaries in vowel space are known to differ across languages, on top of natural variation. This is borne out by the results in Table 5.

The 'front–back' contrast 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]. The results in Table 6 show a much better cross-language consistency compared with vowel height, pre-

Table 7: *Retroflex vs. non-retroflex consonants.*

	Indo-Aryan				Dravidian				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
I-A	20.7	14.3	8.3	8.4	22.6	28.1	25.6	21.2	20.9	0.0
Drav	34.4	13.9	9.2	11.5	7.8	9.4	13.2	9.3	27.2	0.0
M-P	22.6	49.5	51.3	25.2	51.7	50.0	45.9	40.1		0.0

Table 8: *High specified (±) vs. unspecified (∅).*

	Indo-Aryan				Dravidian				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
bn	**1.5**	6.1	5.5	6.1	4.3	6.8	4.8	3.5	9.2	6.6
gu	3.8	**2.8**	3.8	4.0	4.2	5.7	5.1	2.4	10.1	7.3
mr	3.7	4.6	**2.1**	4.3	4.4	5.8	5.4	2.6	10.3	7.8
si	7.2	6.0	5.6	**1.7**	4.9	6.3	6.4	3.5	11.7	8.8
kn	3.4	5.3	4.4	3.5	**1.7**	4.7	4.2	2.2	8.5	6.4
ml	3.5	5.4	4.7	4.2	3.2	**2.7**	3.6	2.3	9.1	7.3
ta	3.5	5.5	5.0	4.3	3.4	4.5	**1.8**	2.2	9.7	7.1
te	3.4	5.2	4.2	4.0	3.6	4.5	3.8	**1.2**	9.8	7.8
jv	4.3	6.9	6.7	6.5	5.1	6.8	6.2	4.3	**4.0**	4.1
su	4.3	8.3	7.9	6.5	5.9	8.4	6.6	5.5	6.2	**1.6**

sumably because central vowels are excluded here.

The concept of retroflex consonants is expressed indirectly via the PHOIBLE features [−ANTERIOR, −DISTRIBUTED]. The union of thus defined retroflex phonemes from all ten languages under study forms the positive class, along with a manually added Sinhala prenasalized stop /ⁿɖ/ [40]. The negative class consists of all the dental and alveolar phonemes corresponding to their retroflex counterparts in the positive class. For example, alveolar /d/ is a non-retroflex phoneme contrasting with retroflex /ɖ/. Due to data sparsity, we trained on groups of languages from the same family. Results appear in Table 7. Among the Malayo-Polynesian languages considered, only Javanese has retroflex consonants, which it acquired through loanwords from Indo-Aryan or Dravidian languages. When training on these languages we get mostly random results (50% AOC). Even within the other two language groups, this contrast is generally not very consistent. The absence of retroflex stops in Sundanese is however predicted perfectly.

The class 'high-specified' is comprised of phonemes with PHOIBLE feature ±HIGH (either positive or negative). The corresponding 'high-unspecified' negative class consists of all the phonemes with PHOIBLE feature ∅HIGH (not applicable). This contrast, unlike the others, partitions the union of the phoneme inventories for all languages in this study. We chose this contrast specifically with the expectation that it would not be very meaningful: the positive class is heterogeneous, consisting of vowels and velar consonants. Contrary to expectations, Table 8 shows that this contrast is very robust across languages.

The 'delayed release' contrast only applies to the consonants. Those with the PHOIBLE feature +DELAYED RELEASE are assigned to the positive class, the remaining consonants to the negative class. This contrast is naturally heterogeneous, since it puts fricatives and non-lateral fricatives in the positive class, and all other consonants, including lateral fricatives, in the negative class. Table 9 shows it to be robust.

The 'anterior' contrast is defined by the opposition of PHOIBLE features +ANT vs. −ANT. This is based on the place of articulation of consonants: labials to alveolars are in the positive class, consonants further back than alveolars are in the negative class. This contrast is clear and homogenous, but difficult to predict, even within the same language, per Table 10.

The AUC% within a language can be interpreted as a measure of consistency of our data for a given language. We are not surprised to find that the less consistent a contrast is within a

Table 9: *Delayed release vs. non-delayed release consonants.*

	Indo-Aryan				Dravidian				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
bn	**0.8**	1.1	1.6	0.7	1.3	1.5	0.3	0.9	1.1	2.2
gu	2.2	**0.6**	1.3	0.5	1.2	1.1	0.2	0.6	1.0	1.1
mr	2.3	1.0	**0.2**	0.8	1.4	1.2	0.5	0.8	1.1	1.4
si	3.0	1.7	2.6	**0.7**	1.4	1.5	0.7	0.9	1.3	2.1
kn	2.8	1.2	2.2	0.5	**0.7**	1.2	0.3	0.6	1.1	2.0
ml	2.6	1.3	1.9	0.5	1.2	**0.5**	0.5	0.8	1.5	1.9
ta	3.1	1.7	2.4	0.7	1.3	1.3	**0.4**	0.8	1.2	1.7
te	2.9	1.6	2.1	0.6	1.3	1.4	0.5	**0.7**	1.2	2.2
jv	4.1	1.9	2.5	0.7	1.6	2.2	0.8	0.8	**0.7**	2.2
su	3.4	1.5	2.0	0.7	1.3	1.3	0.5	0.9	1.0	**0.3**

Table 10: *Anterior vs. non-anterior.*

	Indo-Aryan				Dravidian				M-P	
	bn	gu	mr	si	kn	ml	ta	te	jv	su
bn	**4.6**	22.5	25.1	23.0	31.9	36.0	33.7	23.2	30.2	23.6
gu	14.5	**7.3**	13.4	10.8	18.5	20.7	22.3	13.5	29.2	27.0
mr	18.0	13.8	**4.4**	13.1	16.9	20.6	24.4	13.7	28.1	26.3
si	16.6	16.2	17.3	**2.8**	21.4	24.5	28.0	17.1	25.9	21.7
kn	18.6	11.7	12.6	11.5	**5.4**	15.2	18.9	12.4	27.7	24.8
ml	24.7	16.7	16.4	15.9	14.9	**8.6**	17.6	13.2	34.8	31.5
ta	18.6	15.9	18.2	15.9	15.6	15.5	**9.5**	13.1	35.2	31.9
te	16.7	12.5	13.5	11.3	13.2	13.7	15.4	**4.7**	30.2	27.9
jv	25.7	23.4	23.2	15.7	29.3	33.5	36.7	26.3	**9.8**	11.7
su	24.4	24.6	25.2	17.7	32.7	37.2	37.6	29.3	15.5	**2.5**

langugage, the less consistent it generally is across languages. While these contrasts vary significantly in terms of average cross-linguistic consistency, many of them are equally robust between languages within the same family and across families. Some contrasts, for example the front–back distinction in vowels and the delayed release distinction are nearly universal. Even some heterogeneous feature contrast, such as 'high-specified', hold up well cross-linguistically. Some contrasts are inherently more difficult to predict. For example, distinguishing close vowels from non-close (including close-mid) vowels is naturally difficult. The 'anterior' feature from PANPHON is arguably homogenous, but divides the consonants along a boundary that runs between alveolars and post-alveolars, which turns out to be difficult to predict reliably. This justifies further research into the design of feature inventories that are highly consistent in multilingual settings.

4. Conclusion

In this study we have investigated an empirical approach to recognizing phonological contrasts both within a language and across the language family boundaries. We proposed a classifier that detects the presence (or lack) of a particular linguistic contrast given the sequence of acoustic frames. We demonstrated the effectiveness of the proposed approach on detecting several typical yet interesting phonological contrasts in languages from three distinct language families. In addition, the proposed approach serves as a tool for critically assessing the usefulness of individual articulatory features in a multilingual setting. Our approach can potentially form a building block of a system for inducing the phonological representations from the speech data. This, in turn, will facilitate the development of speech technologies in low-resource settings, where no such reliable representations are readily available.

5. Acknowledgements

The authors thank Isin Demirsahin for various phonology insights, Anna Katanova for helping with the experiments and the anonymous reviewers for many useful suggestions.

6. References

[1] G. N. Clements, *Contemporary Views on Architecture and Representations in Phonology*, ser. Current Studies in Linguistics. MIT Press, 2009, vol. 48.

[2] N. Chomsky and M. Halle, *The Sound Pattern of English.* New York: Harper & Row, 1968.

[3] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1872–1891, 2002.

[4] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[5] A. W. Black and T. Schultz, "Speaker clustering for multilingual synthesis," in *Proc. ITRW on Multilingual Speech and Language Processing*, Stellenbosch, South Africa, 2006.

[6] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *Proc. ICASSP 2018*. IEEE, 2018, pp. 4779–4783.

[8] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 555–566, 2000.

[9] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP 2002*, Denver, USA, Sep. 2002.

[10] R. A. Bates, M. Ostendorf, and R. A. Wright, "Symbolic phonetic features for modeling of pronunciation variation," *Speech Communication*, vol. 49, no. 2, pp. 83–97, 2007.

[11] S. Stüker and A. Waibel, "Porting speech recognition systems to new languages supported by articulatory feature models," *Speech and Computer, SPECOM*, 2009.

[12] B. E. Dresher, "The phoneme," in *The Blackwell Companion to Phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds. Wiley, 2011, ch. 11.

[13] M. Ostendorf, "Moving beyond the beads-on-a-string model of speech," in *Proc. IEEE ASRU Workshop*, 1999, pp. 79–84.

[14] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual Articulatory Features," in *Proc. ICASSP 2003*, vol. 1, 2003, pp. 144–147.

[15] I. Demirsahin, M. Jansche, and A. Gutkin, "A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech," in *Proc. SLTU 2018*, 2018, pp. 80–84.

[16] S. Moran, D. McCloy, and R. Wright, *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014, http://phoible.org/.

[17] D. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors." in *Proc. COLING 2016*, Japan, December 2016, pp. 3475–3484.

[18] Y. Tsvetkov, S. Sitaram, M. Faruqui, G. Lample, P. Littell, D. Mortensen, A. W. Black, L. Levin, and C. Dyer, "Polyglot Neural Language Models: A Case Study in Cross-Lingual Phonetic Representation Learning," *arXiv preprint arXiv:1605.03832*, 2016.

[19] E. M. Ponti, H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen, "Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing," *arXiv preprint arXiv:1807.00914*, 2018.

[20] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[21] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. ICASSP 2008*. IEEE, 2008, pp. 4261–4264.

[22] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proc. ICASSP 2014*. IEEE, 2014, pp. 2594–2598.

[23] A. Phillips and M. Davis, "BCP 47 – Tags for Identifying Languages," *IETF Trust*, 2009.

[24] M. Emeneau, "India as a Linguistic Area," *Language*, vol. 32, no. 1, pp. 3–16, 1956.

[25] C. Johny and M. Jansche, "Brahmic Schwa-Deletion with Neural Classifiers: Experiments with Bengali," in *Proc. SLTU 2018*, 2018, pp. 259–263.

[26] J. A. E. Wibawa, S. Sarin, C. F. Li, K. Pipatsrisawat, K. Sodimana, O. Kjartansson, A. Gutkin, M. Jansche, and L. Ha, "Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech," in *Proc. LREC 2018*, 7-12 May 2018, Miyazaki, Japan, 2018, pp. 1610–1614.

[27] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[28] A. Gutkin, M. Jansche, and T. Merkulova, "FonBund: A Library for Combining Cross-lingual Phonological Segment Data," in *Proc. LREC 2018*, Miyazaki, Japan, May 2018, pp. 2236–2240.

[29] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," in *Proc. SPECOM*, vol. 1, 2005, pp. 191–194.

[30] B. H. Repp, "Categorical Perception: Issues, Methods, Findings," in *Speech and Language: Advances in Basic Research and Practice*. Elsevier, 1984, vol. 10, pp. 243–335.

[31] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 2006.

[32] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[33] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[34] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016, pp. 265–283.

[35] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley, "Google Vizier: A Service for Black-Box Optimization," in *Proc. 23rd ACM SIGKDD 2017*, 2017, pp. 1487–1495.

[36] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–4.

[37] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1212.html#abs-1212-5701

[39] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489*, 2017.

[40] A. Wasala and K. Gamage, "Research report on phonetics and phonology of Sinhala," Language Technology Research Laboratory, University of Colombo School of Computing, Tech. Rep. 35, 2005, Working Papers 2004–2007.