MultiVerse: Efficient and Expressive Zero-Shot Multi-Task Text-to-Speech

Taejun Bak[1*], **Youngsik Eom**[2], **Seungjae Choi**[2], **Young-Sun Joo**[2],
[1]SK Telecom, Republic of Korea
[2]Audio AI Lab., NC Research, NCSOFT Corp., Republic of Korea
taejun.bak@sk.com
{yseom, seung, ysjoo555}@ncsoft.com

Abstract

Text-to-speech (TTS) systems that scale up the amount of training data have achieved significant improvements in zero-shot speech synthesis. However, these systems have certain limitations: they require a large amount of training data, which increases costs, and often overlook prosody similarity. To address these issues, we propose MultiVerse, a zero-shot multi-task TTS system capable of performing TTS and speech style transfer in zero-shot and cross-lingual conditions, while requiring much less training data than traditional data-driven approaches. To ensure zero-shot performance even with limited data, we leverage source-filter theory-based disentanglement, utilizing the prompt for modeling filter-related and source-related representations. Additionally, to further enhance prosody similarity, we adopt a prosody modeling approach combining prompt-based autoregressive and non-autoregressive methods. Evaluations demonstrate the remarkable zero-shot multi-task TTS performance of MultiVerse and show that MultiVerse not only achieves zero-shot TTS performance comparable to data-driven TTS systems with much less data, but also significantly outperforms other zero-shot TTS systems trained with the same small amount of data. In particular, our innovative prosody modeling technique enables Multiverse to generate speech with a high degree of prosody similarity to the given prompts.

1 Introduction

Deep learning-based text-to-speech (TTS) has advanced to the point of synthesizing human-like speech (Wang et al., 2017; Ren et al., 2019). However, recent research has been extended beyond this scope, exploring various ways of broadening the application of speech synthesis models. Representative tasks include synthesizing unseen speaker's speech, known as **zero-shot TTS** (Jia et al., 2018; Cooper et al., 2020), generating speech in a language that the monolingual target speaker has not seen, referred to as **cross-lingual TTS** (Cho et al., 2022; Zhang et al., 2019; Xin et al., 2021), and transferring the prosody of a speech reference to a target speaker, known as **speech style transfer** (Lee et al., 2021; Huang et al., 2022). Furthermore, recent TTS research has integrated the zero-shot task with cross-lingual or style transfer tasks (Casanova et al., 2022; Zaïdi et al., 2022).

To expand TTS applications in zero-shot conditions, it is crucial to ensure generalization across various speech components, such as content, style, and speaker identity. Disentangled representations facilitate this by enabling the system to capture interpretable and controllable features, thereby improving generalization of TTS systems through the learning of these individual components (Bengio et al., 2013; Lipton, 2018). However, due to the often entangled nature of these components, separately learning their general characteristics remains challenging.

Data-driven methods are a representative approach to learning generalized acoustic components from large-scale speech datasets. Specifically, these methods scale up the TTS system to train on massive datasets, making them robust under unseen conditions (Wang et al., 2023; Zhang et al., 2023; Le et al., 2023). Additionally, Jiang et al. (2023, 2024); Shen et al. (2023); Ju et al. (2024) adopt disentangled modeling to separately learn acoustic components. Disentangled modeling contributes to ensure generalization in each component by independently encapsulating interpretable elements. However, a significant amount of training dataset is required as the decoder needs to learn the relationships between disassembled elements. Preparing large-scale data is especially challenging for minority languages. Moreover, even with large-scale training data, there is still potential room for improvement in prosody similarity in zero-shot sce-

[*]Work performed at NCSOFT.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9130–9147
November 12-16, 2024 ©2024 Association for Computational Linguistics

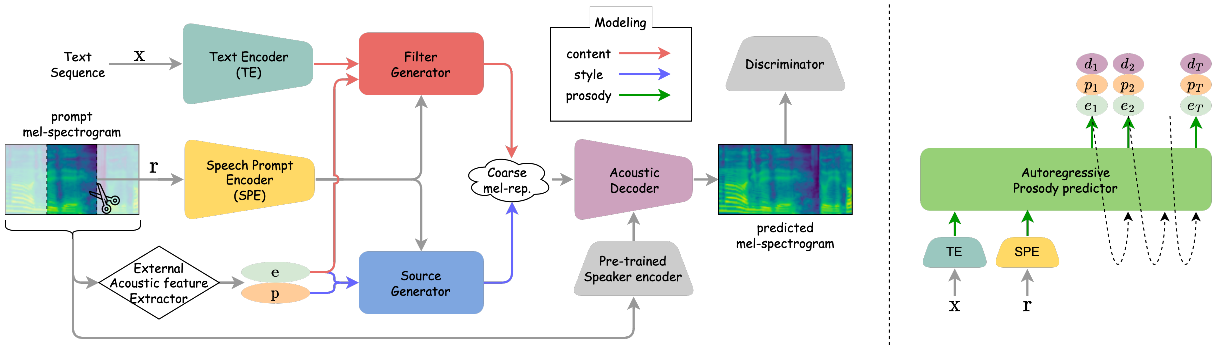


Figure 1: Overall structure of MultiVerse. The acoustic model and the autoregressive prosody predictor are on the left and right side of the figure, respectively. During training, overall modules are trained together, except the pre-trained speaker encoder. Multi-task TTS can be accomplished by varying input conditions.

narios.

In this paper, we introduce a multi-task TTS system, called MultiVerse, enabling speech synthesis and speech style transfer in zero-shot and cross-lingual conditions, requiring significantly less data compared to the data-driven approaches and featuring enhanced prosody modeling. MultiVerse enhances training efficiency by source-filter theory-based decomposed modeling (Fant, 1970). Specifically, MultiVerse decomposes speech generation into filter- and source-related representation generation, with prompt speech utilized in the modeling of each representation. Notably, both representations result in features that have a similar distribution to the mel-spectrogram (Bak et al., 2021). Therefore, it is suitable for the decoder to learn the interdependent relationship between the representations even with small data. Furthermore, MultiVerse introduces an effective style modeling method. While several zero-shot models have modeled either the acoustic features (Shen et al., 2023) or prosody latents to capture speech style (Jiang et al., 2023), MultiVerse utilize both autoregressive (AR) based acoustic feature modeling and Non-AR based prosody modeling.

We evaluate MultiVerse's performance in zero-shot TTS tasks under various conditions. Regarding language, we evaluate both intra- and cross-lingual speech synthesis and measure naturalness and similarity in neutral and expressive speech style. Additionally, we compare our model with a large-scale TTS model using data-driven methods and a speech style transfer model. Evaluation results demonstrate that MultiVerse has the following advantages: (1) Zero-shot intra- and cross-lingual TTS is achievable with a small amount of training data. MultiVerse can achieve similar zero-shot synthesis in both timbre and prosody with only $\frac{1}{60}$ of the training data compared to VALL-E (Wang et al., 2023). (2) The proposed prosody modeling is also effective in reflecting conditions in various tasks of speech synthesis. Even without requiring information about the content or phoneme duration of the prompt speech, it can reflect prosody similar to the prompt in both intra-lingual and cross-lingual settings.

2 MultiVerse

MultiVerse models speech by disassembling it into two components: filter and source. The proposed model comprises three main modules: (1) an acoustic model based on the source-filter theory (Fant, 1970) that generates mel-spectrograms given text and speech prompts, (2) an AR prosody predictor that predicts prosody-related acoustic features (duration, pitch, energy) from input conditions, and (3) a discriminator for adversarial training.

2.1 Source-Filter Theory Based Decomposed Modeling

Inspired by the source-filter theory, which explains the response between the vocal tract filter and the sound source, the proposed acoustic model generates mel-spectrogram through the filter and source generators. The filter generator is tasked with producing the vocal tract filter-related representation, while the source generator outputs the source-related representation. We name these two representations as filter representation and source representation, respectively. Both representations are obtained from feed forward transformer-based generator (Ren et al., 2019).

Filter representation We consider filter representation as hidden states that contain information related to speech content, pronunciation, and speaker identity, but is less dependent on prosody.

This representation, obtained by the filter generator, is modeled by taking phoneme representation as input, along with the energy embedding. Both input features are upsampled by the gaussian upsampling (Shen et al., 2020).

Source representation We consider source representation primarily as hidden states that contain prosodic information, such as intonation, rhythm, and stress patterns, with low dependence on content. The source generator produces source representation from frame-wise upsampled phoneme-level pitch and energy embeddings. During training, it generates the source representation from ground-truth acoustic feature embeddings, while during inference, it utilizes predicted acoustic feature embeddings. We provide further experimental analysis on both representations in Appendix B.

Choi et al. (2021); Bak et al. (2021) also utilized source-filter based speech disentanglement. Unlike existing methods, our approach adopts prompt-based modulation in modeling both representations. Since both vocal tract filter and sound source are influenced by speaker characteristics, the prompt speech is reflected in both generators. The mel-spectrogram of prompt speech serves as the input for obtaining hidden states from the speech prompt encoder. Parameters for the FiLM (Perez et al., 2018) layer are predicted from the cross-attention output between the input of generators and the output of speech prompt encoder as in (Shen et al., 2023). Subsequently, the FiLM layer modulates the representations within the generators.

2.2 Increasing Filter Capacity of the Acoustic Decoder

The intermediate representation formed by combining both representations, referred to as coarse mel-representation, resembles the interaction between the vocal tract filter and the sound source (Bak et al., 2021). Since this fusion follows the frequency response in the source-filter model, the coarse mel-representation is closely related to a high-dimensional features of speech. Consequently, the acoustic decoder has an advantage in learning the interdependent relationship between the filter and source representations.

To produce mel-spectrograms while preserving various types of information such as speech content and style within the coarse mel-representation, it is necessary to increase the filter capacity of the acoustic decoder. To increase the filter capacity, the acoustic decoder's transformer block replaces convolution layers with sample-adaptive kernel selection-based convolution layers (Kang et al., 2023). It aims to find suitable convolution filters for the speech prompt. Specifically, learnable filters of each convolution layer are weighted sums based on predicted weights from the global style embedding. The aggregated filter is then modulated and de-modulated (Karras et al., 2020). The global style embedding is derived from the pre-trained speaker encoder. More detail information about sample-adaptive kernel selection is described in Appendix C.

2.3 Two-stage Prosody Modeling

MultiVerse's prosody modeling consists of two stages: first, the prosody predictor models the acoustic features autoregressively; second, the source generator models prosody in the latent space non-autoregressively using these acoustic features.

2.3.1 Autoregressive Prosody Modeling

The proposed AR prosody predictor models the time-varying distribution of acoustic features (duration, pitch, energy) as a conditional language modeling task. The goal of the AR prosody predictor is prediction of acoustic units suitable for the given text and prompt conditions. Due to the time-dependent nature of prosody and the need to model large-variations in prosody, we adopt an AR approach (Kharitonov et al., 2022). The prosody predictor is trained to predict acoustic units $\mathbf{c}_t = \{\mathbf{d}_t, \mathbf{p}_t, \mathbf{e}_t\}$ corresponding to phoneme sequences $\mathbf{x} = \{x_1, x_2, ..., x_T\}$, where $\mathbf{d}$, $\mathbf{p}$, $\mathbf{e}$ are the duration, pitch, and energy unit sequences, respectively. These unit sequences of which each value corresponds to an index are obtained by quantizing normalized acoustic sequences. Prosody modeling, which is conditioned on the speech prompt $\mathbf{r}$ and the phoneme sequence $\mathbf{x}$, is written as the following equation:

$$p(\mathbf{c}|\mathbf{x}, \mathbf{r}; \theta_{ARP}) = \prod_{t=0}^{T} p(\mathbf{c}_t|\mathbf{c}_{<t}, \mathbf{x}, \mathbf{r}; \theta_{ARP}), \quad (1)$$

where θ_{ARP} represents the parameters of AR prosody predictor. To model the prosody using a prompt-based in-context learning, we utilize the phoneme sequence and the prompt as a prefix, a similar approach to Wang et al. (2023).

The AR approach is also utilized in the prosody latent language model (P-LLM) in Mega-TTS (Jiang et al., 2024), which autoregressively models

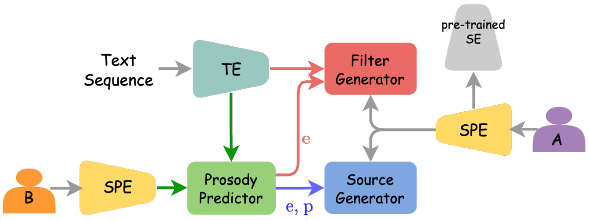


Figure 2: Style transfer process to transfer speech style from speaker B to speaker A. The acoustic decoder is omitted for simplicity.

vector-quantized codebook (Van Den Oord et al., 2017) of prosody hidden states. However, the performance of vector quantization depends on the quantity and diversity of training data (Gersho and Gray, 2012). In contrast, the AR prosody predictor, which models acoustic feature units, is data-efficient. We provide further analysis on the limitation of VQ based modeling in Appendix D. Additionally, the P-LLM is limited to specific prompt speech, such as the particular languages present in the training data, and requires alignment information of the prompt. Conversely, the AR predictor does not impose restrictions on the utilization of prompt speech.

2.3.2 Non-Autoregressive Prosody Modeling

Non-AR prosody modeling refines prosody at the frame-level from time-dependent prosody features. In this process, the source-filter generator converts acoustic feature embeddings into the source representation, reflecting the prosody characteristics of the prompt by the attention mechanism and the modulation.

2.4 Learning Objectives

The learning objectives consist of three components: reconstruction loss, adversarial loss, and acoustic feature loss. The reconstruction loss is the L1 loss between the generated and the ground-truth's mel-spectrogram. The adversarial loss utilizes LSGAN (Mao et al., 2017), incorporating a multi-window discriminator (Chen et al., 2020; Ye et al., 2022) with 2D patch unit lengths $\{32, 64, 128\}$. The acoustic feature loss computes the sum of cross-entropy losses for each acoustic feature, comparing the output units of the prosody predictor with the ground-truth acoustic units.

2.5 Multi-Task TTS

The proposed model can perform multiple tasks according to the input condition. First, zero-shot TTS takes an unseen speaker's prompt as input. Second, cross-lingual TTS is accomplished by using different languages for the speech prompt and input text. Additionally, the speech style transfer enters two different prompts into different modules, as illustrated in Figure 2. These three tasks can be combined with each other. For example, zero-shot cross-lingual TTS or zero-shot style transfer, and even all three tasks can be performed at once, namely zero-shot cross-lingual speech style transfer. Detailed inference process are provided in Appendix F.

3 Experimental Environments

3.1 Datasets

Training datasets Training datasets consist of English and Korean speech datasets. We used the open datasets LibriTTS (train-clean-100, train-clean-360) (Zen et al., 2019) and VCTK (Yamagishi et al., 2019) and an internal dataset as the English dataset. As the Korean dataset, we used the open dataset AI-Hub[1] with multi-style and an internal dataset. Specification for datasets are provided in Appendix E. We re-sampled all speech data to a 22.05 kHz sampling rate, and obtained an 80-band mel-spectrogram as the acoustic feature. The Short-Time Fourier Transform (STFT) parameters included a bin size of 1024, with window size of 1024 and a hop sizes 256.

Evaluation datasets We diversified the composition of the evaluation dataset. The dataset for evaluating zero-shot performance is made up of utterances from speakers not included in the training. Specifically, it is divided by language (English and Korean) and speaking style (neutral and expressive). For the English dataset, the neutral style is represented by LibriTTS dev-clean and the expressive style is represented by EXPRESSO (Nguyen et al., 2023). Four styles (confused, enunciated, happy, and sad) were selected from the various styles available in EXPRESSO. The Korean dataset includes the neutral style from the AI-Hub multi-speaker dataset and the expressive style from an internal dataset. The expressive internal dataset includes voices in emotional and theatrical styles.

3.2 Experimental Setup

Baselines To evaluate speech generation performance under the same training data conditions, we trained GANSpeech (Yang et al., 2021) and

[1] https://aihub.or.kr

YourTTS (Casanova et al., 2022) as baselines. GANSpeech is a Non-autoregressive transformer-based multi-speaker TTS model, which adopts adversarial training using speaker conditioned discriminator for securing generalization. We configured GANSpeech to facilitate zero-shot and cross-lingual tasks by using a pre-trained speaker encoder (Chung et al., 2020), referred to as GANSpeech+. YourTTS, a conditional variational autoencoder-based end-to-end model, incorporates a learning objective for consistent speaker modeling and handles both cross-lingual and zero-shot tasks. We utilized the official implementation of YourTTS[2], but did not use the official pre-trained model of YourTTS because it has not been trained with Korean speech data. For evaluation of speech style transfer performance, we conducted a comparison with Daft-exprt (Zaïdi et al., 2022), a speech style transfer model. Daft-exprt transfers style with extracted prosody information from the reference voice. We used the official implementation of Daft-exprt[3].

Model configuration The MultiVerse's encoders, generators, and decoder are all constructed with the transformer blocks. Except for the AR prosody predictor, all modules operate in a non-AR manner. The global style embedding is obtained from the output of an open-source pre-trained ResNet-based speaker recognition model (Chung et al., 2020). Detailed configurations for each module of MultiVerse and the corresponding hyper-parameters are described in Appendix G. Since the proposed model and the GANSpeech+ output mel-spectrogram, we utilized a pre-trained Avocodo (Bak et al., 2023) which is an universal neural vocoder. In the case of YourTTS, a neural vocoder was not employed because it directly generates waveforms.

Training To obtain the phoneme sequence, we performed grapheme-to-phoneme (G2P) processing on the audio transcripts. For English, we utilized the IPA-based open-source tool[4], and for Korean, an internal G2P tool was employed. The acoustic features included duration, fundamental frequency (F_0), and energy. Duration extraction employed the Montreal Forced Aligner 2.0 (McAuliffe et al., 2017) and the internal Forced Aligner for English and Korean speech, respectively. Praat toolkit[5] was used for F_0 extraction. Detailed acoustic feature pre-processing is described in the Appendix H. The training batch size was 48 and 8 NVIDIA A100 GPUs were used in training. The optimization employed the ADAMW optimizer (Loshchilov and Hutter, 2019), with the parameters (β_1, β_2) set as $(0.8, 0.99)$, and a NOAM learning rate scheduler (Vaswani et al., 2017). The training of the AR prosody predictor started at the peak learning rate to stabilize training. A random segment of the reference speech was selected as the prompt for each training iteration, and the reconstruction loss was not computed for that segment (Shen et al., 2023). To prevent the model from excessively mimicking the prosody of the prompt, a shorter segment was used for the prosody predictor. MultiVerse and GANSpeech+ were trained for 600k iterations, while YourTTS was trained for 500k iterations.

Metrics For objective evaluation, we measured speech intelligibility, speaker and prosody similarity between the prompt and the synthesized speech. The evaluation of speech intelligibility involves comparing the Character- and Word-error-rate (CER, WER) measured by automatic speech recognition (ASR) using the pre-trained Whisper[6] model (Radford et al., 2023). The speaker embedding cosine similarity (SECS) using open-source voice encoder[7] measures speaker similarity between the prompt and the synthesized speech. For prosody similarity evaluation, F_0 Pearson correlation coefficient (F_0 PCC) between F_0 of the prompt and synthesized speech is calculated. For evaluating style transfer performance, we measured the similarity in pitch and duration between the synthesized speech and the style prompt. Dynamic Time Warping was employed to measure the F_0 similarity (F_0 DTW). SECS was also used to evaluate similarity between the target speaker's audio sample and the style-transferred speech. Subjective evaluation included three Mean Opinion Score (MOS) tests: The Naturalness-MOS (N-MOS) test evaluates the naturalness and intelligibility of the speech. The Similarity-MOS (S-MOS) test assesses the speaker similarity and the Prosody Similarity-MOS (PS-MOS) test evaluates the prosody similarity between the synthesized and the prompt speech for expressive style speech only. A total of 10 native speakers evaluated 15 English audio samples, and

[2] https://github.com/Edresson/YourTTS

[3] https://github.com/ubisoft/ubisoft-laforge-daft-exprt

[4] https://github.com/rhasspy/gruut-ipa

[5] https://github.com/YannickJadoul/Parselmouth

[6] https://huggingface.co/openai/whisper-large

[7] https://github.com/resemble-ai/Resemblyzer

Table 1: The objective and subjective evaluation results in a zero-shot scenario using speech prompts with variations in style (Neutral (N) and Expressive (E)) and language (English (ENG) and Korean (KOR)).

Prompt Style	Prompt Language	Method	CER (↓)	WER (↓)	SECS (↑)	F_0 PCC (↑)	N-MOS	S-MOS	PS-MOS
		Intra-lingual							
N	ENG	Ground-truth	0.89	2.84	0.820	-	4.17±0.13	4.12±0.16	-
		GANSpeech+	**0.76**	**2.61**	0.736	0.037	3.83±0.15	3.72±0.18	-
		YourTTS	3.50	7.78	0.810	0.021	2.88±0.15	3.83±0.16	-
		MultiVerse	0.89	2.70	**0.852**	**0.073**	**3.87±0.14**	**4.42±0.12**	-
	KOR	Ground-truth	4.15	21.08	0.845	-	4.15±0.11	4.63±0.07	-
		GANSpeech+	**3.23**	**17.10**	0.740	0.069	**4.29±0.09**	3.19±0.12	-
		YourTTS	6.56	26.16	0.790	0.037	3.69±0.10	3.72±0.12	-
		MultiVerse	3.76	18.94	**0.834**	**0.147**	3.91±0.10	**4.12±0.11**	-
E	ENG	Ground-truth	1.52	6.37	0.806	-	3.71±0.15	4.45±0.11	4.26±0.15
		GANSpeech+	**1.51**	**5.10**	0.700	0.064	**3.62±0.14**	3.30±0.20	3.17±0.21
		YourTTS	4.49	10.62	0.755	0.047	2.99±0.16	3.76±0.16	3.51±0.18
		MultiVerse	1.79	6.05	**0.811**	**0.100**	3.27±0.16	**4.27±0.13**	**4.08±0.13**
	KOR	Ground-truth	6.66	19.89	0.836	-	4.65±0.08	4.47±0.10	4.05±0.11
		GANSpeech+	**5.16**	**17.95**	0.733	0.028	**4.43±0.08**	3.48±0.13	3.78±0.11
		YourTTS	8.07	24.77	0.793	0.027	3.96±0.10	4.06±0.10	4.05±0.10
		MultiVerse	5.22	18.54	**0.830**	**0.066**	4.28±0.08	**4.24±0.09**	**4.29±0.09**
		Cross-lingual							
N	ENG (to KOR)	GANSpeech+	4.72	22.55	0.667	0.025	**4.20±0.08**	2.69±0.10	-
		YourTTS	10.40	36.07	0.780	0.011	3.04±0.10	3.32±0.11	-
		MultiVerse	**4.43**	**21.98**	**0.780**	**0.043**	4.09±0.09	**3.56±0.10**	-
	KOR (to ENG)	GANSpeech+	**0.28**	**1.18**	0.656	0.001	**3.72±0.15**	2.45±0.19	-
		YourTTS	5.46	10.14	0.761	0.015	2.60±0.15	**3.70±0.19**	-
		MultiVerse	0.39	1.34	**0.782**	**0.048**	3.63±0.14	3.12±0.19	-
E	ENG (to KOR)	GANSpeech+	4.99	23.28	0.637	0.089	**3.92±0.10**	2.68±0.11	3.15±0.12
		YourTTS	10.22	34.71	0.735	0.072	2.90±0.11	3.08±0.12	3.16±0.11
		MultiVerse	**4.44**	**22.45**	**0.758**	**0.122**	3.32±0.10	**3.20±0.11**	**3.56±0.11**
	KOR (to ENG)	GANSpeech+	0.47	**1.37**	0.645	0.060	3.79±0.15	2.61±0.19	3.21±0.18
		YourTTS	6.31	11.46	0.763	0.053	2.47±0.15	**3.78±0.18**	**3.77±0.19**
		MultiVerse	**0.42**	1.40	**0.773**	**0.083**	**3.84±0.13**	2.89±0.19	3.31±0.19

24 native speakers evaluated 10 Korean audio samples. A detailed description about the MOS tests are described in Appendix J.

4 Experimental Results

4.1 Zero-shot Scenario

We conducted objective and subjective evaluations for speech synthesis using unseen speaker's or unseen language's speech prompt. We categorized experiments into intra-lingual and cross-lingual experiments based on the language of the speech prompt and the target language. For the evaluation, 400 audio samples were synthesized using speech prompt and unpaired input text. Both the speech prompt and the text were selected randomly from the evaluation datasets. The speech prompt consisted of randomly sliced audio segments, each lasting 3 seconds. Experimental results are presented in Table 1, respectively[8]. In Appendix J, we additionally conducted an ablation study to assess the individual impact of proposed methods used in MultiVerse.

Intra-lingual The evaluation results show that the MultiVerse outperforms across languages and speech styles. In particular, it synthesized speech maintaining speaker or prosody similarity of the prompt. Lower SECS and higher F_0 PCC results indicate that the robust style and prosody modeling of MultiVerse enables it to generate speech highly similar to the prompt, even in expressive styles. Subjective evaluation results also confirmed that the proposed model generates speech that is aurally natural and similar to the prompt. Notably, it excels in speaker and prosody similarity compared to baselines. In the naturalness test (N-MOS), GANSpeech+ achieves slightly higher scores than the proposed model. However, due to the low generalization performance of GANSpeech+, it synthesized speech with the speaking style and speaker identity far from prompt speech. The similarity tests (S- and PS-MOS) indicate that GANSpeech+ relies on learned features rather than reflecting information from the prompt speech.

[8] Audio samples are available at `https://nc-ai.github.io/speech/publications/multiverse/index.html`.

Table 2: The results of subjective evaluation for the comparison with data-driven large-scale models.

Prompt Language	Model	N-MOS	S-MOS
ENG	Ground-truth	4.31±0.11	4.21±0.14
	VALL-E	3.56±0.14	**4.36±0.11**
	MultiVerse	**3.85±0.14**	4.30±0.13
CHN (to ENG)	VALL-EX	3.05±0.27	3.09±0.31
	MultiVerse	**3.51±0.26**	**3.33±0.30**

Table 3: The results of objective evaluation for the comparison with data-driven models. The training data size in hours for each model is indicated in parentheses.

Method	CER	WER	SECS	F_0 PCC
Intra-lingual				
Mega-TTS (20k)	0.000	0.000	0.800	**0.269**
MultiVerse	0.000	0.000	**0.835**	0.215
NaturalSpeech2 (44k)	0.002	0.008	0.814	0.079
MultiVerse	0.002	0.008	**0.826**	**0.091**
Voicebox (60k)	0.016	0.052	**0.779**	0.076
MultiVerse	**0.014**	**0.048**	0.718	**0.187**
Cross-lingual				
Mega-TTS (20k)	**0.007**	**0.058**	**0.747**	0.095
MultiVerse	0.013	0.077	0.681	**0.164**
Voicebox (50k)	**0.001**	**0.013**	**0.812**	0.063
MultiVerse	0.002	0.017	0.692	**0.130**

Meanwhile, YourTTS achieves higher speaker similarity than GANSpeech+, but suffers from lower quality. These results show that both baseline models have limitations in similarity and speech robustness.

Cross-lingual The evaluation results of MultiVerse's cross-lingual task show a similar tendency to the intra-lingual task. The S-MOS and PS-MOS scores in 'KOR (to ENG)', which synthesizes English speech by inputting a Korean expressive prompt, are relatively lower than 'ENG (to KOR)' because of a data imbalance caused by the absence of expressive style in the English training dataset. However, when amount of expressive style speech in the English training datasets are prepared, similarity performance may improve even under the combination of unseen language, speaker, and style conditions. In N-MOS, MultiVerse occasionally received lower scores than GANSpeech+. We speculate that this may be due to the disparity between the language of the prompt and the input text. GANSpeech+ generally exhibits high naturalness but synthesizes speech with low similarity because of synthesizing speech with low speaker similarity and a neutral style, ignoring the prompt.

4.2 Comparison with Data-Driven Models

We compared performance between our proposed model and data-driven large-scale models. We selected VALL-E (Wang et al., 2023) and VALL-EX (Zhang et al., 2023) as baselines for subjective evaluation. VALL-E and VALL-EX were trained on over 60k hours of English speech data and over 70k hours of English and Chinese speech data, respectively. For a fair comparison, we synthesized speech using MultiVerse based on identical prompts and scripts used in publicly available speech samples of each model[9]. Table 2 presents the results of the N-MOS and S-MOS tests. Despite using relatively small amounts of training data, MultiVerse is able to synthesize speech that is not only more natural but also comparable in similarity to the prompt when compared to large-scale models. Moreover, even though MultiVerse was not exposed to a Chinese dataset during training, it outperforms VALL-EX. For objective evaluation, Mega-TTS (Jiang et al., 2023)[10], NaturalSpeech2 (Shen et al., 2023)[11], and Voicebox (Le et al., 2023)[12] were selected as baselines. In the cross-lingual task, Mega-TTS used Chinese prompts, and Voicebox used Spanish, French, German, Portuguese, and Polish prompts to generate English speech in the style of the prompts. Table 3 describes the results which demonstrate that MultiVerse achieved comparable zero-shot performance to large-scale TTS models with only about 1.2k hours of training data. Notably, the F0 PCC results confirm that MultiVerse can generate voices with higher prosody similarity. However, in the cross-lingual task, MultiVerse generated speech with lower speaker similarity than the baselines. We speculate that the performance degradation is due to the prompt speech being in a language not included in the training data.

4.3 Speech Style Transfer

Table 4 presents objective and subjective evaluation results for scenarios in which the input text aligns with the content of the style prompt (same-text) and scenarios in which the target text differs from the one of the style prompt (different-text). All models were trained using Korean datasets, and the full

[9] https://www.microsoft.com/en-us/research/project/vall-e-x/

[10] https://mega-tts.github.io/demo-page/

[11] https://speechresearch.github.io/naturalspeech2/

[12] https://voicebox.metademolab.com/

Table 4: Evaluation results of speech style transfer.

Task	Model	F_0 DTW(↓)	Dur. RMSE(↓)	CER(↓)	WER(↓)	SECS(↑)	N-MOS
same-text	Daft-Exprt	0.370	3.480	8.03	24.71	0.859	2.63±0.21
	MultiVerse	**0.348**	**1.500**	**3.25**	**17.94**	**0.865**	**3.69±0.22**
different-text	Daft-Exprt	**0.438**	-	3.51	14.02	0.862	2.82±0.21
	MultiVerse	0.440	-	**2.27**	**11.55**	**0.868**	**3.27±0.18**

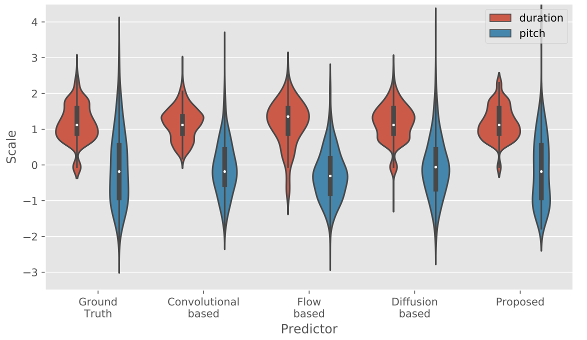


Figure 3: Violin plot describing duration and pitch distributions.

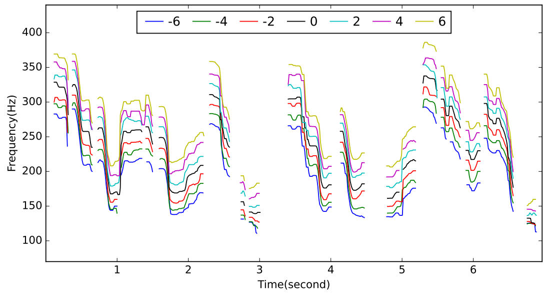


Figure 4: F_0 contours of pitch-shifted synthesized speech whose predicted pitch units are manipulated with $\{-6, -4, -2, 2, 4, 6\}$.

frame of the prompt speech was employed during inference. Objective evaluations were conducted on a total of 225 audio samples, and the naturalness of 10 audio samples of each model was assessed by 9 Korean participants in the listening test. Object evaluation results demonstrate that MultiVerse achieves speech style transfer in both same- and different-text environments. Please note that MultiVerse can model prosody suitable to input text and prompt speech by in-context learning, while Daft-exprt takes extracted acoustic features for transferring speech style. Nevertheless, MultiVerse effectively transfers the style of prompt, especially rhythm and speed. The SECS result demonstrates that MultiVerse, which utilize speaker information from prompt speech, preserves speaker information similarly to Daft-exprt, which uses deterministic speaker ID. In the N-MOS test, participants rated MultiVerse as generating more natural and clear speech than Daft-exprt. It is attributed to low intelligibility and restricted style reflection on the input text, as Daft-exprt directly transfers style using extracted acoustic features.

4.4 Acoustic Feature Modeling

We compared the acoustic feature modeling performance of the AR prosody predictor with convolutional layer-based, Flow-based and Denoising Diffusion Probabilistic Model-based predictors. Specifically, we selected the convolutional layer-based predictor from FastSpeech2 (Ren et al., 2021), the Flow-based duration predictor and pitch predictor from YourTTS and VarianceFlow (Lee et al., 2022), and DDPM-based predictor from Li et al. (2023), respectively. Each predictor was trained with Korean data. The distributions of predicted duration and pitch according to each predictor are illustrated in Figure 3. Acoustic features of expressive speech have a distribution close to multi-modal or has large variation. The convolutional layer and Flow-based predictors that do not consider the time-dependency, underfit close to the average. Diffusion based predictor also fail to model the target distribution, while the AR prosody predictor approximates the target distribution relatively closely. It indicates that the prompt-based AR structure effectively models the time-dependent characteristics of prosody. Additionally, the proposed AR prosody predictor can control prosody by adjusting acoustic feature indices. Figure 4 illustrates the F_0 contour of synthesized speech generated by manipulating pitch index sequences to constant values. The change in the acoustic index corresponds to a variation in the F_0 contour of the synthesized speech.

5 Conclusion

This paper introduces MultiVerse, an efficient and expressive zero-shot multi-task TTS system designed to address the limitations of existing zero-shot TTS systems that depend on large-scale training datasets. MultiVerse employs a structure that disentangles speech components into filter and source representations: this structure contributes to achieving zero-shot TTS performance comparable to data-driven TTS approaches, even with a small amount of data. Additionally, it enhances prosody similarity through a hybrid prosody modeling that combines both autoregressive and non-autoregressive mechanisms. Quantitative and qualitative evaluations across various language and style demonstrate that MultiVerse excels in zero-shot, cross-lingual TTS, and speech style transfer.

6 Limitations

MultiVerse, which generates mel-spectrograms, requires a system to convert the mel-spectrograms into waveforms, such as a neural vocoder. Therefore, the performance of the vocoding system can potentially affect the performance of MultiVerse. Additionally, utilizing pre-processed acoustic features, i.e., duration, pitch, and energy, becomes more costly as the amount of training data increases. Hence, one of the next goals of this research could be to incorporate unsupervised modeling methods for acoustic features.

7 Broader Impacts

We aimed to enhance the versatility of deep-learning-based TTS models. While speech generative models offer valuable support for creating digital content, concerns arise about their potential misuse for fraud and crime. This study is designed to minimize these potential negative impacts and effectively support TTS models for content creators. We emphasize ethical considerations, especially regarding data privacy, by ensuring that all voice data used in training and evaluation is sourced from publicly available datasets or internal datasets with the explicit consent of the speakers. Moreover, the study looks ahead to future societal implications, striving to expand the capabilities of TTS models in a manner that aligns with ethical and social responsibilities related to content creation.

8 Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of Co-Pilot technology for automatic completion of generative AI-based 3D Webtoon, Project Number: RS-2024-00400004, Contribution Rate: 25%)

References

Taejun Bak, Jae-Sung Bae, Hanbin Bae, Young-Ik Kim, and Hoon-Young Cho. 2021. FastPitchFormant: Source-Filter Based Decomposed Modeling for Speech Synthesis. In *Proc. Interspeech*, pages 116–120.

Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. 2023. Avocodo: Generative adversarial network for artifact-free vocoder. In *Proc. AAAI Conference on Artificial Intelligence*, volume 37, pages 12562–12570.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. 2021. SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model. In *Proc. Interspeech*, pages 3645–3649.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu. 2020. Hifisinger: Towards high-fidelity neural singing voice synthesis. *arXiv preprint arXiv:2009.01776*.

Hyunjae Cho, Wonbin Jung, Junhyeok Lee, and Sang Hoon Woo. 2022. SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech. In *Proc. Interspeech*, pages 1–5.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in neural information processing systems*, 34:16251–16265.

Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han.

2020. In defence of metric learning for speaker recognition. In *Proc. Interspeech*.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6184–6188.

G. Fant. 1970. *Acoustic theory of speech production*. 2. Walter de Gruyter.

Allen Gersho and Robert M Gray. 2012. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media.

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in neural information processing systems*, 35:10970–10983.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun MA, and Zhou Zhao. 2024. Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis. In *The Twelfth International Conference on Learning Representations*.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al. 2023. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In *Conference on Computer Vision and Pattern Recognition*, pages 10124–10134.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Conference on Computer Vision and Pattern Recognition*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 8666–8681.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*.

Neeraj Kumar, Srishti Goel, Ankur Narang, and Brejesh Lall. 2021. Normalization driven zero-shot multi-speaker speech synthesis. In *Proc. Interspeech*, pages 1354–1358.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2021. STYLER: Style Factor Modeling with Rapidity and Robustness via Speech Decomposition for Expressive and Controllable Neural Text to Speech. In *Proc. Interspeech*, pages 4643–4647.

Yoonhyung Lee, Jinhyeok Yang, and Kyomin Jung. 2022. Varianceflow: High-quality and controllable text-to-speech using variance information via normalizing flow. In *International Conference on Acoustics, Speech and Signal Processing*, pages 7477–7481.

Xiang Li, Songxiang Liu, Max W. Y. Lam, Zhiyong Wu, Chao Weng, and Helen Meng. 2023. Diverse and Expressive Speech Prosody Prediction with Denoising Diffusion Probabilistic Model. In *Proc. INTERSPEECH 2023*, pages 4858–4862.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proc. International Conference on Computer Vision*, pages 2794–2802.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech*, pages 498–502.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proc. AAAI Conference on Artificial Intelligence*, volume 32.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. 2020. Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301*.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Atli Thor Sigurgeirsson and Simon King. 2023. Do prosody transfer models transfer prosody? In *International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech*, pages 4006–4010.

Detai Xin, Tatsuya Komatsu, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2021. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual tts. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6608–6612. IEEE.

Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari. 2020. Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space. In *Proc. Interspeech*, pages 2947–2951.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.

Jinhyeok Yang, Jae-Sung Bae, Taejun Bak, Young-Ik Kim, and Hoon-Young Cho. 2021. GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis. In *Proc. Interspeech*, pages 2202–2206.

Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu. 2022. Syntaspeech: Syntax-aware generative adversarial text-to-speech. In *Proc. International Joint Conference on Artificial Intelligence*, pages 4468–4474.

Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, and Lawrence Carin. 2020. Improving zero-shot voice style transfer via disentangled representation learning. In *International Conference on Learning Representations*.

Julian Zaïdi, Hugo Seuté, Benjamin van Niekerk, and Marc-André Carbonneau. 2022. Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis. In *Proc. Interspeech*, pages 4591–4595.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*, pages 1526–1530.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. Interspeech*, pages 2080–2084.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

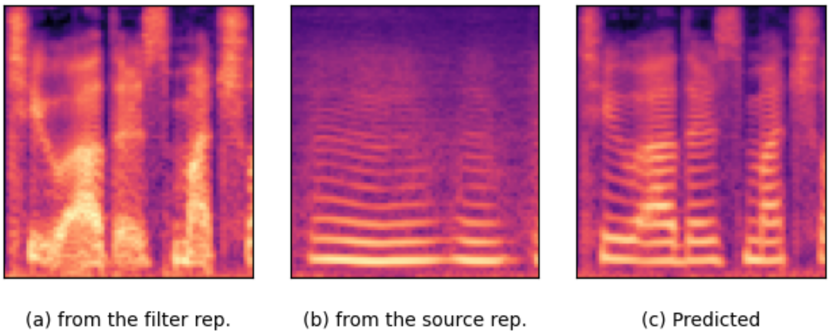


Figure 5: Mel-spectrograms of synthesized audio sample: (a) generated from the filter representation only. (b) generated from the source representation only. (c) generated from the coarse mel-representation.

Table 5: Analysis results on filter and source representations.

Task	Model	CER(↓)	WER(↓)	SECS(↑)	F_0 PCC(↑)
Intra-lingual	MultiVerse	**3.11**	**12.02**	**0.831**	**0.111**
	filter representation	3.26	13.23	0.668	0.023
	source representation	-	-	0.574	0.090
Cross-lingual	MultiVerse	**2.65**	**12.42**	**0.775**	**0.089**
	filter representation	3.03	14.56	0.622	0.020
	source representation	-	-	0.544	0.074

A Related Works

Zero-shot TTS Zero-shot TTS synthesizes speech for unseen speakers not present in the training dataset. Learning general speech features, like speaker identity and speaking style, is crucial in this task. Zero-shot TTS models often incorporate a pre-trained speaker encoder for modeling speaker information (Jia et al., 2018; Kumar et al., 2021; Cooper et al., 2020). Some employ specific objectives to improve speaker similarity (Casanova et al., 2021, 2022). However, challenges arise when generating voices significantly different from the training data, impacting similarity and naturalness. Recent advances in language models have prompted exploration of data-driven methods in speech synthesis (Wang et al., 2023; Borsos et al., 2023; Kharitonov et al., 2023; Shen et al., 2023), enhancing generalization to unseen voices. Despite this, acquiring large speech datasets is costly and challenges persist in obtaining diverse data for various languages. Existing models have primarily focused on speaker similarity, leaving the issue of prosody similarity unresolved.

Cross-lingual TTS Cross-lingual TTS aims to generate speech in a language that is different from the monolingual speaker while preserving speaker's voice. However, training on multilingual multi-speaker data may entangle language and speaker information, resulting in diminished similarity. Therefore, disentangling the language from the speaker becomes crucial in cross-lingual TTS. Zhang et al. (2019); Xin et al. (2020) propose adversarial layers to disentangle speaker and language information. Xin et al. (2021) utilize mutual information minimization to decouple the information. Despite the objective of disentanglement, these models suffer from unstable training, creating a trade-off between disentanglement and speaker similarity (Zhang et al., 2019). Recently, data-driven methods have also been applied to improve the generalization in cross-lingual TTS (Zhang et al., 2023).

Speech style transfer Speech style transfer aims to synthesize speech with a speaking style similar to the reference speech, notwithstanding differences like identity or content. Style transfer models learn to model inherent elements of voice style, disentangling these elements from content and speaker identity (Lee et al., 2021; Zaïdi et al., 2022; Huang et al., 2022; Yuan et al., 2020). Lee et al. (2021) disentangles acoustic features (duration, prosody, energy) by encoding them separately.

Daft-Exprt (Zaïdi et al., 2022) uses domain adversarial training to separate prosody and speaker information. Huang et al. (2022) proposes mix-style layer normalization to remove style information from filter representation. While these studies enhance style transfer performance, their limited disentanglement hinders style transfer in out-of-domain environments (Sigurgeirsson and King, 2023).

B Analysis on Filter and Source Representations

MultiVerse adopts source-filter theory-based decomposed modeling to learn disentangled representations which are divided into filter and source. As described in Section 2, the filter generator produces the filter representation related to speech content, pronunciation and accent. On the other hand, the source generator produces the source representation that contains prosodic information that is less correlated to the content, such as intonation, rhythm, and stress patterns.

Figure 5 presents mel-spectrograms generated from the each representations. To analyze these representations in detail, we conducted objective evaluation. In this experiments, audio samples were generated by passing both representations individually through the acoustic decoder, not forming the coarse mel-representation. Table 5 demonstrates the evaluation results. CER and WER results indicate that the synthesized speech from the filter representation has comparable intelligibility to that generated by MultiVerse. This means that the filter representation primarily contains linguistic information of input text. Additionally, SECS result shows that the filter representation is more related to the speaker identity than the source representation. Meanwhile, the synthesized speech from the source representation is sounding like mumble sounds “Hmmmm, Mmmmmm ..."; ASR model failed to recognize the speech.However, it has more similar pitch distribution than that from the filter representation because the source representation, generated from acoustic features, learns the prosodical patterns included in the prompt speech.

C Detailed on Acoustic Decoder

The proposed acoustic decoder employs sample-adaptive kernel selection (Kang et al., 2023) to learn convolutional filters suitable for the speech prompt. This approach generates a mel-spectrogram while preserving the information of the coarse mel-representation. Figure 6 depicts the detailed structure of sample-adaptive kernel selection.

The specific process is as follows: the mapping network maps the style vector from a global style embedding and random noise sampled from a normal Gaussian distribution. The K-bank convolutional filters of each sample-adaptive convolution layer are aggregated by a weighted sum based on the weights predicted for each kernel by the style vector. Subsequently, the aggregated filter undergoes modulation and demodulation by scale, where the scale is obtained from the output of an affine layer with the style vector as input (Karras et al., 2020).

The proposed acoustic decoder replaces the convolutional layer of the feed-forward transformer block with a sample-adaptive kernel selection-based convolutional layer. ReLU activation function is used for non-linearity between the two sample-adaptive convolutional layers. Detailed parameter descriptions for sample-adaptive kernel selection are provided in Table 7.

D Comparison with Vector Quantization based Prosody Modeling

The P-LLM of Mega-TTS also leverages autoregressive language model-based prosody modeling. However, as mentioned in Section 2.3.1, the approach using vector quantization-based codebooks is influenced by the quantity of training data. To verify this, a simple comparison was conducted. We trained a model, referred as MultiVerse(VQ), by replacing MultiVerse's prosody modeling with Mega-TTS's VQ encoder. Both MultiVerse and MultiVerse(VQ) synthesized speech from the test dataset, using ground-truth mel-spectrogram as prompts and ground-truth phoneme durations. This experimental setup aimed to observe the performance of vector quantization-based models with limited training data.

We observed that the synthesized speech by MultiVerse(VQ) occasionally resulted in distorted audio, as depicted in Figure 8. MultiVerse(VQ) produced more distorted speech when targeting expressive style. It also fails to reflect the prosody of the prompt speech, indicating that expressive style was not effectively modeled by vector quantization, likely due to the relatively small number of expressive style audio samples. In contrast, Multi-

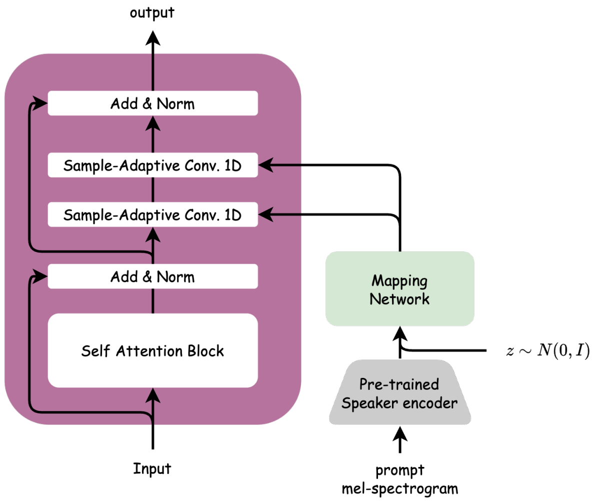
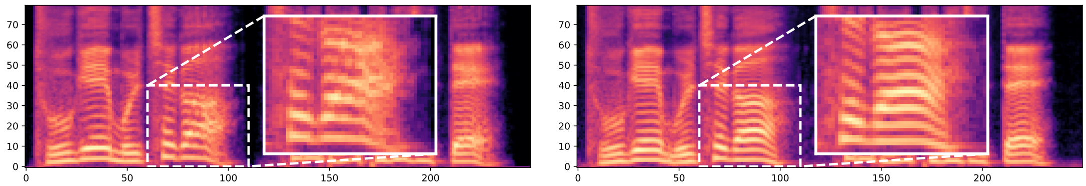


Figure 6: The diagram of the proposed acoustic decoder with the sample-adaptive kernel selection.



(a) MultiVerse w/o sample-adaptive kernel selection.

(b) MultiVerse.

Figure 7: Mel-spectrograms of synthesized audio sample: MultiVerse and MultiVerse w/o sample-adaptive kernel selection.

Table 6: Detailed dataset information.

Dataset	Language	Number of Speakers	Time(h)
LibriTTS	ENG	1133	262
VCTK	ENG	101	
Internal	ENG	42	
AiHub	KOR	46	969
Internal	KOR	229	

Verse demonstrated relatively robust synthesis of expressive style speech.

E Detailed Dataset Information

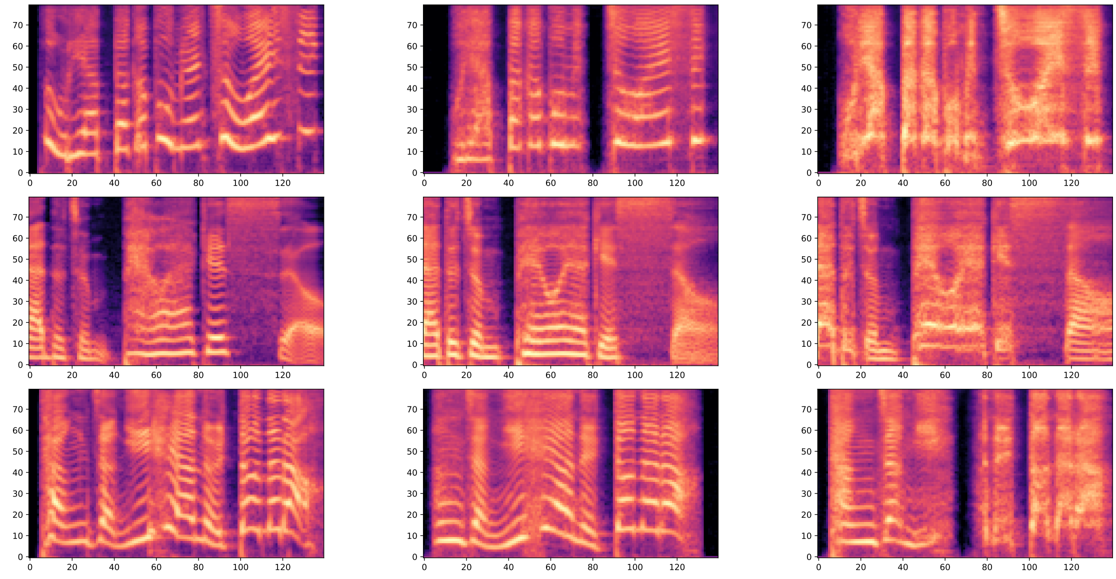
Table 6 describes detailed dataset information. The English training set was constructed from approximately 262 hours of speech data. The LibriTTS (Zen et al., 2019), VCTK (Yamagishi et al., 2019), and internal datasets were recorded from 1133, 101, and 42 speakers, respectively[13]. The speech styles included both neutral narration and conversational styles. The Korean dataset consisted of a total of 969 hours of speech data, with the AI-Hub multi-speaker[14] and internal datasets recorded from 46 and 229 speakers, respectively. It included various speech styles, such as narration and acting. 10% of all datasets were excluded from the training set for testing purposes.

F Inference Details

In this section, we provide the inference process in detail. For zero-shot TTS, both input text and a prompt voice are required. In this case, a mismatch between the text and prompt is permissible. Whether the input text and prompt voice's language match determines if it is intra-lingual or cross-lingual TTS. At first, the input text is processed to obtain the output x from the text encoder. Next, the prompt speech is used to obtain the hidden representation r from the speech prompt encoder, and the global style embedding r_g is ob-

[13] License: CC-BY-4.0

[14] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=542



(a) Ground Truth (b) MultiVerse (c) MultiVerse(VQ)

Figure 8: Mel-spectrograms of (a) ground-truth, (b) generated speech from MultiVerse and (c) generated speech from MultiVerse with vector quantization (VQ) based prosody modeling.

tained from the pre-trained speaker encoder. The concatenated x and r along the time axis serve as a prefix for the AR prosody predictor to autoregressively decode the phoneme acoustic features, including duration $\mathbf{d}$, pitch $\mathbf{p}$, and energy $\mathbf{e}$. The predicted phoneme-specific duration is utilized for upsampling to the frame level. The filter generator outputs the filter representation using the upsampled version of x and $\mathbf{e}$. Simultaneously, the source generator produces the source representation from the upsampled version of $\mathbf{p}$ and $\mathbf{e}$, with r utilized in generating both representations. These two representations are combined and transformed into a mel-spectrogram by the acoustic decoder, where r_g is used to determine the decoder's filter.

For style transfer, input text and two audio samples are required: prompt speech for the target speaker, denoted as y_s, and prompt speech for the target style, represented as y_p. The flow of generation is the same as in zero-shot TTS, but the application of prompts differs. From the speech prompt encoder, we obtain the hidden representation for the speaker r_s and the hidden representation for style r_p for each prompt speech. The global style embedding is obtained from y_s. r_p is utilized for predicting phoneme acoustic features, while r_s is used for applying the prompt. Consequently, it is possible to generate a voice that reflects the style of y_p with the speaker identity of y_s. Moreover, prosody-controllable TTS is facilitated by controlling the index of the predicted acoustic features.

G Model Configuration

Table 7 describes detailed information of hyperparameters of modules in MultiVerse.

H Acoustic Feature Pre-Processing

To obtain sequences of acoustic units for training, acoustic features, such as duration, fundamental frequency (F_0), and energy, were pre-processed using the following procedures. Each acoustic feature sequence was calculated, followed by normalization and quantization, to be transformed into acoustic unit sequences. Specifically, the procedures for each acoustic feature is as follows: For duration, to obtain the duration per phoneme, we used both the internal forced aligner and an external aligner (McAuliffe et al., 2017) for Korean and English, respectively. The duration sequences were already in integer values, so we used them directly as duration unit sequences without additional normalization and quantization. We set the maximum duration value to 32. F_0 extraction from the speech signal utilized a Praat-based extractor[15]. The F_0 sequences, extracted in Hertz units, were averaged per phoneme, then were normalized using speaker-specific statistical information. The normalized F_0 sequence values were quantized into 64 values within a certain range to obtain the pitch unit sequence. In this paper, the normalized F_0 sequence was clipped to the range from -4 to +4. For energy,

[15] https://github.com/YannickJadoul/Parselmouth

Table 7: Detailed hyper-parameters of MultiVerse.

Module	**Configuration**	**Value**
Text Encoder	Encoder Layers	6
	Feed-forward dim	2048
	Hidden dim	512
	Kernel size	3
	Number of heads	8
Speech Prompt Encoder	Encoder Layers	3
	Feed-forward dim	2048
	Hidden dim	512
	Kernel size	9
	Number of heads	8
AR Prosody Predictor	Encoder Layers	3
	Feed-forward dim	2048
	Hidden dim	512
	Number of heads	8
	Duration codebook, dim	32, 512
	Pitch codebook, dim	64, 512
	Energy codebook, dim	64, 512
Filter / Source Generator	Encoder Layers	3
	Feed-forward dim	2048
	Hidden dim	512
	Kernel size	3
	Number of heads	8
Acoustic Decoder	Encoder Layers	3
	Feed-forward dim	2048
	Hidden dim	512
	Kernel size	3
	Number of heads	8
Sample-Adaptive Kernel Selection	Mapping network depth	4
	Mapped style dim	256
	Noise dim	64
	Global style dim	512
	Size of kernel bank	4
Multi-Window Discriminator	Number of discriminators	3
	Window size	32, 64, 128
	Conv2d size	3
	Hidden size	128
Total Number of Parameters		260.62M

frame-wise energy of the speech signal was calculated as the magnitude of the linear spectrogram and averaged per phoneme. The normalized energy sequence was quantized into 64 values within the range of -5 to +5.

I Details on the Subjective Evaluation

For the subjective evaluation of samples, we performed three kinds of listening tests: the naturalness (N-MOS), speaker similarity (S-MOS), and prosody similarity (PS-MOS). Amazon Mechanical Turk (MTurk)[16] and internal evaluation

[16] https://www.mturk.com/

tools were used for the evaluation of the English and Korean samples, respectively. For each task (intra-lingual neutral/expressive, cross-lingual neutral/expressive), 15 test samples for English and 10 test samples for Korean were randomly selected per model. 10 native English participants and 24 native Korean participants rated the audio samples. Evaluation scores were evaluated at 0.5-point intervals from 1 to 5 points. We guided participants to evaluate audio samples by focusing only on the evaluation factor while ignoring other factors. In N-MOS test, the quality of the sound is ignored and only the naturalness of the speech is evaluated. For S-MOS test, we emphasized that participants should concentrate only on determining how closely related synthesized speech and prompt speech, disregarding content and prosody. Meanwhile, we requested participants to assess how similar the prosody, including rhythm and stress patterns, between the synthesized and prompt speech, disregarding content and timbre. The actual evaluation screen and contents used are shown in Figure 9.

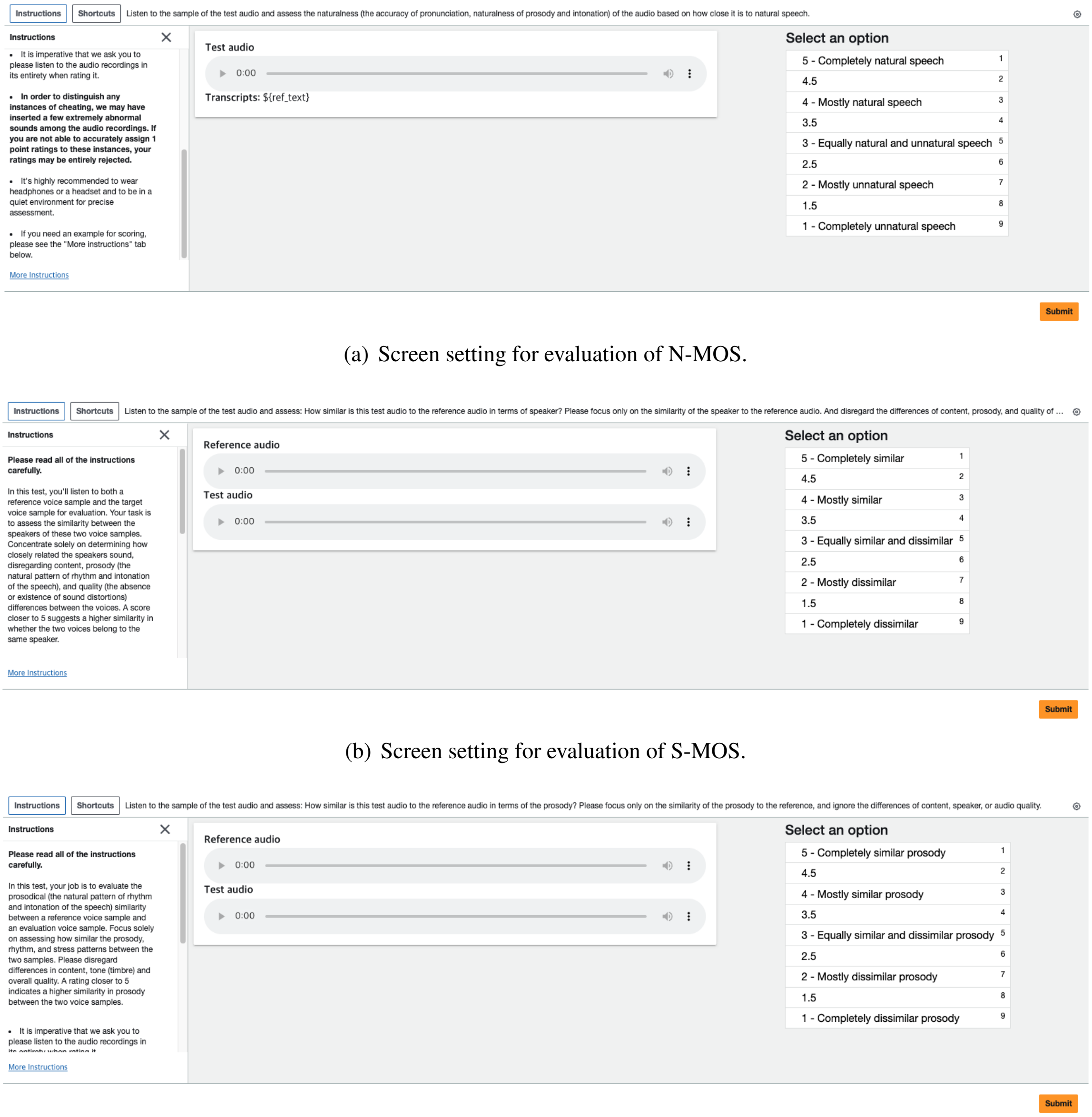
J Ablation Study

To examine the specific impacts of the proposed methods, we compared models by sequentially removing individual components from the baseline architecture of our proposed model, namely the source-filter, sample-adaptive kernel selection, and the FiLM layer. Since the AR prosody predictor has already been compared with other predictors in Section 4.4, we did not include it here. Evaluations were conducted for both intra- and cross-lingual tasks, similar to the zero-shot scenario. CER, WER, and SECS were used as evaluation metrics. The ablation evaluation results are detailed in Table 8.

The evaluation results for CER and WER demonstrate that the FiLM layer, among the proposed methods, enhances speech robustness. No other method, excluding the FiLM layer, improved speech robustness. On the other hand, performance improved when the source-filter method was not used, which is related to the trade-off between speaker similarity and speech robustness, as observed in the objective evaluation results in Section 4.1. All proposed methods influenced the enhancement of speaker similarity. Even when only one of the proposed methods was not utilized, SECS deteriorated. This indicates that the proposed methods contribute to improving the MultiVerse's generalization performance to learn speaker characteristics. We also observed that sample-adaptive kernel selection helps the acoustic decoder to generate a higher-quality mel-spectrogram. The difference in mel-spectrograms between MultiVerse with and without sample-adaptive kernel selection is depicted in Figure 7. MultiVerse without sample-adaptive kernel selection synthesized mel-spectrogram with decreased quality, resulting in lower intelligibility or the presence of artifacts.

Table 8: The result of the ablation study.

Task	Model	CER(↓)	WER(↓)	SECS(↑)
Intra-Lingual	MultiVerse	**3.11**	**12.02**	**0.831**
	w/o source-filter	2.78	11.07	0.818
	w/o sample-adaptive kernel selection	3.01	11.53	0.817
	w/o FiLM layer	3.26	11.83	0.817
Cross-Lingual	MultiVerse	**2.65**	**12.42**	**0.775**
	w/o source-filter	2.49	12.15	0.760
	w/o sample-adaptive kernel selection	2.66	12.47	0.760
	w/o FiLM layer	2.65	12.34	0.758



Listen to the sample of the test audio and assess the naturalness (the accuracy of pronunciation, naturalness of prosody and intonation) of the audio based on how close it is to natural speech.

Instructions

- It is imperative that we ask you to please listen to the audio recordings in its entirety when rating it.
- **In order to distinguish any instances of cheating, we may have inserted a few extremely abnormal sounds among the audio recordings. If you are not able to accurately assign 1 point ratings to these instances, your ratings may be entirely rejected.**
- It's highly recommended to wear headphones or a headset and to be in a quiet environment for precise assessment.
- If you need an example for scoring, please see the "More instructions" tab below.

More Instructions

Test audio

0:00

Transcripts: ${ref_text}

Select an option

5 - Completely natural speech	1
4.5	2
4 - Mostly natural speech	3
3.5	4
3 - Equally natural and unnatural speech	5
2.5	6
2 - Mostly unnatural speech	7
1.5	8
1 - Completely unnatural speech	9

(a) Screen setting for evaluation of N-MOS.

Listen to the sample of the test audio and assess: How similar is this test audio to the reference audio in terms of speaker? Please focus only on the similarity of the speaker to the reference audio. And disregard the differences of content, prosody, and quality of ...

Instructions

Please read all of the instructions carefully.

In this test, you'll listen to both a reference voice sample and the target voice sample for evaluation. Your task is to assess the similarity between the speakers of these two voice samples. Concentrate solely on determining how closely related the speakers sound, disregarding content, prosody (the natural pattern of rhythm and intonation of the speech), and quality (the absence or existence of sound distortions) differences between the voices. A score closer to 5 suggests a higher similarity in whether the two voices belong to the same speaker.

More Instructions

Reference audio

0:00

Test audio

0:00

Select an option

5 - Completely similar	1
4.5	2
4 - Mostly similar	3
3.5	4
3 - Equally similar and dissimilar	5
2.5	6
2 - Mostly dissimilar	7
1.5	8
1 - Completely dissimilar	9

(b) Screen setting for evaluation of S-MOS.

Listen to the sample of the test audio and assess: How similar is this test audio to the reference audio in terms of the prosody? Please focus only on the similarity of the prosody to the reference, and ignore the differences of content, speaker, or audio quality.

Instructions

Please read all of the instructions carefully.

In this test, your job is to evaluate the prosodical (the natural pattern of rhythm and intonation of the speech) similarity between a reference voice sample and an evaluation voice sample. Focus solely on assessing how similar the prosody, rhythm, and stress patterns between the two samples. Please disregard differences in content, tone (timbre) and overall quality. A rating closer to 5 indicates a higher similarity in prosody between the two voice samples.

- It is imperative that we ask you to please listen to the audio recordings in its entirety when rating it

More Instructions

Reference audio

0:00

Test audio

0:00

Select an option

5 - Completely similar prosody	1
4.5	2
4 - Mostly similar prosody	3
3.5	4
3 - Equally similar and dissimilar prosody	5
2.5	6
2 - Mostly dissimilar prosody	7
1.5	8
1 - Completely dissimilar prosody	9

(c) Screen setting for evaluation of PS-MOS.

Figure 9: Screen settings for evaluation of N-MOS, S-MOS, and PS-MOS.