Interspeech 2024
1-5 September 2024, Kos, Greece

Meta Learning Text-to-Speech Synthesis in over 7000 Languages

Florian Lux[1], *Sarina Meyer*[1], *Lyonel Behringer*[2], *Frank Zalkow*[2], *Phat Do*[3], *Matt Coler*[3], *Emanuël A. P. Habets*[2], *Ngoc Thang Vu*[1]

[1]University of Stuttgart, Germany
[2]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[3]University of Groningen, The Netherlands

florian.lux@ims.uni-stuttgart.de

Abstract

In this work, we take on the challenging task of building a single text-to-speech synthesis system that is capable of generating speech in over 7000 languages, many of which lack sufficient data for traditional TTS development. By leveraging a novel integration of massively multilingual pretraining and meta learning to approximate language representations, our approach enables zero-shot speech synthesis in languages without any available data. We validate our system's performance through objective measures and human evaluation across a diverse linguistic landscape. By releasing our code and models publicly, we aim to empower communities with limited linguistic resources and foster further innovation in the field of speech technology.

Index Terms: speech synthesis, multilingual, low-resource

1. Introduction

The field of text-to-speech (TTS) synthesis offers a crucial component across a variety of applications and research fields, including accessibility features for the visually impaired, medical applications, language learning tools, language revitalization, voice privacy, literary studies, personal assistants, and entertainment. However, out of the over 7000 languages in the world[1], only a few communities currently have access to a high-quality, controllable TTS system in their native language.

Prior work on massively multilingual TTS (i.e., dealing with hundreds of languages) is sparse. The MMS models [1] cover 1107 languages by combining self-supervised pretraining with supervised finetuning, resulting in one single-speaker monolingual model per language with remarkable quality. Similarly, the authors of the CMU Wilderness dataset [2] train 699 monolingual models in a fully supervised manner. Virtuoso [3] is a 101-language model trained in a semi-supervised manner that does not need paired data, but still requires unpaired adaptation data. Other works on multilingual and low-resource TTS, while operating on a smaller scale, explore transfer learning [4,5], dual transformation [5], meta learning [6,7], or separating the semantic level from the acoustic level [8]. Mismatches in phoneme sets are handled by employing specialized representations as the input, such as bytes [9] or linguistically-motivated features [6, 10].

In this work, we present the first TTS system that can synthesize speech in a total of 7212 languages, covering nearly all spoken languages cataloged in Glottolog [11]. We achieve this by pretraining a TTS model on a massive scale of 462 languages with a total of over 18,000 hours of paired data, which we collected from publicly available sources. The underlying TTS model is designed to be language agnostic except for a language embedding, which is used as a conditioning signal. While collecting this data and training such a model is already challenging, the resulting system still covers less than 6.4 % of all considered languages, illustrated in Figure 1. For the remainder, we leverage the embeddings of supervised languages to approximate those of unseen languages, sharing an otherwise language-agnostic model across all languages. To achieve this, we make use of meta learning under the learn-to-compare framework, similar to Siamese nets [12]. Using these predicted language embeddings during inference, our system can generate speech even for unseen languages.

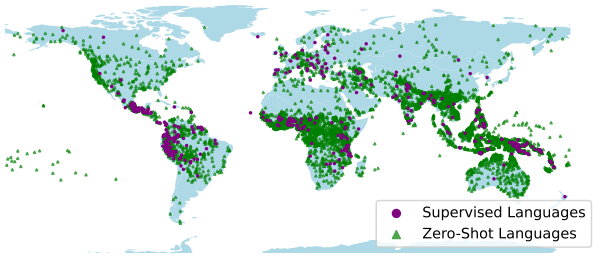

Figure 1: *An overview of the coverage of supervised (462) and zero-shot (6750) languages in our work on the world map.*

Summarizing our contributions, we propose 1) a reproducible data collection that includes paired text and speech data in 462 languages, 2) a pipeline and architecture that allows for scaling TTS to an arbitrary number of languages while being highly controllable, 3) a novel loss function that enforces a semantically meaningful structure in the language embedding space, and 4) a procedure that combines meta learning with zero-shot inference enabling the model to synthesize speech in languages for which no data is available. We evaluate our contributions using objective measures and human evaluation on a set of high-, medium-, and low-resource languages that exhibit a wide range of typological properties. Our code, models, demos, and data are available under an open source license[2].

2. Proposed Methods

2.1. Massively Multilingual Synthesis

2.1.1. Data Acquisition and Cleaning

To start, we collected a large corpus of publicly available datasets with paired text and speech across various languages, containing over 50,000 hours of data spoken by thousands of speakers. Since such excessive amounts of highly diverse data from many different sources require careful cleaning, we can

[1]According to Glottolog: http://glottolog.org

[2]https://github.com/DigitalPhonetics/IMS-Toucan/releases/tag/v3.0

10.21437/Interspeech.2024-1335

Table 1: *Multilingual datasets and the subset of hours we used. * refers to a new dataset generated from eBible and MMS TTS*

Dataset	# Languages	Hours Used
Bible-MMS*	371	1230
Fleurs [13]	90	377
Snow Mountain [14]	15	269
African Voices [15]	11	5
CSS10 [16]	8	105
Multilingual LibriSpeech [17]	8	16,298
Indian TTS [18]	6	4.9
Zambezi Voice [19]	3	0.9
Living Audio Dataset [20]	3	1.6

only use subsets of the full datasets. After a cleaning procedure, which we explain in the following, we end up with around 18,000 hours of data. An overview of the multilingual datasets we used is shown in Table 1, and an overview of the monolingual datasets in Table 2. Most of these datasets are not intended for TTS training and contain noisy recordings, audios with multiple speakers speaking, errors in their labels, and other problems. Therefore, we filtered the audio samples using 1) an open-source speaker diarization system [21] to retain only excerpts that contain a single speaker, 2) reference-free speech quality metrics to filter out the samples with too much noise, and 3) the loss of our aligner and TTS (see Section 2.1.2) to find out which samples may have erroneous labels.

To ensure that the vector space of the language embeddings spans the entirety of possible language embeddings, which we later approximate (see Section 2.2.3), our pretraining set should be as diverse as possible. Hence, we generate speech using the subsets of the eBible dataset [22] that are under free licenses as the text input to the MMS TTS models [1]. By generating 2000 sentences of synthetic speech per language for 371 languages, we substantially increased the linguistic diversity of our data.

2.1.2. Synthesis Pipeline Design and Training

Our pipeline consists of modular, exchangeable blocks. The general approach is architecture agnostic for most components. We based our implementation on the IMS Toucan toolkit [23]. First, we converted input texts to a sequence of phonemes. We used eSpeak NG[3] for all languages it supports. For the remaining languages, we used transphone [24], which is a zero-shot phonemizer that provides phoneme annotations for all languages in Glottolog. We converted the phonemes into articulatory features (i.e., binary encoded configurations of the vocal tract) [6]. These articulatory feature sequences were then converted to a mel-spectrogram by a FastSpeech-2-like system [25] (50M parameters) that uses FastPitch-style conditioning [26] on pitch and energy per phoneme, to allow for fine-grained controllability of the resulting speech. To obtain the durations needed for this, we made use of a small self-contained aligner [27] that was trained with a phoneme recognition objective. To improve details in high frequencies, we used the post-net proposed in PortaSpeech [28] (40M parameters). This entire synthesis model was conditioned on the outputs of a pretrained speaker-embedding network [29] to allow for zero-shot voice selection. As a secondary conditioning signal, we enriched the input of the encoder with a language embedding that was learned jointly from a lookup table [7]. Everything else was built in a language-agnostic fashion, enabling the zero-shot mechanism described in Section 2.2.3. The spectrogram that was predicted by the model was then converted into a waveform and upsampled from 16 kHz to 24 kHz through the use of a HiFi-GAN vocoder [30] with additional upsampling steps (14M parameters). Finally, to mitigate the potential of harmful uses, we applied an audio watermark that is robust against modifications [31].

[3] https://github.com/espeak-ng/espeak-ng

Table 2: *Selection of the most important monolingual datasets and the subset of hours we used.*

Language	Dataset	Hours Used
English	LibriTTS [32]	236
	HiFi-TTS [33]	111
	VCTK [34]	53
French	Blizzard 2023 [35]	29
	SIWIS [36]	11
German	HUI-Audio-Corpus [37]	190
	Thorsten[4]	34
Spanish	Blizzard 2021 [38]	6
Chinese Mandarin	Aishell-3 [39]	63
Vietnamese	VIVOS [40]	15
Javanese	Javanese ASR [41]	59
Persian	ShEMO [42]	3.1
Arabic	ClArTTS [43]	10
Amharic	ALFFA Amharic [44]	2.5
Swahili	ALFFA Swahili [45]	12
Ukrainian	Lada[5]	6

Training the synthesis model on all data at once, however, is not trivial and fails to converge or has problems with information leakage between language and speaker embeddings. This is likely caused by the 371 synthetic datasets derived from MMS all being single-speaker data, resulting in a high correspondence between language and speaker. To remedy this issue, we employed a training curriculum. First, we trained on a subset of data that consists of only multi-speaker datasets for 40,000 steps. Then, we continued training using all data for a further 120,000 steps with balanced amounts of samples per language per batch. Using eight A6000 GPUs, this training took four days to complete with a combined batch size of 152. Further implementation details can be inferred from our open-source code.

2.2. Approximating Synthesis in Unseen Languages

2.2.1. Metrics for Language Similarity

For our zero-shot inference mechanism, we need to measure the phonetic distance between languages. Following [46], [24], and [10], we select three metrics on which we base this distance measure: 1) the distance between nodes over youngest common ancestor, normalized by branch depth in the phylogenetic language tree as not all branches have the same granularity, 2) the distance on the world map, using the ellipsoid distance between language locations according to Glottolog, and 3) the angular similarity of phoneme sets of languages (ASP) based on phonepiece [47]. Their effectiveness is shown in Section 3.1.

2.2.2. Language Embedding Space Structure Loss

We constrain our language embedding space to follow the metrics described in Section 2.2.1 by introducing a Language Embedding Space Structure (LESS) loss function $\mathcal{L}_{\text{LESS}}$ which makes the distance between two language embeddings $e(l_1)$ and $e(l_2)$ similar to the average distance between the languages according to the previously discussed metrics with a normalized

[4] https://doi.org/10.5281/zenodo.5525342
[5] https://doi.org/10.5281/zenodo.7396774

value range. The exact loss function is given by

$$\mathcal{L}_{\text{LESS}} = \Delta\left(\Delta\big(e(l_1), e(l_2)\big), \frac{1}{|M|}\sum_{m\in M} m(l_1, l_2)\right), \quad (1)$$

where Δ denotes the Euclidean distance and M is the set of metrics consisting of the normalized tree distance, the normalized map distance, and the inverse ASP. In preliminary tests, we found that $\mathcal{L}_{\text{LESS}}$ greatly reduces the chances of the synthesis model diverging by simply adding it to the TTS training loss.

2.2.3. Meta Learning Unseen Language Representations

Since our pipeline produces phoneme sequences for any language, to specify the target language, all we need to change within our TTS model is the language embedding. Hence, we can synthesize speech in an unseen language by simply approximating the corresponding language embedding. To achieve this with the limited number of data points we have available (462), we choose to employ a meta-learning technique. Similar to the idea behind Siamese networks [12], we want to cluster the language embeddings in a latent space, to determine which supervised languages a given language is similar to. We train a three-layer perceptron (96 parameters) as a scoring function that we call Meta Learner (ML) to map pairs of languages, defined by their distance metrics from the set M, to approximated language embedding distances. Using the definitions from Section 2.2.2, we achieve this by optimizing ML towards fulfilling

$$\Delta\big(e(l_1), e(l_2)\big) = \text{ML}\big(m(l_1, l_2) \text{ for } m \in M\big). \quad (2)$$

Using ML as a learned distance function between languages, we find the k nearest neighbors from our supervised set for an unseen language and average them to approximate the target embedding. We empirically find the best performance at $5 \leq k \leq 25$. Neighbors beyond the minimum are added if their distance falls below a threshold, which we define to be the median distance of the 25th-nearest neighbors across all languages.

3. Experiments

In our evaluation of the TTS model, we differentiate between high-resource languages (data is abundant), mid-resource languages (some data is available), and low-resource languages (no data is available). We evaluate two languages for each of these categories and aim for a good spread across the world map and language families. For the high-resource set, we choose English (eng) as a Germanic language and French (fra) as a Romance language to ground the performance of our model in these well-explored settings. For the medium-resource set, we choose Welsh (cym) as a Celtic language and Vietnamese (vie) as an Austroasiatic language, which is also a tonal language. For the low-resource language set, we choose Breton (bre), the only Celtic language spoken on the European mainland, and Aymara (aym), an Amerindian isolate language spoken mainly in Peru and Bolivia. We choose to evaluate our approach using real low-resource languages rather than simulating a low-resource scenario by limiting data, despite its availability, believing this offers a more accurate reflection of the system's potential impact. While this choice narrows our evaluation scope, it aligns our work closely with real-world applications and challenges.

3.1. Language Embedding Approximation

To evaluate different techniques for approximating language embeddings, we calculated the mean squared error (MSE) between the actual language embeddings from our supervised set and their approximations, obtained by averaging their nearest embeddings selected according to different metrics (see Section 2.2.3). Figure 2 demonstrates that our learned metric outperforms any of the individual metrics, as well as their average distance. Furthermore, all metrics perform better than using randomly chosen languages as the nearest neighbors, indicating that they are all effective measures of phonetic language similarity to some extent.

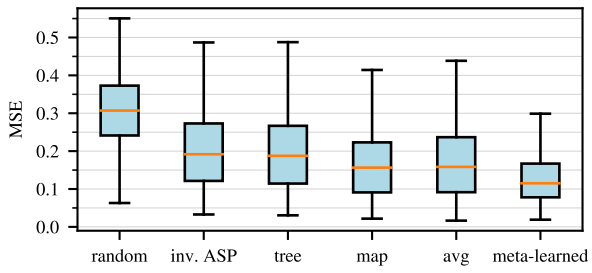

Figure 2: *Reconstruction error for approximating the 462 language embeddings from our supervised set using their k nearest neighbors, which are determined either at random, via distance metrics (inverse ASP, tree distance, map distance), their average (avg), or our meta-learned distance function.*

We further conducted a small internal pilot study to determine when the reconstruction error reaches an acceptable level perceptually. We asked participants to listen to simulated zero-shot languages (i.e., using a generated embedding instead of the ground-truth one, despite it being available) in their native language and rate whether the resulting speech sounds natural to them. From this, we find that our proposed metric is the only one which is perceived as sufficiently natural, demonstrating the utility of meta learning for estimating language embeddings.

Analyzing the selection of nearest neighbors qualitatively shows that the learned metric mostly behaves similar to the map distance, however it pivots to following the tree distance or the ASP if either of them is close to zero or close to one. E.g., Breton is approximated using just five languages: French, Dutch, Hungarian, English and Latin. Notably, Welsh is not used despite being the closest in terms of both map and tree distance, likely due to its higher inverse ASP. Hence the metric seems to be able to generalize to this multi-step policy.

3.2. Objective Evaluation of the Synthesis

We computed objective measures for each selected language, with the exception of Aymara, since we lack sufficient quantities of reference speech recordings and appropriate models to measure performance. We based our objective measures on 1000 test samples per language. To evaluate speech intelligibility, we computed the word error rate (WER) between ground truth and automatic transcriptions, obtained by the state-of-the-art automatic speech recognition system Whisper [48] (version "large-v3"). We additionally computed the phoneme error rate (PER) for the phoneme transcripts, which we obtained by applying the phonemizers described in Section 2.1.2 to the ground truth and automatic transcriptions. The PER is less affected by phonetically similar sounding mistakes than the WER, serving as a secondary indicator for intelligibility. To estimate speech quality, we used WV-MOS, which is a fine-tuned wav2vec2.0 model that aims to predict mean opinion score (MOS) ratings [49].

Table 3: *Objective evaluation measures: Word error rate (WER, ↓), phoneme error rate (PER, ↓), and WV-MOS scores (↑). Aymara is excluded for a lack of evaluation resources.*

		HighRes		MidRes		LowRes
		eng	fra	vie	cym	bre
Ours	WER	0.1 ± 0.1	0.2 ± 0.2	0.3 ± 0.5	0.7 ± 0.2	1.0 ± 0.3
	PER	0.0 ± 0.0	0.0 ± 0.1	0.1 ± 0.2	0.2 ± 0.1	0.7 ± 0.4
	WV-MOS	4.4 ± 0.2	3.9 ± 0.3	4.0 ± 0.3	3.6 ± 0.2	4.0 ± 0.3
MMS	WER	0.2 ± 0.2	0.2 ± 0.2	0.3 ± 0.2	0.4 ± 0.2	N/A
	PER	0.0 ± 0.1	0.0 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	N/A
	WV-MOS	3.9 ± 0.3	4.0 ± 0.3	3.5 ± 0.4	3.6 ± 0.3	N/A
Ref	WER	0.1 ± 0.1	0.2 ± 0.2	0.1 ± 0.1	0.5 ± 0.3	1.0 ± 0.4
	PER	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.1	0.1 ± 0.1	0.5 ± 0.3
	WV-MOS	4.1 ± 0.5	3.9 ± 0.3	3.2 ± 0.4	1.9 ± 1.4	2.9 ± 0.9

Table 3 shows the objective performance for our system, the MMS system, and a reference obtained by vocoding real-world recordings. Note that the MMS system does not support Breton. The WER and PER scores show that our and the MMS system are highly intelligible for English, French, and Vietnamese. The high error rates for Welsh and especially Breton (also for the reference) might point towards the low performance of Whisper for these under-resourced languages. Further investigations of the aspect of intelligibility for under-resourced languages remain for future work. The WV-MOS scores indicate that our system's synthesis quality is on par with or better than MMS. Note that the low scores for the references of some languages are due to the low quality of the reference recordings available.

3.3. Subjective Evaluation of the Synthesis

We additionally conducted subjective listening tests, engaging native speakers of the mid- and low-resource languages. These tests were facilitated through an online study using the web-MUSHRA framework [50], reaching out to native speakers via research networks and community contacts, ensuring respectful and meaningful engagement with each language community. In this study, we asked the listeners to rate how similar a presented audio sample sounds to someone speaking the respective language as a native speaker on a scale from 1 ("foreign speaker imitating the language without any training") to 5 ("native speaker of the language").

We received 450 ratings from 15 raters for Vietnamese, 390 ratings from 13 raters for Welsh, 200 ratings from 10 raters for Breton, and 180 ratings from 9 raters for Aymara, each evenly spread across all systems. Note that all participants self-identified as being native speakers of the respective language. Figure 3 shows boxplots for the obtained listening test scores. The median score of our system is 4 for all four languages. The MMS system was rated with a median score of 4 as well in the two languages it supports. We conducted a Mann-Whitney U test [51] for both Vietnamese and Welsh and found no significant difference between the ratings for our system and MMS. This indicates that our system performs on par while supporting nearly seven times as many languages and offering various controllability options. Some ratings for the references are lower than expected due to the limited quality of those recordings and differences between language varieties.

4. Conclusion

We presented a TTS system that is scalable to an arbitrary number of languages, achieving zero-shot inference on unseen languages through massively multilingual pretraining and a meta-learning approach to approximate language conditioning signals. In this way, we created the first TTS system, which can be used for languages where absolutely no data is available, not even for semi-supervised or transfer learning. The system proves effective across varying resource levels in both objective and subjective evaluation. While the amount of languages covered by the evaluation is a limitation, the linguistic diversity in their selection, as well as the high quality of the ratings by the native speakers, helps ensure the reliability of the results. In the future, it would be interesting to explore if fine-tuning our universal model to language-specific expert models, like MMS, could lead to improvements in those languages.

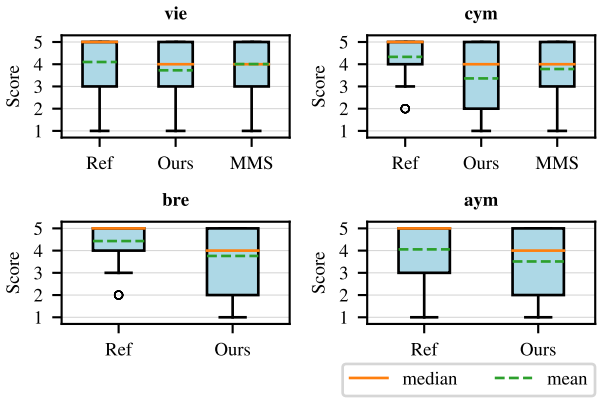

Figure 3: *Boxplots for the listening test results.*

5. Ethical Considerations

Given the risk of misuse of synthetic voice generation, we emphasize that our system is designed for positive applications such as education, accessibility, and cultural preservation. The synthesis is distinguishable from human speech, especially through the use of audio watermarking, as described in Section 2.1.2. We further acknowledge the sensitivities associated with using indigenous languages, especially those that communities wish to keep un-documented or limited to specific uses. Our approach involves engaging with community representatives to seek guidance and, where applicable, consent before integrating any language into our system. We are committed to excluding any language from our system upon request from its community, reinforcing our commitment to technology that serves rather than exploits.

6. Acknowledgments

We thank Edwin Banegas-Flores, the voice in the Aymara recordings, and Ruben Hilari-Jilalu, who helped disseminate the Aymara listening test. We thank Anaïs Scornet (*Mignoned ar brezhoneg*) for helping with the Breton test. Lastly, we thank the *Canolfan Bedwyr* center (Bangor University) and the *Mentrau Iaith Cymru* association for their help with the Welsh test.

7. References

[1] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, 2024.

[2] A. W. Black, "CMU Wilderness Multilingual Speech Dataset," in *ICASSP*. IEEE, 2019.

[3] T. Saeki, H. Zen, Z. Chen, N. Morioka, G. Wang *et al.*, "Virtuoso: Massive Multilingual Speech-Text Joint Semi-Supervised Learning for Text-to-Speech," in *ICASSP*. IEEE, 2023.

[4] 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," *Interspeech*, 2019.

[5] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li *et al.*, *LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition*. ACM, 2020.

[6] F. Lux and N. T. Vu, "Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features," in *ACL*, 2022.

[7] F. Lux, J. Koch, and N. T. Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS," in *AACL*, 2022.

[8] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin *et al.*, "Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision," *ACL*, 2023.

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

[10] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "Strategies in Transfer Learning for Low-Resource Speech Synthesis: Phone Mapping, Features Input, and Source Language Selection," in *ISCA Speech Synthesis Workshop*, 2023.

[11] H. Hammarström, "Glottolog: a free, online, comprehensive bibliography of the world's languages," in *Linguistic and Cultural Diversity in Cyberspace*. UNESCO, 2015.

[12] J. Bromley, I. Guyon, Y. LeCun *et al.*, "Signature verification using a "Siamese" time delay neural network," *NeurIPS*, 1993.

[13] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod *et al.*, "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech," in *SLT*. IEEE, 2023.

[14] K. Raju, V. Anjaly, R. A. Lish, and J. Mathew, "Snow Mountain: Dataset of Audio Recordings of The Bible in Low Resource Languages," *arXiv:2206.01205*, 2022.

[15] P. Ogayo, G. Neubig, and A. W Black, "Building African Voices," in *Interspeech*, 2022.

[16] K. Park and T. Mulc, "CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages," *Interspeech*, 2019.

[17] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020.

[18] F. He, S.-H. C. Chu, O. Kjartansson, C. Rivera, A. Katanova *et al.*, "Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems," in *LREC*, 2020.

[19] C. Sikasote, K. Siaminwe, S. Mwape, B. Zulu, M. Phiri *et al.*, "Zambezi Voice: A Multilingual Speech Corpus for Zambian Languages," in *Interspeech*, 2023.

[20] D. A. Braude, M. P. Aylett, C. Laoide-Kemp, S. Ashby, K. M. Scott *et al.*, "All Together Now: The Living Audio Dataset," in *Interspeech*, 2019.

[21] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Interspeech*, 2023.

[22] V. Akerman, D. Baines, D. Daspit, U. Hermjakob *et al.*, "The eBible Corpus: Data and Model Benchmarks for Bible Translation for Low-Resource Languages," *arXiv:2304.09919*, 2023.

[23] F. Lux, J. Koch, S. Meyer, T. Bott, N. Schauffler *et al.*, "The IMS Toucan system for the Blizzard Challenge 2023," in *Proc. Blizzard Challenge Workshop*. Speech Synthesis SIG, 2023.

[24] X. Li, F. Metze, D. Mortensen, S. Watanabe, and A. Black, "Zero-shot Learning for Grapheme to Phoneme Conversion with Language Ensemble," in *ACL*, 2022.

[25] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *ICLR*, 2020.

[26] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP*. IEEE, 2021.

[27] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *SLT*. IEEE, 2023.

[28] Y. Ren, J. Liu, and Z. Zhao, "Portaspeech: Portable and high-quality generative text-to-speech," *NeurIPS*, 2021.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell *et al.*, "SpeechBrain: A General-Purpose Speech Toolkit," 2021.

[30] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*, 2020.

[31] R. San Roman, P. Fernandez, H. Elsahar, A. Défossez, T. Furon *et al.*, "Proactive Detection of Voice Cloning with Localized Watermarking," *arXiv:2401.17264*, 2024.

[32] H. Zen, V. Dang, R. Clark *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.

[33] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi Multi-Speaker English TTS Dataset," in *Interspeech*, 2021.

[34] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[35] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *Blizzard Challenge Workshop*, 2023.

[36] J. Yamagishi, P.-E. Honnet, P. Garner, and A. Lazaridis, "The SIWIS French Speech Synthesis Database," 2016.

[37] P. Puchtler, J. Wirth, and R. Peinl, "HUI-audio-corpus-German: A high quality TTS dataset," in *German Conference on Artificial Intelligence*, 2021.

[38] Z.-H. Ling, X. Zhou, and S. King, "The Blizzard Challenge 2021," in *Blizzard Challenge Workshop*, 2021.

[39] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker Mandarin TTS corpus and the baselines," *arXiv:2010.11567*, 2020.

[40] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese Speech Recognition System," in *WLSI/OIAF4HLT*, Y. Murakami, D. Lin, N. Ide, and J. Pustejovsky, Eds., 2016.

[41] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali," in *SLTU*, 2018.

[42] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," *LREC*, 2019.

[43] A. Kulkarni, A. Kulkarni, S. A. M. Shatnawi, and H. Aldarmaki, "ClArTTS: An Open-Source Classical Arabic Text-to-Speech Corpus," in *Interspeech*, 2023.

[44] S. T. Abate, W. Menzel, and B. Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *Interspeech*, 2005.

[45] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of Swahili resources for an automatic speech recognition system," in *Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2012.

[46] P. Wu, J. Shi, Y. Zhong, S. Watanabe, and A. W. Black, "Cross-Lingual Transfer for Speech Processing Using Acoustic Language Similarity," in *ASRU*. IEEE, 2021.

[47] X. Li, F. Metze, D. R. Mortensen, A. W. Black, and S. Watanabe, "Phone Inventories and Recognition for Every Language," in *LREC*, 2022.

[48] A. Radford, J. W. Kim, T. Xu, G. Brockman *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," *ICML*, 2023.

[49] P. Andreev, A. Alanov, O. Ivanov, and D. P. Vetrov, "HiFi++: A Unified Framework for Bandwidth Extension and Speech Enhancement," in *ICASSP*. IEEE, 2023.

[50] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal *et al.*, "webMUSHRA – A Comprehensive Framework for Web-based Listening Tests," *Journal of Open Research Software*, 2018.

[51] A. Rosenberg and B. Ramabhadran, "Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores," in *Interspeech*, 2017.