A Study of Japanese Mixed Emotional Speech Synthesis Based on an End-to-End Emotional Speech Synthesis Model

Issei Sakata* and Tetsuo Kosaka*
* Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: t242327m@st.yamagata-u.ac.jp Tel: +81-238-263368

***Abstract*—In this study, we examined mixed emotion speech synthesis for Japanese using Emotional-VITS, an emotional speech synthesis model based on VITS, which is an end-to-end speech synthesis model. On the dataset of pretrained models, we compared models trained in two languages (Japanese and Chinese) and models trained only in Japanese. Regarding the quality of the synthesized speech, we confirmed that the model trained only in Japanese has higher naturalness than the model trained in Japanese and Chinese in both subjective and objective evaluation experiments, and that the input text is accurately synthesized. We also examined the emotional features used in speech synthesis and confirmed that the average of the emotional features of other speakers, other than the speaker to be synthesized, yields a higher emotion recognition rate than the average of the emotional features extracted from the target speech or the average of the emotional features of the speech of the speaker to be synthesized. Furthermore, it was found that even in mixed emotion speech synthesis, in which two types of emotions are mixed, the emotion recognized by the subject changes consistently depending on the mixing ratio, and it was found that emotions close to human sensation are possible.**

I. Introduction

Deep learning-based text-to-speech (TTS) technology has traditionally used a two-step process, whereby text is converted into an intermediate representation, such as a mel-spectrogram, which is then synthesized into speech[1]. However, in this two-stage process, the second-stage model is trained using the intermediate representations generated in the first stage, making the performance of the second-stage model dependent on that of the first stage model. This makes it difficult to optimize both models simultaneously, thereby hindering the overall improvement in speech synthesis quality. To address this issue, end-to-end speech synthesis models have been proposed in recent years. This approach enables the direct synthesis of speech from text without an intermediate representation, achieving speech synthesis that is closer to human emotional expression. In particular, VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech)[2] uses Variational AutoEncoder (VAE)[3] to connect two-stage modules via hidden variables, enabling efficient end-to-end learning and achieving high-quality speech synthesis. With this development, research is being conducted on emotional speech synthesis that mimics human emotional expression.

Plutchik's wheel of emotions[4], proposed by the American psychologist Plutchik, is a psychological theory regarding human emotions. This theory posits that there are eight basic human emotions: anger, disgust, fear, joy, sadness, surprise, anticipation, and trust. Other emotions are considered to be either mixed or derived from these basic emotions. In particular, emotions that arise from a combination of basic emotions are referred to as mixed emotions. For example, a mixture of joy and surprise produces a mixed emotion of excitement, while a mixture of sadness and surprise produces a mixed emotion of rejection. By incorporating Plutchik's theory of the wheel of emotions, particularly the idea that mixing two types of emotions produces a different emotion, into a speech synthesis model, speech synthesis that closely resembles human emotional expression is possible.

As TTS backbones in mixed emotion speech synthesis research, encoder-decoder-based models[5] and diffusion-based models[6]–[8] have been proposed. However, mixed emotion speech synthesis using an end-to-end model based on VITS has not yet been proposed. Furthermore, no mixed emotion speech synthesis models using deep learning have been proposed for Japanese to date.

Therefore, in this study, we examine mixed emotion speech synthesis targeting Japanese using a VITS-based emotion speech synthesis model. Furthermore, we verify the emotional control of speech synthesis using the emotional characteristics of speakers different from the synthesis target, as well as whether control is possible according to the mixture ratio of emotions.

II. Mixed Emotion Speech Synthesis Using Emotional-VITS

A. VITS

VITS uses VAE to generate speech directly from text using hidden variables, thereby realizing end-to-end speech synthesis. Additionally, by incorporating a normalization flow[9] and adversarial learning, high-quality speech is generated. Furthermore, to address the "one-to-many problem" in which multiple speech variations (such as pitch and rhythm) exist for a given text input, it introduces a probabilistic duration predictor, enabling the synthesis of speech with diverse rhythms and intonations from the same text.

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

B. Speech Emotion Recognition Model using wav2vec2.0

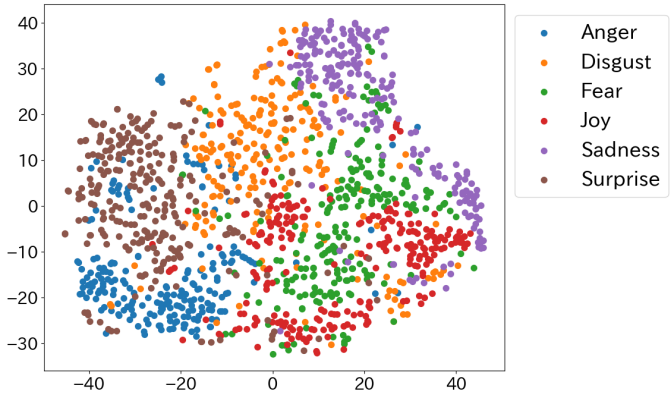

Fig. 1. Distribution of emotional features extracted from JVNV corpus speech using t-SNE analysis

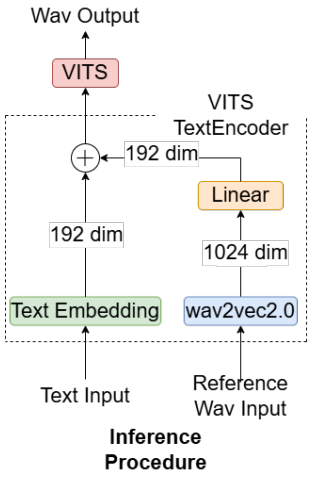

Fig. 2. Overall architecture of Emotional-VITS

wav2vec2.0[10] is a framework that uses transformers[11] for self-supervised learning using speech and is designed to capture speech representations from large amounts of unlabeled speech. In this study, we used w2v2-L-robust-12[12] as a speech emotion recognition model to extract emotional features from emotional speech. The model reduces wav2vec2-Large-Robust[13] to 12 transformer layers from its original size, pretrained on approximately 63,000 h of English speech (including telephone conversations), and fine-tunes it on the emotional speech corpus MSP-podcast[14] to extract 1,024-dimensional emotion features. To verify whether this model can appropriately capture emotion features in Japanese emotion speech, which differs from the language it was trained on, we performed emotional feature extraction on recordings from the JVNV Japanese emotional speech corpus[15]. To visualize the distribution of the extracted emotional features, we reduced them from 1024 dimensions to 2 dimensions using t-SNE analysis[16] and created the scatter plot shown in Fig. 1. From this figure, it can be seen that clusters are formed for each emotion and that w2v2-L-robust-12, which was trained on English emotion speech, appropriately captures emotion features in Japanese emotion speech.

C. Emotional-VITS

In this study, we used Emotional-VITS[17] as the VITS-based emotional speech synthesis model. Emotional-VITS is an emotional speech synthesis model that integrates the end-to-end TTS model VITS and the speech emotion recognition model w2v2-L-robust-12. The architecture used during inference is shown in Fig. 2. In Emotional-VITS, only the Text Encoder of VITS was modified. Specifically, the 192-dimensional text features extracted from the text are added to the emotion features extracted from the speech by w2v2-L-robust-12 and reduced to 192 dimensions by a linear layer, which is used as the output of the Text Encoder. During training, the text, emotion features extracted from the corresponding speech, and linear spectrograms are used. In synthesis, the text and emotion features extracted from the reference speech are used as input to generate speech that expresses the emotion. When synthesizing basic emotion speech, emotion features extracted from speech expressing the target emotion are used. When synthesizing mixed emotion speech, emotion features extracted from speech expressing the two emotions to be mixed are linearly combined, and the resulting features are used.

III. Evaluation Experiments

A. Experimental Setup

In these experiments, we compared two types of pretrained models. One is trained on Japanese and Chinese speech data (hereinafter referred to as the Japanese+Chinese model), and the other is trained on Japanese speech data only (hereinafter referred to as the Japanese model). The Japanese+Chinese model is released on Hugging Face and is readily available[18]. Although the details of the training data for this model are unknown, it has a large number of 804 trained speakers and is expected to have high versatility in synthesizing emotional speech of unspecified speakers. However, because this study focuses on Japanese, the Japanese+Chinese model may not fully capture the phonetic characteristics that are unique to Japanese. Therefore, we constructed a Japanese model trained solely on Japanese speech using the JVS corpus[19]. The JVS corpus consists of recorded speech from 100 male and female native Japanese speakers. In these experiments, we used approximately 23.7 h of data for training and approximately 2.6 h for verification. However, because this Japanese model has fewer speakers (100) than the Japanese+Chinese model and is expected to have less total training data, it is necessary to consider whether it will lead to improved performance. Both pre-trained models were trained with a sampling frequency of 22.05 kHz, a batch size of 64, and a learning rate of 2×10^{-4}. In addition, while the number of training steps for the Japanese+Chinese model is unknown, the Japanese model was trained for 400,000 steps.

TABLE I
RESULTS OF MOS, UTMOSV2, MCD AND CER EVALUATION EXPERIMENTS

Configuration		MOS(↑)	UTMOSv2(↑)	MCD[dB](↓)	CER[%](↓)
Ground Truth		4.909	2.514	-	-
Japanese + Chinese model	Target Speech	2.485	1.298	8.071	11.38
	Speaker's Ave.	2.523	1.419	7.889	10.29
	Other's Ave.	2.644	1.313	8.259	11.55
Japanese model	Target Speech	3.000	2.068	7.805	8.078
	Speaker's Ave.	3.083	**2.397**	**7.732**	7.176
	Other's Ave.	**3.295**	2.017	7.884	**6.771**

In addition, we used the JVNV corpus as emotional speech data to fine-tune both pretrained models. The JVNV corpus is a Japanese emotional speech corpus that includes both linguistic and non-linguistic speech, such as interjections. Each utterance text was automatically generated using ChatGPT and exhibited high emotional recognition performance. This corpus contains six emotions–anger, disgust, fear, joy, sadness, and surprise–expressed in speech by four speakers (two men and two women), covering six of the eight basic emotions in Plutchik's wheel of emotions, excluding expectation and trust. In these experiments, approximately 2.73 h of data were used as training data, approximately 0.3 h as validation data, and approximately 0.3 h as test data. Training was conducted for 40,000 steps with a sampling frequency of 22.05 kHz, batch size of 32, and learning rate of 2×10^{-4}.

Furthermore, we used three types of emotion vectors: target speech vector, speaker's average vector, and other person's average vector. The target speech vector is extracted from the emotional features of the emotional speech corresponding to the text content to be synthesized, and is characterized by a high degree of consistency with the text content. The speaker's average vector is computed by averaging the emotional vectors from all utterances expressing the target emotion for the speaker being synthesized. This representation is designed to be invariant to inter-speaker differences. The other person's average vector is obtained by averaging the emotional vectors from all utterances of the target emotion across all speakers excluding the one being synthesized. While it has low consistency with the text content being synthesized, it benefits from a large amount of reference data for computing the emotional representation.

B. Evaluation Experiments on the Quality of Synthesized Speech

We conducted subjective evaluation experiments using the Mean Opinion Score (MOS) and objective evaluation experiments using UTMOSv2[20], Mel Cepstral Distortion (MCD), and Character Error Rate (CER) to evaluate the quality of synthetic speech produced by the Japanese+Chinese and Japanese models. For the evaluation speech, we selected one male and one female speaker from the JVNV corpus and used six emotions recorded in the JVNV corpus with three types of emotion vectors to generate synthetic speech. In the subjective evaluation experiments on the naturalness of synthetic speech using MOS, 11 native Japanese speakers evaluated the naturalness of the synthetic speech on a 5-point scale, with 1 being the lowest and 5 being the highest. UTMOSv2 is a speech quality evaluation method that uses a model to predict the MOS of synthesized speech on a 5-point scale from 1 to 5, with 1 being the lowest and 5 being the highest score. MCD calculates the Mel-Cepstral distance between recorded speech and synthesized speech. A smaller value indicates that the synthesized speech has a greater acoustic similarity to the recorded speech. CER calculates the character-by-character error rate between the input text of the speech synthesis model and the transcription results of the speech recognition model for synthesized speech. A smaller value indicates that the synthesized speech accurately reproduces the input texts. In the CER experiment, we used the large-v3 Whisper model[21] as the speech recognition model. The results are shown in Table I.

From these results, we found that the Japanese model had higher naturalness than the Japanese+Chinese model in all emotion vectors, based on the MOS and UTMOSv2. In addition, a t-test conducted at the 5% level showed a significant difference between the Japanese+Chinese and Japanese models. In the comparison of each emotion vector, we found that the speech synthesized using the other person's average vector obtained the highest score in the MOS, whereas the speech synthesized using the speaker's average vector obtained the highest score in the UTMOSv2. The difference between the results of MOS and UTMOSv2 is thought to be due to the fact that UTMOSv2 was insufficiently trained with emotion speech data, and there are still issues with the evaluation accuracy of emotion speech. The training datasets for UTMOSv2 are BVCC[22], BC2008[23], BC2009[24], BC2010[25], BC2011[26], SOMOS[27], and sarulab-data [28]. These datasets mainly consist of text-to-speech synthesis and voice conversion synthesis, and do not include emotional speech data. In these experiments, we placed emphasis on the MOS results, which are a reliable indicator evaluated directly by humans, and determined that the synthesized speech using the other person's average vector was more natural. Next, regarding MCD, the Japanese model obtained lower values than the Japanese+Chinese model for all emotion vectors, indicating that the Japanese model was able to synthesize speech with an acoustic similarity closer to the recorded speech. Finally, regarding CER, the Japanese model obtained lower values than the Japanese+Chinese model for all emotion vectors, indicating that the Japanese model was more accurate in synthesizing the input text. This is because the pre-trained model of the

TABLE II
RESULTS OF SPEECH EMOTION RECOGNITION EXPERIMENTS USING TEXT OF JVNV-V CORPUS [%]

	GT	Target Speech	Speaker's Ave.	Other's Ave.
Anger	96.3	85.2	77.8	74.1
Disgust	63.0	70.4	59.3	77.8
Fear	81.5	66.7	66.7	77.8
Joy	92.6	66.7	92.6	81.5
Sadness	92.6	74.1	63.0	77.8
Surprise	88.9	88.9	96.3	77.8
Overall	85.8	75.3	75.9	**77.8**

TABLE III
RESULTS OF SPEECH EMOTION RECOGNITION EXPERIMENTS USING PLAIN TEXT [%]

	Speaker's Ave.	Other's Ave.
Anger	33.3	29.2
Disgust	25.0	37.5
Fear	37.5	25.0
Joy	62.5	**66.7**
Sadness	45.8	**50.0**
Surprise	29.2	**45.8**
Overall	38.9	**42.4**

Japanese model has higher text-to-speech alignment accuracy than the pre-trained model of the Japanese+Chinese model.

C. Evaluation Experiments on Synthesized Speech of Basic Emotions

Before evaluating mixed emotion speech synthesis, we first conducted an evaluation of basic emotion speech synthesis. To evaluate how accurately the Japanese model can express emotions in synthetic speech, we conducted subjective emotion recognition experiments using synthetic speech corresponding to the six basic emotions defined in the JVNV corpus. When selecting the text, the emotion recognition rate of the recorded speech in the JVNV corpus was 94.21% with interjections and 87.37% without interjections, and it has been reported that the presence of interjections increases the emotion recognition rate[15]. Therefore, in this study, we used text from the JVNV-V corpus, which does not include interjections, to eliminate this influence. This evaluation experiment was conducted with 14 participants.

However, in this evaluation experiment, we found that the subjects tended to predict emotions based solely on the text content. Therefore, to evaluate the conditions that eliminate the influence of text content, additional evaluation experiments were conducted using text (plain text) generated by ChatGPT, from which it is difficult to predict emotions. Additional experiments were conducted with 6 participants. Additionally, because there is no ground truth speech for text-to-speech synthesis using plain text, we didn't use the target speech vector.

In both experiments, the participants were asked to select which of the six basic emotions they perceived in the synthesized speech presented to them. The results of each experiment are shown in Table II and III, respectively.

These results indicate that synthetic speech using the other person's average vector achieved higher emotion recognition rates than synthetic speech using the speaker's average vector, regardless of the text used. The reason why synthetic speech using the other person's average vector achieved higher emotion recognition rates is thought to be because the other person's average vector uses a larger amount of data for calculation than the speaker's average vector, enabling a more accurate expression of emotional characteristics. Furthermore, when using text from which emotions are difficult to predict, the emotion recognition rate was approximately 35% lower compared to text with more predictable emotional content. This suggests that the semantic content of the text significantly affects the accuracy of participants' emotion recognition in synthetic speech.

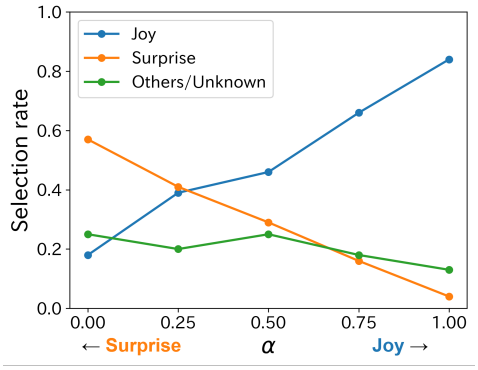

Fig. 3. Results of emotion selection ratio for mixed emotions (joy and Surprise)

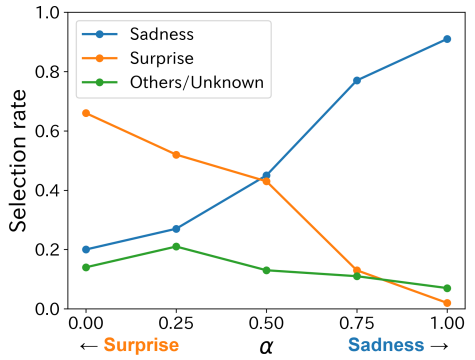

Fig. 4. Results of emotion selection ratio for mixed emotions (sadness and Surprise)

D. Evaluation Experiments on Synthesized Speech of Mixed Emotions

We conducted the subjective evaluation experiments on emotion recognition for synthetic speech with mixed emotions using a Japanese model. In these experiments, we evaluated whether emotions were correctly recognized according to the mixing ratio by varying the mixing ratio of two types of emotions and mixing them into the synthetic speech. The emotions evaluated were joy, sadness, and surprise, which were particularly high in emotion recognition rates in the basic emotion recognition experiments using plain text in

Section III-C, using the other person's average vector. We used two types of mixtures: joy and surprise, and sadness and surprise. We used a linear combination method that assumed linear emotional intensity as a method for mixing emotions. For example, the mixture of emotion A and emotion B was calculated using the mixture ratio α, which was defined as $\alpha\times$ emotion A vector $+ (1-\alpha)\times$ emotion B vector. The mixture ratio α was set to five values in the range from 0-1: 0, 0.25, 0.5, 0.75, and 1. 14 subjects were asked to select the emotion they heard from the mixed emotion synthesis speech from three options: emotion A, emotion B, and others/unknown. The results are shown in Fig. 3 and 4.

From these results, we found that changing the ratio of emotions consistently changed the emotion recognition rate and that emotion control was performed correctly in mixed emotion synthetic speech combining joy and surprise and sadness and surprise. The selection ratio of emotions tended to reverse at $\alpha = 0.25$ for the mixture of joy and surprise, and at $\alpha = 0.50$ for the mixture of sadness and surprise. In general, when two emotions are mixed at $\alpha = 0.50$, it is assumed that each is selected in equal proportions. However, in the case of the mixture of joy and surprise, there was a bias toward joy.

IV. Conclusion

In this study, we examined mixed emotion speech synthesis targeting Japanese using Emotional-VITS, an emotion speech synthesis model based on the end-to-end speech synthesis model VITS. We verified the emotional control of speech synthesis using the emotional characteristics of speakers different from the synthesis target and whether control is possible according to the mixed emotion ratio. Regarding the quality of the synthesized speech, we conducted subjective and objective evaluation experiments and confirmed that the Japanese model had higher naturalness than the Japanese+Chinese model in pre-trained models, accurately synthesized the input text, and achieved higher speech synthesis quality using a model trained with Japanese-only speech data. Additionally, for mixed emotion synthesis speech combining joy and surprise or sadness and surprise, we confirmed that the emotions recognized by the participants consistently changed according to the mixing ratio and that emotion control could be properly performed with these mixed emotion synthesis speech. Furthermore, it was shown that using a vector averaged from the emotional characteristics of speakers other than the synthesized speaker yielded the highest emotion recognition rate, and it was confirmed that using a vector averaged from a larger number of emotion vectors yielded a higher emotion recognition rate. In the future, we will aim to achieve speech synthesis that is more closely aligned with human emotional expression by adopting TTS that can control pitch and energy, which are important elements of emotional speech. In addition, when two types of emotions are mixed according to Plutchik's wheel of emotions, we examine whether the subjects can recognize the mixed emotions from the synthesized speech.

Acknowledgment

We thank Mr. Tianyi LI (currently at the Institute of Science Tokyo) for his help with the Japanese+Chinese model experiments.

References

[1] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.

[3] D. P. Kingma, M. Welling, *et al.*, *Auto-encoding variational bayes*, 2013.

[4] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*, Elsevier, 1980, pp. 3–33.

[5] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2022.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Emomix: Emotion mixing via diffusion models for emotional speech synthesis," *arXiv preprint arXiv:2306.00648*, 2023.

[8] R. Chevi and A. F. Aji, "Daisy-TTS: Simulating wider spectrum of emotions via prosody embedding decomposition," *arXiv preprint arXiv:2402.14523*, 2024.

[9] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*, PMLR, 2015, pp. 1530–1538.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[13] W.-N. Hsu, A. Sriram, A. Baevski, *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[14] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[15] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "JVNV: A corpus of Japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

[16] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17] *Innnky/emotional-vits*, [Online; accessed 2025-05-21]. [Online]. Available: https : / / github . com / innnky / emotional-vits.

[18] *Vits-uma-genshin-honkai - a Hugging Face Space by zomehwh*, [Online; accessed 2025-05-21]. [Online]. Available: https://huggingface.co/spaces/zomehwh/vits-uma-genshin-honkai.

[19] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[20] 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 *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 818–824.

[21] 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*, PMLR, 2023, pp. 28 492–28 518.

[22] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" *arXiv preprint arXiv:2105.02373*, 2021.

[23] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *The Blizzard Challenge 2008*, 2008, pp. 1–18. DOI: 10.21437/Blizzard.2008-1.

[24] S. King and V. Karaiskos, "The Blizzard Challenge 2009," 2009.

[25] A. W. Black, S. King, and K. Tokuda, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010.

[26] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, vol. 2011, 2011, pp. 1–10.

[27] G. Maniati, A. Vioni, N. Ellinas, *et al.*, "SOMOS: The Samsung open MOS dataset for the evaluation of neural text-to-speech synthesis," *arXiv preprint arXiv:2204.03040*, 2022.

[28] *sarulab-speech/VMC2024-sarulab-data*, [Online; accessed 2025-08-26]. [Online]. Available: https://github.com/sarulab-speech/VMC2024-sarulab-data.