The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech

Işın Demirşahin, Martin Jansche, Alexander Gutkin

Google AI, London, United Kingdom

{isin,mjansche,agutkin}@google.com

Abstract

We present a multilingual phoneme inventory and inclusion mappings from the native inventories of several major South Asian languages for multilingual parametric text-to-speech synthesis (TTS). Our goal is to reduce the need for training data when building new TTS voices by leveraging available data for similar languages within a common feature design. For West Bengali, Gujarati, Kannada, Malayalam, Marathi, Tamil, Telugu, and Urdu we compare TTS voices trained only on monolingual data with voices trained on multilingual data from 12 languages. In subjective evaluations multilingually trained voices outperform (or in a few cases are statistically tied with) the corresponding monolingual voices. The multilingual setup can further be used to synthesize speech for languages not seen in the training data; preliminary evaluations lean towards good. Our results indicate that pooling data from different languages in a single acoustic model can be beneficial, opening up new uses and research questions.

Index Terms: multilingual text-to-speech, South Asian languages, phonology

1. Introduction

Learning to speak is a significant investment. It takes considerable time and resources to bootstrap a text-to-speech (TTS) voice in a given language for the first time. Training TTS acoustic models requires nontrivial amounts of recorded speech in each language. When bootstrapping multiple languages, a naive approach that treats all language-specific work as independent results in additive resource requirements, which scales linearly in the number of languages. Can we do better?

Our work on TTS is part of a larger effort to build out language and speech technology for many diverse languages. Research and development on any single language does not happen in isolation, but in the context of past and/or ongoing work on other languages. With TTS models already available in several languages, work on additional languages does not have to be an independent effort which starts from zero. Instead, we want to be able to leverage existing data.

Parametric approaches to TTS, mediated by explicit acoustic models, are arguably less brittle than nonparametric concatenative approaches. In previous work we have built parametric TTS voices from monolingual multi-speaker data [1] and from multilingual multi-speaker data which combines closely related languages with highly overlapping phoneme inventories [2]. Our present goal is to pool training data from languages with somewhat less overlap in their sound systems. We focus on languages of South Asia, spanning two very different language families, namely Indo-Aryan [3] and Dravidian [4]. Table 1 shows a superset of the languages we are concerned with, including the variety of orthographies used, as well as modern words related to the Sanskrit word for *culture*. These languages exhibit considerable variation within each group, but also such similarities across groups that they are often considered a textbook example of the concept of *Sprachbund* [5]. The phonological variation is nontrivial and needs to be modeled in some fashion in any multilingual approach.

Table 1: *Descendants of* saṁskṛti (culture) *across languages.*

Language	Orthography	Transliteration	Pronunciation
Bengali	সংস্কৃতি	saṁskṛti	ʃɔŋskriti
Gujarati	સંસ્કૃતિ	saṁskṛti	sənskruti
Hindustani	سنسکرتی संस्कृति	snskrty saṁskṛti	sənskriti:
Kannada	ಸಂಸ್ಕೃತಿ	saṁskṛti	samskruti
Malayalam	സംസ്കൃതി	saṁskṛti	samskrədi
Marathi	संस्कृती	saṁskṛtī	sənskruti
Nepali	संस्कृति	saṁskṛti	səmskriti
Odia	ସଂସ୍କୃତି	saṁskṛti	ʃɔŋʃkruti
Punjabi	سنسکرتی ਸੰਸਕ੍ਰਿਤੀ	snskrty saṃsakritī	sənskriti:
Sinhala	සංස්කෘතිය	saṁskṛtiya	saŋskrutiyə
†Tamil	சமசுகிருதம்	camacukirutam	samasukirudam
Telugu	సంస్కృతి	saṁskṛti	samskruti

†Tamil does not use a cognate of saṁskṛti to express the concepts of *culture* or *civilization*. The Tamil word given above is the name of the Sanskrit language.

In a training dataset for TTS acoustic models a labeled example is a transcribed utterance, i.e., a tuple consisting of a linguistic transcription, a speech waveform, as well as metadata about the speaker and language. Here we are mainly concerned with the input features of the acoustic model, which are derived from the linguistic transcription and from metadata. In our case the transcription is simply a sequence of phonemes in each language. To arrive at a uniform representation of input features for the acoustic model, all transcriptions are projected into the International Phonetic Alphabet [6], and these intermediate IPA transcriptions are further mapped to broadly articulatory feature vectors that form part of the model inputs, using available phonological resources [7].

To allow the acoustic models to capture shared regularities across languages, we argue that the IPA representation of transcriptions should be matched in a deliberate way. For example, the Dravidian languages and Sinhala[1] tend to have at least word-internal length contrasts among most or all of their vowels. In Malayalam, say, there is a robust distinction between short /e/ and long /eː/. Among the Indo-Aryan languages considered here, a comparable contrast is only present in Sinhala. It is absent from Hindi, Bengali, etc. This raises the question of how the [e]-like vowel of Hindi should be transcribed: In a multilingual speech corpus, do we want the Hindi vowel to correspond to Malayalam /e/, or to /eː/, or neither, or some more

[1]Though of Indo-Aryan origin, Sinhala tends to behave phonologically more like its long-time Dravidian neighbor Tamil than, say, its distant Indo-Aryan relative Bengali.

10.21437/SLTU.2018-17

complex option? The IPA conventions [6, p. 159–60] would encourage us to transcribe the Hindi phoneme as /e/, since there is no applicable length contrast in Hindi. However, as we shall see below, we decided to transcribe this Hindi phoneme as /eː/ instead, since it tends to pattern more with the Hindi-internal long vowels (there is a word-internal length contrast in the high vowels, for example) as well as with the long mid vowels of the languages that have a length contrast here.[2]

The novelty of this work is in the phonological inventory design (described in detail in Section 2) that is significantly more phonemic than some of the existing orthography-influenced work [9].

2. Inventory design

Our aim is to combine datasets from similar languages while training a single acoustic model to minimize the data required to synthesize new languages. Findings in [10, 11, 2] demonstrate that joint training with other languages outperforms monolingual systems. In order to be able to use the data from different languages we need a unified underlying representation. In this work, 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 in the multilingual phoneme inventory.

Shared inventories Where possible, we use the same inventory for phonologically close languages. For example we use an identical phoneme inventory for Telugu and Kannada (te-IN in Table 2), and West Bengali and Odia (bn-IN in Table 2). The phoneme inventory for Gujarati is identical to Marathi, (mr-IN in Table 2) except that the breathy sonorants in Marathi (/nʰ/[3], /mʰ/, /jʰ/, /lʰ/, /ʊʰ/) are not in the phoneme inventory for Gujarati.

Consonants In almost all of the languages in this work there was no phonological contrast between /pʰ/ and /f/. We choose /f/ as the default representation for this sound since the /f/ sounds in English loan words are almost always pronounced as [f], whereas the /pʰ/ sounds in native words can be pronounced as either [pʰ] or [f]. Exceptions are Urdu and West Bengali where there is a claimed contrast between the two sounds, so this language has both phonemes in its inventory. Urdu in particular has a clear orthographic distinction between پھا /pʰaː/ and فا /faː/.

All these languages make a distinction between an alveolar sibilant fricative /s/ and a post-alveolar, palatal, or retroflex sibilant fricative. We use /ʃ/ as the generic post-alveolar sibilant fricative for the multilingual inventory, regardless of the canonical representation for the sound in the literature. The languages that have a phonological contrast between a palatal/post-alveolar /ʃ/ and a retroflex fricative have an additional /ʂ/. As none of these languages have a phonological contrast between alveolo-palatal /ɕ/ and a post-alveolar /ʃ/, we conflated these into the generic post-alveolar fricative /ʃ/.

All languages in this work have a labial nasal /m/, and most have a dental nasal.[4] For the remaining nasals, the phoneme inventories are compiled based on the frequency in the data and the contrast perceived by the native speakers, instead of the orthography or the textbook definitions of the phonemes. Specifically, the velar nasal /ŋ/ was added to all phoneme inventories because the data for this phoneme is robust thanks to English loanwords. Conversely, the putative aspirated retroflex nasal /ɳʰ/ in Marathi was conflated with the aspirated dental nasal /nʰ/ because, although the former phoneme may theoretically exist, it is very rare, and not attested in our dataset.

Alveolar stops /t̪/ and /d̪/ are only phonologically contrastive in Malayalam. Although the voiced alveolar stop [d̪] may arguably be observed in Tamil in the /n̪ r/ sequences, pronounced as [n̪ d̪ r], since the distribution is complementary, we can treat [d̪ r] as an allophone of /r/.

Diphthongs Diphthongs pose several interrelated questions for our set of languages. Unfortunately their phonological status is often obscured by orthography: The Brahmic scripts tend to inherit letters for only the two historical Sanskrit diphthongs, and Perso-Arabic script leaves many vowels and diphthongs systematically ambiguous. The phonological reality is often much more complex than the orthography would suggest. Masica [3,

Table 2: *Unified phoneme inventory.*

mul	bn BD	bn IN	ml IN	mr IN	si LK	ta IN	te IN	ur PK
ə		✓	✓	✓	✓			✓
əː					✓			
a			✓		✓	✓	✓	
aː	✓	✓	✓	✓	✓	✓	✓	✓
æ	✓	✓	✓	✓	✓		✓	
æː					✓			
i	✓	✓	✓	✓	✓	✓	✓	✓
iː			✓		✓	✓	✓	✓
u	✓	✓	✓	✓	✓	✓	✓	✓
uː			✓		✓	✓	✓	✓
e			✓		✓	✓	✓	
eː	✓	✓	✓	✓	✓	✓	✓	✓
ɛ								✓
o			✓		✓	✓	✓	
oː	✓	✓	✓	✓	✓	✓	✓	✓
ɔ	✓	✓		✓				✓
i̪	✓	✓						
e̪	✓	✓						
u̪	✓	✓						
o̪	✓	✓						
~		✓						✓
k	✓	✓	✓	✓	✓	✓	✓	✓
q								✓
kʰ	✓	✓		✓			✓	✓
x								✓
g	✓	✓	✓	✓	✓	✓	✓	✓
gʰ	✓	✓		✓			✓	✓
ɣ								✓
ŋ	✓	✓	✓	✓	✓	✓	✓	✓
ᵑg					✓			
tʃ	✓	✓	✓	✓	✓	✓	✓	✓
ts				✓				
tʃʰ	✓	✓		✓			✓	✓
dʒ	✓	✓	✓	✓	✓	✓	✓	✓
dz				✓				
z		✓						✓
dʒʰ	✓	✓		✓			✓	✓
dzʰ				✓				
ʒ								✓
ɲ			✓		✓	✓	✓	

mul	bn BD	bn IN	ml IN	mr IN	si LK	ta IN	te IN	ur PK
ʈ	✓	✓	✓	✓	✓	✓	✓	✓
ʈʰ	✓	✓		✓			✓	✓
ɖ	✓	✓	✓	✓	✓	✓	✓	✓
ɽ		✓						✓
ɖʰ	✓	✓		✓			✓	✓
ɽʰ								✓
ɳ			✓	✓		✓	✓	
ⁿɖ					✓			
ʈ̪			✓					
ɖ̪			✓					
n̪			✓			✓		
t	✓	✓	✓	✓	✓	✓	✓	✓
tʰ	✓	✓		✓			✓	✓
d	✓	✓	✓	✓	✓	✓	✓	✓
dʰ	✓	✓		✓			✓	✓
n	✓	✓	✓	✓	✓	✓	✓	✓
nʰ				✓				
ⁿd					✓			
p	✓	✓	✓	✓	✓	✓	✓	✓
pʰ		✓						✓
f	✓	✓	✓	✓	✓	✓	✓	✓
b	✓	✓	✓	✓	✓	✓	✓	✓
bʰ	✓	✓		✓			✓	✓
m	✓	✓	✓	✓	✓	✓	✓	✓
mʰ				✓				
ᵐb					✓			
j	✓	✓	✓	✓	✓	✓	✓	✓
jʰ				✓				
r	✓	✓	✓	✓	✓	✓	✓	✓
rʰ				✓				
ɾ			✓			✓		
l	✓	✓	✓	✓	✓	✓	✓	✓
lʰ				✓				
ɭ			✓	✓		✓	✓	
ɻ			✓			✓		
ʊ		✓	✓	✓	✓	✓	✓	✓
ʊʰ				✓				
ʃ	✓	✓	✓	✓	✓	✓	✓	✓
ʂ			✓				✓	
s	✓	✓	✓	✓	✓	✓	✓	✓
h	✓	✓	✓	✓	✓	✓	✓	✓

[2] This raises a foundational question of what we are describing here. The notion of a *phoneme* has had many subtly or starkly different historical meanings [8]. We generally aim for a broad phonemic level of transcription within each language, but in a cross-language setting most notions of phoneme are not meaningful. Our goal here is to unify cross-language sound classes in terms of phonetic similarity. The resulting units are still used for broad transcriptions within each language, hence we continue to refer to them as *phonemes* for simplicity.

[3] /ɳʰ/ is conflated with /nʰ/ as the former is very rare.

[4] Since the prototypical nasals and stops in these languages are dental rather than alveolar, we use unmarked symbols for dental, e.g. /n/ and use the retrograde diacritic to represent the alveolars, as in /n̪/.

p. 115f.] describes the New Indo-Aryan diphthongs in great detail. We use Bengali for a brief illustration.

Bengali has a large number of diphthongs that give rise to a diversity of opinions in the published literature about their precise number and nature (estimates range from 8 to 38). There is considerable agreement in the literature that only non-syllabic versions of /i/, /u/, /e/, and /o/ can appear as off-glides in diphthongs (the height contrast in the Bengali mid vowels is neutralized). We therefore have semi-vowels /i̯/, /e̯/, /u̯/, and /o̯/ in our unified inventory.[5] We transcribe the Bengali diphthongs as a combination of one of the 7 nuclear vowels followed by an off-glide, for a theoretical maximum number of $7 \times 4 = 28$ falling diphthongs.[6] In particular we can transcribe all 16 diphthongs mentioned in [12]. By transcribing diphthongs as vowel-plus-glide sequences, we keep the basic inventory of symbols small and leave open the possibility of discovering unanticipated combinations in the course of transcribing the recordings or lexicon.

Nasal vowels Several Indo-Aryan languages (including Hindustani and West Bengali) employ contrastive nasalization among vowels. Adding separate units for nasal vowels would potentially double the vowel inventory; instead we adopt a non-standard nasalization marker /~/. A phoneme sequence such as /u:~/ indicates a nasalized vowel sound [ũ:], for example in Hindi जूँ /ʤu:~/, which contrasts with जून /ʤu:n/.

Vowel length In Marathi and Bengali vowel length is not phonologically contrastive, so a single phoneme was used for Marathi इ, ई /i/ and उ, ऊ /u/. In some languages vowel length is neutralized under certain conditions, e.g. for some word-final vowels in Hindustani and Sinhala. We chose to use short vowel sounds in all of those situations, but this is a preliminary decision that should be investigated empirically.

In other situations we have opted to transcribe vowels without a length contrast as long. This is generally the case for Indo-Aryan /a:/, /e:/, and /o:/, corresponding to the independent letters आ, ए, and ओ in Hindi and Marathi. This is especially appropriate in Marathi where the additional vowels ॲ /æ/ and ऑ /ɔ/ are short.

Related work Our approach shares the goals and desiderata of recent work by Ramani and others [9], in particular the need for a common phoneset and feature encoding. While Ramani et al. target six major languages using a label set designed for 13 languages of India, the set of languages we are dealing with is more diverse and extends beyond India. Crucially, our unified inventory is much more phonemic compared with the heavily spelling-influenced approach of [9]. For example, Ramani et al. transcribe Bengali ণ the same as Marathi retroflex ण /ɳ/, but phonemically Bengali has no retroflex nasal [12]: both ণ and ন are read as /n/ (dental). The same could be said about Hindustani: Urdu has no letter for a retroflex nasal at all (unlike Sindhi, which uses ڻ); and though Hindi has retained the letter ण, it has arguably lost the phoneme /ɳ/.

None of the languages under discussion have "vocalic r" as a phonemic unit. Sanskrit ⟨ऋ⟩ has reflexes in most modern Brahmic scripts (except Tamil), but is usually read /ri/, /ru/, or even /ir/ or /ur/. For us these are simply sequences of ordinary phonemes, whereas Ramani et al. use a special symbol for the various vocalic-r graphemes, which systematically conflates words on the basis of graphemic similarity, despite pronunciation differences: consider the diverging pronunciations of the common prefix ⟨vr̥tt⟩ in the word वृत्त /ʋritt/ in Hindi, వృత్తము /ʋruttamu/ in Telugu, and වෘත්තය /ʋurtəjə/ in Sinhala. Recovering from an over-eager conflation is difficult during acoustic training, because the source grapheme is relatively rare.

Cross-linguistically, Ramani et al. transcribe Hindi and Marathi अ as /a/ and unify that with the /a/ sound of Dravidian. That correspondence is not phonetically accurate and creates problems in our larger set of languages, in particular for Sinhala which uses both /a/ and /ə/. We transcribe Hindi and Marathi अ as /ə/ and identify it with Sinhala /ə/, whereas Sinhala /a/ matches Dravidian /a/ and has no counterpart in Hindi, Marathi, or Bengali.

[5] Because of the presence of the additional non-high semivowels /e̯/ and /o̯/, we write /i̯/ rather than IPA [j], and /u̯/ rather than [w], but treat each pair as interchangeable. We avoid using /j/ in transcriber-facing tools, since it is too easily confused with the letter ⟨j⟩ which stands for the sound /ʤ/ in many customary transliteration schemes.

[6] Published inventories with more than 28 diphthongs include rising diphthongs involving on-glides in addition.

Table 3: *Language dataset details.*

Language	Code	Male		Female	
		N	*S*	*N*	*S*
Bengali	bn-IN	5,000	10	6,500	23
Gujarati	gu-IN	2,053	18	2,219	18
Kannada	kn-IN	2,214	36	2,186	24
Malayalam	ml-IN	2,023	18	2,103	24
Marathi	mr-IN	4,800	1	5,600	11
Tamil	ta-IN	1,956	25	2,335	25
Telugu	te-IN	2,154	23	2,294	24
Urdu	ur-PK	4,000	1	4,500	1

Table 4: *Additional datasets used in multilingual training.*

Language	Code	Gender	*N*	*S*	Public
Bengali (Bangladesh)	bn-BD	male	1,892	6	✓
Hindi	hi-IN	female	>4,000	1	
Nepali	ne-NP	female	2,065	18	✓
Sinhala	si-LK	female	2,065	12	✓

3. Experiments

3.1. Dataset details

Summaries of datasets for 16 language-plus-gender combinations from Indo-Aryan and Dravidian families are shown in Table 3 alongside the BCP 47 language and region code [13], the number of utterances (*N*) and the number of speakers (*S*). The recordings for most of the languages have been crowd-sourced using the methodology and equipment described in [2]. The only exceptions are the two Urdu speakers, and two West Bengali and two Marathi speakers who were recorded in a professional studio. Additional datasets that we used in our experiments are shown in Table 4 alongside the gender codes, number of utterances and number of available speakers. They correspond to four additional languages, which we include during the training of the multilingual acoustic model. We hypothesize that the presence of these languages will benefit the model. Out of the four datasets, the Bangladeshi Bengali, Nepali and Sinhala are crowd-sourced databases that we previously made publicly available[7]. The audio for all the speakers was recorded at 48 kHz. There is a representational imbalance between the Indo-Aryan and Dravidian language families in our training data. Out of the twelve available languages only four belong to the Dravidian family (Kannada, Malayalam, Tamil and Telugu). The Indo-Aryan languages also dominate in terms of the number of training utterances.

[7] http://www.openslr.org/resources.php

Table 5: *Subjective Mean Opinion Scores (MOS).*

Language	Male		Female	
	Mono-	Multilingual	Mono-	Multilingual
Bengali	3.83 ± 0.09	**3.94** ± 0.08	3.40 ± 0.13	**3.63** ± 0.09
Gujarati	3.34 ± 0.11	**3.77** ± 0.09	2.97 ± 0.11	**3.37** ± 0.11
Kannada	3.75 ± 0.12	**3.90** ± 0.10	3.77 ± 0.11	**3.79** ± 0.10
Malayalam	**4.14** ± 0.11	4.10 ± 0.12	3.89 ± 0.14	**4.03** ± 0.12
Marathi	**4.32** ± 0.09	4.25 ± 0.11	3.63 ± 0.14	**3.68** ± 0.14
Odia		3.80 ± 0.09		
Tamil	3.45 ± 0.12	**3.61** ± 0.10	3.36 ± 0.11	**3.62** ± 0.10
Telugu	3.01 ± 0.11	**3.79** ± 0.08	3.72 ± 0.10	**4.15** ± 0.06
Urdu	3.92 ± 0.11	**4.03** ± 0.12	**3.66** ± 0.13	3.63 ± 0.13

3.2. System configurations and evaluation methodology

Given the language datasets described above, we experiment with constructing parametric text-to-speech systems for the eight languages shown in Table 3. Each system consists of a linguistic front-end followed by LSTM-RNN acoustic model, the technical details of which are described in [11]. For each voice (language plus gender), we experiment with two scenarios: In the first scenario, we construct monolingual configurations, one for each gender, based on the available monolingual corpus only. In the second scenario, we construct multilingual configurations, one for each gender, that employ a single multilingual acoustic model trained using the data from all the speakers, genders and languages displayed in Tables 3 and 4. This results in four configurations for each language. Because most of our corpora are multi-speaker, the speaker identity input feature that guides the acoustic model during synthesis was selected with the help of native speakers.

In addition, we constructed a configuration for Odia, for which we have built a linguistic front-end but have no audio recordings available. In this experiment we investigate how well Odia can be synthesized using only the data from other languages. The speaker feature for Odia was selected to correspond to our best Marathi male speaker. In this study we only investigated this particular Odia male configuration.

Each of the configurations was evaluated by a rater pool of native speakers using subjective Mean Opinion Score (MOS) listening test. While the background of the listeners was previously shown to affect their perception [14], for reasons of anonymity no background information about the native speakers is available to us. For each test we used 100 sentences not included in the training data. Each rater was asked to evaluate a maximum of 100 stimuli. Each item was required to have at least 5 ratings. The raters used headphones. After listening to a stimulus, the raters were asked to rate the naturalness of the stimulus on a 5-point scale (1: "Bad", 2: "Poor", 3: "Fair", 4: "Good", 5: "Excellent"). Each participant had one minute to rate each stimulus. The rater pool for each language included at least 5 raters. For each language, all configurations were evaluated in a single experiment.

3.3. Results and analysis

Table 5 shows the results of subjective listening tests for 17 voices (16 voices for the eight core languages plus a male voice for held-out Odia) where, for each language plus gender, a monolingual and a multilingual configuration described in the previous section was tested. Each mean opinion score is shown along with the corresponding confidence interval statistics at 95% confidence level [15] computed using the recommendations in [16]. The highest scores for each language plus gender are shown in bold.

As can be seen from the results, the multilingual male configurations for most of the languages outperform the monolingual male ones, with the exception of Malayalam and Marathi, where the results are slightly lower. This, however, does not pose statistically significant difference because the confidence intervals between the two scores for the respective language configurations overlap heavily. There are small improvements over the monolingual male configurations in Bengali, Kannada, Tamil and Urdu, while the improvements in Gujarati and Telugu are very significant. The multilingual female voices also seem to mostly perform better than their monolingual counterparts, with the exception of Kannada, Marathi and Urdu, where the differences are not statistically significant. The biggest improvements in female voices are observed for Gujarati, Tamil and Telugu.

The result for Odia, the language for which we have no training data, is very encouraging. The Odia raters rated this voice as leaning towards "Good" on the MOS scale. Overall, the results demonstrate that pooling the data from different languages of India and re-using the shared phonemes across languages within a single acoustic model is definitely beneficial. Also, the speaker identity and gender features seems to play an important part in the perceived quality of the resulting voice – for some languages, like Tamil and Telugu, a female voice is preferred, while for the rest of the languages the raters prefer the male voices.

4. Conclusion and future work

In this paper we presented a unified multilingual phoneme inventory specifically designed to support multilingual text-to-speech for a variety of South Asian languages. The principle features of the design include maximum degree of sharing between phoneme inventories of phonologically close languages and reliance on the available data, as well as human-perceived phonological contrast when introducing new phonemes, as opposed to blindly following the canonical orthography or textbook-derived definitions. We have tested the resulting phoneme inventory in a real-world scenario by building a text-to-speech system based on a multilingual acoustic model trained on eleven languages. We evaluated male and female voices for eight languages represented in the training data (Bengali, Gujarati, Kannada, Malayalam, Marathi, Tamil, Telugu and Urdu) as well as one voice for a language not seen during training (Odia). We showed that for 13 out of 16 in-domain voices, the multilingual system outperforms the corresponding monolingual system. We also showed that synthesizing an unseen language resulted in the voice that was acceptable to the native speakers. Future work will involve more experiments with optimizing the current inventory, for which we have made a few judgment calls that need more thorough investigation. In addition, we would like to investigate the influence of phoneme inventory on the linguistic features automatically derived from publicly available typological databases [7], especially in the context of synthesizing low-resource languages for which little or no data is available, and increase the coverage of our representation to significantly more languages.

5. Acknowledgements

The authors thank Archana Amberkar, Cibu Johny, Debasri Chakrabarti, Prabakaran Pakkiri, Ravikumar Ragam and Taif Wali Mohd for providing native speaker expertise. We would also like to thank Anna Katanova, Knot Pipatsrisawat and Supheakmungkol Sarin for their help in preparing the training data required for building the multilingual models.

6. References

[1] A. Gutkin, L. Ha, M. Jansche, O. Kjartansson, K. Pipatsrisawat, and R. Sproat, "Building statistical parametric multi-speaker synthesis for Bangladeshi Bangla," in *5th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU '16)*, 2016, pp. 194–200.

[2] J. A. E. Wibawa, S. Sarin, C. 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. of the 11th International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 1610–1614.

[3] C. P. Masica, *The Indo-Aryan Languages*. Cambridge University Press, 1991.

[4] B. Krishnamurti, *The Dravidian Languages*. Cambridge University Press, 2003.

[5] N. Trubetzkoy, "Proposition 16," in *Actes du Premier Congrès International de Linguistes, a La Haye, du 10–15 Avril 1928*. A. W. Sijthoff, 1930, pp. 17–18.

[6] International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[7] A. Gutkin, M. Jansche, and T. Merkulova, "FonBund: A library for combining cross-lingual phonological segment data," in *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 2236–2240.

[8] 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.

[9] B. Ramani, S. Christina, G. A. Rachel, V. S. Solomi, M. K. Nandwana, A. Prakash, A. Shanmugam S, R. Krishnan, S. P. Kishore, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *8th ISCA Speech Synthesis Workshop*, 2013, pp. 291–296.

[10] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proc. of Interspeech*, 2016, pp. 2468–2472.

[11] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," in *Proc. of Interspeech 2017*, Sweden, August 2017, pp. 2183–2187.

[12] S. u. D. Khan, "Bengali (Bangladeshi standard)," *Journal of the International Phonetic Association*, vol. 40, no. 2, pp. 221–225, 2010.

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

[14] C. Woehrling and P. B. d. Mareüil, "Identification of regional accents in French: perception and categorization," in *Proc. of 9th International Conference on Spoken Language Processing (Interspeech)*, Pittsburgh, USA, 2006, pp. 1511–1514.

[15] T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics*. Wiley, 1990, vol. 5.

[16] Recommendation ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," *International Telecommunication Union*, July 2012.