

Applying Phonological Feature Embeddings for Cross-Lingual Transfer in Text-to-Speech

Johannes Abraham Louw
Voice Computing Research Group
Council for Scientific and Industrial Research
Pretoria, South Africa
jalouw@csir.co.za

Zenghui Wang
Department of Electrical Engineering
University of South Africa
Pretoria, South Africa
wangz@unisa.ac.za

Abstract—In this work, we build upon our previous research where we introduced phonological features as input to text-to-speech systems. While the use of phonological features is not a novel concept in our research, our focus in this study is on the comprehensive analysis of the embeddings produced by the encoder model, which we believe offers novel insights into the model’s ability to capture and generalize phonological patterns across languages. Cross-lingual transfer experiments are conducted using both a resource-rich and a resource-constrained language to explore the model’s cross-lingual transfer capabilities across different linguistic families. The analysis of the embedding vectors produced by the encoder model is conducted using cluster maps to visualize the hierarchical clusters obtained using a clustering procedure. This analysis reveals the model’s learning patterns and provides insights into how phonological features contribute to the model’s ability to handle linguistic diversity and data scarcity.

Keywords—cross-lingual; phonological feature; resource-scarce; text-to-speech

I. INTRODUCTION

Modern non-autoregressive text-to-speech (TTS) neural architectures, such as FastSpeech2 [1] and VITS [2], have significantly advanced the field by enhancing the naturalness and expressiveness of synthesized speech. These advancements represent a departure from traditional paradigms, such as concatenative synthesis [3] and statistical parametric speech synthesis (SPSS) [4], which were limited by their reliance on handcrafted features and statistical models. Nevertheless, despite these advancements, there remains a substantial dependency on training data. This data requirement can pose challenges, especially in resource-scarce languages. Typically, more data reduces the reliance on domain-specific knowledge during the construction of a TTS model, whereas a scarcity of data increases the necessity for such specialized knowledge.

The recent advancements and successes in Large Language Models (LLMs) like the Generative Pre-trained Transformer (GPT) [5] have hinted that similar breakthroughs can be achieved in other sequence-to-sequence problems, such as TTS, by leveraging the capabilities of these powerful neural architectures. TTS systems like TorToiSe [6], VALL-E [7] and SPEAR-TTS [8] utilise LLMs as part of their architectures, requiring orders more data than is available for most languages (both SPEAR-TTS and VALL-E were trained with 60k hours of English speech), creating even more of a divide between

what is possible using data from a well resourced language versus a resource-scarce language. Several techniques have been proposed to address the resource-scarce challenges and enhance the quality of TTS synthesis in such contexts, these include: (a) transfer learning, which allows TTS models to leverage knowledge learned from related tasks or languages with more abundant data; (b) zero-shot and few-shot learning, techniques that aim to train TTS models using very limited annotated data; and (c) semi-supervised learning, a combination of a small amount of annotated data with a larger pool of unlabeled or weakly labeled data to train TTS models.

In this work, we employ transfer learning techniques to train TTS models in languages with severe resource constraints. To address the challenges posed by limited data availability in these languages, we introduce a novel approach that leverages phonological features (PFs) as input to the TTS system, diverging from conventional character or phoneme-based systems. The contributions of this research are threefold: Firstly, we propose a model architecture capable of accepting International Phonetic Alphabet (IPA) input from any language and generating corresponding phonological feature vectors for TTS input. This architecture serves as a universal framework adaptable to diverse linguistic contexts. Secondly, we conduct cross-lingual transfer experiments using both a resource-rich and a resource-constrained language. These languages are selected to represent a significant divergence in the language tree, allowing for an exploration of the model’s cross-lingual transfer capabilities across different linguistic families. Lastly, the research includes a comprehensive analysis of the embedding vectors produced by the encoder model. This analysis provides insights into the learned representations, shedding light on the model’s ability to capture and generalize phonological patterns across languages.

II. RELATED WORK

Phonological features are a combination of articulatory and auditory/acoustic features, providing a more abstract representation of speech sounds. This abstraction offers several benefits: it simplifies data by focusing on higher-level linguistic features, ensuring language independence and facilitating easier cross-lingual adaptation. PFs also exhibit robustness to variations in speech sounds, accommodating diverse accents,

dialects, and speaking styles. Furthermore, they enhance generalization and out-of-vocabulary handling within the TTS system, allowing for adaptability to unseen or uncommon words.

Related work has explored the utility of PFs in various contexts. Prior studies have focused on cross-lingual phone inventory mapping or direct integration of PFs in TTS models. For instance, [9] employed PFs for cross-lingual TTS but primarily concentrated on speaker adaptation, utilizing a training dataset two orders of magnitude larger than the present study. Similarly, [10] investigated PFs for cross-lingual TTS with a larger dataset and an autoregressive model. Another study by [11] delved into low-resource multilingual TTS but employed a different neural model and lacked a comparative analysis across varying dataset sizes for resource-scarce languages.

In previous work [12] on the use of phonological features for cross-lingual TTS transfer we reported on our methodology, data, experiments and results, which is summarized in the next section. In this work we examine in detail what happens on the embedding level of the text encoder.

III. METHOD, DATA, EXPERIMENTS AND RESULTS

In the proposed method, detailed in [12], the model architecture is derived from the VITS [2] model, a non-autoregressive end-to-end TTS system. The VITS model utilizes a conditional variational autoencoder (VAE) for waveform generation from text inputs, incorporating normalizing flows and adversarial training to enhance synthesized speech quality. Major modifications made to the vanilla VITS architecture include: (a) the original *Text Encoder* in VITS is replaced with a specialized phonological features encoder capable of processing binary phonological feature vectors; (b) the original *Decoder* in VITS is substituted with a Multi-Band inverse Short-Time Fourier Transform (MB-iSTFT) decoder.

A baseline and proposed model were developed. The baseline model adopts the standard multi-speaker MB-iSTFT variant of VITS, using phonemess as input. This model is chosen for its alignment with state-of-the-art methodologies, computational efficiency, and relatively minor degradation in naturalness compared to VITS. It incorporates a phoneme embedding lookup table with a dimension of 384. In the proposed model the text encoder is replaced with a phonological encoder. This encoder employs a sequential forward model with nonlinear activation, allowing the model to learn embeddings for multidimensional input features. The resulting output from the phonological encoder is then fed into the VITS architecture’s encoder. A dimension of 384 is selected to maintain comparability with the baseline.

English serves as the resource-rich source language, while isiXhosa was chosen as the resource-scarce target language. The speech data for both languages were accompanied by text transcriptions and sourced from a subset of the LibriTTS [13] corpus, focusing on male speakers to align with the composition of the resource-scarce datasets. To manage computational load during training, utterances containing 20 or fewer tokens are selected, resulting in a dataset of 12,000 utterances suitable

TABLE I. MEAN MUSHRA SUBJECTIVE NATURALNESS SCORES (0-100) WITH A 95% CONFIDENCE LEVEL AND OBJECTIVE RESULTS: MEL-CEPSTRAL DISTORTION (MCD), LOGARITHMIC SCALE, FUNDAMENTAL FREQUENCY (f_0) ROOT MEAN SQUARED ERROR (RMSE), LINEAR SCALE, FUNDAMENTAL FREQUENCY (f_0) MEAN ABSOLUTE ERROR (MAE), LINEAR SCALE, FUNDAMENTAL FREQUENCY (f_0) VOICING CLASSIFICATION PERCENTAGE ERROR (VCE), AND PEARSON CORRELATION COEFFICIENT OF FUNDAMENTAL FREQUENCY (f_0) VOICING CLASSIFICATION (PCC) (FROM [12]).

Voice	MUSHRA	MCD (dB) ↓	f_0			
			RMSE ↓ (Hz)	MAE ↓ (Hz)	VCE ↓ (%)	PCC ↑
xho5min_ph	49.36 ± 4.79	6.45	12.45	7.65	11.55	0.601
xho5min_pf	56.34 ± 4.1	6.01	11.09	6.66	11.76	0.692
xho15min_ph	60.65 ± 3.97	6.04	10.53	6.16	10.93	0.760
xho15min_pf	63.40 ± 3.58	5.70	10.68	6.15	10.75	0.771
xho30min_ph	60.99 ± 3.68	5.80	10.18	5.92	11.26	0.798
xho30min_pf	70.20 ± 3.78	5.44	9.77	5.62	11.05	0.818
xhoall_ph	65.52 ± 3.78	5.57	10.18	5.59	10.75	0.811
xhoall_pf	70.18 ± 3.4	5.24	10.89	6.11	11.01	0.813

for training on a single GPU. For the resource-scarce isiXhosa data, subsets were extracted from the Lwazi corpora [14], again focusing on male speakers and filtering for utterances with 20 or fewer tokens. To simulate varying degrees of resource scarcity, four subsets were chosen for each language, comprising 5-, 15-, 30 minutes and all speech data (2.5 hours for isiXhosa).

Test and validation sets were constructed by randomly selecting 10 test utterances and 20 validation utterances from each language, ensuring these are distinct from the training data. The validation set was used to assess model performance. Table I gives the main subjective and objective results from the experiments from [12], with a *Multiple Stimuli with Hidden Reference and Anchor* (MUSHRA) [15] test used for the subjective evaluation. We also investigated Afrikaans as a resource-scarce language, but in this work we focus on isiXhosa because the languages are not related at all and this fact highlights the results. As can be seen in the results, the proposed models (“*_pf”) outperform the baseline models (“*_ph”) in both the subjective and objective results, across all duration segments.

IV. EMBEDDINGS

In this work, we re-synthesized the selected test utterances for isiXhosa. To facilitate cross-lingual transfer, both the baseline and proposed models were initially trained on English data and subsequently fine-tuned using limited resources of isiXhosa data. The models under examination included: a baseline model trained on English data and fine-tuned with varying amounts of isiXhosa data (5-, 15-, 30 minutes, and 2.5 hours). Model I utilized phonemes as input, while Model II employed phonological features.

Of particular interest are consonant clicks, of which nine are modeled in our isiXhosa phonemene set: velarized dental click ([ǀ]), velarized alveolar lateral click ([ǁ]), post alveolar click ([ǃ]), dental click ([!]), aspirated dental click ([ǂ]), alveolar

upper front teeth or the alveolar ridge behind them. A velarized dental click combines the features of both velarization and dental articulation. Alveolar clicks are produced with the tongue against the alveolar ridge (the bump behind the upper front teeth). The acoustic properties of clicks differ from those of English consonants. Clicks tend to have a sharp onset and a relatively flat spectrum, making them easy to distinguish from other speech sounds. English consonants, on the other hand, have a more complex acoustic structure, with variations in intensity, duration, and spectral shape depending on the context in which they occur.

Our first investigation into the embeddings produced by the phoneme and phonological features models are given in Figure 1. We synthesized the isiXhosa tests sets with both Model I and II, trained with just English data. We captured the 384-dimension embeddings produced by each input token (phoneme or phonological feature vector), after which we standardized the values around their mean and standard deviation. We then plotted the cluster maps (a combination of a dendrogram and a heatmap) of the two models, with all the consonants compared to the clicks. Figure 1(a) gives the cluster map for Model I and 1(b) for Model II. Since Model I has not seen any clicks at this point in time, the clicks correlate with themselves because of the model initialisation, i.e. these values are still the initialized values and have not changed. Model II shows some smearing across a range of phonemes due to the inputs not being the actual phonemes, but rather their constituting phonological features. Most of these features have seen some samples in the English data and therefore the smearing effect where the clicks are initially a combination of different phonological features. The top of the dendrograms show the hierarchical clusters, obtained with a weighted average linkage clustering (WPGMA) procedure. In Figure 1(a) we can see that the clusters are random with regards to the clicks and no structure has yet been learned because the model has not seen any clicks yet. On the other hand, the hierarchical cluster in Figure 1(b) already shows some structure, with for example $[\text{||}]$ and $[\text{||}^h]$ clustered together and $[\text{g}^h]$ and $[\text{!}]$ clustered together.

For our second investigation we created cluster maps for the embeddings of the models trained on English data and fine-tuned on 5 minutes of isiXhosa data, given in Figure 2. We again synthesized the isiXhosa tests sets with both Model I and II and captured the 384-dimension embeddings produced by each input token (phoneme or phonological feature vector). We standardized the values around their mean and standard deviation. All the consonants are compared to the clicks. Figure 2(a) gives the cluster map for Model I and 2(b) for Model II. We can start observing the smearing effect in Figure 2(a), as was evident in Figure 1(b), due to the model starting to learn the data, and surprisingly, combining different elements from different phonemes for each click. The smearing effect is much more pronounced in Figure 2(b), with the model starting to learn which phonological features make up each individual click. In the hierarchical cluster of Figure 2(a) we can see that the model has clustered $[\text{g}^h]$ and $[\text{||}]$ together, and $[\text{g}^h]$

and $[\text{!}^h]$ together. The model is learning clusters from the raw data, even though no intrinsic pattern has been supplied as input (like in the case of the phonological features). In Figure 2(b) we can see larger clusters, $[\text{||}^h]$, $[\text{||}]$, and $[\text{g}^h]$ are clustered together, and $[\text{!}]$, $[\text{g}^h]$ and $[\text{!}^h]$ are clustered together, and finally $[\text{!}^h]$, $[\text{!}]$ and $[\text{g}^h]$ are clustered together. This is to be expected, as the model is getting a lot more fine grained information from the phonological features, which is easier to learn and apply across phonemes.

Our third investigation was into the evolution of the dendrograms of the embeddings as the amount of fine-tuning isiXhosa data grows, from 5-minutes, 10-minutes, 15-minutes, 30-minutes and lastly 2.5 hours. In these dendrograms we used all of the phonemes (including the vowels). Figure 3 (a) gives this evolution for Model I and Figure 3 (b) for Model II. From the dendrograms we can clearly see that Model II learns quicker than Model I, as Model II has already clustered all the clicks into three sub-groups after only 5-minutes of training data. At 15-minutes of data Model I is getting better, but the click clusters are still disjointed, whereas Model II has three click clusters of 3 clicks each, which then combine as the hierarchy moves up. By 30-minutes of fine-tuning data Model I has not improved much in comparison with 15-minutes, $[\text{||}^h]$ and $[\text{!}]$ are still not clustered with other clicks. The final 2.5 hour dendrogram for Model I shows an improvement in that all the clicks are clustered together, but they have not yet formed larger clusters up in the hierarchy as with Model II. At 30-minutes of fine-tuning data Model II's clusters have moved around, but the clicks still form 3 clusters of 3, we can also see plosive clusters and rudimentary vowel clusters forming. By 2.5 hours of fine-tuning data for Model II, we can see well established vowel, click and nasal clusters.

V. CONCLUSION

This work has presented a comprehensive analysis of TTS models utilizing phonological features as input, building upon our previous research in this area. While the concept of using PFs is not novel in our research trajectory, our in-depth analysis of the encoder model's embeddings offers novel insights into the model's ability to generalize phonological patterns across languages. Our experiments, conducted with both resource-rich English and resource-scarce isiXhosa, have demonstrated the effectiveness of the proposed PF-based models. The objective and subjective evaluations consistently showed that the PF-based models outperform the baseline models across various duration segments. However, there are several avenues for future work that warrant exploration. One potential direction is to investigate the incorporation of additional linguistic features alongside PFs to further improve model performance. Additionally, exploring the adaptability of the proposed model to other language families and linguistic contexts could broaden its applicability. Furthermore, while our analysis provides valuable insights, it also reveals some limitations. The reliance on phonological features may introduce complexities in handling certain linguistic nuances, and the model's performance may vary depending on the quality

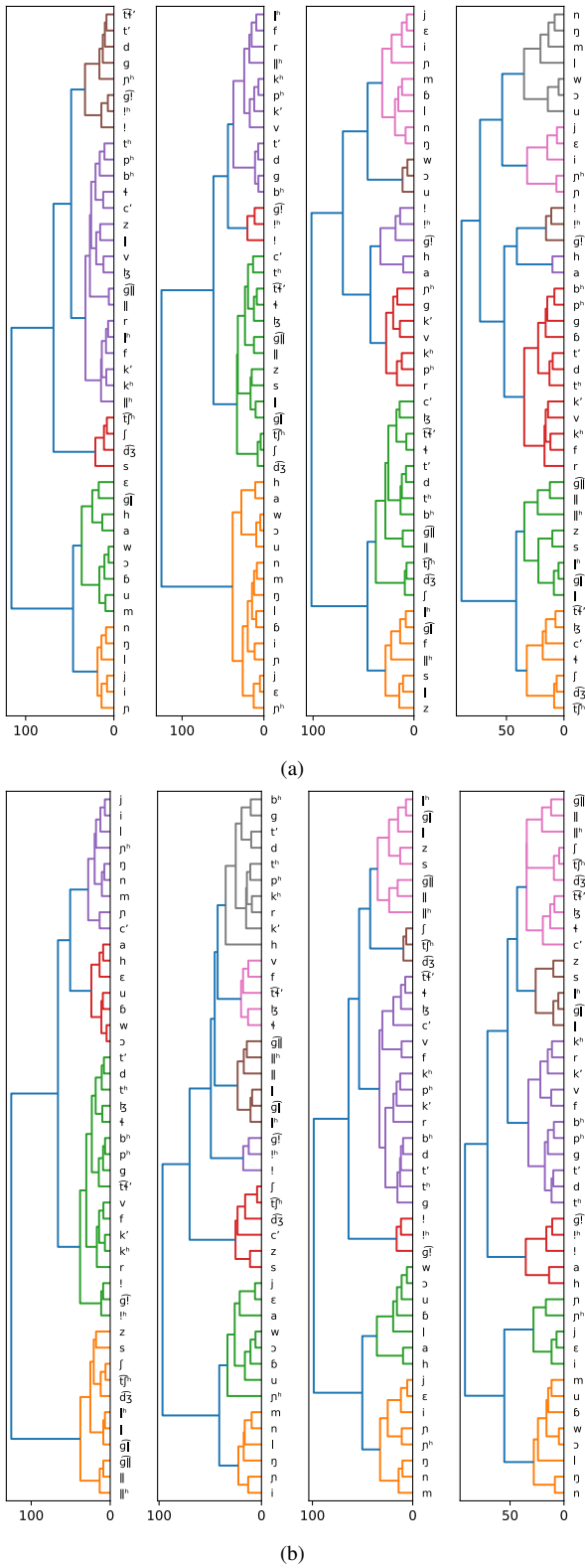


Fig. 3. The evolution of the dendrograms of the embeddings: trained on English and fine-tuned on 5-minutes, 15-minutes, 30 minutes and 2.5 hours of isiXhosa data. (a) Model I, phoneme input model; (b) Model II, phonological features input model.

and diversity of the input data. Addressing these limitations and improving the model’s robustness to linguistic variations are essential for its broader deployment in multilingual TTS applications. In summary, this work contributes to the growing body of research on TTS synthesis in resource-scarce languages by showcasing the potential of phonological features and providing a foundation for future studies in this domain. Despite the promising results, continuous efforts are needed to refine the model, address its shortcomings, and explore new avenues for enhancing multilingual TTS systems.

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 (ICLR)*, 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,” in *International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 5530–5540.
- [3] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Atlanta, Georgia, USA: IEEE, May 1996, pp. 373–376.
- [4] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [5] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. Xin Zhao, and J. Zhu, “Pre-trained models: Past, present and future,” *AI Open*, vol. 2, pp. 225–250, 2021.
- [6] J. Betker, “Better speech synthesis through scaling,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.07243>
- [7] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,” 2023. [Online]. Available: <https://arxiv.org/abs/2301.02111>
- [8] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, “Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.
- [9] 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 *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Sep. 2021, pp. 1594–1598.
- [10] D. Wells and K. Richmond, “Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis,” in *Proceedings of the 11th Speech Synthesis Workshop*, Budapest, Hungary, Aug. 2021, pp. 160–165.
- [11] 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*, Nov. 2022, pp. 741–751.
- [12] J. A. Louw, “Cross-lingual transfer using phonological features for resource-scarce text-to-speech,” in *Proceedings of the 12th Speech Synthesis Workshop*, Grenoble, France, Sep. 2023.
- [13] 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,” in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 1526–1530.
- [14] K. V. Calteaux, F. De Wet, C. Moors, D. Van Niekerk, B. McAlister, A. S. Grover, T. Reid, M. Davel, E. Barnard, and C. Van Heerden, “Lwazi II Final Report: Increasing the impact of speech technologies in South Africa,” CSIR, Tech. Rep., 2013.
- [15] R. ITU-R, “BS. 1534-1. Method for the subjective assessment of intermediate sound quality (MUSHRA),” *International Telecommunications Union, Geneva*, 2001.