Comparative Analysis of Interlingual Relationship in Multilingual Speech Synthesis

Youngjae Kim 20222255 | **Subin Kim** 20232933 | **Daehee Kim** 20232586 | **Sangwon Ryu** 20232130

Abstract

The recent advancements in the field of Text-to-Speech (TTS) have shown promising results. However, due to the limited availability of datasets, the challenge of synthesis in a low-resource still remains unresolved. To address this, researchers have explored joint training approaches that consider the relationships between languages. While previous studies have investigated criteria such as language family and phonetic similarity, consensus on their relative effectiveness is lacking. This project addresses these gaps by examining the significance of criteria in determining interlingual relationships. We conduct experiments with nine languages from diverse language families, adopting a joint training approach. We conduct an analysis of the interlingual relationships across nine languages from various language families by measuring the similarity at the n-gram level. Our analysis reveals that phoneme similarity is effective only within the same language family.

1 Introduction

Recent advancements in Text-to-Speech (TTS) have been substantial, yet achieving high-quality synthesized speech still necessitates extensive datasets. Obtaining a substantial amount of high-quality data is challenging, particularly in low-resource scenarios. In response to this challenge, researchers have explored approaches such as joint training [7] or cross-lingual transfer learning [6, 5, 1, 11] to enhance performance.

Previous studies have generally explored the effectiveness of criteria such as language family, based on phylogenetic considerations and phonetic similarity using phonemes to determine which criterion contributes more to performance. While there has been research suggesting that phonetic similarity is more helpful than language family in general, recent studies have not definitively clarified which criterion is more beneficial.

In this project, we aim to address the limitations of prior research and investigate which criterion holds the most significance in determining interlingual relationships. We conducted experiments with nine languages from a broader range of language families compared to previous studies. To better highlight the relationships between languages, we adopted a joint training approach rather than an unstable setting like low-resource scenarios. Furthermore, we suggest a consecutive phonemes approach to capture the intricate combinations of various phonetic units, where these phonemes collectively contribute to the formation of a distinct sound.

As mentioned above, we systematically analyze the impact of interlingual relationships on multilingual text-to-speech. We further investigate a new approach not covered in previous studies. To identify whether phylogenetic or phonetic aspects serve as more crucial indicators and reveal insights, we utilized several metrics to calculate the similarity between languages.

As a result, we find that the consecutive phonemes approach proved to be effective, particularly in the context of language family relationships. Through this project, we suggest a proper approach for efficient learning in multilingual speech synthesis.

2 Methodology

In our study, we analyze interlingual relationships using both phylogenetic and phonetic features as criteria for analysis. Furthermore, unlike previous research, we proposed a method that takes into consideration the features of speech by utilizing consecutive phonemes.

Table 1: Grouping the datasets for each language into a language family

Family	Language	Dataset
Germanic	*English*	[9]
	German	[13]
	Dutch	[13]
Romance	*Spanish*	[13]
	Portuguese	[3]
	French	[13]
Slavic	*Russian*	[13]
	Ukrainian	[15]
	Polish	[16]

2.1 Criteria

Phylogenetic features. As shown in Table 1, we chose Spanish, Portuguese, and French from the Romance family, English, German, Dutch from the Germanic family, and Russian, Ukrainian, and Polish from the Slavic family. In previous studies, language families were not evenly distributed, and the comparison involved few languages. We aim to achieve a more generalized result by including a broader range of languages in our analysis. We analyze the performance differences based on whether the languages learned together belong to the same language family. To classify language families, we consult the Glottolog database [12].

Phonetic features. For phonetic analysis, we utilize a total of five diverse similarity metrics, including Cosine, Jaccard, Kendall Tau, Spearman, and the Angular similarity employed in previous studies [5, 1].

2.2 Consecutive phone approach

In previous studies focusing on phonetic features, the approach primarily calculates similarity for a single symbol. Given the characteristics of speech, where different combinations of symbols yield distinct results, we believe that considering multiple symbols together would be more beneficial. Therefore, we group each symbol using an n-gram approach. When deriving the preceding phonetic features, we consider not only uniphones but also diphones and triphones.

3 Experiments and Analysis

3.1 Experimental Details

Dataset preparation. We collect open-source single-speaker TTS datasets for a total of nine languages in Table 1. To ensure a balanced representation, the datasets are adjusted to align with a 10-hour duration for each language. Additionally, we preprocess all audio samples to a sample rate of 22050. To calculate the similarity for all languages, we follow previous research and convert the text of each language into a common set of phonetic symbols using the International Phonetic Alphabet (IPA). For IPA conversion, we utilize the phonemizer library [2].

Model training. We employ the End-to-End non-auto-regressive model, VITS [10], as our speech synthesis model. VITS is a model designed for single-language TTS. To enhance adaptability for multilingual TTS, we integrate additional modules, including language embeddings, inspired by multilingual TTS models based on VITS [8, 4]. For training, we standardize all speech datasets to a 22,050 sample rate and utilized 8 A5000 GPUs. The training was conducted with 16 batch sizes and 1000 epochs.

Output evaluation. We conduct training by grouping two languages from nine languages. For evaluating the pronunciation of the synthesized speech, we calculate the Character Error Rate (CER). To obtain generalized results, we extract outcomes from datasets not used in training across five random seeds. The whisper-large-v2 model from OpenAI [14] is utilized for CER computation.

Table 2: The CER performance for all language pairs. The results represent the mean of 5 random seeds ($\bar{x}$) with standard deviations (σ) all below 0.02.

	English	German	Dutch	Spanish	Portuguese	French	Russian	Ukrainian	Polish
English	$0.021 \pm (0.003)$	$0.025 \pm (0.003)$	$0.027 \pm (0.005)$	$0.02 \pm (0.002)$	$0.024 \pm (0.003)$	$0.029 \pm (0.009)$	$0.021 \pm (0.004)$	$0.019 \pm (0.004)$	$0.025 \pm (0.004)$
German	$0.031 \pm (0.002)$	$0.035 \pm (0.003)$	$0.039 \pm (0.004)$	$0.032 \pm (0.005)$	$0.039 \pm (0.001)$	$0.039 \pm (0.007)$	$0.031 \pm (0.002)$	$0.033 \pm (0.003)$	$0.028 \pm (0.004)$
Dutch	$0.101 \pm (0.013)$	$0.106 \pm (0.012)$	$0.103 \pm (0.013)$	$0.107 \pm (0.013)$	$0.143 \pm (0.015)$	$0.144 \pm (0.014)$	$0.11 \pm (0.007)$	$0.108 \pm (0.008)$	$0.157 \pm (0.054)$
Spanish	$0.022 \pm (0.003)$	$0.026 \pm (0.003)$	$0.036 \pm (0.002)$	$0.025 \pm (0.004)$	$0.026 \pm (0.001)$	$0.029 \pm (0.003)$	$0.022 \pm (0.003)$	$0.029 \pm (0.003)$	$0.025 \pm (0.003)$
Portuguese	$0.089 \pm (0.019)$	$0.088 \pm (0.01)$	$0.113 \pm (0.016)$	$0.084 \pm (0.017)$	$0.084 \pm (0.014)$	$0.096 \pm (0.02)$	$0.072 \pm (0.012)$	$0.077 \pm (0.016)$	$0.077 \pm (0.014)$
French	$0.136 \pm (0.019)$	$0.096 \pm (0.006)$	$0.129 \pm (0.008)$	$0.107 \pm (0.01)$	$0.099 \pm (0.008)$	$0.096 \pm (0.01)$	$0.084 \pm (0.007)$	$0.114 \pm (0.006)$	$0.093 \pm (0.006)$
Russian	$0.11 \pm (0.018)$	$0.118 \pm (0.011)$	$0.145 \pm (0.016)$	$0.106 \pm (0.013)$	$0.144 \pm (0.015)$	$0.124 \pm (0.012)$	$0.102 \pm (0.011)$	$0.108 \pm (0.016)$	$0.158 \pm (0.015)$
Ukrainian	$0.165 \pm (0.012)$	$0.176 \pm (0.013)$	$0.191 \pm (0.015)$	$0.176 \pm (0.009)$	$0.174 \pm (0.011)$	$0.192 \pm (0.008)$	$0.163 \pm (0.011)$	$0.163 \pm (0.007)$	$0.174 \pm (0.012)$
Polish	$0.056 \pm (0.007)$	$0.035 \pm (0.006)$	$0.068 \pm (0.01)$	$0.042 \pm (0.009)$	$0.038 \pm (0.009)$	$0.043 \pm (0.007)$	$0.06 \pm (0.006)$	$0.038 \pm (0.002)$	$0.035 \pm (0.007)$

Table 3: Correlation between Character Error Rate (CER) and Similarity Metrics for Different n-grams, Analyzed Across Language Family Contexts. The features with the strongest correlation for each similarity are bolded.

	All			Only Family			Only non-Family		
	1-gram	2-gram	3-gram	1-gram	2-gram	3-gram	1-gram	2-gram	3-gram
Jaccard	0.091	0.224	0.153	**-0.18**	-0.02	-0.004	0.168	0.337	0.335
Spearman	0.134	0.107	0.036	**-0.267**	-0.236	-0.268	0.368	0.311	0.248
Cosine Similarity	-0.021	-0.138	-0.029	-0.464	**-0.473**	-0.296	0.208	0.005	0.111
Angular Similarity	-0.043	-0.142	-0.037	-0.455	**-0.468**	-0.292	0.206	0.005	0.111
Kendal tau	0.143	0.114	0.042	-0.237	-0.211	**-0.247**	0.384	0.313	0.246

3.2 Results

The CER results for all language pairs are displayed in Table 2. Since CER performance across language pairs alone does not provide a clear correlation, we delve deeper into the data for each criterion to conduct a more comprehensive analysis.

Phylogenetic features. When assessing the performance difference in CER based on language family, we observe a slight improvement. However, the p-value remains unusually high, making it difficult to consider the result statistically significant and casting doubt on its reliability.

Phonetic features. When examining the correlation between phonetic similarity and CER performance, as shown in Table 3, no significant relationships are observed for all languages. However, when divided by language family, there was a notable negative correlation between phonetic similarity and CER performance within the same language family. Additionally, the approach based on consecutive phonemes showed a tendency to be beneficial.

3.3 Discussion

As evidenced by the results, it becomes apparent that relying solely on phylogenetic features does not yield significant correlations. However, when analyzing based on phonetic features, it is observed that meaningful indicators emerge when considering similarity within language families. This underscores the importance of evaluating similarity based on language families in the context of phonetic features. Furthermore, the similarity based on consecutive phonemes demonstrated a stronger correlation, suggesting that considering consecutive phonemes, as opposed to the previous study's single phoneme-based approach, was more meaningful.

4 Conclusion

In this study, we analyze the significant impact of phylogenetic and phonetic features on performance in multilingual text-to-speech. Previously, there was a prevailing conclusion that phonetic features were more crucial, but recent studies have not definitively identified what exactly is significantly important. Our analysis of the performance differences based on the presence of each factor reveals that considering both phonetic and phylogenetic features together rather than separately is meaningful. Furthermore, by considering the similarity based on consecutive phonemes instead of the uniphone frequency used in previous studies, we are able to obtain more meaningful correlations. Therefore, we propose a more efficient approach for multilingual text-to-speech. In future work, we plan to conduct additional research to investigate the effectiveness of this scenario in low-resource settings.

References

[1] Strategies in transfer learning for low-resource speech synthesis: Phone mapping, features input, and source language selection. In: 12th ISCA Speech Synthesis Workshop (SSW2023). pp. 21–26. ISCA (Aug 2023). https://doi.org/10.21437/ssw.2023-4, 12th ISCA Speech Synthesis Workshop (SSW2023) ; Conference date: 26-08-2023 Through 28-08-2023

[2] Bernard, M., Titeux, H.: Phonemizer: Text to phones transcription for multiple languages in python. Journal of Open Source Software **6**(68), 3958 (2021). https://doi.org/10.21105/joss.03958, https://doi.org/10.21105/joss.03958

[3] Casanova, E., Junior, A.C., Shulby, C., Oliveira, F.S.d., Teixeira, J.P., Ponti, M.A., Aluísio, S.: Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese. Language Resources and Evaluation pp. 1–13 (2022)

[4] Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A.: Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In: International Conference on Machine Learning. pp. 2709–2720. PMLR (2022)

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

[6] Do, P., Coler, M., Dijkstra, J., Klabbers, E.: A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages. In: Proc. Interspeech 2021. pp. 16–20. ISCA (Aug 2021). https://doi.org/10.21437/Interspeech.2021-1565, interspeech 2021 ; Conference date: 30-08-2021 Through 03-09-2021

[7] Gutkin, A., Sproat, R.: Areal and phylogenetic features for multilingual speech synthesis. In: Proc. of Interspeech 2017. pp. 2078–2082. August 20–24, 2017, Stockholm, Sweden (2017), https://www.isca-speech.org/archive/Interspeech_2017/pdfs/0160.PDF

[8] Hyunjae Cho, Wonbin Jung, J.L.S.H.W.: Sane-tts: Stable and natural end-to-end multilingual text-to-speech. In: International Conference on Machine Learning (2022)

[9] Ito, K., Johnson, L.: The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/ (2017)

[10] Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. pp. 5530–5540. PMLR (2021)

[11] Lux, F., Koch, J., Vu, N.T.: Low-resource multilingual and zero-shot multispeaker TTS. In: He, Y., Ji, H., Li, S., Liu, Y., Chang, C.H. (eds.) 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 (Volume 1: Long Papers). pp. 741–751. Association for Computational Linguistics, Online only (Nov 2022), https://aclanthology.org/2022.aacl-main.56

[12] Nordhoff, S., Hammarström, H.: Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources (01 2011)

[13] Park, K., Mulc, T.: Css10: A collection of single speaker speech datasets for 10 languages. Interspeech (2019)

[14] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)

[15] Smoliakov, Y.: Lada: Ukrainian High-Quality Female Text-to-Speech Dataset (Dec 2022). https://doi.org/10.5281/zenodo.7396774, https://doi.org/10.5281/zenodo.7396774

[16] Solak, I.C.: The m-ailabs speech dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/ (2019)