Multilingual Speech Editing

Antonis Asonitis [1] **Luca A. Lanzendörfer** [1] **Frédéric Berdoz** [1] **Roger Wattenhofer** [1]

Abstract

We present MuSE, an autoregressive model for multilingual speech editing at both phoneme and word-level granularity. We train MuSE using a three-stage curriculum learning approach, progressing from text-phoneme grounding to full bidirectional audio infilling. To support training at this scale, we release *multilingual-audio-alignments*, the largest publicly available word-and-phoneme-level aligned speech corpus to date, spanning over 32k hours across 13 languages. We demonstrate state-of-the-art performance on the established RealEdit benchmark and introduce *MultiLingualEdit*, a multilingual extension with handcrafted editing examples in 13 languages. The codebase and datasets are made publicly available.

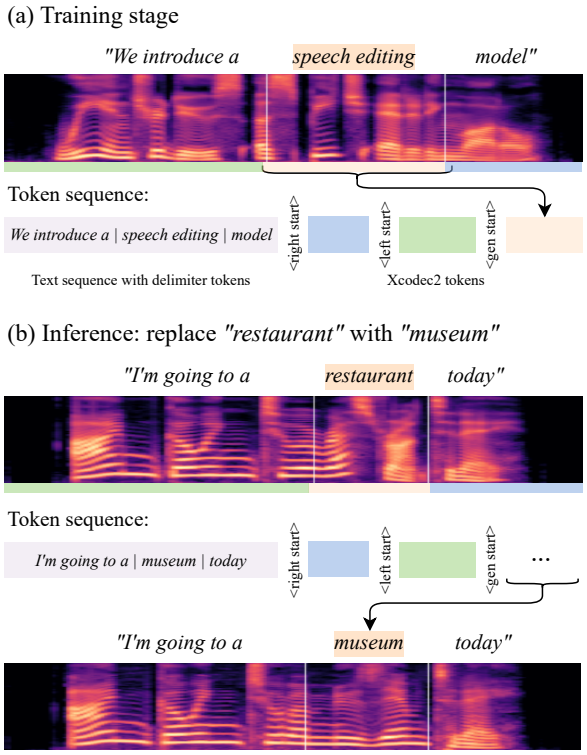

Figure 1. Overview of MuSE. **(a)** Training: MFA word boundaries define a span to mask. MuSE learns to reconstruct the masked audio given the target words and surrounding audio context. The cross-entropy loss is computed only over the predicted audio tokens. **(b)** Inference: the user edits the transcript and MuSE generates audio tokens for the new text, which is then stitched into the original recording.

1. Introduction

Speech editing modifies recorded audio by altering its text transcript, such as correcting a mispronunciation or replacing a word, without re-recording. Unlike text-to-speech (TTS), where the model generates an entire utterance from scratch, editing requires the synthesized segment to match the surrounding unedited audio in speaker identity, prosody, and acoustics.

Early methods achieved text-guided speech insertion and substitution by combining a single-speaker TTS model with a voice conversion model, concatenating the generated segment with the unedited audio (Ren et al., 2026). However, because generation was not conditioned on the surrounding speech context, these approaches suffered from unnatural prosody mismatches and boundary artifacts. To address this, subsequent research framed speech editing as a contextual audio infilling problem. Models evolved from unidirectional LSTMs with bidirectional fusion (Tan et al., 2021) to masked reconstruction objectives with convolutional or Transformer architectures (Wang et al., 2022; Bai et al., 2022). More recently, diffusion and flow-matching models such as FluentSpeech (Jiang et al., 2023) and Voicebox (Le et al., 2023) have achieved stronger local acoustic consistency by iteratively refining the mel-spectrogram or latent representations of the edited region.

Parallel to these efforts, a new paradigm has emerged for speech generation, moving from predicting continuous mel-spectrograms to autoregressively modeling discrete audio tokens. Models such as VALL-E (Chen et al., 2025a), CosyVoice (Du et al., 2025), and more recently Qwen3-

[1]ETH Zurich, Switzerland. Correspondence to: Antonis Asonitis <aasonitis@ethz.ch>.

Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

TTS (Hu et al., 2026) leverage autoregressive decoding methods to achieve state-of-the-art results in text-to-speech generation. Similarly, in speech editing, models such as VoiceCraft (Peng et al., 2024) use Transformers to predict discrete speech tokens given bidirectional context and a target transcript. Framing speech synthesis as a next-token prediction task allows these approaches to leverage the Transformer's inherent strength at capturing long-range dependencies. Evaluation of these systems has largely centered on the RealEdit benchmark (Peng et al., 2024), which evaluates word-level insertions, substitutions, and deletions.

Two key gaps have limited progress on speech editing. First, phoneme-level editing requires phoneme-level alignments, yet most public speech corpora are either monolingual, too small, or lack the phoneme annotations needed for training. Second, the standard benchmark RealEdit only covers English, and no multilingual speech editing benchmark has been previously proposed. Moreover, existing speech editing models operate at a single granularity, either words or G2P-derived phonemes, reducing control when different granularities are desired.

To this end, we propose MuSE (**Mu**ltilingual **S**peech **E**dit), a model built on Llasa (Ye et al., 2025), as illustrated in Figure 1. MuSE introduces bidirectional audio context, expands its vocabulary with IPA phoneme tokens, and uses dedicated delimiter tokens to separate the left context, right context, and the edited region. After aligning a large-scale speech corpus at both the word and phoneme levels, we train MuSE with a three-stage curriculum that progresses from phoneme-text grounding through phoneme-conditioned speech generation to bidirectional speech editing using words, phonemes, or a mix of both.

Our contributions can be summarized as follows:

- We propose **MuSE**,[1] the first multilingual speech editing model that supports both word-level and phoneme-level editing. Unlike prior work, MuSE accepts raw text, IPA phonemes, or a mix of both as conditioning input, trained via a three-stage curriculum.
- To train MuSE, we release **multilingual-audio-alignments**, a 32.5k hour corpus spanning 13 languages, force-aligned at the phoneme level with the Montreal Forced Aligner, resulting in the largest such public resource.
- We evaluate MuSE against previous work using **MultiLingualEdit**, a multilingual extension of RealEdit with 20 handcrafted editing examples per language, including insertions, substitutions, and deletions, across 13 languages.

[1]Samples available at https://lucala.github.io/muse_samples/

2. Methodology

2.1. Data and Alignments

Phoneme-conditioned editing requires timestamp alignments between phonemes and their acoustic frames, however, such data is scarce. We therefore curate the multilingual-audio-alignments dataset, a corpus of 32.5k hours across 13 languages, aggregated from 42 public dataset sources. These sources span a range of speaking styles, including read speech from audiobooks, crowd-sourced and synthetic read sentences, parliamentary proceedings, podcasts, spontaneous conversation, broadcast media, and entertainment dialogue.

All transcripts were first lowercased and stripped of punctuation. Where available, we applied the NVIDIA NeMo text normalizer (Zhang et al., 2021) to handle numbers, abbreviations, and other non-standard tokens. We then ran the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) using the pretrained per-language acoustic model and IPA phoneme dictionary for each of the 13 languages. Because each language has its own phoneme inventory, we rely on MFA's per-language pretrained models, which map to the corresponding IPA symbol set. The aligned phonemes therefore correspond to the canonical dictionary pronunciation of each word in the given language, not to the speaker's specific acoustic realisation. Utterances for which MFA failed to produce a valid alignment or generated unknown phoneme tokens were discarded. The average acceptance rate across all languages was approximately 85%: high-resource languages such as English, French, and German retain close to 100% of utterances, while lower-resource languages such as Portuguese and Polish see acceptance rates of only 30-40%, primarily due to noisier transcripts and less robust acoustic models. The processing of this data required a combined 75k CPU hours.

2.2. Model Architecture

MuSE extends Llasa (Ye et al., 2025), a recent text-to-speech framework that builds on a pre-trained Llama 3.2 (Grattafiori et al., 2024) 1B language model to jointly process text and discrete audio tokens. Llasa represents speech with XCodec2 (Ye et al., 2025), a single-codebook quantizer that encodes waveforms at a rate of 50 tokens per second. By training on approximately 250k hours of speech data, Llasa achieves high-fidelity synthesis while inheriting the long-range dependency modeling capabilities of the underlying Transformer. We use the multilingual Llasa 1B checkpoint as our starting point.

To adapt Llasa for speech editing, we introduce three modifications. First, we extend the vocabulary with IPA phoneme tokens covering all thirteen target languages, derived from the per-language MFA (McAuliffe et al., 2017) IPA dictio-

naries used during forced alignment (Section 2.1), together with a set of delimiter tokens that demarcate the left audio context, right audio context, target transcript span, and generated output region. Second, whereas existing editors such as VoiceCraft condition exclusively on phoneme sequences derived via grapheme-to-phoneme conversion, MuSE accepts raw text, IPA phonemes, or an arbitrary mixture of the two within a single edit, providing fine-grained control over the generation granularity. Third, we restructure the input sequence to provide bidirectional audio context. At inference time the model receives the right and left audio segments as prefixes and autoregressively predicts the tokens corresponding to the edited region.

2.3. Training

We found that fine-tuning Llasa for speech editing fails to converge efficiently when using randomly initialized phoneme embeddings. To address this, we propose three architectural adaptations that result in significant training stability and accelerated convergence. A comparative analysis of our proposed strategy against the baseline approach is detailed in Figure 2.

Informed Embedding Initialization. Extending a pre-trained vocabulary with new tokens typically requires random initialization, which can destabilize early training. To mitigate this, we initialize each IPA phoneme embedding from the embedding of acoustically related tokens already present in the large language model. Specifically, we use Gemini 3 Flash (Anil et al., 2023) to map each IPA symbol to the closest English syllable or word (e.g., tʃ → *ch*, aɪ → *eye*). A small amount of Gaussian noise ($\sigma = 0.001$) is added to break symmetry.

Instruction-Based Training Format. Llasa is pre-trained with natural-language instruction prefixes (e.g. *"Convert the text to speech: ..."*). To maximally leverage this instruction-following capability, every training sample in MuSE has a natural-language instruction prepended to the token sequence.

Three-Stage Curriculum Training. We decompose training into three stages of increasing complexity, with the aim to progressively adapt the model from text-to-speech to speech editing capabilities. In *Stage 1* the model is trained on bidirectional conversion between orthographic text and IPA phoneme sequences (grapheme-to-phoneme and phoneme-to-grapheme), grounding the newly added phoneme embeddings in the large language model's existing text representations. In *Stage 2* the model is trained on autoregressive speech generation conditioned on phonemes or mixtures of phonemes and words, essentially extending Llasa to perform phoneme-to-speech. A curriculum mixing schedule gradually shifts the conditioning from text to phonemes, with the phoneme ratio ramping from 0.1 to 1.0 over the first half of the training set. In *Stage 3* the model is trained on bidirectional speech editing: given left and right audio context together with a target phoneme or text sequence, it generates the missing middle segment. Transition margins of 0.1s on each side of the edit boundary ensure smooth acoustic continuity, particularly helpful for deletion edits where the model regenerates the boundary segments of the newly concatenated audio. Cross-entropy loss is computed only over text and phoneme outputs in Stage 1 and over generated XCodec2 audio tokens in Stages 2 and 3.

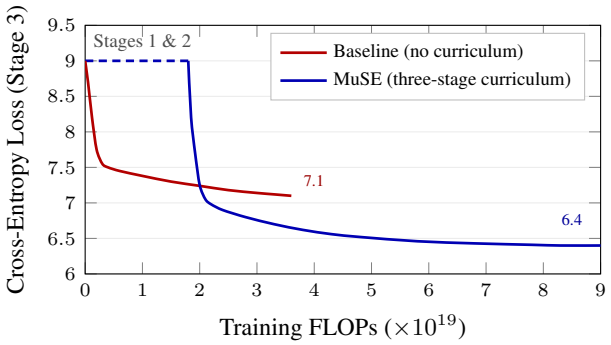

Figure 2. Cross-entropy loss on the Stage 3 editing objective versus total training FLOPs (1B-parameter model). The dashed line indicates the compute budget allocated to Stages 1 and 2 ($\approx 1.8 \times 10^{19}$ FLOPs). Directly fine-tuning on Stage 3 without curriculum pre-training or informed embedding initialization plateaus near 7.1, whereas the full three-stage curriculum converges to 6.4. We note that the baseline produced unintelligible speech, mostly silence.

The peak learning rate is 1×10^{-4} for Stage 1 and 5×10^{-5} for Stages 2 and 3. We use a per-device batch size of 4 with gradient accumulation over 5 steps and clip gradients at a max norm of 1.0. All training is conducted in BF16 mixed precision on 4×H200 GPUs.

2.4. Evaluation Benchmark

To our knowledge, no multilingual speech editing benchmark exists. We therefore construct **MultiLingualEdit**, comprising 20 handcrafted editing examples per language across all 13 supported languages (260 examples total). Each language contains the same balanced split of 10 substitutions, 5 insertions, and 5 deletions. Source utterances were drawn from established public datasets including HiFi-TTS (Bakhturina et al., 2021b), MSP-Podcast (Busso et al., 2025), Common Voice (Ardila et al., 2020), Expresso (Nguyen et al., 2023), Emilia (He et al., 2024), MLS (Pratap et al., 2020), Multilingual TEDx (Salesky et al., 2021), JSUT (Sonobe et al., 2017), ReazonSpeech (Yin et al., 2023), Zeroth Korean (Jo et al., 2018), Russian LibriSpeech (Bakhturina et al., 2021a), and CML-TTS (Oliveira et al., 2023), with community datasets on Hugging Face filling in Mandarin, Thai, Turkish, and Portuguese. For each

Table 1. Word-level editing results on the RealEdit benchmark (LibriSpeech clean test, 25 samples).

Model	WER ↓	CER ↓	UTMOS ↑
VoiceCraft	0.062	0.033	3.795
VoiceCraft-X	0.078	0.041	3.713
F5-TTS	0.213	0.153	3.604
MuSE	**0.037**	**0.023**	**3.823**

Table 2. Multilingual editing results on MultiLingualEdit. W.R. indicates head-to-head win rate during the human evaluation. (*) WER is omitted due to unsegmented scripts.

Language	Model	WER ↓	CER ↓	UTMOS ↑	W.R. ↑
English	VoiceCraft	0.044	0.015	3.783	40.2%
	VoiceCraft-X	0.070	0.040	3.688	**57.9%**
	MuSE	**0.039**	**0.014**	**3.789**	52.5%
French	VoiceCraft-X	0.271	0.266	2.526	18.8%
	MuSE	**0.208**	**0.148**	**2.685**	**81.2%**
German	VoiceCraft-X	**0.126**	0.070	2.853	33.1%
	MuSE	0.148	**0.061**	**3.028**	**66.9%**
Spanish	VoiceCraft-X	0.114	**0.045**	2.605	45.8%
	MuSE	**0.079**	0.049	**2.814**	**54.2%**
Polish	VoiceCraft-X	**0.175**	0.047	2.586	25.0%
	MuSE	0.205	**0.038**	**3.060**	**75.0%**
Portuguese	VoiceCraft-X	0.104	0.061	2.176	31.25%
	MuSE	**0.082**	**0.047**	**2.453**	**68.75%**
Italian	VoiceCraft-X	0.258	0.106	2.652	45.6%
	MuSE	**0.252**	**0.079**	**2.908**	**54.4%**
Japanese*	VoiceCraft-X	–	0.260	2.774	**61.9%**
	MuSE	–	**0.241**	**2.995**	38.1%
Mandarin*	VoiceCraft-X	–	0.147	2.629	31.9%
	MuSE	–	**0.138**	**2.783**	**68.1%**
Korean*	VoiceCraft-X	–	0.406	1.961	45.0%
	MuSE	–	**0.306**	**1.987**	**55.0%**
Russian	**MuSE**	0.233	0.094	2.886	–
Turkish	**MuSE**	0.382	0.257	2.160	–
Thai	**MuSE**	0.577	0.471	2.162	–

utterance, we used Gemini 3 Flash (Anil et al., 2023) to propose plausible edits (changing a word or short phrase) which were then reviewed and corrected by proficient language speakers to ensure grammatical naturalness. MultiLingualEdit is constructed as a disjoint set from the multilingual-audio-alignments training data, ensuring no sample overlap.

3. Experiments and Results

We evaluate MuSE on substitution, insertion, and deletion edits across two benchmarks: English RealEdit following the VoiceCraft evaluation protocol, and the new MultiLingualEdit benchmark. We compare against VoiceCraft (Peng et al., 2024), VoiceCraft-X (Zheng et al., 2025), and F5-TTS (Chen et al., 2025b) using their official public checkpoints at 16 kHz with default inference settings. No public speech-editor competitor supports Russian, Turkish, or Thai, so we report MuSE numbers in those languages without baselines, as the first such reference points. We report word error rate (WER), character error rate (CER), and UTMOS (Saeki et al., 2022) for speech quality, alongside subjective win rates (W.R.) from head-to-head listening tests with 20 native speakers per language, each rating 8 randomly sampled head-to-head trials that include the original ground-truth utterance (180 listeners total).

A property of MuSE is that it can be conditioned at either phoneme or word granularity. To make each comparison head-to-head and fair, we evaluate MuSE using *each baseline's own native input granularity*: phonemes on RealEdit to match VoiceCraft, and words on MultiLingualEdit to match VoiceCraft-X. Reported gains therefore reflect modeling improvements rather than a difference in input modality.

3.1. Results

With MuSE conditioned on phonemes to match VoiceCraft's native input, we measure WER and CER using the original transcripts and transcripts obtained from Whisper-Large-v3 (Radford et al., 2023) on the edited audio. MuSE outperforms existing models across all metrics on RealEdit (Table 1), including speech naturalness (UTMOS). Evaluating multilingual speech editing on MultiLingualEdit, with MuSE conditioned on words to match VoiceCraft-X's native input, MuSE achieves lower CER and higher UTMOS than VoiceCraft-X across most languages, with particularly strong gains on French, Spanish, Mandarin, and Korean (Table 2). Subjective win rates confirm these trends, although preferences vary on English and Japanese, where VoiceCraft-X remains highly competitive. Taken together with the RealEdit results, MuSE wins at each baseline's own native granularity. We additionally report MuSE's performance on Turkish, Russian, and Thai. Since no other speech editing model supports these languages, we cannot conduct a direct and fair comparison. However, these results establish a necessary baseline for future multilingual research.

4. Conclusion

In this work, we presented MuSE, a multilingual speech editing model that operates at both word and phoneme granularity. Built on Llasa and trained via a three-stage curriculum, progressing from text-phoneme grounding through phoneme-to-speech synthesis to bidirectional context audio editing, MuSE achieves state-of-the-art results on RealEdit and on the new MultiLingualEdit benchmark. We additionally release *multilingual-audio-alignments*, a 32.5k-hour phoneme-aligned corpus spanning 13 languages and 42 sources, the largest such public dataset, as well as *MultiLingualEdit*, the first multilingual speech editing benchmark. These contributions lower the barrier to multilingual speech editing research and encourage future work on fine-grained prosodic control, speaker adaptation, and extension to additional languages.

Impact Statement

This work aims to advance multilingual speech editing by enabling word- and phoneme-level corrections across languages, with potential benefits for accessibility, localization, education, and audio post-production. Speech editing can be misused to alter recordings, impersonate speakers, or create misleading audio. Because edited speech may preserve speaker identity and surrounding acoustic context, deployments should include safeguards against deceptive or malicious use.

References

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., et al. Gemini: A Family of Highly Capable Multimodal Models, 2023. arXiv:2312.11805.

Ardila, R., Branson, M., Davis, K., Kohler, M., Henretty, M., Meyer, J., et al. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.

Bai, H., Zheng, R., Chen, J., Ma, M., Li, X., and Huang, L. A3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing. In *International Conference on Machine Learning (ICML)*, 2022.

Bakhturina, E., Lavrukhin, V., and Ginsburg, B. A Toolbox for Construction and Analysis of Speech Datasets. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks (NeurIPS)*, 2021a.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., and Zhang, Y. Hi-Fi Multi-Speaker English TTS Dataset. In *Proceedings of Interspeech*, 2021b.

Busso, C., Lotfian, R., Sridhar, K., Salman, A. N., Lin, W.-C., Goncalves, L., et al. The MSP-Podcast Corpus, 2025. arXiv:2509.09791.

Chen, S., Wang, C., Wu, Y., Zhang, Z., Zhou, L., Liu, S., et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:705–718, 2025a.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., et al. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025b.

Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., et al. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. In *International Conference on Learning Representations (ICLR)*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. The Llama 3 Herd of Models, 2024. arXiv:2407.21783.

He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., et al. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., et al. Qwen3-TTS Technical Report, 2026. arXiv:2601.15621.

Jiang, Z., Yang, Q., Zuo, J., Ye, Z., Huang, R., Ren, Y., and Zhao, Z. FluentSpeech: Stutter-Oriented Automatic Speech Editing with Context-Aware Diffusion Models. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

Jo, J. et al. Zeroth-Korean: An Open-Source Korean Speech Corpus, 2018. URL https://www.openslr.org/40/. Accessed February 2026.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., et al. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech*, 2017.

Nguyen, T. A., Hsu, W.-N., D'Avirro, A., Shi, B., Gat, I., Fazel-Zarani, M., et al. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Proceedings of Interspeech*, 2023.

Oliveira, F. S., Casanova, E., Cândido Júnior, A., Soares, A. S., and Galvão Filho, A. R. CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages. In *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*, 2023.

Peng, P., Huang, P.-Y., Li, S.-W., Mohamed, A., and Harwath, D. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proceedings of Interspeech*, 2020.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, 2023.

Ren, Y., Yi, J., Tao, J., Wen, Z., and Wang, T. Edit Content, Preserve Acoustics: Imperceptible Text-Based Speech Editing via Self-Consistency Rewards, 2026. arXiv:2602.00560.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Proceedings of Interspeech*, 2022.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., et al. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proceedings of Interspeech*, 2021.

Sonobe, R., Takamichi, S., and Saruwatari, H. JSUT Corpus: Free Large-Scale Japanese Speech Corpus for End-to-End Speech Synthesis, 2017. arXiv:1711.00354.

Tan, D., Deng, L., Yeung, Y. T., Jiang, X., Chen, X., and Lee, T. EditSpeech: A Text Based Speech Editing System Using Partial Inference and Bidirectional Fusion. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Wang, T., Yi, J., Fu, R., Tao, J., and Wen, Z. CampNet: Context-Aware Mask Prediction for End-to-End Text-Based Speech Editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2241–2254, 2022.

Ye, Z., Zhu, X., Chan, C.-M., Wang, X., Tan, X., Lei, J., et al. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis, 2025. arXiv:2502.04128.

Yin, Y., Mori, D., and Fujimoto, S. ReazonSpeech: A Free and Massive Corpus for Japanese ASR. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing (ANLP)*, 2023.

Zhang, Y., Bakhturina, E., Gorman, K., and Ginsburg, B. NeMo (Inverse) Text Normalization: From Development to Production. In *Proceedings of Interspeech*, 2021.

Zheng, Z., Peng, P., Diwan, A., Huynh, C. P., Sun, X., Liu, Z., et al. VoiceCraft-X: Unifying Multilingual, Voice-Cloning Speech Synthesis and Speech Editing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.