Cross-lingual Low Resource Speaker Adaptation Using Phonological Features

Georgia Maniati[1]*, Nikolaos Ellinas[1]*, Konstantinos Markopoulos[1], Georgios Vamvoukakis[1],
June Sig Sung[2], Hyoungmin Park[2], Aimilios Chalamandaris[1], Pirros Tsiakoulis[1]

[1]Innoetics, Samsung Electronics, Greece
[2]Mobile Communications Business, Samsung Electronics, Republic of Korea
{g.maniati, n.ellinas} @samsung.com

Abstract

The idea of using phonological features instead of phonemes as input to sequence-to-sequence TTS has been recently proposed for zero-shot multilingual speech synthesis. This approach is useful for code-switching, as it facilitates the seamless uttering of foreign text embedded in a stream of native text. In our work, we train a language-agnostic multispeaker model conditioned on a set of phonologically derived features common across different languages, with the goal of achieving cross-lingual speaker adaptation. We first experiment with the effect of language phonological similarity on cross-lingual TTS of several source-target language combinations. Subsequently, we fine-tune the model with very limited data of a new speaker's voice in either a seen or an unseen language, and achieve synthetic speech of equal quality, while preserving the target speaker's identity. With as few as 32 and 8 utterances of target speaker data, we obtain high speaker similarity scores and naturalness comparable to the corresponding literature. In the extreme case of only 2 available adaptation utterances, we find that our model behaves as a few-shot learner, as the performance is similar in both the seen and unseen adaptation language scenarios.

Index Terms: cross-lingual, multilingual, speaker adaptation, speech synthesis, low resource

1. Introduction

Text-to-speech (TTS) systems have traditionally used sequences of discrete symbols as inputs. Recently proposed neural architectures [1, 2] have shown that an efficient end-to-end acoustic model is possible by directly consuming text characters. The inputs to state-of-the-art TTS systems consist of either text characters (graphemes) or phonemes, with the superiority of phoneme-based systems recently quantified [3]. In multilingual TTS, these inputs may originate from various speakers and languages introducing variable factors in the model's logic.

Synthesizing speech from multiple speakers with the use of learnable speaker embeddings has been thoroughly examined from the very start of neural TTS [4] up to most recent efforts [5]. Controlling language with learnable embeddings is also straightforward [6, 7] and recently, the concept of meta-learning has been shown effective for this purpose [8]. In order to avert the inherent problem of language-dependent speaker representations, domain adaptation has been utilized [9].

It is common for the input phoneme representations to be mapped into trainable embeddings, which can be shared across phonemes in the multilingual setting [6]. For this purpose, the International Phonetic Alphabet (IPA) [10] can be used [11]. However, tasks such as code-switching and low-resource language TTS introduce the need for multi-valued representations

that will allow the learning of shared qualities across phonemes enabling generalization to previously unseen combinations.

1.1. Related work

When investigating multilingual TTS, the input linguistic sequence plays an important role as it incorporates all the distinct language characteristics. Gutkin *et al.* [12–15] use phonological features (PFs) combining them with phonemes as inputs to multilingual neural TTS models and show improvements in intelligibility across seen and even unseen languages [12]. Effectively, approaches which concatenate PFs to phonemes do not allow synthesis of unseen phonemes without further training. To this end, Unicode-bytes-based multispeaker multilingual models have been proposed [16, 17]. This alternative approach allows unseen characters to be synthesized without entailing any model changes, but since bytes only encode typographical relations, transfer learning of phonological information cannot be achieved for unseen byte combinations. Staib *et al.* [18] train a multispeaker variant of Tacotron 2 solely on PFs and show that their model remains unchanged across seen and unseen languages, while enabling the approximation of sounds absent in the training set. As they aim at code-switched speech, they only train a monolingual and a small multilingual model. Our work extends the idea of utilizing phonological features to achieve cross-lingual speaker adaptation. In addition, we present extended experimental results that investigate the effect of language phonological similarity as well as the effect of the adaptation data size. Similar features have been used in a feed-forward acoustic model for cross-lingual speaker adaptation using ground truth target phoneme durations [19].

Cross-lingual speaker adaptation can leverage the benefits of a fixed phoneme representation such as IPA. In [11], such a model is fine-tuned to the voice of a speaker using 20 minutes of data, while in [20], cross-lingual cloning is achieved without further training by utilizing x-vectors extracted from a pretrained system and a common ARPABET phoneme set augmented with stress and tone embeddings. Language-dependent phones can also be used, as in [21] where a transformer-based model is trained on a large set of 50 language locales following data imbalance strategies and allowing extensions to new languages with as few as 6 minutes of data. In the low-resource setting, [22] apply different language encoders on language-dependent phones and [23] show that a learnable linguistic embedding trained in a VAE-like structure can generalize to other languages and adapt to new speakers with few data.

1.2. Proposed method

In this paper, we apply handcrafted phonological features in cross-lingual TTS and speaker adaptation with very few data. We follow prior work [24] in order to train a multilingual

---

*Equal contribution

end-to-end model [6] without the addition of language embeddings, since we aim at a model independent of input or output language identity. First, we investigate if cross-lingual TTS based on phonological features can be improved by using additional training data from typologically related and unrelated languages, and explore the relation of the ratio of unseen phonemes and the perceived intelligibility and naturalness of cross-lingual speech. Second, we experiment with cross-lingual speaker adaptation and the amount of adaptation data irrespective of the target speakers' native language. Our results demonstrate that as little as 6 s of adaptation data suffice to achieve synthesized speech highly similar to the target speaker's voice, which is a notable advancement to phoneme-based cross lingual adaptation models requiring 20 min of data for the same task [11]. The limited data scenario combined with the absence of restrictions to either the target language or the language of the target speaker enables applications such as multilingual TTS of a low-resource language speaker and personalized TTS as long as IPA definitions of the language's phonemes are available.

2. Method

2.1. Feature set

According to PFs' theories, each phoneme of a language can be decomposed into a bundle of simultaneous features. Jakobson *et al.* [25,26] were the first to introduce a small set of acoustically-defined universal distinctive features. These were later replaced by articulatory features emphasizing their innate nature [27].

In our work, features are derived from the articulatory-inspired IPA definitions [10], as categorical multi-valued features. Each feature is encoded into a varied number of dimensions, resulting in an initial 23-dimensional PFs vector. The features are 1-hot encoded, except for vowel *openness* and *frontness* which assume continuous values. All phonemes are split into semi-phonemes before they are fed to the acoustic model to account for cases of changing phoneme quality (diphthongs, affricates). Each semi-phoneme is mapped to the corresponding PFs vector, resulting in 46 dimensions per phoneme. For monophthongs, the PFs vector is replicated twice. Each phoneme is then appended with: a binary *duration* feature differentiating between diphthongs/long vowels/double consonants and regular phonemes; a binary *stress* feature for primary and secondary stress; 7 dimensions for 1-hot encoded punctuation, word boundary, padding and end-of-sequence tokens.

2.2. Acoustic model architecture

The acoustic model follows an attention-based sequence-to-sequence architecture which converts the input linguistic sequence into a sequence of acoustic frames for the LPCNet vocoder [28]. In our case, the input text of any given language is transformed into internal phonemes by the corresponding frontend module, then mapped to its phonological features' representation using a dictionary lookup, first to IPA phonemes and then to the PFs vectors. We leverage the benefits of the reduced dimensionality of the LPCNet features together with a stable Mixture-of-Logistics (MoL) attention module in order to construct a robust model with near natural speech quality [24].

Since we are working on a multispeaker setup, each speaker is assigned a learnable speaker embedding, which is used to condition the decoder at each step. A speaker classifier which predicts the identity of each speaker from the encoder outputs is also used during training as introduced in [6]. This classifier is trained by utilizing the concept of domain adversarial

Table 1: *Training dataset details*

Language	Code	Hours	Speakers	Phonemes
US English	en	125.9	3	49
German	de	107.0	4	58
French	fr	84.7	4	40
Spanish	es	89.4	4	33
Italian	it	97.2	4	57
Korean	ko	164.6	3	46

training [29] in order to introduce a degree of disentanglement between the linguistic representations and the speaker identity. The fact that language identity is absent from the model makes this module even more necessary. Finally, the model is augmented with a residual variational encoder [6,30] which aims to encode latent factors of audio and increase the naturalness and robustness of the model in the cross-lingual transfer setting.

2.3. Speaker adaptation

We want to test the feasibility of adapting the model to an unseen speaker with limited data and enabling them to speak an arbitrary number of languages, regardless of the languages contained in the training set or the speaker's native language. Our choice of PFs allows the model to be language-independent and as a result there are no restrictions to which speaker or language are compatible with the model without applying any modifications. We select a random same-gender speaker-id which is assigned to the target speaker and fine-tune the model for a small number of iterations in a single speaker setting. During the initial training the encoder has learned meaningful representations of the input PFs and the attention module has learned to align these representations with the acoustic frames. Since the target speaker's data are limited, we freeze these modules' weights in order to preserve them from forgetting their generic targets. We found this method helps in terms of pronunciation and nativeness of the target language.

3. Experiments and results

3.1. Data and training setup

To train our models, we use an internal multilingual multi-speaker dataset comprising 668.8 hours of speech in 6 languages, 22 speakers and 134 unique phonemes. Details about the dataset are shown in Table 1. Varied language configurations' chunks are drawn from this dataset and used as training data throughout our experiments in Sections 3.3 and 3.4. We use 24 kHz audio data in order to extract the output 22-dimensional acoustic features which consist of 20 Bark-scale cepstral coefficients, the pitch period and the pitch correlation. The model parameters are trained using the Adam optimizer [31], a batch size of 64 and an initial learning rate of 10^{-3}, which linearly decays to $3 \cdot 10^{-5}$ in 600K iterations. For speaker adaptation the learning rate is kept stable for another 5K iterations.

3.2. Formal evaluation

Our models were assessed against naturalness, intelligibility and speaker similarity. Naturalness was evaluated via mean opinion score (MOS) ranging from 1 to 5, with 1 indicating unnatural and 5 natural speech. No natural sample of the target language was available for the speakers, as all experiments are cross-lingual, and inserting natural samples of different speakers in the single-speaker tests might affect scores reflecting voice preference. Speaker similarity evaluation was based on

Table 2: *Cross-lingual TTS setup and results of naturalness (MOS with 95% confidence interval) and intelligibility (WER)*

Setup			Results	
Train	**Test**	**UPR%**	**WER%**	**MOS**
de	en	18.6 ±4.7	23.41	3.47±0.10
	fr	20.3 ±7.4	11.97	3.23±0.11
	it	38.3 ±6.6	3.19	2.30±0.08
de+es	en	11.3 ±3.9	15.71	3.68±0.06
	fr	6.6 ±3.9	11.68	3.06±0.08
	it	7.4 ±4.1	4.95	2.12±0.06
de+es+ko	en	9.1 ±3.8	13.90	3.59±0.05
	fr	6.6 ±3.9	12.11	3.09±0.06
	it	7.4 ±4.1	3.05	2.78±0.05
es+ko	en	30.3 ±6.3	32.87	3.05±0.07
	fr	30.4 ±8.5	33.27	1.70±0.06
	it	11.2 ±5.4	5.80	2.06±0.06
en+es	de	27.1 ±5.9	17.77	2.48±0.09
	fr	15.1 ±6.0	15.69	2.56±0.09
	it	6.9 ±4.1	3.10	2.72±0.06
en+fr	de	25.8 ±4.7	26.86	2.21±0.10
	es	5.2 ±3.2	19.03	2.58±0.06
	it	13.8 ±3.7	6.11	2.35±0.06

a Likert scale, ranging from "1: Sounds like a totally different person" to "5: Sounds like exactly the same person" compared to the reference sample. For intelligibility, listeners were asked to transcribe the generated samples as accurately as possible, and the word error rate (WER) of their responses was computed.

All formal evaluations were conducted online via Amazon Mechanical Turk [32]. Only native speakers of the target language were recruited. Every audio sample was evaluated by 20 unique participants. In MOS tests, a validation sample was inserted in each test page to control for potential spurious participants: for naturalness, listeners were instructed to select one response from 1 to 5; for speaker similarity, a different gender voice was used in one of the samples, and the listener was expected to select 1. After excluding test pages where participants failed to pass validation, and the responses of participants whose WER was over 80%, the responses of 260 subjects were analyzed for MOS, 174 for intelligibility and 144 for similarity.

For each test language, 300 sentences were randomly sampled from conversational corpora and Wikipedia articles. As we were eager to conduct the models' evaluation in the most challenging stress-test setups, we subsequently converted the sentences into phonemes and used the corpus selection tool introduced in [33], so as to sort them in descending phonetic coverage order. The top 35 phonologically diverse sentences per language comprised each language's test set for all evaluations.

3.3. Cross-lingual text-to-speech

In our first set of experiments, we attempted cross-lingual text-to-speech; that is, synthesis in languages that are unseen in our models' training data, using combinations of the data described in Table 1. PFs of unseen phonemes are derived from the IPA. Due to the nature of our method, the resulting speech is accented and may retain prosodic characteristics of the source languages. We started with a *de* monolingual multispeaker model, and gradually augmented the training data with *es* and then *ko*, while we evaluated the models' performance in synthesizing *en*, *fr* and *it*. As we could only conduct formal evaluations for a limited number of language configurations, in this experiment

we opted to investigate whether the addition of data from: (i) a language phylogenetically close to the target language can favour cross-lingual synthesis (the addition of *es* data for *fr* and *it*), (ii) a typologically diverse language (*ko*) can degrade synthesis of a cross-lingual model based on PFs. Subsequently, we trained two models on English and one Romance language (*en+es, en+fr*) as well as an *es+ko* model which we aimed to compare to *de+es+ko* with regards to the effect of the proportion of the distanced language data within the training data.

For each training setup, we conducted the experiments using as many test speakers as are the languages in the training setup, and for each language, one female speaker was selected as test voice. That is, for the *de+es* setup, the *en, fr, it* samples have been generated by a *de* female speaker as well as an *es* female speaker.[1] We follow this protocol aiming to control a potential effect of the proximity of the source speaker's language to any of the target languages. As no significant variation was observed, the results presented in Table 2 are the averaged values among all test speakers.

The intelligibility evaluation of the models was conducted first. For any given test speaker and target language combination, we created 1-4 intelligibility tests, as many as the number of training setups in which the speaker's data were contained. Each of the tests comprised the 35 test sentences, where we randomly drew samples from all training setups for the target language, such that each listener would evaluate several models without listening to each sentence more than once. We conducted minimal processing of the responses prior to analysis, i.e. removed punctuation and normalized case, while we could not resolve homographs or correct spelling mistakes. In the second test, we evaluated how natural the cross-lingual samples were. As we are interested in the relationship between the count of unseen phonemes and the target language intelligibility and naturalness, we calculate the unseen phoneme rate (UPR) for each train-test setup. The UPR is computed per test utterance, as the number of phonemes not present in the training language(s), divided by the total number of phonemes in the utterance. We report the mean and standard deviation of the UPR over all test language's utterances.

Table 2 shows that augmenting the training data of a monolingual or multilingual PFs model with diverse language data can improve intelligibility and naturalness in an unseen language. This improvement is apparent for *en*, first generated from a monolingual, then a small multilingual and finally an augmented model with *ko*. Interestingly, the addition of a typologically diverse language in the latter setup improves intelligibility, while MOS is not significantly affected. We observe that when the unrelated language data exceed in quantity the data of the related language, the PFs model's intelligibility is severely affected; *es+ko*, where the size of *ko* data is almost double the size of *es*, performs poorly in intelligibility in all test languages. Replacing *ko* with *en* results in a more intelligible model. We note that naive listeners assign low MOS scores to accented speech, even if the speech is highly intelligible. This is the case especially for *it*, where WER is the lowest across all languages but listeners consistently assign low values across setups. This could also be linked to cultural factors of perception of accented speech. For most target languages, we observe that MOS is negatively correlated to UPR and WER is correlated to UPR, suggesting that in high UPR stress test scenarios, the approximation of unseen phonemes cannot attenuate the ef-

---

[1]Samples at: https://innoetics.github.io/publications/phonological-features/index.html

Table 3: *Cross-lingual speaker adaptation setup and MOS results of naturalness and speaker similarity with 95% confidence interval*

Adaptation Setup			Naturalness (MOS)				Speaker Similarity (MOS)			
	utt	dur(s)	de	fr	it	es	de	fr	it	es
en	32	217	3.63±0.14	3.72±0.11	3.22±0.09	3.30±0.08	4.15±0.12	4.28±0.11	3.82±0.10	4.28±0.09
(seen)	8	41	3.78±0.15	3.84±0.11	3.38±0.08	3.38±0.07	3.99±0.14	4.22±0.12	3.83±0.10	4.13±0.09
	2	6	3.25±0.18	3.11±0.13	2.88±0.10	2.89±0.09	3.60±0.17	3.68±0.14	2.99±0.11	3.44±0.12
gr	32	262	3.04±0.15	3.22±0.12	2.98±0.09	3.07±0.09	3.82±0.18	4.05±0.12	3.56±0.12	4.04±0.09
(unseen)	8	78	3.45±0.13	3.54±0.12	3.17±0.09	3.14±0.09	3.84±0.17	4.21±0.11	3.63±0.12	4.13±0.09
	2	20	3.20±0.13	3.22±0.12	2.97±0.09	2.91±0.09	3.80±0.17	4.13±0.12	3.48±0.12	4.00±0.10

fects of pronunciation errors in quality. Our informal evaluation showed that although in most cases the approximations made by the model are to neighbouring phonemes, in *de+es→it* some phonemes collapse to inappropriate phonemes, affecting results for this model. Notably, its performance is improved with the addition of the unrelated language data, even if UPR remains unchanged (*de+es+ko* outperforms *de+es* for *it*).

3.4. Cross-lingual speaker adaptation

Our PFs model is trained on the entire dataset (Table 1) for 600K iterations, as diverse language data have been shown to ameliorate its performance. Then, it is fine-tuned to the speaker adaptation data for 5K iterations. An informal evaluation of adaptation configurations is conducted by alternatively freezing the weights of the attention, the encoder and both. We conclude that the frozen encoder is key to preserving the pronunciation of the target language, while the frozen attention contributes to the stability of the model, preventing end-of-sentence attention failures. We keep both modules frozen in our adaptation setup.

Due to our choice of PFs and the language-independent nature of the model, there are no restrictions to either the target language the model can generate speech in, or to the language of the speaker adaptation data (adaptation language). As we aim to investigate how the quality of the adapted speech synthesized by the PFs model is affected by the presence of the adaptation language in the training data, we select 2 male voices from an internal dataset, a native American English speaker (*en*) and a Greek native speaker (*gr*). Our corpus selection tool [33] was used to sort each voice's corpus on the basis of phonetic coverage of the language, and the 32 most phonologically rich sentences were selected as the adaptation corpus per speaker. In our first set of experiments, we fine-tune the model using the speaker adaptation data from the seen language (*en*), while in the second, we use the adaptation data from the unseen language (*gr*). Furthermore, within each language setup, we experiment with decreasing the amount of adaptation data. We are primarily interested in the extent to which we can limit the adaptation data for the multilingual PFs model. Also, we are eager to examine whether there is any difference in the quantity of adaptation data required to achieve comparable quality results in case of a seen and an unseen adaptation language. We formally evaluate models fine-tuned on 32, 8, 2 top utterances, as sorted by the corpus selection tool. Since these sentences in each of the languages differ in length, and the voices differ in speaking rate, the setups may vary in audio duration. We formally evaluate all models' performance in test languages unseen in the adaptation data, i.e. *de, fr, it, es*, against naturalness and speaker similarity to the original speaker. For the similarity tests, the speaker's reference sample is provided in the original language.

As expected, the adapted models presented here perform better than cross-lingual TTS models (Table 2) as the test languages are seen during training and thus the output speech is not accented. The results in Table 3 show a tendency of slight deterioration of the quality of the adapted speech with the decrease in the amount of adaptation data, across languages. However, for most languages and setups the differences are not statistically significant and fall within the confidence interval range of the mean, showing that the proposed PFs model is robust to very limited data. For the seen adaptation language, the decrease from 32 to 8 utterances, i.e to 41 s of speech, bears no significant changes in perceived naturalness or speaker similarity of the adapted speech, across all languages. The extreme *en-2* scenario, i.e. adaptation with 6 s of speech deteriorates the model. For the unseen adaptation language, the same pattern is observed. Notably, *gr-8* appears significantly better in naturalness compared to few setup-test language combinations (*gr-32:de, gr-2:it, gr-2:es*), but given the generalized tendency of *en-8* and *gr-8* models to perform better, we suspect that models adapted on 8 utterances were favoured by the number of fine-tuning iterations used throughout the adaptation experiments (5K). Speaker similarity is retained notably high across data quantity setups without statistically significant differences. Comparing the two adaptation language setups, we conclude that the use of an unseen language for adapting the PFs model does not entail more data for achieving comparable results to adapting with a seen language.

4. Conclusions

In this work, we train a language-agnostic multispeaker Tacotron-based model conditioned on a set of IPA-derived phonological features. The model can perform cross-lingual TTS in any language and is evaluated in high unseen phoneme rate scenarios of various source-target language configurations. We find that augmenting the training data with diverse language data can improve intelligibility and naturalness in an unseen language. We observe that the cross-lingual speech quality is negatively correlated to the ratio of unseen phonemes. Subsequently, we show that cross-lingual speaker adaptation with very few data is possible by fine-tuning the model on a new speaker of a seen or unseen language. We experiment with the effect of the size of the adaptation data on speech quality and find that speaker similarity is retained notably high across data quantity setups. With as few as 32 and 8 utterances of target speaker data, we achieve high speaker similarity scores and naturalness comparable to similar works. In the extreme case of 2 utterances, performance is similar for seen and unseen adaptation languages, showing that our PFs model is robust to very limited data. In future work, we plan to form a better understanding of how the speaker-id assigned to the unseen adaptation language's speaker affects the model. Moreover, we plan to investigate alternative phonological features as well as whether the benefits of such representations can be leveraged for monolingual speaker adaptation in extreme low resource scenarios.

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[2] 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, pp. 4779–4783.

[3] J. Fong, J. Taylor, K. Richmond, and S. King, "A comparison between letters and phones as input to sequence-to-sequence models for speech synthesis," in *Proc. ISCA Speech Synthesis Workshop*, 2019, pp. 223–227.

[4] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, 2018.

[5] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[6] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. Interspeech 2019*, 2019, pp. 2080–2084.

[7] Z. Cai, Y. Yang, and M. Li, "Cross-lingual multispeaker text-to-speech under limited-data scenario," *arXiv preprint arXiv:2005.10441*, 2020.

[8] T. Nekvinda and O. Dušek, "One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech," in *Proc. Interspeech*, 2020, pp. 2972–2976.

[9] 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 *Proc. Interspeech*, 2020, pp. 2947–2951.

[10] International Phonetic Association and International Phonetic Association Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[11] H. Hemati and D. Borth, "Using IPA-based tacotron for data efficient cross-lingual speaker adaptation and pronunciation enhancement," *arXiv preprint arXiv:2011.06392*, 2020.

[12] A. Gutkin, M. Jansche, and T. Merkulova, "Fonbund: A library for combining cross-lingual phonological segment data," in *Proc. LREC*, 2018, pp. 2236–2240.

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

[14] I. Demirsahin, M. Jansche, and A. Gutkin, "A unified phonological representation of south asian languages for multilingual text-to-speech," in *Proc. SLTU*, 2018, pp. 80–84.

[15] A. Gutkin and R. Sproat, "Areal and phylogenetic features for multilingual speech synthesis," in *Proc. Interspeech*, 2017, pp. 2078–2082.

[16] B. Li, T. Sainath, W. Chan, Y. Wu, and Y. Zhang, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. ICASSP*, 2019, pp. 5621–5625.

[17] M. He, J. Yang, and L. He, "Multilingual byte2speech text-to-speech models are few-shot spoken language learners," *arXiv preprint arXiv:2103.03541*, 2021.

[18] 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, pp. 2942–2946.

[19] I. Himawan, S. Aryal, I. Ouyang, S. Kang, P. Lanchantin, and S. King, "Speaker adaptation of a multilingual acoustic model for cross-language synthesis," in *Proc. ICASSP*, 2020, pp. 7629–7633.

[20] Z. Liu and B. Mak, "Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment," in *Proc. Interspeech*, 2020, pp. 2932–2936.

[21] J. Yang and L. He, "Towards universal text-to-speech," in *Proc. Interspeech*, 2020, pp. 3171–3175.

[22] M. de Korte, J. Kim, and E. Klabbers, "Efficient neural speech synthesis for low-resource languages through multilingual modeling," in *Proc. Interspeech*, 2020, pp. 2967–2971.

[23] H.-T. Luong and J. Yamagishi, "Latent linguistic embedding for cross-lingual text-to-speech and voice conversion," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 150–154.

[24] N. Ellinas, G. Vamvoukakis, K. Markopoulos, A. Chalamandaris, G. Maniati, P. Kakoulidis, S. Raptis, J. S. Sung, H. Park, and P. Tsiakoulis, "High quality streaming speech synthesis with low, sentence-length-independent latency," in *Proc. Interspeech*, 2020, pp. 2022–2026.

[25] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press, 1951.

[26] R. Jakobson and M. Halle, *Fundamentals of Language*. Mouton, 1956.

[27] N. Chomsky and M. Halle, *The sound pattern of English*. ERIC, 1968.

[28] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[29] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.

[30] W. N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, 2019.

[31] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[32] K. Crowston, "Amazon mechanical turk: A research tool for organizations and information systems scholars," in *Shaping the future of ICT research. Methods and Approaches*. Springer, 2012, pp. 210–221.

[33] A. Chalamandaris, P. Tsiakoulis, S. Raptis, and S. Karabetsos, "Corpus design for a unit selection tts system with application to bulgarian," in *Proc. 4th Conference on Human language technology: challenges for computer science and linguistics*, 2009, pp. 35–46.