A Systematic Review and Analysis of Multilingual Data Strategies in Text-to-Speech for Low-Resource Languages

Phat Do[1], Matt Coler[1], Jelske Dijkstra[12], Esther Klabbers[3]

[1]University of Groningen, Campus Fryslân, Leeuwarden, the Netherlands
[2]Fryske Akademy/Mercator Research Centre, Leeuwarden, the Netherlands
[3]ReadSpeaker, Driebergen-Rijsenburg, the Netherlands

`{t.p.do, m.coler, j.e.dijkstra}@rug.nl, esther.judd@readspeaker.com`

Abstract

We provide a systematic review of past studies that use multilingual data for text-to-speech (TTS) of low-resource languages (LRLs). We focus on the strategies used by these studies for incorporating multilingual data and how they affect output speech quality. To investigate the difference in output quality between corresponding monolingual and multilingual models, we propose a novel measure to compare this difference across the included studies and their various evaluation metrics. This measure, called the Multilingual Model Effect (MLME), is found to be affected by: acoustic model architecture, the difference ratio of target language data between corresponding multilingual and monolingual experiments, the balance ratio of target language data to total data, and the amount of target language data used. These findings can act as reference for data strategies in future experiments with multilingual TTS models for LRLs. Language family classification, despite being widely used, is not found to be an effective criterion for selecting source languages.

Index Terms: text-to-speech, speech synthesis, low-resource languages, multilingual synthesis, cross-lingual synthesis

1. Introduction

1.1. Low-resource languages and multilingual TTS models

Low-resource languages (LRLs) are languages that lack sufficient data for research and development of language tools. In the domain of speech synthesis or text-to-speech (TTS), this translates to a lack of data for training acoustic models, which are usually high-quality annotated recordings of human speakers. This shortage in training data hinders both research and applications of TTS for LRLs. While this has been an issue since the era of traditional statistical parametric speech synthesis, it is worsened in the current TTS research climate, due to the dominance of "data-hungry" deep neural networks (DNNs). The recent introduction and increasing popularity of sequence-to-sequence (S2S) TTS models further highlight the problem: they output much more natural speech, but also require significantly large amounts of training data. Thus, these latest TTS models bring an opportunity for languages with more abundant data, but exacerbate the data scarcity challenge for LRLs.

To solve this issue of insufficient data in TTS for LRLs, two approaches have been proposed. The first is to use models that can be trained on multi-speaker data, i.e., speech data from many speakers who each contribute a small amount of recordings, but constitute an adequate amount of data when combined. This approach has been fruitfully explored by many studies (e.g., [1], [2], [3], etc.), but unfortunately such sources of multi-speaker data are not always available for LRLs. In such

cases, the second approach may hold more promise: using models that can be trained on multilingual data, i.e., data that comes not only from the language to be synthesized (target language, usually an LRL), but also from other languages with more readily available data (source or auxiliary languages), to make up for the small amount of data in the target language.

Research in TTS using multilingual data has been done since the early 2000s (e.g., [4]), spanning across the architecture paradigm from Hidden Markov Models (HMM)-based models to the latest S2S models. However, to the best of our knowledge, there has yet to be a systematic review that provides a general insight into the effect of using multilingual TTS models for LRLs. In addition, given the aforementioned data challenge for LRLs, it is helpful to have an analysis that focuses on the data augmentation strategies of past studies, i.e., how they incorporated multilingual training data into their TTS experiments. By finding correlations between these particularities in data augmentation strategy and the synthesized speech quality, such an analytical review can be useful for future research in TTS for LRLs that use multilingual data, especially regarding the efficiency in using such data.

1.2. Objectives and research questions

This contribution is thus dedicated to identifying studies that use multilingual TTS models for LRLs, and providing a systematic and analytical review using the experimental designs and reported results of such studies. For the latter, we identified two research questions:

RQ1 Using the same limited amount of LRL data, how does the output quality of multilingual TTS models compare to that of monolingual models?

RQ2 What factors in the data augmentation strategy influence the effect of using multilingual TTS models on output quality, and to what extent do they affect it?

Section 2 details the study selection process and the selected studies. Sections 3 and 4 provide analyses and discussions targeted at each of the research questions. Section 5 offers conclusions and directions for future research.

2. Study selection

We searched for relevant studies using the Scopus database and the Online Archive of the International Speech Communication Association (ISCA)[1] for more recent publications not yet part of Scopus. The included studies had to meet the following criteria:

---

[1]`https://www.isca-speech.org/iscaweb/index.php/online-archive`

1) delineate the design, evaluation, and training data used in the TTS experiment being reported, 2) discuss multilingual or cross-lingual TTS models, and 3) deal with LRLs or similar situations in which target language data was lacking. The search resulted in a total of 23 studies that met these criteria.

2.1. Characteristics of the included studies

Table 1 details the 23 included studies and their acoustic model architectures, while Table 2 shows the output speech evaluation metrics used in these studies. We extracted all reported values of output speech quality from these studies, along with corresponding experimental details, e.g., name and family of the target language(s) and source language(s), amount of speech data used for target and source language(s), etc. These resulting values ($n = 880$) were used for analysis.

Table 1: *Included studies and model architectures*

Qty.	Studies	Architecture
1	[5]	Unit selection synthesis
7	[6], [7], [8], [9], [10], [11], [12]	Hidden Markov Model synthesis (HMM)
7	[12], [13], [14], [15], [16], [17], [18]	Neural network (non-S2S) synthesis (DNN)
9	[19], [20], [21], [22], [23], [24], [25], [26], [27]	Sequence-to-sequence synthesis (S2S)

2.2. Notable studies

From the chosen studies, notable studies include [18] (hereafter Study A), the latest included DNN-based study, which used 5 different evaluation metrics for TTS in Tibetan (with Mandarin as the source language). For S2S-based studies, among the latest are [27] (Study B), which explored TTS for Indic LRLs in the Indo-Aryan and Dravidian families, and [25] (Study C), which investigated strategies for using Dutch and other European languages to aid a limited amount of English data.

3. Multilingual vs. monolingual models with the same target data

3.1. Multilingual Model Effect (MLME)

For a straightforward comparison between multilingual and monolingual models, we focused only on groups of experiments that: 1) came from the same study, 2) shared everything in research design, especially target language data, and 3) differed from each other only in whether or not they were also trained on source language data (i.e., multilingual vs. monolingual).

From these studies, we wanted to obtain a measure that captured the difference in reported values for output quality between multilingual and monolingual models. Since we wanted to compare across the included studies, which used a wide range of evaluation metrics (as shown in Table 2), we needed a cross-study, cross-metric measure. As such a measure did not yet exist, we created one. This measure, hereafter referred to as the *MultiLingual Model Effect (MLME)*, was derived from the reported results as follows:

$$MLME(\%) = \frac{v_{multi} - v_{mono}}{v_{mono}} * 100 * \begin{cases} 1, \text{if } m \sim \text{quality (*)} \\ -1, \text{otherwise} \end{cases}$$

Table 2: *Output evaluation metrics used in included studies*

Metric	Group	Studies
MCD (Mel-Cepstral Distortion)	Acoustics	[18], [21], [27]
F0 RMSE (Root-Mean-Square Error)	Acoustics	[12], [18]
V/UV (Voiced/UnVoiced error)	Acoustics	[12], [18]
BAP (Band APeriodicities distortion)	Acoustics	[18]
DTWMCD (Dynamic Time Warping Mel-Cepstral Distortion)	Acoustics	[19]
L2 NSE (L2 Norm-Squared on log-Mel spectrogram)	Acoustics	[24]
LF0 RMSE (Log F0 Root-Mean-Square Error)	Acoustics	[11]
LSD (normalized Log Cepstral Distance)	Acoustics	[12]
MGC RMSE (Mel-Generalized Cepstrum coefficients Root-Mean-Square Error)	Acoustics	[11]
MSE (Mean-Square Error)	Acoustics	[13]
CER (Character Error Rate)	Intelligibility	[21], [26]
Intelligibility % (percentage of intelligible sentences)	Intelligibility	[7]
SUS-Wacc (Word accuracy in Semantically Unpredictable Sentences)	Intelligibility	[23]
WER (Word Error Rate)	Intelligibility	[20]
MOS (Mean Opinion Score)	Naturalness/ Quality	[6], [8], [9], [14], [15], [16], [17], [18], [21], [23]
A/B Preference (preference rate b/w test & control)	Quality	[5], [10], [11], [12], [13]
MUSHRA (MUltiple Stimuli with Hidden Reference & Anchor)	Naturalness	[25]
DMOS (Degradation MOS)	Similarity	[18], [22], [27]

where v_{multi} and v_{mono} are the reported values of output quality from the corresponding multilingual and monolingual models, respectively, and (*) is the scenario in which the metric m positively correlates with general output quality (the higher, the better, e.g., MOS, MUSHRA, etc.), as opposed to the opposite correlation (the lower, the better, e.g., MCD, WER, etc.).

3.2. Results

We obtained MLME values from the following studies: [6], [12], [13], [14], [16], [17], [18], [20], [22], [25], [26], and [27], and reported them in Table 3, both as a whole and in specific groups of evaluation metrics, in the form of median (*M*) and interquartile range (*IQR*). Also reported are the *p*-values of the corresponding one-sample Wilcoxon signed rank tests for the hypothesis that the median MLME values are larger than 0.

3.3. Discussion

Given the same amount of target language data, there were statistically significant improvements in output quality of multilingual models over monolingual models, both in general and group-specific cases. Despite the wide variation in the reported medians, pairwise comparisons using Wilcoxon rank sum tests did not indicate significant differences in the MLME among the four groups of evaluation metrics. In other words, multilingual models had better output quality than their monolingual coun-

Table 3: *Multilingual-model effect (MLME) analysis*

Metric group	n	MLME (%)		Wilcoxon test
		M	IQR	(M > 0)
All	159	3.6	19.5	$p < .001$
Acoustics	79	0.7	11.4	$p = .002$
Intelligibility	13	20.5	20.5	$p = .03$
Naturalness/Quality	55	4.6	24.8	$p = .02$
Similarity	12	12.1	16.7	$p = .01$

terparts, and this improvement did not differ based on the type of evaluation metric used. Given the rather low median MLME (3.6%), it was necessary to investigate the data augmentation strategies of the included studies in order to find relevant factors contributing to the low MLME values.

4. Influential factors in data augmentation strategy for multilingual models

Statistical models were fitted to the collected data set, with the MLME as the dependent variable and five proposed independent variables, which are described in the section below.

4.1. Proposed independent variables

4.1.1. Ratio of target language data in multilingual vs. monolingual models (tgt.data.ratio)

Section 3 only used MLME values calculated from groups of experiments with the same amount of target language data in corresponding multilingual and monolingual models. However, some included studies also conducted experiments with different amounts of target language data between the two. Including such experiments ($n = 87$) to the data set would have two benefits: 1) increasing the sample size considerably, and 2) adding a potential predictor of the MLME: the relative difference in the amounts of target language data between corresponding multilingual and monolingual models. Thus, MLME values of these experiments were calculated (also following 3.1) and included in the data set for RQ2. The predictor mentioned above was then the first independent variable, a ratio calculated as follows:

$$tgt.data.ratio(\%) = \frac{tgt.data_{multi} - tgt.data_{mono}}{tgt.data_{mono}} * 100$$

where $tgt.data_{multi}$ and $tgt.data_{mono}$ are the amounts of target language data in the corresponding multilingual and monolingual models, respectively. This variable was added to the data set not only for the experiments discussed above, but also for other experiments that used the same amount of target language data in their corresponding multilingual and monolingual models. In such cases, this variable would be equal to 0.

4.1.2. Acoustic model architecture (arch.)

This represents the choice of acoustic model architecture (shown in Table 1) in the experiments. Due to a lack of comparisons between multilingual and monolingual models, no unit-selection experiments and only 7 HMM-based experiments were eligible for inclusion. Upon inspection, MLME values from the latter also turned out to be outliers (resulting from exploratory experiments with unusually high MLME values). Considering this and the small sample size, these HMM-based experiments were excluded, leaving 164 DNN-based and 76

S2S-based experiments for the analyses. Output quality of S2S models is generally said to be sensitive to training data quantity (discussed in 1.1), so it was important to include this variable.

4.1.3. Amount of target language data (tgt.data)

This stands for the amount of training data used for the target language and was included to test for its potential effect on the MLME. Since all studies (except for [19] and [26]) mentioned this only in either time duration or number of utterances, an average estimation was needed to convert it to a common measurement. Following the descriptions provided in [10], [11], [12], [16], [17], [20], [21], [22], [23], [24], and [27], we obtained an average utterance length of the speech data sets used in these studies: 6.1 seconds. This was then used to convert all the training data quantities to the corresponding estimated number of utterances.

4.1.4. Ratio of balance between target language data and total data (balance.ratio)

This variable represents the "balance" between target language data and total training data, and was calculated as follows:

$$balance.ratio(\%) = \frac{tgt.data}{tgt.data + src.data} * 100$$

where $tgt.data$ and $src.data$ are the amounts of training data (converted to number of utterances) for the target language and source language(s), respectively. By including this variable and testing its effect on the MLME, we aimed to investigate the role of this balance in experiments with multilingual TTS models.

4.1.5. Ratio of source data from the same language family to total source data (same.family.ratio)

A popular strategy is to use data from source languages that are close to the target language, so that the model can learn better from the combined data (due to more similarities between these languages). However, this "closeness" and its effect, to the best of our knowledge, are still arbitrary and there has been no conclusive evidence to date. Many of the included studies chose source languages that were from the same language family as the target language, implying the use of language family classification as an indicator of the mentioned "closeness". To investigate this, we included this variable, calculated as follows:

$$same.family.ratio(\%) = \frac{same.family.src.data}{src.data} * 100$$

where $same.family.src.data$ is the amount of source data that comes from languages in the same family as the target language, and $src.data$ is the total amount of source data. The concept of language family here follows the "phylogenetic features" used in [14] and [15], which used classifications provided by [28]. For each of the languages in the data set, its classification was obtained from [28] and the third level of subgroup[2] (or the lowest level for languages with less than three levels) was used to determine the language family. For example, Bengali's classification is Indo-European > Indo-Iranian > Indo-Aryan > Outer Languages > Eastern > Bengali-Assamese, so it was considered as belonging to the "Indo-Aryan" language family.

---

[2]Different subgroup levels were tested and this resulted in the best match with the classifications used in the included studies.

4.2. Predictive model for MLME

From the final data set ($n = 240$), we set out to find a good predictive model of the MLME from the independent variables described in 4.1. As many experiments came from the same studies, linear mixed effect models ([29]) were used to account for the by-study variability. To test for the significance of the effects, p-values were obtained by likelihood tests of the proposed models and "null" models (those without the effect being tested for). One by one, the independent variables were added to the model, together with interactions among them (where suitable), and statistically significant variables would be kept (except ones that led to warnings of singular fits). For reference, the step-down model-building approach provided by [30] was also used, resulting in the same model as the manual approach.

4.3. Results

Acoustic model architecture had a significant effect on the MLME ($p < .001$), resulting in a large difference (33.9 percentage points, pp) between the mean MLME values of the two groups. Considering 1) this large difference in mean values, 2) the better output quality of S2S models relative to that of DNN models, and 3) the different orders of importance in the effects, these two architectures will be reported separately. In Table 4 and Table 5, the effects are reported using their regression coefficients (B) and standard errors (SE), accompanied by significance codes from their corresponding p-values, and in the order of decreasing importance based on their standardized regression coefficients (β). Their standard deviations (SD) are also reported for effect size comparison.

Table 4: *Effects of data strategy on MLME in DNN models*

Effect (unit)	B $\pm$ SE (pp)	SD	β
tgt.data (100 utt.)	-0.06 $\pm$ 0.02 (*)	83.92	-0.19 (-5.34)
balance.ratio (1 pp)	0.66 $\pm$ 0.4 (*)	8.71	0.12 (3.37)
tgt.data.ratio (1 pp)	0.02 $\pm$ 0.12 (***)	44.08	0.07 (1.97)
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05			

DNN models had a mean MLME value of -0.35% ($\pm$ 9.7%). From this "base" value, corresponding MLME values for specific experiment details could be predicted using the reported values in Table 4. For example, using the B for *tgt.data*, when all other effects are held constant, an increase of 100 utterances in *tgt.data* decreases the MLME by 0.06 pp (from the "base" value). Using the β, when *tgt.data* goes up by 1 SD (8,392 utterances), the MLME goes down by 0.19 SD (5.34 pp). Similar interpretations can be done for the remaining effects.

Table 5: *Effects of data strategy on MLME in S2S models*

Effect (unit)	B $\pm$ SE (pp)	SD	β
tgt.data.ratio (1 pp)	0.47 $\pm$ 0.10 (***)	39.65	0.72 (20.23)
balance.ratio (1 pp)	-0.52 $\pm$ 0.26 (.)	15.90	-0.22 (-6.18)
tgt.data (100 utt.)	-0.06 $\pm$ 0.02 (*)	83.92	-0.19 (-5.34)
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1			

S2S models had a mean MLME value of 33.56% ($\pm$ 7.64%) and the effects reported in Table 5 can be interpreted similarly to those in DNN models.

4.4. Discussion

For DNN models: The amount of target data is the most important factor. The more target language data there is available, the less potential increase there is in output quality when using multilingual models. This is in line with expectations. This effect does not differ for S2S models.

The MLME also positively correlates with the data balance ratio. In Study A (mentioned in 2.2), this ratio was 28.75% (the highest in all included DNN studies) in its experiments with the top 25% MLMEs. These experiments had a median (M) MLME of 7.14%, higher than the third quartile (Q_3) of the MLME in DNN studies (5.58%). Thus, for DNN models, when adding data from source languages, the target data - total data balance ratio should still be kept high to maintain data efficiency.

For S2S models: Despite the small effect size in DNN models, here the ratio of target data difference (between corresponding multilingual and monolingual models) affects the MLME the most. Thus, in TTS for LRLs using S2S models, it is better to use as much target language data as there is available (as opposed to, e.g., replacing part of the target data by source data to avoid increasing the total amount of training data).

Noticeably, compared to DNN models, the data balance ratio has a reverse effect on the MLME in S2S models. The highest MLME value from Study C had the lowest data balance ratio, and the top 50% MLME values of Study B ($M = 26.18\%$, higher than the Q_3 of S2S studies - 24.63%) had a median data balance ratio of only 2.04%, almost as low as the Q_1 of S2S studies - 1.37%. This could be a sign that, for S2S models, having more training data is always beneficial, and this is even more important than keeping a high ratio of target data over total data.

For both: The ratio of source data from the same language family over total source data did not significantly affect the MLME. Therefore, language family, or at least its classification method used here, is not viable as a criterion for selecting source languages. For future research, more elaborate methods should be tested, e.g., those that involve measurements of shared phonetic space or phoneme inventories of the languages.

5. Conclusions

Using data from previous studies, we investigated the change in speech output quality between multilingual models and corresponding monolingual models, given the same limited amount of data from a low-resource target language. We designed a novel metric to measure this change in output quality across the included studies and the wide range of evaluation metrics used by them. Via this measure, we confirmed the improvement in output speech quality in the multilingual models over their monolingual counterparts.

Going further, we set out to find factors in the strategies for incorporating multilingual data that affected this improvement. We found that it was affected by the ratio of target language data between corresponding multilingual and monolingual models, the balance ratio of target language data over total training data, and the amount of target language data, but in different orders of importance and manners depending on the acoustic model architecture used. These effects should be empirically tested in future research and, if proven, can be used for reference in multilingual text-to-speech models for low-resource languages.

Language family classification, despite being used by many included studies, proved ineffective as a criterion for selecting source languages. Future research is intended to further investigate this topic.

6. References

[1] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[2] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[3] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and V. Klimkov, "Effect of data reduction on sequence-to-sequence neural tts," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7075–7079.

[4] A. W. Black and K. A. Lenzo, "Multilingual text-to-speech synthesis," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–761.

[5] N.-H. Samsudin and M. Lee, "Building text-to-speech systems for resource poor languages," in *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2012, pp. 3327–3334.

[6] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[7] M. Mustafa, R. Ainon, R. Zainuddin, Z. Don, and G. Knowles, "A cross-lingual approach to the development of an HMM-based speech synthesis system for Malay," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, 2011, pp. 3197–3200.

[8] M. Mustafa, Z. Don, R. Ainon, R. Zainuddin, and G. Knowles, "Developing an HMM-based speech synthesis system for malay: A comparison of iterative and isolated unit training," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 5, pp. 1273–1282, 2014.

[9] H. Yang, K. Oura, H. Wang, Z. Gan, and K. Tokuda, "Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9927–9942, 2014.

[10] T. Justin, F. Mihelič, and J. Žibert, "Towards automatic cross-lingual acoustic modelling applied to hmm-based speech synthesis for under-resourced languages," *Automatika*, vol. 57, no. 1, pp. 268–281, 2016.

[11] S. Sarfjoo and C. Demiroglu, "Cross-lingual speaker adaptation for statistical speech synthesis using limited data," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, 2016, pp. 317–321.

[12] Q. Yu, P. Liu, Z. Wu, S. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, 2016, pp. 5545–5549.

[13] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, 2016, pp. 2468–2472.

[14] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, 2017, pp. 2183–2187.

[15] A. Gutkin and R. Sproat, "Areal and phylogenetic features for multilingual speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, 2017, pp. 2078–2082.

[16] I. Demirsahin, M. Jansche, and A. Gutkin, "A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech," in *SLTU*, 2018.

[17] J. Wibawa, S. Sarin, C. Li, K. Pipatsrisawat, K. Sodimana, O. Kjartansson, A. Gutkin, M. Jansche, and L. Ha, "Building open Javanese and Sundanese corpora for multilingual text-to-speech," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, pp. 1610–1614.

[18] W. Zhang, H. Yang, X. Bu, and L. Wang, "Deep learning for Mandarin-tibetan cross-lingual speech synthesis," *IEEE Access*, vol. 7, pp. 167 884–167 894, 2019.

[19] P. Baljekar, S. Rallabandi, and A. Black, "An investigation of convolution attention based models for multilingual speech synthesis of Indian languages," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, 2018, pp. 2474–2478.

[20] Y. Lee and T. Kim, "Learning pronunciation from a foreign language in speech synthesis networks," *ArXiv*, 2018.

[21] 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," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, 2019, pp. 2075–2079.

[22] A. Prakash, A. L. Thomas, S. Umesh, and H. Murthy, "Building Multilingual End-to-End Speech Synthesisers for Indian Languages," in *Proc. of 10th ISCA Speech Synthesis Workshop (SSW'10)*, 2019, pp. 194–199.

[23] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages," *IEEE Access*, 2020.

[24] S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura, "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis," *arXiv preprint arXiv:2011.02128*, 2020.

[25] M. d. Korte, J. Kim, and E. Klabbers, "Efficient Neural Speech Synthesis for Low-Resource Languages Through Multilingual Modeling," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2967–2971.

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

[27] A. Prakash and H. A. Murthy, "Generic Indic Text-to-Speech Synthesisers with Rapid Adaptation in an End-to-End Framework," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2962–2966.

[28] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World. Twenty-fourth edition*. SIL International, 2021. [Online]. Available: http://www.ethnologue.com.

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[30] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.