Cross-lingual transfer using phonological features for resource-scarce text-to-speech

Anonymous submission to SSW 2023

Abstract

In this work we explore the use of phonological features in cross-lingual transfer in resource-scarce settings. We modify the architecture of VITS to take a phonological feature vector as input, instead of phonemes or characters. Subsequently, we train multispeaker base models using data from LibriTTS and then fine-tune on single speaker Afrikaans and isiXhosa datasets of varying sizes, representing the resource-scarce setting. We evaluate the synthetic speech both objectively and subjectively and compare it to models trained with the same data using the standard VITS architecture. In our experiments, the proposed system using phonological features as input converges significantly faster and with less data than the base system. We show that the model using phonological features is able to produce sounds in the target language that were unseen in the source language, even in languages with significant linguistic difference and with only 5 minutes of data in the target language.

Index Terms: text-to-speech, resource-scarce, phonological features, cross-lingual

1. Introduction

Modern non-autoregressive text-to-speech (TTS) neural architectures like FastSpeech2 [1] and VITS [2] make it possible to build natural sounding TTS voices using an order of magnitude less data than the previous commercially dominant concatenative synthesis or statistical parametric speech synthesis (SPSS) techniques [3]. However, these models still require more training data than may necessarily be available in the target language, that is the language of the voice being built, if it is a resource-scarce language. In general the more data you have the less domain knowledge is required to build a TTS model, and vice versa.

The term resource-scarce (or low-resource) is not well defined and has many dimensions to it, including: linguistic resources, acoustic resources, human capital and computing resources. In this work we use the term resource-scarce for languages where there is a level of word segmentation, text normalization and grapheme-to-phoneme (G2P) conversion functions available, as well as limited recordings of relatively high-quality, 30 minutes or less.

Techniques that have emerged to alleviate the data-scarcity problems include self-supervised learning and cross-lingual transfer. With self-supervised learning (SSL) a representation of speech that contains low-level acoustic events, lexical knowledge, syntactic and semantic information is captured [4]. These representations can also be used to reconstruct the speech signal [5] and are therefore suitable as an intermediate speech representation in a TTS model. This allows one to develop a two stage TTS model, where the first stage is the conversion of text to an SSL representation, trained using relatively low-quality, but usually abundant, automatic speech recognition (ASR) data. The second stage is then the conversion of the SSL representation to speech, which can be trained on a smaller high-quality TTS style corpora. Recent examples of these models include [5], [6], [7] and [8]. These models seem suitable for building voices with limited data but to our knowledge have not yet been applied for cross-lingual voices.

In cross-lingual transfer a TTS model is pre-trained on data from a resource-rich language ("source language"), in order to assist the learning of the mappings between text and speech in resource-scarce languages ("target language"). The reasoning being that human languages share similar pronunciations and semantic structures [9]. Some of the challenges that need to be addressed include differing phone inventories and representations as well as language similarity. In general the differing phone inventories or input representations can be modelled by either using *unified representations* or *separate representations*, which was investigated in [10] where it was found that a unified input representations can improve the accent and naturalness of multilingual TTS models.

In this work we compare a direct unified representation to phonological features as input to a cross-lingual TTS model. The question we ask is whether it is advantageous to use phonological features in cross-lingual transfer learning, specifically in resource-scarce scenarios. Our contributions in this study are:

1. We adapt a state-of-the-art TTS model to accept phonological features as input, instead of the more traditional character or phoneme based schemes.
2. We train various TTS voices, simulation resource-scarce scenarios, and show that using PF one can train a cross-lingual model with as little as five minutes of data and achieve naturalness on par or better than statistical parametric speech synthesis (SPSS) models.

2. Related Work

2.1. Mappings

Different languages have different phone inventories, which complicate cross-lingual transfer. Many systems use some form of phone mapping function in order to map the target language phone inventory to the source language phone inventory. After this mapping the model trained on the source language can simply be fine-tuned on the target language data. During inference time this same mapping needs to be applied in order to generate appropriate phone, where the trained model will convert the phonetic sequences into speech. A trainable input space mapping between the linguistic symbols of a high-resourced language and a low-resourced language was proposed in [11], but

the results seem to degrade significantly on languages which are not relatively similar. A method where the speaker encoder adapts the speaker embedding of different languages to the same space, thereby creating a language-independent speaker space was proposed in [12]. This technique does not apply in the resource-scarce case as the model has relatively high data requirements for both the source and target languages. A promising technique employed by [13] was to use language independent phonological features instead of linguistic symbols, thereby having the model learn a mapping between the phonological and acoustic features. In [14], PHOIBLE [15], a large database of phonological inventories of over 2000 languages, is used. The phonological binary features defined in PHOIBLE are used to guide the phone mapping function from the source to target languages.

Other input representations have also been used to address both the differing input scripts and phoneme inventories of different languages. In [16] input texts that are UTF-8 encoded are segmented into their byte representations, which are then used as the input tokens to the TTS model.

Some other common representations that have been used for cross-lingual TTS include the International Phonetic Alphabet (IPA). This alphabet has been designed to be universal and a language independent phone (and phoneme) inventory. [17] found that language-unique IPA symbols could leak the speaker information, and therefore are not ideal for cross-lingual TTS.

2.2. Phonological features

Phonological features (PFs) consists of a mix of articulatory and auditory or acoustic features [18] although sometimes the terms phonological features and articulatory features are used interchangeably [19]. According to [19], PF can be used as a replacement for character models, where characters are used as a mapping to phonemes or phones. Many recent studies [20, 13, 17, 21, 22] have focused on using phonological features, either as a tool to perform a cross-lingual phone inventory mapping, or as direct input into a TTS model. Phonological features are attractive as TTS input due to the possibility of sharing features between phone identities and the universal nature of the features.

Various schemes for encoding PFs have been investigated, including using binary features such as voicing: *voiced/unvoiced*, mouth shape: *rounded/unrounded* [23] and multivalued features such as vowel height (*high, mid, low*) [24]. In this work we use a phonological feature set based on the work in [21], with some modifications which are discussed in detail in section 3.

Our work is similar to the low-resource multilingual work done in [25], but we use a different neural model, much less training data for the resource-rich source language and we also objectively and subjectively compare different sizes of resource-scarce datasets. [13] also used PFs for cross-lingual TTS, but their focus was on speaker adaptation, and they used 2 orders more training data. In [26] they also investigated PFs for cross-lingual TTS, but they use much more training data and an autoregressive model.

3. Proposed Method

3.1. Model architecture

3.1.1. Baseline

Our baseline model is the multi-band, multi-speaker and inverse short-time Fourier transform (iSTFT) variant [1][27] of the VITS [2] architecture, using phonemes as input. We chose this model primarily because it is relatively lightweight, has very fast inference due to the iSTFT and only a relatively small degradation in naturalness in comparison to VITS. The model has a phoneme embedding lookup table with a dimension of 96.

3.1.2. Proposed

The baseline architecture is modified by replacing the text encoder with a phonological encoder consisting of a sequential forward model with a non-linear activation. This allows the model to learn embeddings for multidimensional input features, instead of the standard one dimensional character or phoneme based input. The output of the phonological encoder feeds into the encoder of the VITS architecture in the same manner as the text encoder of the baseline model. We chose the dimension of this output to be 96 in order so be comparable to the baseline.

3.2. Phonological Feature set

Our phonological feature set is based on the work in [21] with minor changes, we have nine broad categories, eight categories of which correspond directly with the categories as defined in the IPA. All features are binary, and in total the combined number of features in all the categories are 60, 5 of which are "modifiers", i.e. they either modify the preceding or proceeding phone feature vector. For example, for a vowel that is lengthened, the "vowel" feature as well as the "lengthened" feature will both be one. The resulting binary feature vector of each phone then has a length of 55.

We defined an additional "non-speech marker" category, for features that are either syntactic or may influence the overall prosody of the utterance under investigation. Table 1 gives the broad categories as well as the detailed features used in this work.

4. Experiment and Results

4.1. Data

The resource-rich language data used in this work is a subset of the LibriTTS [28] corpus, more specifically the speech of the male speakers are used because the resource-scarce datasets are also male. We filtered all the utterances and selected a subset that contained 20 or less tokens per utterance, the reason being that it should decrease the computation load during training. A total of 12000 utterances were selected at random from this filtered set in order to keep the data manageable on a single GPU.

The resource-scarce data is represented with subsets from the Afrikaans and isiXhosa male speakers Lwazi [29] corpora. All the utterances were also first filtered and only utterances with 20 or less tokens were selected. To represent extreme resource scarceness we selected 3 subsets for each of the Afrikaans and isiXhosa from this filtered set: 5 minutes, 15 minutes and 30 minutes of speech data. Each of these sets were selected at random and independently from the others from each

[1] https://github.com/MasayaKawamura/MB-iSTFT-VITS

Table 1: *Phonological features set used in this work.*

			Articulation			Mouth		
Modifier	**Category**	**Non-speech marker**	**Place**	**Manner**	**Tongue position**	**Openness**	**Shape**	**Voicing**
lengthened	consonant	silence	dental	plosive	central	mid	rounded	unvoiced
half-length	vowel	padding	postalveolar	nasal	back	close-mid	unrounded	voiced
shortened	phoneme	question mark	velar	approximant	front central	close		
aspirated		exclamation mark	palatal	trill	front	open-mid		
ejective		start of sentence	glottal	flap	central back	close close-mid		
		end of sentence	uvular	fricative		open-mid open		
		syllable boundary	labiodental	lateral-approximant		open		
		word boundary	labial-velar	implosive				
		phrase boundary	alveolar	vibrant				
			bilabial	click				
			alveolopalatal	lateral-fricative				
			retroflex	lateral-flap				
			pharyngal					
			epiglottal					
			labial-palatal					

corpus. The details of the data used are given in Table 2. For both the Afrikaans and isiXhosa datasets 10 utterances were selected at random to present the test set, and 20 utterances to present the validation set. These test and validation utterances from each language do not make part of any training set of table 2.

Table 2: *Details of training datasets.*

#	Language	Speaker	Source	Duration
1	English	124 males	LibriTTS	3:01:15.56
2	Afrikaans	1 male	Lwazi	0:05:00.15
3				0:15:05.28
4				0:30:00.74
5				3:33:26.53
6	isiXhosa	1 male	Lwazi	0:05:01.40
7				0:15:02.24
8				0:30:02.12
9				2:46:04.77

The text annotations were tokenized and normalised for each utterance and the linguistic descriptions consisted of the phoneme strings with syntactic markup (syllable breaks, word breaks and phrase breaks) inserted into the phoneme strings. These are the same markers as in the "non-speech marker" category in Table 1. These phoneme strings were used as the input to the baseline model.

The input to the proposed model is the phoneme strings with syntactic markup of the baseline model which are then transformed into a PFs vector using the defined feature set. Each phoneme first maps to an IPA equivalent, which then is transformed into the PFs vector.

For both the baseline and the proposed model we "interspersed" zero valued vectors (zeros in the case of the baseline) into the input vector (or string). This aids the alignment module of the VITS architecture.

For the baseline we defined and used a unified representation between all the languages in question in order to do the cross-lingual transfer.

4.2. Experimental setup

All audio was down-sampled to 16 kHz at 16 bits per sample and each utterance was normalised to the average power level of the dataset (as in Table 2).

4.2.1. Source Language

The LibriTTS English dataset was tokenized, normalised and converted into phoneme strings using inhouse front-end processing tools. Models were trained following the standard VITS multi-speaker setup with the standard hyper-parameters[2], one for the baseline using phonemes as input and one using the converted PFs as input. Both the models were trained for 200k steps until the models converged.

4.2.2. Target Language

Each of the other datasets in Table 2 were tokenized, normalised and converted into phoneme strings as with the baseline LibriTTS dataset, one variant being the phoneme based system and one the PF based system. For each variant and dataset the baseline model was fine-tuned to 100k steps until the model converged and also making sure that the models were not overfitting.

The test sets were also tokenized, normalised and converted into phoneme strings and PF vectors, after which they were synthesized for the objective and subjective evaluations. The reference and synthesized samples can be found here: `https://ghssw2023samples.github.io/experiment/`.

4.3. Evaluation

In order to quantitatively compare the proposed models with the baseline models we used the following objective measures:

- Mel-cepstral distortion (MCD), as defined in [30].
- f_0 Root mean squared error (RMSE), as defined in [31].
- f_0 Mean absolute error (MAE), this is similar to RMSE, but the MAE is less influenced by large outliers than RMSE.
- f_0 Voicing classification error (VCE), as defined in [31].
- Pearson correlation coefficient (CORR).

[2] `https://github.com/MasayaKawamura/MB-iSTFT-VITS`

For subjective evaluation we conducted an online listening test in the form of a web-based *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) [32] framework. The participants were asked to rate the *naturalness* in comparison to a reference sample on a scale from 0 (Very poor) to 100 (Completely natural).

The reference sample was a recording of human speech which was part of the test set. The participants were instructed to make sure that at least one of the rated samples was 100. An HMM-based voice [33] was also built using all the language specific Afrikaans and isiXhosa data. These voices were used as the anchor in the MUSHRA evalution.

A total of 51 people participated in the perceptual evaluation, 27 for Afrikaans and 24 for isiXhosa. Only participants which were either first language speakers or fluent in the language were recruited. Post-screening was done to verify that participants scored the hidden reference stimuli 90 or more for more than 80% of the tests and after post-screening there were 23 people for Afrikaans and 21 for isiXhosa.

The objective and subjective results are given in Table 3. We also calculated the Wilcoxon rank-sum test to determine if the MUSHRA results given in Table 3 are statistically significant between the different voice pairs. The test tests the null hypothesis that two sets of measurements are drawn from the same distribution with the alternative hypothesis that values in one sample are more likely to be larger than the values in the other sample. The results are given in Table 4 for Afrikaans and Table 5 for isiXhosa.

5. Discussion and Conclusion

From Table 3 it is clear that for both Afrikaans and isiXhosa, all the objective and subjective measures tend to show that the proposed model using the phonological features performs better than the baseline model.

For each variant and data subset of the target languages a voice was also trained using the proposed model, except that no pre-training using the LibriTTS data (source language) was performed. These voices and samples were excluded from the evaluations due to the severe lack of intelligibility and noise artifacts. This was unexpected, especially for the Afrikaans and isiXhosa datasets that included all the data of those speakers. This also reaffirms the value of pre-training on a resource-rich dataset, regardless of the language.

From Table 4 it can be seen that there is no statistical difference in naturalness between a voice trained on 15 minutes of data using phonological features as input and a voice trained using all the data (3.5 hours). It is also notable that a voice trained from 5 minutes of data using the proposed phonological features as input is statistically different from an HMM style voice (better when looking at the means of the MUSHRA test from Table 3).

The results for isiXhosa follow the same trend, even though there is a marked difference between the source and target languages. Here a voice built from 30 minutes of speech using the proposed phonological features as input the proposed has been found to not be statistically different to a voice built from all the data (2.75 hours) using the baseline method.

Figures 1 and 2 show samples from the baseline and proposed model built from 5 minutes of data, uttering a dental click. This sound was not present in the pre-training or resource-rich data. As can be seen, it is much more pronounced and identifiable in the sample from the proposed model.

Both the objective and subjective results support the assertion that the proposed model can be used for cross-lingual transfer in resource-scarce scenarios. A voice that is more natural sounding than the previous paradigm statistical parametric speech synthesis models can be built using just 5 minutes of data. Future work will focus on data selection and it's influence on the voice quality, from both the source as well as target languages.

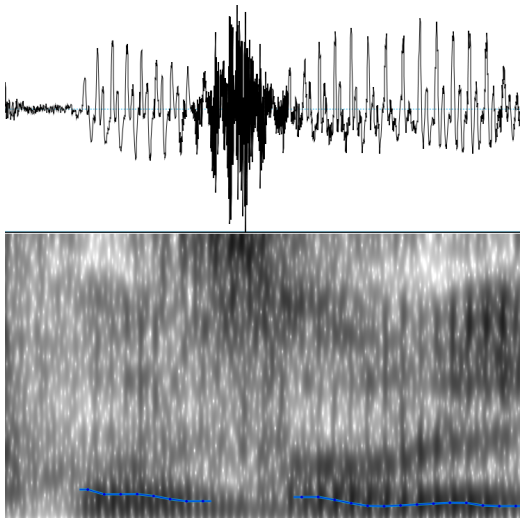

Figure 1: *Sample of baseline isiXhosa voice built from 5 minutes, showing the dental click /!/.*

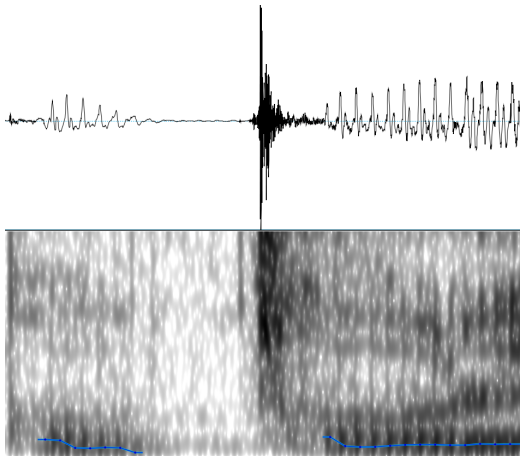

Figure 2: *Sample of proposed model isiXhosa voice built from 5 minutes, showing the dental click /!/.*

6. References

[1] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech," in *International Conference on Learning Representations*, May 2021, URL https://openreview.net/forum?id=piLPYqxtWuA.

[2] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *arXiv preprint arXiv:2106.06103*, 2021.

[3] J. A. Louw, "Neural speech synthesis for resource-scarce languages," in *Proceedings of the South African Forum for Artifi-*

Table 3: *Objective and subjective results for both Afrikaans and isiXhosa. MCD: Mel-cepstral distortion. f_0 RMSE: Fundamental frequency root mean squared error (linear scale). f_0 MAE: Fundamental frequency mean absolute error (linear scale). f_0 VCE %: Fundamental frequency voicing classification percentage error. CORR: Pearson correlation coefficient of voicing classification. MUSHRA: Mean subjective naturalness scores (0-100) with a 95% confidence level.*

Language	**Dataset**	**Variant**	**MCD**	**f_0 RMSE (Hz)**	**f_0 MAE (Hz)**	**f_0 VCE (%)**	**CORR**	**MUSHRA**
Afrikaans	5 min	ph	6.860977	28.94	18.42	15.62	0.401	53.33 ± 4.82
		pf	6.336159	28.08	18.41	15.55	0.411	58.51 ± 3.96
	15 min	ph	6.381576	28.14	17.35	15.71	0.423	60.67 ± 3.97
		pf	6.066329	27.27	17.54	16.00	0.456	74.43 ± 3.59
	30 min	ph	6.243911	27.61	17.49	13.45	0.498	67.29 ± 3.50
		pf	6.131744	26.71	17.2	15.88	0.569	81.56 ± 2.98
	All	ph	5.915963	26.30	16.32	13.61	0.620	76.04 ± 3.42
		pf	5.729585	25.36	15.98	16.52	0.626	86.27 ± 2.36
	HMM		6.705095	27.16	17.23	18.17	0.602	51.30 ± 3.27
	Reference							96.95 ± 0.91
isiXhosa	5 min	ph	6.454278	11.09	6.66	11.76	0.692	49.36 ± 4.75
		pf	6.006995	12.45	7.65	11.55	0.601	56.34 ± 4.13
	15 min	ph	6.040699	10.53	6.16	10.93	0.760	60.65 ± 3.94
		pf	5.704059	10.68	6.15	10.75	0.771	63.40 ± 3.56
	30 min	ph	5.800190	10.18	5.92	11.26	0.798	60.99 ± 3.66
		pf	5.435614	9.77	5.62	11.05	0.818	70.20 ± 3.76
	All	ph	5.566074	10.18	5.59	10.75	0.811	65.52 ± 3.75
		pf	5.235502	10.89	6.11	11.01	0.813	70.18 ± 3.39
	HMM		5.973844	10.32	6.02	13.76	0.800	61.12 ± 2.53
	Reference							97.22 ± 0.84

Table 4: *Statistical significant differences for Afrikaans between the different voice pairs. The emphasised pairs are **not** statistically significantly different ($p < 0.05$).*

		Natural	**HMM**	**All**		**30 min**		**15 min**		**5 min**
				pf	**ph**	**pf**	**ph**	**pf**	**ph**	**pf**
5 min	**ph**	p<0.05	**p=0.46**	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	**p=0.08**	**p=0.17**
	pf	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	**p=0.57**	
15 min	**ph**	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05		
	pf	p<0.05	p<0.05	p<0.05	**p=0.64**	p<0.05	p<0.05			
30 min	**ph**	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05				
	pf	p<0.05	p<0.05	**p=0.08**	p<0.05					
All	**ph**	p<0.05	p<0.05	p<0.05						
	pf	p<0.05	p<0.05							
HMM		p<0.05								

Table 5: *Statistical significant differences for isiXhosa between the different voice pairs. The emphasised pairs are **not** statistically significantly different ($p < 0.05$).*

		Natural	**HMM**	**All**		**30 min**		**15 min**		**5 min**
				pf	**ph**	**pf**	**ph**	**pf**	**ph**	**pf**
5 min	**ph**	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05	**p=0.06**
	pf	p<0.05	**p=0.13**	p<0.05	p<0.05	p<0.05	**p=0.16**	p<0.05	**p=0.24**	
15 min	**ph**	p<0.05	**p=0.95**	p<0.05	**p=0.17**	p<0.05	**p=0.87**	**p=0.40**		
	pf	p<0.05	**p=0.30**	p<0.05	**p=0.49**	p<0.05	**p=0.48**			
30 min	**ph**	p<0.05	**p=0.93**	p<0.05	**p=0.20**	p<0.05				
	pf	p<0.05	p<0.05	**p=0.81**	**p=0.10**					
All	**ph**	p<0.05	**p=0.09**	**p=0.14**						
	pf	p<0.05	p<0.05							
HMM		p<0.05								

cial Intelligence Research, E. Barnard and M. Davel, Eds., Cape Town, South Africa, December 2019, pp. 103–116.

[4] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, October 2022.

[5] H. Siuzdak, P. Dura, P. van Rijn, and N. Jacoby, "Wavthruvec: Latent speech representation as intermediate features for neural speech synthesis," *arXiv preprint arXiv:2203.16930*, 2022.

[6] C. Du, Y. Guo, X. Chen, and K. Yu, "Vqtts: high-fidelity text-to-speech synthesis with self-supervised vq acoustic feature," *arXiv preprint arXiv:2204.00768*, 2022.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 624–16 636, Dec 2022.

[9] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[10] A. Sanchez, A. Falai, Z. Zhang, O. Angelini, and K. Yanagisawa, "Unify and conquer: How phonetic feature representation affects polyglot text-to-speech (TTS)," *arXiv preprint arXiv:2207.01547*, Jul. 2022.

[11] Y.-J. Chen, T. Tu, C.-c. Yeh, and H.-Y. Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pp. 2075–2079, September 2019.

[12] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-Lingual Text-To-Speech Synthesis via Domain Adaptation and Perceptual Similarity Regression in Speaker Space." in *Proceedings of Interspeech*, Oct. 2020, pp. 2947–2951.

[13] G. Maniati, N. Ellinas, K. Markopoulos, G. Vamvoukakis, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Cross-lingual low resource speaker adaptation using phonological features," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brno, Czechia, Sep. 2021.

[14] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning," in *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 2022, pp. 16–22.

[15] S. Moran and D. McCloy, "Phoible 2.0," *Jena: Max Planck Institute for the Science of Human History*, 2019.

[16] M. He, J. Yang, L. He, and F. K. Soong, "Multilingual byte2speech models for scalable low-resource speech synthesis," *arXiv preprint arXiv:2103.03541*, 2021.

[17] H. Zhang, H. Zhan, Y. Zhang, X. Yu, and Y. Lin, "Revisiting ipa-based cross-lingual text-to-speech," *arXiv preprint arXiv:2110.07187*, 2021.

[18] E. S. Flemming, *Auditory representations in phonology*. Routledge, 2013.

[19] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. S. Levin, "Panphon: A resource for mapping IPA segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL, 2016, pp. 3475–3484.

[20] C. Zhang, H. Zeng, H. Liu, and J. Zheng, "Applying phonological features in multilingual text-to-speech," *arXiv preprint arXiv:2110.03609*, 2021.

[21] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological features for 0-shot multilingual speech synthesis," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct 2020, pp. 2942–2946.

[22] C. Zhang, H. Zeng, H. Liu, and J. Zheng, "Applying feature underspecified lexicon phonological features in multilingual text-to-speech," *arXiv preprint arXiv:2204.07228*, 2022.

[23] S. L. Hartman, "A Universal Alphabet for Experiments in Comparative Phonology," *Computers and the Humanities*, vol. 15, pp. 75–82, Aug 1981.

[24] K. Williamson, "Multivalued features for consonants," *Language*, vol. 53, no. 4, pp. 843–871, December 1977.

[25] F. Lux, J. Koch, and N. T. Vu, "Low-resource multilingual and zero-shot multispeaker tts," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, November 2022, pp. 741–751.

[26] D. Wells and K. Richmond, "Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis," in *Proceedings of the 11th ISCA Speech Synthesis Workshop*, Aug 2021, pp. 160–165.

[27] M. Kawamura, Y. Shirahata, R. Yamamoto, and K. Tachibana, "Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform," *arXiv preprint arXiv:2210.15975*, Oct 2022.

[28] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, Apr. 2019.

[29] D. van Niekerk, A. de Waal, and G. Schlünz, "Lwazi II Afrikaans TTS Corpus," https://repo.sadilar.org/handle/20.500.12185/443, November 2015, ISLRN: 570-884-577-153-6.

[30] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Proceedings of Spoken Languages Technologies for Under-Resourced Languages*, May 2008, pp. 63–68.

[31] K. Yu and S. Young, "Continuous f0 modeling for hmm based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2010.

[32] R. ITU-R, "BS. 1534-1. Method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.

[33] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A Hidden Semi-Markov Model-Based Speech Synthesis System," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 825–834, May 2007.