Improving Automatic Speech Recognition Model for Super-Elderly Voice Using Speech Synthesis Model

Ryota Uematsu*, Chee Siang Leow*, Norihide Kitaoka†, and Hiromitsu Nishizaki*
* University of Yamanashi, Japan
E-mail: r_uematsu@alps-lab.org, {leow,hnishi}@yamanashi.ac.jp Tel/Fax: +81-552208361
† Toyohashi University of Technology, Japan
E-mail: kitaoka@tut.jp

***Abstract*—Although speech recognition technology has made remarkable progress, recognizing elderly speech remains challenging due to age-related acoustic characteristics and limited training data. This paper introduces a novel data augmentation framework that utilizes text-to-speech synthesis specifically optimized for elderly voices. Our approach employs Style-Bert-VITS2 to learn from a small elderly speech corpus and generate additional training data while preserving distinctive elderly voice characteristics. Experiments using the EARS corpus demonstrate that our method achieves a statistically significant improvement in recognition accuracy, reducing the Character Error Rate (CER) by 0.31 percentage points compared to conventional approaches.**

I. Introduction

Recent advances in automatic speech recognition (ASR) have greatly expanded the use of speech-based systems in daily life and various industries. Popular applications include virtual assistants such as Apple's Siri and Google's voice services, where users can conveniently query information through speech. This has also motivated the development of automated transcription tools that can generate real-time meeting notes and reduce the burden on human notetakers.

Although speech recognition rates for young people speech have become impressively high, ASR performance on the elderly population remains problematic. Notably, the recognition error rate increases significantly for speakers aged 70 and older compared to younger speakers. Vipperla et al. [1] demonstrated that word error rates (WER) for older voices (60-85 years, mean 68.4) were 10% absolute higher compared to adult voices (30-45 years, mean 41.3). Two primary factors account for this disparity: (1) acoustic changes (e.g., reduced pitch range, slower articulation) and (2) lack of large, annotated corpora of elderly speech. Since these changes do not appear in young peoples' speech data, existing ASR models fail to generalize well to super-elderly speakers.

Furthermore, the issue of aging is pressing in Japan, where the proportion of elderly people continues to rise. Projections estimate that by 2040, nearly 35% of Japan's population will be elderly, increasing the need for ASR systems capable of serving an older demographic. Therefore, improving ASR performance for super-elderly speech has become an urgent research topic.

Several studies have attempted to address the mismatch between young people's and elderly speech characteristics. Chen et al. [3] proposed a transfer learning approach to refine ASR systems for older adults, focusing on adapting models trained on younger speech to better accommodate elderly vocal patterns. Their method demonstrated improved recognition performance by leveraging attention mechanisms and domain adaptation techniques when training data for elderly speakers is limited. Similarly, Suhasini et al. [4] developed specialized ASR systems targeting Tamil-speaking elderly populations, addressing the dual challenge of low-resource languages and age-related speech variations. For Japanese language processing, the EARS (Elderly Adults Read Speech) corpus [5], [6] was specifically designed to collect utterances from speakers in their 70s to 90s, although the corpus contains only 12 hours of speech data in total.

Fukuda et al. [6], who developed this corpus, conducted a comparative analysis between the EARS corpus and a widely utilized speech corpus for acoustic model training. Their findings revealed a significant trend: as speaker age increased, the greater the decline in speech recognition accuracy.

Various approaches have been attempted to address data scarcity [7]–[9]. Among these, text-to-speech (TTS) synthesis has emerged as a promising solution for ASR data augmentation across multiple domains. Do et al. [10] demonstrated significant improvements in accented English recognition through unsupervised TTS synthesis, while Matsuzaka et al. [11] successfully applied similar approach to dysarthric speech adaptation. While TTS-based data augmentation has been extensively explored across various domains, its potential for enhancing ASR performance on elderly speech has received limited attention.

More sophisticated approaches have explored closed-loop architectures that co-train TTS and ASR components. Tjandra et al. [12] proposed such a framework that enables training on concatenated labeled and unlabeled data, demonstrating that ASR and TTS can improve performance by teaching each other using only unpaired data. However, these approaches typically require parallel training of both systems and may not fully capture the unique prosodic patterns of elderly speech. To address synthesis quality concerns, Ueno et al. [13] developed phone-informed mel-spectrogram refinement techniques for efficient data augmentation in automatic speech recognition systems.

Specifically targeting elderly speech characteristics, Chen et al. [14] developed a human-in-the-loop framework for elder-

979-8-3315-7206-8/25/$31.00

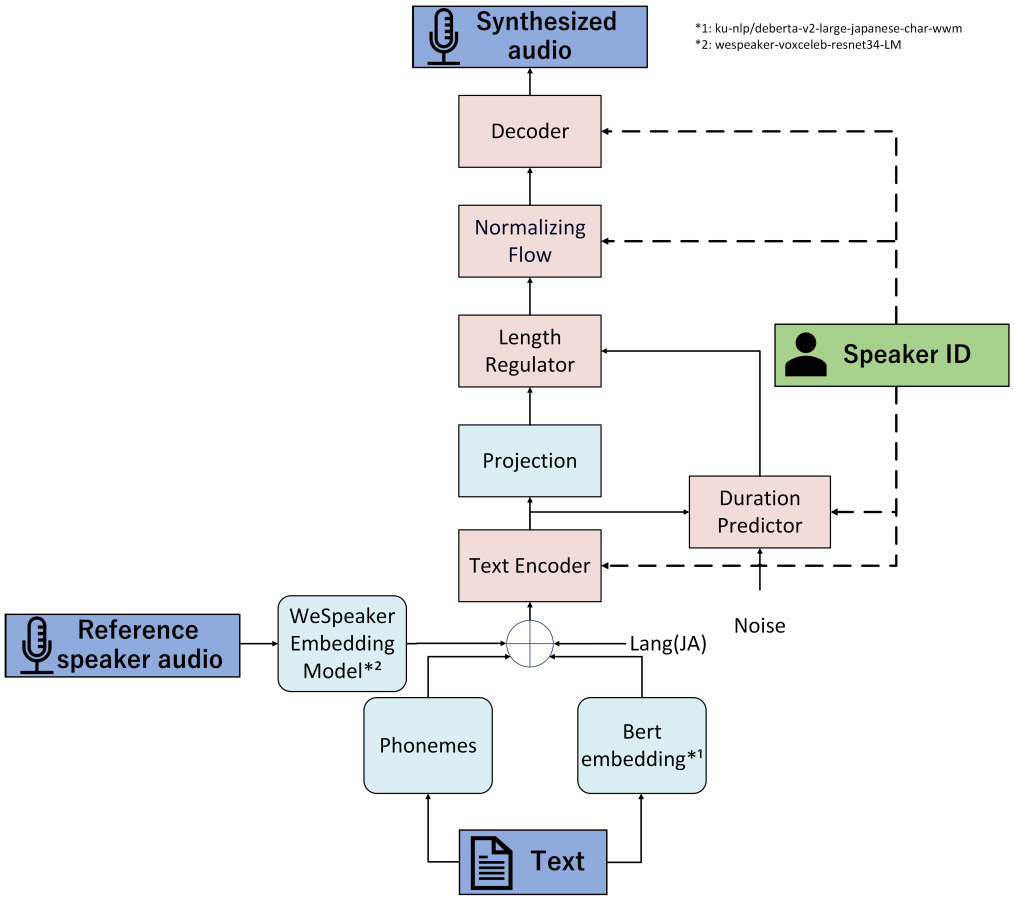


Fig. 1: Overview of the Style-Bert-VITS2 [2] model architecture used for elderly speech synthesis

facing TTS that automatically selects speaking styles preferred by older listeners, achieving slower speaking rates through iterative preference optimization. This work demonstrates the potential for age-specific TTS adaptations in addressing the unique challenges of elderly speech synthesis.

Various Text-to-Speech models are currently being developed [15]–[18], the growing role of TTS in speech processing leads speaker embedding models have been increasingly utilized to enhance the quality. Traditionally, speaker embedding models have predominantly relied on generative approaches, with the i-vector framework based on Gaussian Mixture Models (GMM) being widely used. However, the advent of x-vector [19], a discriminative method leveraging deep learning, marked a significant improvement in speaker recognition accuracy over i-vector [20], thereby accelerating the transition toward deep learning-based approaches. This shift has been particularly impactful in the context of TTS-driven data augmentation, where high-quality speaker embeddings enable more realistic and diverse synthetic speech. Following this trend, numerous deep learning-based speaker embedding models, such as Deep Speaker [21] by Li et al. and RawNet [22] by Jung et al., have been proposed, further improving robustness.

To overcome these issues, this paper proposes a data augmentation framework for super-elderly ASR that uses a dedicated TTS model—trained on a small set of real super-elderly speech—to generate new synthetic utterances. While existing approaches for elderly speech recognition have explored GAN-based adversarial methods for both dysarthric and elderly speech [23] and self-supervised learning model integration techniques [24], TTS-based data augmentation specifically tailored for elderly ASR remains unexplored. Although TTS-based augmentation has shown significant promise in dysarthric speech recognition [25], its specific application to super-elderly speech recognition represents a focused research opportunity in the field. Our synthetic data reflects unique elderly acoustic characteristics, including slower speaking rates, reduced high-frequency energy, and natural prosodic variations specific to super-elderly speakers.

Our key contributions are:

- Proposing a method that trains a high-quality TTS model (Style-Bert-VITS2) using only a small set of super-elderly speech, thus eliminating the need for large-scale elderly corpora.
- Demonstrating that an ASR system (Whisper) trained on a combination of real and TTS-generated super-elderly speech achieves improved recognition accuracy over a baseline trained only on the existing EARS corpus.

- Providing an in-depth analysis of how well synthesized speech preserves elderly acoustic characteristics and how it impacts ASR performance when added as training data.

II. Proposed Method

A. Speech Synthesis via VITS and Its Extensions

Our approach relies on an advanced TTS architecture that synthesizes elderly-like speech from any given text. The core model is VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [26], which directly converts text into waveforms. VITS uses a variational autoencoder to create latent representations of audio and a generative adversarial network to produce high-fidelity signals. It also employs normalizing flow models combined with a monotonic alignment method to align text tokens with audio frames. An enhanced version known as VITS2 [27] offers further improvements by replacing the original duration predictor with a more efficient, GAN-based approach, incorporating Transformer blocks to capture longer-range dependencies, embedding speaker information in the text encoder for multi-speaker scenarios, and refining alignment through Gaussian noise injection during early training. Bert-VITS2 [28] introduces a BERT-style text encoder for richer lexical and semantic features, while Style-Bert-VITS2 [2] adds continuous style control and speaker identification embedding to provide more flexible tone and prosody control. In this study, Style-Bert-VITS2 is selected because it consistently produces realistic Japanese speech, including expressive voices that match an elderly style. The Style-Bert-VITS2 model architecture is shown in Figure 1.

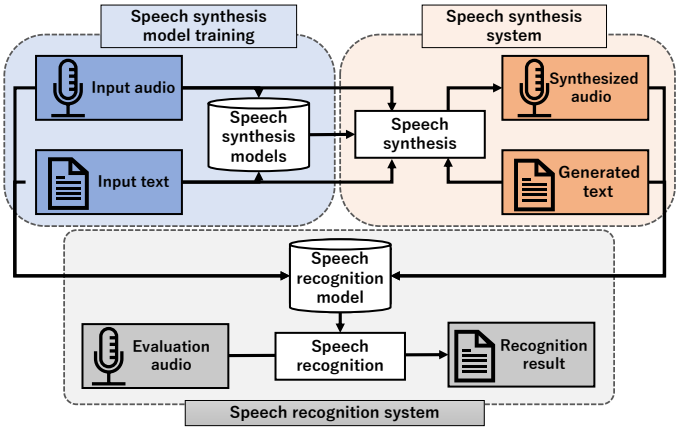


Fig. 2: Overall pipeline for improving super-elderly ASR using TTS-based data augmentation

B. System Architecture for Super-Elderly ASR

Figure 2 provides an overview of our pipeline for improving super-elderly ASR. We begin by preparing a small dataset of real super-elderly utterances, typically from speakers aged 70–99, to train Style-Bert-VITS2. Although the corpus is relatively small, the model learns notable properties such as lower pitch and slower articulation. Once trained, the TTS system can generate substantial amounts of synthetic speech using textual prompts, for instance from news sources, while applying an elderly reference style. This synthetic data exhibits reduced high-frequency energy and other age-related vocal traits. We then merge the newly created utterances with existing super-elderly recordings, forming an augmented corpus of real and synthetic speech. The resulting corpus is used to fine-tune OpenAI's Whisper [29], a Transformer-based encoder-decoder model that benefits from extensive multilingual pretraining. By exposing Whisper to a broader range of elderly-like speech patterns, we improve recognition accuracy on real super-elderly utterances.

1) TTS Model Training: To prepare the Style-Bert-VITS2 model for elderly voice generation, we collect a small set of super-elderly audio segments from the EARS corpus, excluding any data intended for final testing. Each audio sample is paired with its corresponding transcription. During training, the model maps phonetic sequences to time-aligned segments in the waveform, characterizing pitch, timbre, and prosody specific to older speakers, while adversarial mechanisms help achieve realistic audio quality. As training converges, the model becomes adept at reproducing core features of elderly speech, including reduced pitch range and subtle changes in harmonic structure.

2) Synthesis of Elderly Speech: Once the TTS model is trained, we generate a large number of synthetic utterances to supplement the real elderly data. These synthesized utterances are derived from various text sources, such as newspaper articles, with an elderly speaker embedding applied to maintain consistent style. As a result, the generated audio exhibits weaker high-frequency components commonly associated with advanced age. Barring cases of severe artifacts or empty output, most TTS results are retained. This step multiplies the available training data without requiring extensive new recordings from older speakers.

3) Whisper Fine Tuning: The final stage integrates the synthetic and real super-elderly data into a unified corpus for fine-tuning Whisper [29]. The model, originally trained on massive multilingual data, is adapted to the super-elderly domain by updating all of its Transformer layers at a modest learning rate. Both real and synthetic audio are converted into log-mel spectrograms, and the corresponding transcripts serve as targets in a standard supervised setup. By observing numerous variations of elderly-like speech during training, Whisper learns to handle the distinctive acoustic and prosodic characteristics often missing in younger-oriented ASR data.

III. Experiments and Results

This section describes the overall experimental setup and presents the outcomes of incorporating synthesized super-elderly speech into the ASR pipeline. All evaluations were designed to compare the recognition accuracy of different training configurations and to confirm whether synthesized data can effectively compensate for the scarcity of real super-elderly recordings.

A. Experimental Conditions

The training data for our TTS model and ASR system is summarized in Table I. We used 140 super-elderly utterances (approximately 0.33 hours) from the EARS corpus, consisting of 14 speakers, ensuring no overlap with the final test set. After training a Style-Bert-VITS2 model, we synthesized an additional 3,584 elderly-like utterances (approximately 8.15 hours) using text prompts from the Mainichi Shimbun [30]. For comparison, we also use GPT-SoVITS TTS model to generates 3,584 utterances (approximately 7.41 hours). To train and evaluate Whisper, we combined either the baseline EARS (0.33 hours) or EARS plus one of the synthetic sets EARS(Style-Bert-VITS2 Synthesized) or EARS(GPT-SoVITS Synthesized)), and we reserved a separate 11.5-hour portion (5,472 utterances, consisting of 109 speakers) of EARS for testing. This test portion contained super-elderly speech that did not appear in any training set, allowing us to gauge out-of-sample performance.

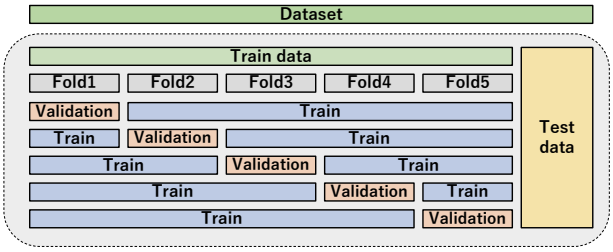


Fig. 3: Data partitioning strategy: hold-out validation and 5-fold cross-validation for training and testing

TABLE I: Training data for TTS and ASR models

Dataset	Hours	# Utter.	Age
EARS	0.33	140	70–99
EARS (Style-Bert-VITS2 Synth.)	8.15	3,584	70–99
EARS (GPT-SoVITS Synth.)	7.41	3,584	70–99
EARS (Test)	11.5	5,472	70–99

To configure Style-Bert-VITS2, we used the open-source repository settings with a batch size of 4, 100 training epochs, and a learning rate of 0.0001. We also supplied an elderly style reference, drawn from the small training subset of super-elderly speech. For the ASR experiments, we employed the `whisper-large-v3` model in the Hugging Face `transformers` library. All layers were unfrozen, and training used a batch size of 7, 100 epochs, the AdamW optimizer with a warmup period of 10% of total training steps., and a learning rate of 1×10^{-7}. However, early stopping was employed during training including TTS to prevent overfitting. We conducted five-fold cross-validation, alongside a hold-out approach, to rigorously evaluate the impact of data augmentation. Figure 3 illustrates how data was partitioned for validation and testing.

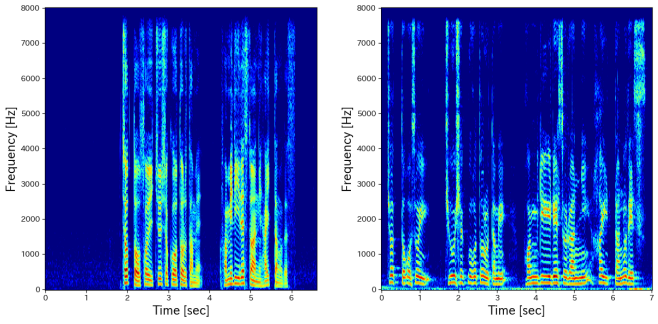
Recognition performance was assessed using the character error rate (CER), defined as

$$\mathrm{CER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

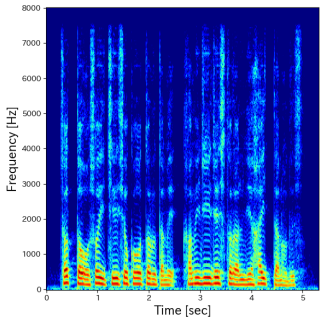
where S is substitution errors, D is deletions, I is insertions, C is correct characters, and N is the total number of characters in the reference text.

B. Speech Synthesis Analysis

We examined whether Style-Bert-VITS2 accurately captured the characteristics of super-elderly voices by comparing the spectral properties of synthesized audio with those of younger-people speech from the JNAS corpus [31], [32], which primarily contains recordings from speakers aged 20-50 years, and real super-elderly speech from EARS in Figure 4. The generated waveforms generally showed weaker mid-to-high frequency energy, resembling actual elderly speakers. However, there were occasional deviations such as faster articulation or fewer pauses, suggesting that the model does not fully replicate the slower and sometimes more disfluent patterns typical among older individuals. Despite these shortcomings, the synthetic data was sufficiently realistic to offer a notable benefit when used to augment the training of our ASR system.



(a) young people's speech (b) super-elderly speech



(c) synthesized elderly speech

Fig. 4: Spectral comparison of (a) young people's speech, (b) super-elderly speech, and (c) synthesized elderly speech

IV. ASR Results and Discussion

Table II presents the cross-validation CER results under three training configurations: using only the EARS corpus

TABLE II: Cross-Validation CER [%] results comparing baseline and augmented training approaches against EARS (test) dataset

Training Data	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**	**Mean**
EARS only (baseline)	10.05	10.06	9.85	10.15	10.03	10.03
EARS + EARS(GPT-SoVITS Synthesized)	9.89	9.75	9.64	9.66	9.68	9.72
EARS + EARS(Style-Bert-VITS2 Synthesized)	**9.71**	**9.79**	**9.68**	**9.66**	**9.75**	**9.72**

(baseline), augmenting EARS with 3,584 synthesized utterances from Style-Bert-VITS2 (EARS + EARS(Style-Bert-VITS2 Synthesized)), and augmenting EARS with 3,584 synthesized utterances from GPT-SoVITS (EARS + EARS(GPT-SoVITS Synthesized)). The baseline system achieves a mean CER of 10.03%. When Style-Bert-VITS2 data is added, the CER decreases to 9.72%, an absolute improvement of 0.31 percentage points. Improvements across individual folds range from 0.17 to 0.48, and statistical tests (two-sided t-test with Bonferroni correction, $p < 0.01$) indicate significance in three out of five folds, while the remaining two folds showed significance at the uncorrected level $p < 0.05$). These findings confirm that TTS-based data augmentation yields tangible benefits for super-elderly speech recognition.

By contrast, incorporating synthetic data from GPT-SoVITS shows comparable performance at 9.72%. This outcome suggests that both GPT-SoVITS and Style-Bert-VITS2 perform similarly in capturing elderly prosody, demonstrating that the choice of TTS model has negligible impact on the target speaker profile representation.

Overall, these experiments show that producing large-scale training utterances through TTS is advantageous for super-elderly ASR, provided that the synthetic speech matches real elderly characteristics. Although we successfully modeled important features such as reduced high-frequency energy and pitch range, fully recreating the slower, more disfluent speaking styles typical of super-elderly voices remains challenging. Further refinements, such as enhanced style control or the manual filtering of subpar synthetic audio, could improve future performance.

V. Conclusion and Future Work

This paper demonstrated a speech-synthesis-based data augmentation approach for super-elderly ASR. By using a small set of real elderly audio to train Style-Bert-VITS2 and generating large amounts of synthetic speech, we improved accuracy in a Whisper-based recognition system compared to training on limited real data alone. While synthesis captured many elderly speech traits, it did not fully mimic slower speech rates or hesitations due to decline in vocal cord function. Future work includes filtering low-quality TTS outputs, modeling disfluencies, and tailoring gender-specific models so the TTS model can generate more realistic elderly speech audio for training ASR models.

In future work, several extensions and refinements can be pursued. First, some TTS outputs may be poor quality or insufficiently elderly-like; an automatic or semi-automatic filtering method could remove unnatural samples and further boost ASR results. Second, using gender-specific TTS models might capture more nuanced voice characteristics, given potential differences in pitch, timbre, and speaking style. Third, incorporating slower, more disfluent patterns into the TTS pipeline (e.g., modeling hesitations or filler words) could yield even more realistic training data.

Acknowledgment

This research was supported by the JSPS Grant-in-Aid JP23H00493 and JP25H00566.

References

[1] R. Vipperla, S. Renals, and J. Frankel, "Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, p. 525 783, 2010.

[2] Litagin02, *Style-Bert-VITS2: Bert-VITS2 with more controllable voice styles*, https://github.com/litagin02/Style-Bert-VITS2, 2024.

[3] L. Chen and M. Asgari, "Refining Automatic Speech Recognition System for Older Adults," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7003–7007.

[4] S. Suhasini and B. Bharathi, "ASR TAMIL SSN@ LT-EDI-2024: Automatic Speech Recognition System for Elderly People," in *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, Association for Computational Linguistics, Mar. 2024, pp. 294–298.

[5] N. Kitaoka, *Elderly Adults Read Speech Corpus (EARS)*, Speech Resources Consortium, National Institute of Informatics. (dataset), 2023. https://doi.org/10.32130/src.EARS.

[6] M. Fukuda, R. Nishimura, H. Nishizaki, K. Yamamoto, and N. Kitaoka, "A new speech corpus of super-elderly Japanese for acoustic modeling," *Computer Speech & Language*, vol. 77, p. 101 424, 2023.

[7] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6820–6824.

[8] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams," in *Interspeech*, 2019, pp. 2843–2847.

[9] A. M. Alhumud, A.-Q. Muhammad, Y. O. Alomar, A. Alzahrani, and R. Souissi, "Improving automated speech recognition using retrieval-based voice conversion," in *The second tiny papers track at ICLR 2024*, 2024.

[10] C.-T. Do, S. Imai, R. Doddipatla, and T. Hain, "Improving Accented Speech Recognition Using Data Augmentation Based on Unsupervised Text-to-Speech Synthesis," in *2024 32nd European Signal Processing Conference*, 2024, pp. 136–140.

[11] Y. Matsuzaka, R. Takashima, C. Sasaki, and T. Takiguchi, "Data Augmentation for Dysarthric Speech Recognition Based on Text-to-Speech Synthesis," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies*, 2022, pp. 399–400.

[12] A. Tjandra, S. Sakti, and S. Nakamura, "Machine Speech Chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020.

[13] S. Ueno and T. Kawahara, "Phone-Informed Refinement of Synthesized Mel Spectrogram for Data Augmentation in Speech Recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8572–8576.

[14] X. Chen, Q. Huang, X. Wu, Z. Wu, and H. Meng, "HILvoice:Human-in-the-Loop Style Selection for Elder-Facing Speech Synthesis," in *2022 13th International Symposium on Chinese Spoken Language Processing*, 2022, pp. 86–90.

[15] W. Ping, K. Peng, A. Gibiansky, *et al.*, "Deep Voice 3: 2000-Speaker Neural Text-to-Speech," in *International Conference on Learning Representations*, 2018.

[16] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech 2017*, 2017, pp. 4006–4010.

[17] Y. Ren, Y. Ruan, X. Tan, *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[18] S. Kim, K. Shih, r. badlani rohan, *et al.*, "P-Flow: A Fast and Data-Efficient Zero-Shot TTS through Speech Prompting," in *Advances in Neural Information Processing Systems*, vol. 36, Curran Associates, Inc., 2023, pp. 74 213–74 228.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[20] Dehak, Najim and Kenny, Patrick J. and Dehak, Réda and Dumouchel, Pierre and Ouellet, Pierre, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[21] C. Li, X. Ma, B. Jiang, *et al.*, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[22] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification," in *Interspeech 2019*, 2019, pp. 1268–1272.

[23] Z. Jin, M. Geng, J. Deng, *et al.*, "Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 413–429, 2024.

[24] S. Hu, X. Xie, M. Geng, *et al.*, "Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 3561–3575, 2024.

[25] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," in *Interspeech*, 2024.

[26] 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*, ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 5530–5540.

[27] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design," in *Interspeech 2023*, 2023, pp. 4374–4378.

[28] FishAudio, *Bert-vits2*, https://github.com/fishaudio/Bert-VITS2, 2024.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 28 492–28 518.

[30] Mainichi Shimbun, *Mainichi shimbun CD-ROM edition*, CD-ROM database, Oct. 1995. [Online]. Available: https://www.nichigai.co.jp/cgi-bin/nga_search.cgi?KIND=EBCD1&ID=A8039.

[31] *The Acoustical Society of Japan (2006): ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS)*, Speech Resources Consortium, National Institute of Informatics. (dataset). https://doi.org/10.32130/src.JNAS.

[32] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.