CROSS-LINGUAL F5-TTS: TOWARDS LANGUAGE-AGNOSTIC VOICE CLONING AND SPEECH SYNTHESIS

Qingyu Liu[1,3], Yushen Chen[1,2], Zhikang Niu[1,2], Chunhui Wang[4], Yunting Yang[4],
Bowen Zhang[4], Jian Zhao[4], Pengcheng Zhu[4], Kai Yu[1], Xie Chen[1,2†]

[1]MoE Key Lab of Artificial Intelligence, X-LANCE Lab, School of Computer Science,
Shanghai Jiao Tong University, China [2]Shanghai Innovation Institute, China,
[3]Johns Hopkins University, USA, [4]Geely, China

ABSTRACT

Flow-matching-based text-to-speech (TTS) models have shown high-quality speech synthesis. However, most current flow-matching-based TTS models still rely on reference transcripts corresponding to the audio prompt for synthesis. This dependency prevents cross-lingual voice cloning when audio prompt transcripts are unavailable, particularly for unseen languages. The key challenges for flow-matching-based TTS models to remove audio prompt transcripts are identifying word boundaries during training and determining appropriate duration during inference. In this paper, we introduce Cross-Lingual F5-TTS, a framework that enables cross-lingual voice cloning without audio prompt transcripts. Our method preprocesses audio prompts by forced alignment to obtain word boundaries, enabling direct synthesis from audio prompts while excluding transcripts during training. To address the duration modeling challenge, we train speaking rate predictors at different linguistic granularities to derive duration from speaker pace. Experiments show that our approach matches the performance of F5-TTS while enabling cross-lingual voice cloning.

Index Terms— flow matching, cross-lingual voice cloning

1. INTRODUCTION

Zero-shot Text-to-speech (TTS) aims to generate speech that closely resembles the voice of a given reference speech based on an input text. With the scaling of data and model size, existing TTS systems have shown superior capabilities in mimicking speaker characteristics and synthesizing more expressive speech. The modeling methodologies of zero-shot TTS can be categorized into two types: autoregressive (AR) models [1–6] and non-autoregressive (NAR) models [7–12].

Among NAR models, flow-matching-based approaches have emerged as particularly promising. For example, VoiceBox [7] surpasses AR-based VALL-E [1] in Word Error Rate (WER) and Speech Similarity (SIM) with 20 times faster inference speed. E2 TTS [8] and F5-TTS [9] have achieved remarkable improvements with their simpler architectures. Recent cross-lingual TTS approaches, such as VALL-E X [13], have achieved impressive cross-lingual voice cloning capabilities by training on large-scale multilingual data. However, all these models still rely on audio prompt transcripts, which creates inherent challenges when reference transcripts are unavailable. Different flow-matching-based models employ various duration modeling methods. While some flow-matching-based models use explicit phoneme duration predictors [7] , F5-TTS estimates duration by the ratio of the audio prompt transcript length and target text length. However, this mechanism breaks down in a cross-lingual context, as text-length ratios between different languages may not correspond to speech duration ratios.

In this paper, we propose Cross-Lingual F5-TTS, a novel framework based on F5-TTS that achieves cross-lingual zero-shot voice cloning by dropping audio prompt transcripts. Leveraging the MMS [14] forced alignment, we preprocess our training data to identify word boundaries. To solve the challenge of duration prediction, we introduce dedicated speaking rate predictors at three linguistic granularities including phoneme, syllable, and word levels. These models are trained to estimate speaking rate from the audio prompt, providing a robust, language-independent mechanism for determining the duration of the target utterance. Our approach achieves comparable performance to F5-TTS on LibriSpeech-PC test-clean and Seed-TTS-eval, while successfully extending its capabilities to cross-lingual scenarios with promising results. [1]

2. METHOD

2.1. Preliminary on Flow-matching-based TTS

Flow-Matching-based models [7–9] have recently achieved remarkable performance in TTS tasks. This approach offers significant advantages in terms of model simplicity. Specifically, by leveraging flow-matching, E2-TTS [8] and F5-TTS [9] eliminate additional components such as phoneme duration predictors and complex text encoders, thereby maintaining pipeline simplicity while achieving high-quality synthesis.

The Flow Matching framework aims to learn a time-dependent vector field v_t that matches the probability path p_t between a simple noise distribution p_0 and the data distribution q to generate flows ψ_t for sampled flow step $t \in [0, 1]$. The training objective is formulated as the Conditional Flow Matching (CFM) loss:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,q(x_1),p(x_0)} \|v_t(\psi_t(x)) - \frac{d}{dt}\psi_t(x)\|^2 \qquad (1)$$

This probability path connects a sample $x_0 \sim p(x_0)$ from the Gaussian noise to a data sample $x_1 \sim q(x_1)$ from the training data. Optimal Transport (OT) flow matching provides a particularly effective instantiation of this framework. Under the OT formulation, the flow ψ_t is defined as a straight-line trajectory:

---

† Corresponding author

$$\psi_t(x_0) = (1-t)x_0 + tx_1 \qquad (2)$$

The corresponding velocity field becomes the constant vector $(x_1 - x_0)$, leading to the OT-CFM loss:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,q(x_1),p(x_0)}\|v_t((1-t)x_0 + tx_1) - (x_1 - x_0)\|^2 \quad (3)$$

In this work, we employ F5-TTS as our baseline. F5-TTS is a fully flow-Matching-based TTS system with diffusion transformer (DiT), which is trained on the text-guided speech-infilling task using OT-CFM. The model predicts masked speech $m \odot x_1$ given its surrounding speech $(1-m) \odot x_1$, the noisy speech $(1-t)x_0 + tx_1$ and the extended character sequence z.

2.2. MMS Forced Alignment

To enable transcript-free voice cloning, we leverage the Massively Multilingual Speech (MMS) forced alignment tooling [14] to preprocess our training data and obtain precise word boundary information. The MMS forced alignment system employs a multilingual Wav2Vec2-based acoustic model trained with Connectionist Temporal Classification [15] to generate posterior probabilities for audio frames. For efficient processing of long audio recordings, MMS chunks audio files into 15-second segments to generate posterior probabilities, then concatenates them back into a unified alignment matrix.

We apply MMS forced alignment to preprocess the Chinese and English subsets of Emilia dataset [16]. For each audio sample and its corresponding transcription, we extract the end time of each word. During training, we modify the F5-TTS framework to incorporate this word boundary information as an additional input alongside the original speech and text inputs, as shown in Figure 1. At each training step, we randomly select a word boundary within the transcript and partition the audio sample at this point. The audio segment preceding the selected boundary serves as the audio prompt, while we completely discard its corresponding transcript portion. The remaining audio segment is masked and becomes the target for synthesis.
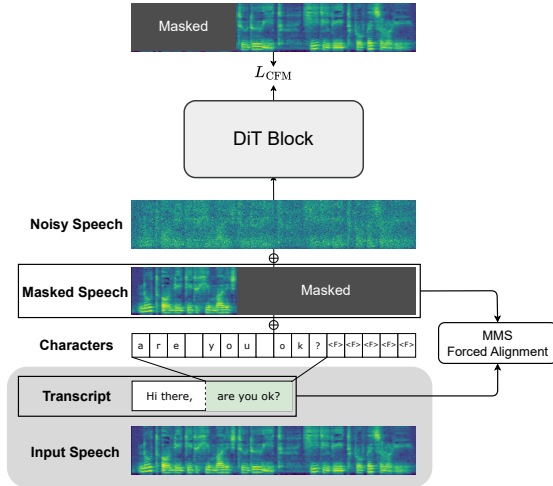

Fig. 1. Training framework of Cross-Lingual F5-TTS. MMS forced alignment produces word boundaries for training data, where the left segment serves as transcript-free audio prompt and the mel spectrogram of the right segment is masked for prediction.

2.3. Speaking Rate Predictor

The elimination of audio prompt transcripts in our transcript-free training approach introduces a critical challenge for duration prediction in speech synthesis. In the original F5-TTS framework, duration is estimated using a simple length-ratio method where duration equals audio prompt duration multiplied by the ratio of target text length to reference text length. However, this mechanism becomes infeasible when audio prompt transcripts are unavailable. To address this challenge, we introduce a dedicated speaking rate predictor that estimates duration directly from the acoustic characteristics of the audio prompt.

We formulate speaking rate prediction as a discrete classification task and train three separate models for different linguistic granularities: phonemes per second, syllables per second, and words per second. We define a category set C with uniform intervals of $\Delta = 0.25$. Specifically, the phoneme-level model uses $C = \{0.25, 0.5, \ldots, 17.75, 18.0\}$ with $N = 72$ classes, while both syllable-level and word-level models use $C = \{0.25, 0.5, \ldots, 7.75, 8.0\}$ with $N = 32$ classes. For a given speaking rate v, the ground truth category c_{gt} is determined by minimum distance mapping: $c_{\text{gt}} = \arg\min_{x \in C} |v - x|$. Each model takes audio features as input and independently predicts the corresponding speaking rate category for its respective linguistic unit.
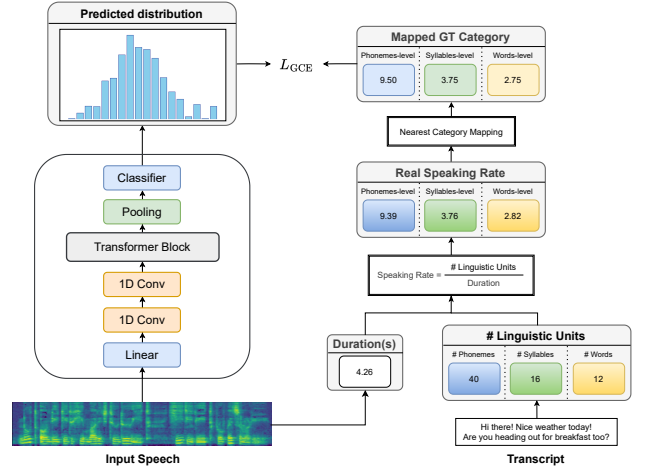
Fig. 2. Training pipeline of the speaking rate predictor. The model predicts discrete rate categories from mel-spectrograms, while ground truth speaking rates are mapped to nearest categories for GCE loss computation.

Our speaking rate predictor employs a transformer-based architecture designed to process mel-spectrogram inputs, as illustrated in Figure 2. The model consists of a mel projection layer that projects input mel-spectrograms to the model's hidden dimension, followed by two 1D convolution layers. Multiple transformer encoder layers process the sequence, and an attention-based sequence pooling mechanism aggregates the temporal information by computing attention weights for each time step and performing weighted averaging to obtain a fixed-size representation. Finally, a classifier outputs class probabilities for the speaking rate categories.

Standard cross-entropy loss treats all classes as independent, which is suboptimal for our ordered speaking rate categories. We introduce a Gaussian Cross-Entropy (GCE) loss that incorporates the ordinal nature of speaking rates:

$$L_{\text{GCE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_c^{\text{soft}} \log(\hat{y}_c) \qquad (4)$$

The soft labels are computed using a Gaussian kernel:

$$y_c^{\text{soft}} = e^{\frac{-(c-c_{\text{gt}})^2}{2\sigma^2}} \qquad (5)$$

where c_{gt} is the ground truth category index, c is the current category index, and σ controls the smoothness of the Gaussian kernel. This formulation assigns higher weights to categories closer to the ground truth, providing tolerance for minor prediction errors while maintaining precision.

During inference, our approach estimates the target generation duration through a straightforward process. The audio prompt is first fed into the speaking rate predictor to estimate the speaker's characteristic pace in terms of phonemes, syllables, or words per second, and the target text is processed to count the corresponding linguistic units. The target audio duration is then calculated as the ratio of the linguistic unit count to the predicted speaking rate. This mechanism enables our Cross-Lingual F5-TTS to perform duration estimation without relying on audio prompt transcripts, thereby facilitating true cross-lingual voice cloning capabilities.

3. EXPERIMENTS

3.1. Dataset

We train our Cross-Lingual F5-TTS on the Emilia dataset [16], a large-scale multilingual speech dataset collected from diverse real-world scenarios. After filtering out transcription failures and misclassified language speech, we retain approximately 95K hours of English and Chinese audio data. To train our speaking rate predictor, we extract a balanced subset containing 500 hours each from the Chinese and English portions of Emilia as training data.

3.2. Preprocessing

We apply MMS forced alignment tooling to extract word boundaries for the Emilia dataset. While the Emilia-pipe [16] employs Whisper-X [17] for transcription generation with considerable success, certain challenges remain in the generated transcriptions. Digit numbers, special symbols, and tokens from other languages occasionally appear in the transcriptions, which cannot be properly recognized during the forced alignment process. To address these cases, we implement specialized preprocessing procedures that skip anomalous tokens and exclude them from word boundary extraction, ensuring robust and accurate alignment results.

3.3. Model Setup

Our baseline is open-sourced F5-TTS-Base [9], which employs a diffusion transformer (DiT) architecture with 22 layers, 16 attention heads, and 1024 dimensional embeddings. The model is trained for 1.2M updates on eight NVIDIA A100 GPUs with a per-GPU batch size of 38,400 audio frames. We use AdamW optimizer [18] with a learning rate that linearly warms up to 7.5×10^{-5} over the first 20k updates, followed by linear decay for the remaining training steps.

Our speaking rate predictor utilizes a transformer-based architecture with 6 layers, 8 attention heads, and 512 dimensional embeddings. Training is conducted on four A100 GPUs for 50k updates with a per-GPU batch size of 38,400 audio frames. The learning rate is warmed up to 2.5×10^{-4} over the first 7.5k updates and then

linearly decayed. For the Gaussian Cross-Entropy loss, we set the standard deviation $\sigma = 1.0$ to balance between tolerance for nearby predictions and precision for the target category.

For inference, we follow the standard F5-TTS settings using Euler ODE solver with 32 function evaluations (NFE = 32), CFG strength 2.0, sway sampling coefficient -1.0, and pre-trained Vocos [19] as the vocoder.

3.4. Evaluation

We follow the evaluation setting of F5-TTS, adopting Seed-TTS-eval and LibriSpeech-PC test-clean as our test set. We also build a multilingual cross-lingual test set with 473 samples of 3-8 second audio prompts from FLEURS [20], covering four languages (German, French, Hindi, Korean) to synthesize both English and Chinese speech. The evaluation is conducted using the following three metrics:

Word Error Rate (WER) measures the intelligibility of synthesized speech by comparing its transcription with the ground truth text. We employ Whisper-large-V3 [21] and Paraformer-zh [22] for automatic recognition and compute WER accordingly.

Speaker Similarity (SIM-o) quantifies the resemblance between the synthesized and the original target speeches. We use WavLM-large-based [23] speaker verification model to extract speaker embeddings and compute the cosine similarity between them.

UTMOS [24] provides an automated assessment of speech naturalness through a pre-trained MOS prediction model. It estimates audio quality without requiring reference recordings or labels. While not an absolute subjective measure, UTMOS provides a practical and efficient way to evaluate the naturalness of synthetic speech.

To evaluate the effectiveness of our speaking rate predictor, we assess its duration prediction accuracy using two complementary metrics:

Mean Relative Error (MRE) measures the relative accuracy of duration prediction. It is calculated as the average relative difference between the predicted duration (derived from the number of linguistic units divided by the predicted speaking rate) and the ground truth audio duration.

Mean Absolute Error (MAE) quantifies the absolute deviation between the predicted and ground truth duration.

Table 1. MAE (s) and MRE (%) of predicted durations for different speaking rate predictors.

ID	System	MAE(s)↓	MRE(%)↓
	Librispeech-PC *test-clean*		
M1	Phonemes-level predictor	0.759	**11.932**
M2	Syllables-level predictor	**0.757**	11.945
M3	Words-level predictor	1.171	18.406
	SeedTTS *test-en*		
M1	Phonemes-level predictor	**0.637**	**15.017**
M2	Syllables-level predictor	0.704	16.497
M3	Words-level predictor	0.886	20.031
	SeedTTS *test-zh*		
M1	Phonemes-level predictor	0.845	14.469
M2	Syllables-level predictor	**0.783**	**13.771**
M3	Words-level predictor	0.908	16.156

4. RESULTS

4.1. Speaking Rate Predictor Performance

Table 1 presents the duration prediction accuracy results of our speaking rate predictors across different linguistic granularities. M1 (phoneme-level predictor) outperforms other predictors on LibriSpeech-PC test-clean and shows comparable results to M2 on SeedTTS test-en, demonstrating its effectiveness for English duration modeling. Notably, M2 (syllable-level predictor) performs optimally on SeedTTS test-zh, suggesting that syllables may be more natural linguistic units for duration modeling in Chinese. The consistently inferior performance of M3 (word-level predictor) across all datasets highlights the limitations of coarse-grained linguistic units.

Table 2. Results on LibriSpeech-PC test-clean, SeedTTS test-en, and SeedTTS test-zh.

System	Duration Method	WER(%)↓	SIM-o↑	UTMOS↑
LibriSpeech-PC *test-clean*				
Baseline	Length-ratio	2.205	**0.668**	3.797
CL-F5	**M1**	**2.079**	0.663	3.884
CL-F5	**M2**	2.120	0.658	**3.892**
CL-F5	**M3**	2.894	0.652	3.855
SeedTTS *test-en*				
Baseline	Length-ratio	1.545	**0.676**	3.581
CL-F5	**M1**	**1.513**	0.662	**3.629**
CL-F5	**M2**	1.594	0.660	3.625
CL-F5	**M3**	2.009	0.646	3.593
SeedTTS *test-zh*				
Baseline	Length-ratio	**1.475**	0.762	2.898
CL-F5	**M1**	1.605	0.759	**2.913**
CL-F5	**M2**	1.481	**0.764**	2.887
CL-F5	**M3**	1.616	0.763	2.889

4.2. Intra-lingual Voice Cloning Performance

Table 2 compares our Cross-Lingual F5-TTS (CL-F5) with the original F5-TTS baseline on standard benchmarks. Our method demonstrates competitive performance across all evaluation metrics. On LibriSpeech-PC test-clean, CL-F5 with M1 and M2 achieves superior WER and UTMOS compared to the baseline. This trend continues on SeedTTS test-en, where CL-F5 with M1 delivers better WER and UTMOS than the baseline. Although SIM is slightly lower for CL-F5 compared to the baseline on the English test sets, this modest decrease is an acceptable trade-off for the extension of cross-lingual capability. On SeedTTS test-zh, CL-F5 with M2 demonstrates performance almost identical to the baseline across all metrics. The language-specific preferences observed in Section 4.1 translate to the intra-lingual voice cloning: CL-F5 with fine-grained predictors (M1, M2) consistently outperforms CL-F5 with M3. These results fully validate the effectiveness of our method that CL-F5 retains intra-lingual voice cloning quality while eliminating the dependency on audio prompt transcripts.

4.3. Cross-lingual Voice Cloning Results

Table 3 presents the cross-lingual voice cloning results on our multilingual test set, where the audio prompt and target synthesis languages differ. The cross-lingual results reinforce the language-specific preferences observed in Section 4.1 and Section 4.2: CL-F5

achieves optimal performance with M1 for English targets and M2 for Chinese targets. CL-F5 with M3 shows significant performance degradation in cross-lingual scenarios, with WER dramatically increasing to 16.494% for English targets compared to 2.496% for CL-F5 with M1. This is because M3 produces overly fast speaking rate, leading to compressed temporal patterns that negatively impact intelligibility. These results highlight the critical importance of accurate duration modeling in cross-lingual TTS, as temporal inaccuracies can severely degrade overall synthesis quality.

Remarkably, our approach demonstrates effective cross-lingual voice cloning capabilities despite the speaking rate predictors being trained exclusively on Chinese and English data, showing good generalization to unseen languages without requiring audio prompt transcripts.

Table 3. Cross-lingual voice cloning results of Cross-Lingual F5-TTS. GT length denotes the result obtained by using ground truth total speech length.

Duration Method	WER(%)↓	SIM-o↑	UTMOS↑
Cross-lingual *test-en*			
GT length	**2.462**	0.530	**3.083**
M1	2.496	**0.543**	3.069
M2	4.362	0.518	3.059
M3	16.494	0.486	2.926
Cross-lingual *test-zh*			
GT length	**1.596**	0.558	2.452
M1	2.446	0.555	2.494
M2	1.801	**0.565**	**2.503**
M3	1.946	0.563	2.492

5. CONCLUSION

In this work, we present Cross-Lingual F5-TTS, a novel framework that achieves cross-lingual zero-shot voice cloning by eliminating the dependency on audio prompt transcripts. Our approach addresses a fundamental limitation of existing NAR TTS models by leveraging MMS forced alignment to partition training data at word boundaries. To solve the consequent duration prediction challenge, we introduce dedicated speaking rate predictors that estimate duration directly from acoustic features using phoneme, syllable, or word-level granularities. Experimental results demonstrate that our method maintains comparable performance to the original F5-TTS on LibriSpeech-PC test-clean and Seed-TTS-eval, while successfully enabling cross-lingual voice cloning from speakers of unseen languages that the baseline cannot handle. The speaking rate predictor analysis reveals that finer-grained linguistic units including phoneme and syllable provide more reliable duration estimation. Although our transcript-free approach enables cross-lingual voice cloning, it shows reduced capability in transferring subtle speaker characteristics such as accents and emotions compared to the original F5-TTS. In future work, we plan to explore methods to compensate for the missing linguistic information from audio prompt transcripts, enabling Cross-Lingual F5-TTS to generate more expressive speech while maintaining its cross-lingual capabilities.

6. REFERENCES

[1] C. Wang, S. Chen., Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neu-

ral codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2023.

[2] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, et al., "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[3] D. Jia, Z. Chen, J. Chen, C. Du, J. Wu, J. Cong, X. Zhuang, C. Li, Z. Wei, Y. Wang, and Y. Wang, "DiTAR: Diffusion transformer autoregressive modeling for speech generation," *arXiv preprint arXiv:2502.03930*, 2025.

[4] B. Zhang, C. Guo, G. Yang, H. Yu, H. Zhang, H. Lei, J. Mai, J. Yan, K. Yang, M. Yang, et al., "MiniMax-Speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder," *arXiv preprint arXiv:2505.07916*, 2025.

[5] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu, H. Meng, and F. Wei, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[6] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "CosyVoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[7] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Proc. NeurIPS*, 2023, vol. 36, pp. 14005–14034.

[8] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, Y. Liu, S. Zhao, and N. Kanda, "E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot tts," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 682–689.

[9] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proc. ACL*, 2025, pp. 6255–6271.

[10] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. ICLR*, 2024.

[11] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, et al., "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. ICML*, 2024.

[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[13] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak Foreign Languages with Your Own Voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[14] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, et al., "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[16] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 885–890.

[17] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in *Proc. Interspeech*, 2023.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.

[19] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *Proc. ICLR*, 2024.

[20] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *Proc. IEEE SLT*, 2022.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[22] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *Proc. INTERSPEECH*, 2022.

[23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[24] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," in *Proc. INTERSPEECH*, 2022.