Speak Better via Simultaneous Listening and Speaking in Unified Decoder-Only LLMs

Chung Tran*, Bagus Tris Atmaja*, and Sakriani Sakti*
* Nara Institute of Science and Technology, Japan

***Abstract*—Large Language Models (LLMs) have recently revolutionized text-to-speech (TTS), enabling natural, high-fidelity synthesis. Yet systems like CosyVoice2 and Spark-TTS need over 100,000 hours of training data, limiting low-resource languages. We argue that for TTS to "speak better," the solution is not just adding data but training with its "listening" counterpart—automatic speech recognition (ASR). Prior work on the Machine Speech Chain (MSC) shows joint TTS–ASR training improves low-resource performance but relies on two separate models and complex pipelines. To address this, we propose a Unified Speech Chain (USC), a decoder-only LLM that processes text and speech as token sequences within a single architecture. We further introduce a Multi-Stage Chain Loop (MSCL) that leverages unpaired and unlabeled text and speech for self-improvement. Extensive experiments in low-resource settings show USC significantly improves TTS intelligibility, lowering word and character error rates, while MSCL further exploits unlabeled data, yielding additional gains.**

***Index Terms*—Unified Speech Chain, Low-Resource Languages, Speech Synthesis, Speech Chain**

I. Introduction

Text-to-Speech (TTS) has advanced dramatically over the past decade, evolving from rule-based and statistical models to neural architectures capable of generating natural, expressive, and high-fidelity speech. A key driver of this progress is the rise of Large Language Models (LLMs), which excel at sequence modeling and have recently been extended from text to speech. Models such as VALL-E [1], CosyVoice2 [2], and Spark-TTS [3] demonstrate not only human-like synthesis but also controllability features such as prompt-driven prosody and voice transfer. These advances mark a shift from task-specific pipelines toward general-purpose generative models, where a single architecture can flexibly handle diverse speech generation scenarios.

Most LLM-based TTS systems cast speech generation as an autoregressive sequence prediction task over discrete units—semantic, prosodic, or acoustic—derived from quantized speech representations (e.g., HuBERT [4], EnCodec [5]). This formulation unifies linguistic and acoustic modeling and allows seamless integration with pretrained decoder-only LLMs (e.g., GPT-2 [6], Qwen2.5 [7]), bridging text and speech within a single generative model.

Despite these advances, a major limitation remains: extreme data dependence. State-of-the-art LLM-TTS systems often require more than 100,000 hours of paired speech–text data, which is impractical for low-resource languages and domains. Without such massive supervision, intelligibility and robustness degrade significantly. To address data scarcity, SPEAR-TTS [8] decomposes the task into Reading (text-to-semantic) and Speaking (semantic-to-acoustic) modules, leveraging pretraining and back-translation. Similarly, Align2Speak [9] employs a pretrained ASR model as an external reward to optimize TTS via preference optimization. While effective, this approach still cannot exploit the vast amount of unpaired or unlabeled speech and text, which is much easier to collect.

We argue that for TTS to "speak better," the solution lies not merely in adding more data, but in training together with its "listening" counterpart—Automatic Speech Recognition (ASR). Prior work on the Machine Speech Chain (MSC) [10] demonstrated that coupling TTS with ASR in a self-improving loop can leverage unlabeled data. However, MSC relies on two separate models and complex training strategies (e.g., straight-through estimation via Gumbel-Softmax), which limits simplicity and scalability.

In this research, we simplify the process by adopting a unified decoder-only model, following recent work such as CosyVoice2 [2] and Spark-TTS [3], and multitask models like VoxtLM [11]. Although VoxtLM provides a flexible multitask model, it was not designed to address our central question: whether TTS can "speak better" by explicitly training with its listening counterpart. More importantly, it neither couples listening (ASR) and speaking (TTS) in a loop, nor tackles the challenges of low-resource TTS.

Building on these insights, we propose the Unified Speech Chain (USC), a decoder-only LLM that processes both text and speech as token sequences within a single architecture, jointly optimizing TTS and ASR. To further exploit diverse data sources, we introduce a Multi-Stage Chain Loop (MSCL) method that enables USC to self-improve by iteratively leveraging paired, unpaired, and unlabeled data. By unifying generation and recognition in a single model and incorporating self-improving loops, our approach provides a scalable and low-resource-friendly alternative to existing pipelines. In summary, our work bridges the gap between state-of-the-art LLM-TTS systems and speech chain approaches, introducing a unified and loop-based method tailored to low-resource scenarios.

Our contributions are threefold:

- Unified Speech Chain (USC): We design a decoder-only LLM that integrates TTS and ASR within one architecture, simplifying training while enabling cross-task benefits.
- Multi-Stage Chain Loop (MSCL): We propose a novel training method that exploits unpaired and unlabeled text/speech, enhancing adaptability in low-resource sce-

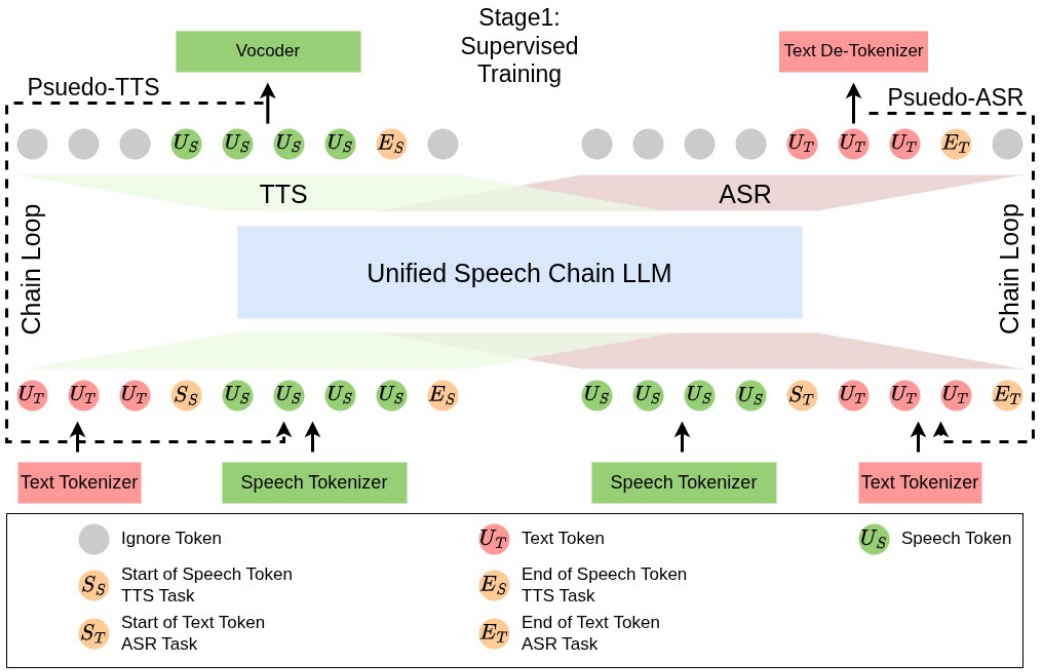


Fig. 1. Unified Speech Chain integrating TTS (Speaking) and ASR (Listening). Both speech and text are tokenized into discrete units, processed in a single Decoder-only SpeechLM, and reconstructed through a vocoder or text de-tokenizer.

narios.

- Comprehensive Evaluation: We demonstrate that USC with MSCL consistently improves TTS intelligibility and robustness across model sizes, data scales, and languages, establishing a new paradigm for low-resource LLM-TTS.

II. Proposed Approach

A. Unified Speech Chain LLM

To enable both ASR and TTS within a single system, we propose the Unified Speech Chain (USC), a model that casts listening and speaking as two directions of the same generative process. As shown in Fig.1, both text and speech are represented as sequences of discrete tokens: text is segmented with Byte Pair Encoding (BPE), while speech is discretized into semantic units derived from HuBERT [4] features clustered with K-means. Special tokens are introduced to mark task boundaries (e.g., start/end of speech for TTS, start/end of text for ASR), allowing the model to distinguish between modalities within a single sequence.

These unified token streams are processed by a decoder-only LLM backbone (Qwen2.5-0.5B), which learns shared cross-modal representations. Within this setup, ASR is formulated as speech-to-text generation, and TTS as text-to-speech generation, both unified under the same generative model. Importantly, the chain loop enables pseudo-labeling: text can be converted into pseudo speech-unit sequences, and speech into pseudo text-unit sequences. This allows USC to learn not only from paired data but also from unpaired and unlabeled inputs, setting the foundation for the Multi-Stage Chain Loop (MSCL) training strategy described in the next section.

B. Training Procedure

To train the model, we construct task-specific input–output sequences, with the overall design illustrated in Figure 1. In both tasks, the model is provided with mixed text–speech sequences enriched with special tokens that mark task boundaries, enabling it to distinguish between modalities within a unified sequence.

- **TTS task (Speaking)**
  - **Input:** $U_T \ \langle S_S \rangle \ U_S \ \langle E_S \rangle$
  - **Target:** $\langle Ig \rangle \ U_S \ \langle E_S \rangle$
- **ASR task (Listening)**
  - **Input:** $U_S \ \langle S_T \rangle \ U_T \ \langle E_T \rangle$
  - **Target:** $\langle Ig \rangle \ U_T \ \langle E_T \rangle$

As shown in Figure 1, U_S and U_T denote sequences of discrete speech and text tokens, respectively. The special tokens $\langle S_S \rangle$ / $\langle E_S \rangle$ and $\langle S_T \rangle$ / $\langle E_T \rangle$ indicate the start and end of speech and text spans, while $\langle Ig \rangle$ marks positions to be ignored when calculating the loss. For multilingual training, $\langle S_S \rangle$ and $\langle S_T \rangle$ are defined per language to provide clear task and language boundaries.

The model is optimized on paired speech–text data according to the procedure outlined in Algorithm 1. Note that both tasks are optimized with a token-level cross-entropy loss computed only over the target span.

C. Multi-Stage Chain Loop

The Multi-Stage Chain Loop (MSCL) extends USC beyond supervised learning by enabling the model to leverage both labeled and unlabeled data. As illustrated in Figure 1, Stage 1 consists of supervised training on paired speech–text data.

Algorithm 1: Training USC-LLM model

Initialize USC-LLM model
while *not converged* **do**
  Sample a batch of data B_{tts} and B_{asr}
  $\mathcal{L}_{tts} \leftarrow$ USC-LLM(B_{tts}) // Speaking
  $\mathcal{L}_{asr} \leftarrow$ USC-LLM(B_{asr}) // Listening
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{tts} + \mathcal{L}_{asr}$
  Back-propagation to update parameters
end

Once this baseline model is trained, MSCL proceeds to Stage 2, where the model generates pseudo-labels from unpaired data, as shown in Figure 2. Specifically, when given text-only inputs, the model operates in TTS mode to synthesize pseudo speech tokens, while for speech-only inputs it operates in ASR mode to produce pseudo text tokens.

In Stage 3, these pseudo-labeled pairs are merged with real supervised data to form enriched training corpora. The model is then retrained in a chain loop, where recognition and generation iteratively reinforce each other. This process allows USC to continuously self-improve, effectively exploiting unpaired and unlabeled resources that would otherwise remain unused.

III. Experimental Setup

A. Datasets and Evaluation Protocol

- **Monolingual Setup:** We use the LJSpeech dataset [12], which contains 24 hours of English speech, and partition it into training, development, and test sets with an 80/10/10 split. The proposed model is trained under multiple configurations, varying model size, training set size, and the use of the MSCL method.
- **Multilingual Setup:** To evaluate generalization and robustness in low-resource conditions, we extend training to a multilingual corpus covering English (EN), Spanish (ES), French (FR), Vietnamese (VI), Chinese (ZH), Japanese (JA), and Indonesian (ID). Specifically, English data is drawn from LJSpeech (24 h); Spanish (24 h), French (19 h), Chinese (6 h), and Japanese (14 h) from CSS10 [13]; Vietnamese (15 h) from VIVOS [14]; and Indonesian (24 h subset) from INDSpeech [15]. Each language is split into training, development, and test sets following the 80/10/10 ratio.
- **Evaluation Metrics:** Intelligibility is evaluated using word error rate (WER) and character error rate (CER), with Whisper-Large V2 [16] for transcription and the JiWER library for error calculation. Speech naturalness is measured using UTMOS [17], a widely used automatic naturalness metric.

B. Implementation Parameters

We employ Qwen2.5-0.5B as the pretrained backbone of USC-LLM. The model is optimized using the AdamW [18] optimizer with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and is trained on a single NVIDIA A100 GPU with a batch size of 8.

For speech tokenization, we adopt the mHuBERT [19] model. This quantizer model maps the features extracted from the 11th layer of mHuBERT into 1,000 discrete tokens. To reconstruct waveforms from token sequences, we employ a Grad-TTS acoustic model trained on the same dataset as USC-LLM, followed by a pretrained HiFi-GAN [20].

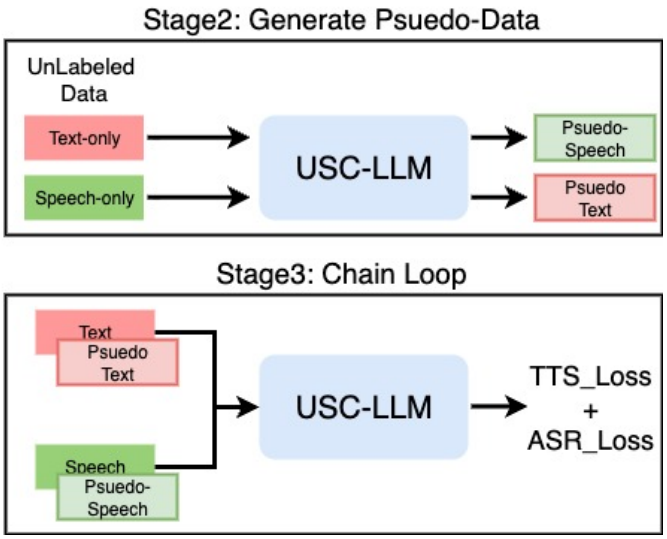


Fig. 2. Pseudo speech and pseudo text are generated from unlabeled data, then combined with real data to optimize USC-LLM model

TABLE I
Performance of TTS-only and USC-LLM on LJSpeech under varying dataset sizes and model depths

Data Size	Layer Depth	TTS-based LLM			USC-LLM		
		WER↓	CER↓	UTMOS↑	WER↓	CER↓	UTMOS↑
50%	12L	**70.08**	**57.43**	-	71.05	58.69	-
	24L	26.35	18.67	-	**20.94**	**14.96**	-
75%	12L	40.95	32.90	-	**30.70**	**24.28**	-
	24L	13.44	8.94	3.90	**12.90**	**8.77**	3.90
100%	12L	24.24	18.75	3.91	**20.51**	**15.98**	3.91
	24L	10.27	6.66	3.91	**10.06**	**6.56**	3.92
100%	VITS [21]	13.74	3.59	3.87	-	-	-

IV. Experimental Results

A. Monolingual USC-LLM

The experimental results in Table I show that the proposed USC model with joint ASR+TTS loss improves performance in most settings. Across different data availability levels (50%, 75%, and 100%), the joint-loss formulation yields lower error rates than training with TTS loss alone, for both the smaller (12-layer) and larger (24-layer) configurations. With the full dataset, the 12-layer USC-LLM reduces WER by 4% and CER by 3% relative to the TTS-only baseline, while the 24-layer variant achieves comparable gains, confirming the robustness of the approach.

TABLE II
IMPACT OF THE MSCL METHOD ON SPEECH SYNTHESIS TASK.

	Labeled Data	PsData	Labeled+ASR PsData		Labeled+TTS PsData		Labeled + ASR-TTS PsData	
			WER↓	CER↓	WER↓	CER↓	WER↓	CER↓
24L (LB)	50%	-	20.94	14.96	20.94	14.96	20.94	14.96
24L	50%	25%	15.50	10.70	22.00	16.35	19.65	14.48
24L	50%	50%	13.76	9.56	20.68	15.27	20.36	15.19
24L (UB)	100%	-	10.06	6.56	10.06	6.56	10.06	6.56
24L (LB)	75%	-	12.90	8.77	12.90	8.77	12.90	8.77
24L	75%	15%	12.20	8.64	11.57	7.83	11.70	7.83
24L	75%	25%	11.58	7.55	10.04	6.64	10.64	6.85
24L (UB)	100%	-	10.06	6.56	10.06	6.56	10.06	6.56

UB: Upper Bound LB: Lower Bound PsData: Pseudo Data

TABLE III
IMPACT OF THE MSCL METHOD ON ASR

DataSize	ASR-TTS PsData	USC-LLM		Multi-model MSC [10]	
		WER↓	CER↓	WER↓	CER↓
50%	-	13.78	7.86	-	-
50%	50%	13.16	7.21	-	3.5
75%	-	10.39	5.22	-	-
75%	25%	9.34	4.41	-	-
100%	-	8.82	4.08	-	3.1

TABLE IV
EVALUATION OF USC-LLM ON MULTILINGUAL DATASETS

Layer Size	Lang	TTS-based LLM		USC-LLM	
		WER↓	CER↓	WER↓	CER↓
12L	ZH	37.62	17.32	**36.65**	**16.28**
	EN	13.64	9.08	**10.58**	**6.70**
	FR	20.38	10.30	**18.00**	**8.39**
	ID	43.03	15.40	**35.62**	**11.98**
	JA	31.03	13.65	**30.31**	**13.36**
	ES	10.75	4.66	**9.09**	**3.84**
	VI	**22.93**	**12.51**	23.82	13.06
24L	ZH	36.54	15.86	**36.12**	**15.53**
	EN	10.97	7.00	**10.15**	**6.19**
	FR	18.04	8.32	**17.82**	**7.97**
	ID	**35.06**	12.78	35.73	**12.28**
	JA	**30.67**	**13.15**	31.59	13.91
	ES	8.92	3.46	**8.40**	**3.35**
	VI	23.76	13.01	**22.64**	**12.53**

We further benchmarked USC-LLM against VITS, a strong TTS baseline. Specifically, the 24-layer USC-LLM achieves lower WER (10.06 vs. 13.74) and slightly higher UTMOS [17] (3.92 vs. 3.87) than VITS, indicating improved intelligibility without sacrificing naturalness. These results demonstrate that USC-LLM effectively enhances speech quality and robustness, particularly under resource-limited conditions.

B. Multi-Stage Chain Loop

Table II reports the results of the proposed MSCL method for TTS. Each row specifies the proportion of labeled and pseudo-labeled data used. For example, a configuration with 50% labeled data may be combined with 50% ASR pseudo data, 50% TTS pseudo data, or both. When only 50% of the data is labeled, incorporating ASR pseudo-labels reduces WER by 5–7% and CER by 4–5%, while TTS pseudo-labels yield limited gains due to model instability. With 75% labeled data, however, both ASR and TTS pseudo-labels improve performance, with TTS contributing an additional 2% reduction in WER and CER.

Table III further shows that MSCL also improves ASR performance, reaching accuracy close to that of the multi-model MSC, though slightly lower. Crucially, USC with MSCL remains unified and substantially easier to train than MSC's dual-model approach.

Overall, these findings demonstrate that MSCL effectively exploits unlabeled and unpaired text and speech, yielding consistent improvements even under low-resource conditions.

C. Multilingual USC-LLM

We further evaluate our approach on multilingual low-resource datasets using both 12-layer and 24-layer model configurations. As shown in Table IV, the ASR+TTS loss improves performance across most languages, yielding absolute reductions of 1–2% in both WER and CER relative to the TTS-only baseline. For instance, with the 12-layer model, the CER for English (EN) and French (FR) decreases by approximately 3% and 2%, respectively. Notably, Spanish (ES) achieves particularly strong results, with WER and CER reaching 9.09% and 3.84%, respectively. Even in more challenging settings such as Chinese, which has only about 6 hours of training data, and Indonesian, which contains repetitive utterances, USC still provides measurable improvements, underscoring its robustness in diverse and low-resource conditions.

V. Conclusion

We presented a Unified Speech Chain (USC) LLM designed for low-resource speech synthesis. Whereas state-of-the-art LLM-TTS models achieve strong performance only with massive paired datasets, and speech chain approaches require multiple models and complex optimization, USC unifies listening (ASR) and speaking (TTS) within a single decoder-only architecture. Together with the Multi-Stage Chain Loop (MSCL), this technique can exploit paired, unpaired, and unlabeled data, yielding consistent improvements in intelligibility and robustness, while maintaining naturalness at least on par with strong baselines.

Beyond these empirical gains, our work highlights a broader perspective: bridging the gap between LLM-TTS and speech chain paradigms through a unified, loop-based method tailored to data-scarce settings. Future directions include enhancing multilingual optimization and extending USC toward expressive speech generation with controllable attributes such as emotion and speaking style.

Acknowledgments Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681, JP24K0296, and JP25H01139, as well as JST NEXUS (JPMJNX25C1).

References

[1] C. Wang, S. Chen, and et al., "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *CoRR*, 2023. DOI: 10.48550/ARXIV.2301.02111. [Online]. Available: https://doi.org/10.48550/arXiv.2301.02111.

[2] Z. Du, Y. Wang, and et al. and, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *CoRR*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.10117.

[3] X. Wang, M. Jiang, and el al., *Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens*, 2025.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, 2021.

[5] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=ivCd8z8zR2.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

[7] Qwen et al., *Qwen2.5 technical report*, 2025.

[8] E. Kharitonov, D. Vincent, Z. Borsos, and et al., "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, 2023.

[9] S. S. Hussain et al., "Align2speak: Improving tts for low resource languages via asr-guided online preference optimization," *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[10] A. Tjandra, S. Sakti, and S. Nakamura, "Machine Speech Chain," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 976–989, 2020. DOI: 10.1109/TASLP.2020.2977776. [Online]. Available: https://doi.org/10.1109/TASLP.2020.2977776.

[11] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, "Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks," in *ICASSP 2024*.

[12] K. Ito and L. Johnson, *The LJ Speech Dataset*, 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/.

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

[14] H.-T. Luong and H.-Q. Vu, *Vivos: Vietnamese speech corpus for asr*, version 1.0, Zenodo, Dec. 2016. DOI: 10.5281/zenodo.7068130. [Online]. Available: https://doi.org/10.5281/zenodo.7068130.

[15] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008. [Online]. Available: https://aclanthology.org/I08-8004.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*.

[17] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *IEEE Spoken Language Technology Workshop (SLT)*.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th ICLR 2019*. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[19] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, "mHuBERT-147: A Compact Multilingual HuBERT Model," in *Interspeech 2024*, 2024.

[20] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[21] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End

Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021. [Online]. Available: https://proceedings.mlr.press/v139/kim21f.html.