LAUGH NOW CRY LATER: CONTROLLING TIME-VARYING EMOTIONAL STATES OF FLOW-MATCHING-BASED ZERO-SHOT TEXT-TO-SPEECH

Haibin Wu[1†], Xiaofei Wang[2], Sefik Emre Eskimez[2], Manthan Thakker[2], Daniel Tompkins[2],
Chung-Hsien Tsai[2], Canrun Li[2], Zhen Xiao[2], Sheng Zhao[2], Jinyu Li[2], Naoyuki Kanda[2]

[1]National Taiwan University, [2]Microsoft Corporation, USA

ABSTRACT

People change their tones of voice, often accompanied by non-verbal vocalizations (NVs) such as laughter and cries, to convey rich emotions. However, most text-to-speech (TTS) systems lack the capability to generate speech with rich emotions, including NVs. This paper introduces EmoCtrl-TTS, an emotion-controllable zero-shot TTS that can generate highly emotional speech with NVs for any speaker. EmoCtrl-TTS leverages arousal and valence values, as well as laughter embeddings, to condition the flow-matching-based zero-shot TTS. To achieve high-quality emotional speech generation, EmoCtrl-TTS is trained using more than 27,000 hours of expressive data curated based on pseudo-labeling. Comprehensive evaluations demonstrate that EmoCtrl-TTS excels in mimicking the emotions of audio prompts in speech-to-speech translation scenarios. We also show that EmoCtrl-TTS can capture emotion changes, express strong emotions, and generate various NVs in zero-shot TTS. See https://aka.ms/emoctrl-tts for demo samples.

Index Terms— zero-shot text-to-speech, emotion control, flow matching, speech-to-speech translation

1. INTRODUCTION

Humans express a wide range of emotions by changing their tone of voice, often accompanied by nonverbal vocalizations (NVs) such as laughter and crying. While current emotional text-to-speech (TTS) systems have made significant advancements [1–12], they still lack the ability to generate emotional speech with fine-grained control (e.g. changing the emotion states within a single generated utterance) and with various types of NVs like laughter and crying. In addition, current emotional TTS systems [1–12] are typically trained on staged datasets with a limited number of speakers; in extreme cases, some are trained on only one speaker. These TTS models often lack the ability to generate emotional speech for any speaker, a feature critical for applications like speech-to-speech translation, that needs to retain both the emotion and speaker characteristics of the source audio when generating the translated speech.

In this paper, we propose EmoCtrl-TTS, an emotion-controllable zero-shot TTS system that can generate highly emotional speech with NVs for any speaker. EmoCtrl-TTS generates the speech by mimicking the voice characteristics and emotion presented by an audio sample, referred to as an audio prompt. EmoCtrl-TTS is based on the flow-matching-based zero-shot TTS [13] and utilizes valence and arousal values to mimic the time-varying characteristics of emotions. In addition, it also utilizes laughter embeddings [14], which we find to be effective for generating not only laughter but also other NVs, including crying. Furthermore, by leveraging over 27k

---
†Work performed during an internship at Microsoft.

Table 1. Comparison of TTS models based on emotion capabilities.

Model	Emotion change	NVs	Emotion data size	Speaker number	Emotion data type
Emo-VITS [1]	✗	✗	N/A	N/A	Staged
ED-TTS [2]	✗	✗	70 hours	1	Staged
EmoDiff [3]	✗	✗	12 hours	10	Staged
EmoMix [4]	✗	✗	∼15 hours	10	Staged
Zhou et al. [5]	✗	✗	∼30 hours	∼100	Staged
QI-TTS [6]	✗	✗	∼15 hours	10	Staged
MsEmoTTS [7]	intensity	✗	22 hours	1	Staged
Shin et al. [8]	✗	✗	57 hours	38	Staged
Lee et al. [9]	✗	✗	21 hours	1	Staged
Li et al. [10]	✗	✗	14 hours	1	Staged
Cai et al. [11]	✗	✗	73 hours	1	Staged
EmoSphere-TTS [12]	✗	✗	∼29 hours	20	Staged
ELaTE [14]	happy↔neutral	laugh	460 hours	N/A	Real
EmoCtrl-TTS	arbitrary	arbitrary	∼27k hours	N/A	Real

hours of highly expressive real-world data through careful data mining, EmoCtrl-TTS achieves significant enhancements in robustness. Comprehensive evaluations demonstrate that EmoCtrl-TTS excels in reproducing the emotions of audio prompts across multiple languages in speech-to-speech translation scenarios. We also show that EmoCtrl-TTS can capture emotion changes, express strong emotions, and generate various types of NVs in zero-shot TTS. Our contributions are summarized as follows: (1) we propose a framework integrating flow-matching-based zero-shot TTS with NV and emotion embeddings; and (2) we conduct comprehensive experiments to evaluate the emotion-controllable zero-shot TTS, demonstrating the superiority of the proposed method.

2. RELATED WORK

2.1. Controlling emotion in TTS

Emotional TTS has undergone substantial advancements in recent years. Table 1 lists various TTS systems from different perspectives.

The first point is whether TTS systems can control the fine-grained emotional attributes within one utterance. Such fine-grained control is ideal for many applications, for example, speech-to-speech translation where nuanced emotional changes need to be transferred to the translated speech. However, as in Table 1, most prior works aimed to control the utterance-level emotion, and only a few works tackled the control of time-varying emotional status. MsEmoTTS [7] used a local emotional strength predictor to estimate syllable-level emotion strength, leveraging it as a condition to control the emotion strength of generated speech. ELaTE [14] leveraged laughter representation to condition the flow-matching-based zero-shot TTS, and showed superior controllability of laughter generation. However, they still lack full controllability of the emotional status.

The second point is the capability to generate NVs. As far as we investigated, most prior emotional TTS works were not able to generate NVs. While ELaTE [14] can generate natural laughter, it
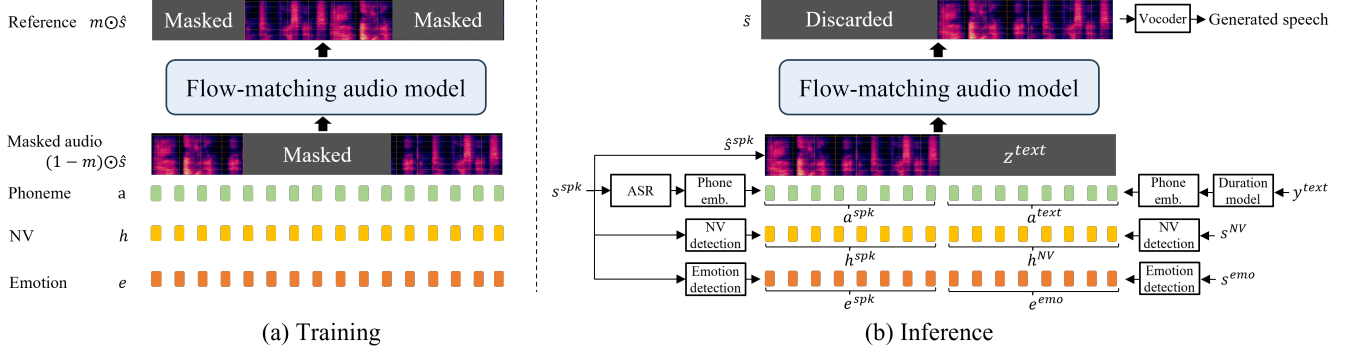
Fig. 1. An overview of (a) training and (b) inference of the audio model of EmoCtrl-TTS.

was not investigated with other NVs such as cries. Our work aims to generate arbitrary types of NVs, including laughter and cries.

The third point is the size of the training data. Due to the difficulty in developing high-quality emotional training data with supervision, most works utilized less than 100 hours of training data. While ELaTE [14] used 460 hours of speech containing laughter, the data scale is still less than 500 hours. To the best of our knowledge, ours is the first to investigate the impact of using large-scale emotional data for TTS training.

The fourth point is the number of speakers. As shown in the table, most of the emotional TTS systems utilized voices from fewer than 100 speakers, with some exceptions where the number of speakers is not available. While the number of speakers in our data is also unavailable due to anonymization, we expect that our training data contains a significantly large variation of speakers given the data scale, which is beneficial for the zero-shot TTS capability.

Finally, the fifth point is whether the emotional training data is staged data or real data. As shown in the table, most existing works utilized staged data for their training, which inevitably limited the variety of the speech. For example, in a speech-to-speech translation scenario, the source language speakers are often not professional actors, and their voice characteristics are different from the staged voice. By using large-scale training data, we aim to achieve highly faithful emotion transfer in the zero-shot TTS scenario.

2.2. Flow-matching-based TTS

2.2.1. Conditional flow matching

Conditional flow-matching [15] is an objective for training generative models. Continuous normalizing flows [16] is employed to convert a simple prior distribution p_0 into a complex distribution p_1 that aligns with the data. To be specific, for a given data point x, a neural network parameterized by θ models a time-dependent vector field $v_t(x;\theta)$. This vector field constructs a flow ϕ_t, which reshapes the prior distribution into the target distribution. Lipman et al. [15] proposed to train such a neural network with the conditional flow-matching objective:

$$\mathcal{L}^{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}||u_t(x|x_1) - v_t(x;\theta)||^2, \quad (1)$$

where q denotes the training data distribution, x_1 represents the random variable for the training data, p_t denotes the probability path at time step t, and u_t is the vector field associated with p_t. Also, Lipman et al. [15] present a conditional flow known as the optimal transport path, defined by the equations $p_t(x|x_1) = \mathcal{N}(x|tx_1, (1 - (1 - \sigma_{\min})t)^2 I)$ and $u_t(x|x_1) = (x_1 - (1 - \sigma_{\min})x)/(1 - (1 - \sigma_{\min})t)$.

2.2.2. Voicebox

Voicebox [13] is the first to utilize conditional flow matching for training zero-shot TTS. It is designed to perform speech-infilling

tasks given audio context and frame-wise phoneme sequence as conditions. Given the promising performance of Voicebox, we also employ conditional flow matching to develop our TTS model.

2.2.3. ELaTE

ELaTE [14] was proposed to generate natural laughing speech with fine-grained controllability. It utilized a frame-level laughter representation derived from the laughter detector [17,18][1] to condition the flow-matching-based zero-shot TTS, showing significantly higher quality and better controllability in generating laughing speech compared to conventional models. However, ELaTE was only tested with laughing speech, and the effects on other NVs, such as crying, have not been investigated. In our preliminary experiment, we found that ELaTE sometimes generates laughing speech even when the audio prompt contains different NVs, such as crying. Our work can be regarded as an extension of ELaTE, where we aim to achieve better emotion controllability as well as the generation of various NVs.

3. EMOCTRL-TTS

3.1. Overview

3.1.1. Model training

Figure 1 (a) illustrates the training procedure of EmoCtrl-TTS. Given a training audio sample s with transcription y, we extract its mel-filterbank features $\hat{s} \in \mathbb{R}^{F \times T}$, where F denotes the feature dimension and T represents the sequence length. Additionally, we employ force alignment and a phoneme embedding layer to obtain a frame-wise phoneme embedding $a \in \mathbb{R}^{D^{phn} \times T}$, where D^{phn} is the phoneme embedding dimension. The phoneme embedding layer is a part of the audio model and is jointly trained. Furthermore, we extract frame-wise embeddings that represent NV $h \in \mathbb{R}^{D^{NV} \times T}$ and emotion $e \in \mathbb{R}^{D^{emo} \times T}$, where D^{NV} and D^{emo} denote the dimensions of NV and emotion embeddings respectively. The embeddings h and e are extracted by using pre-trained NV and emotion detector, respectively, which are discussed in Section 3.2 and 3.3. We leverage the speech infilling task introduced in [13] to train the audio model, focusing on training a conditional flow-matching model to estimate the distribution $P(m \odot \hat{s}|(1 - m) \odot \hat{s}, a, h, e)$, where $m \in \{0, 1\}^{F \times T}$ represents a binary temporal mask, and $\odot$ is the Hadamard product.

3.1.2. Inference

Figure 1 (b) illustrates the inference procedure of EmoCtrl-TTS. During inference, the model takes four inputs: text prompt y^{text}, speaker prompt audio s^{spk}, NV prompt audio s^{NV}, and emotion

---

[1]`https://github.com/jrgillick/laughter-detection`

prompt audio s^{emo}. The text prompt represents the content of the generated speech. Meanwhile, the speaker, NV, and emotion prompts control the characteristics of the speaker, NV, and emotion in the generated speech, respectively. In speech-to-speech translation scenario, we use the source audio for s^{spk}, s^{NV} and s^{emo}, and translated text as y^{text}. This results in the translated speech maintaining the source speaker's voice and emotional characteristics.

The speaker prompt s^{spk} is first converted to the mel-filterbank features $\hat{s}^{spk}$. It is also converted to phoneme embeddings a^{spk} by applying automatic speech recognition (ASR) and then the phoneme embedding layer. The speaker prompt is further converted to NV embeddings h^{spk} and emotion embeddings e^{spk} based on the NV detector and emotion detector, respectively.

Meanwhile, the text prompt y^{text} is converted to text prompt embeddings a^{text} based on the phone duration model [13] followed by the phoneme embedding layer. The NV prompt embedding h^{NV} and the emotion prompt embedding e^{emo} are extracted from the NV detector and emotion detector, respectively. Note if the lengths of h^{NV} and h^{emo} are different from that of a^{text}, we apply linear interpolation to h^{NV} and h^{emo} to match their lengths to that of a^{text}.

The flow-matching-based audio model will then generate mel-filterbank features $\tilde{s}$ based on the learned distribution of $P(\tilde{s}|[\hat{s}^{spk}; z^{text}], [a^{spk}; a^{text}], [h^{spk}; h^{NV}], [e^{spk}; e^{emo}])$, where z^{text} is an all-zero matrix with a shape of $F \times T^{text}$, and $[;]$ denotes concatenation operation in the time dimension. The generated part of $\tilde{s}$ is then converted to the speech signal using a vocoder.

3.2. NV embeddings

For our proposed framework, it is essential to figure out a suitable embedding that can represent the characteristics of various NVs. In ELaTE [14], an embedding obtained from an off-the-shelf laughter detection model [17, 18][1] was used to control laughter in zero-shot TTS.

One of our findings in this work is that this laughter detector-based embedding actually captures a broader range of NV types than just laughter. By appropriately using the laughter-detector-based embedding, we have successfully generated various NVs such as crying and moaning. Therefore, throughout this paper, we use a 32-dimensional embedding from the laughter detection model as the NV embedding.

3.3. Emotion embeddings

Based on the Russell's circumplex model of emotion, emotions can be represented in two major ways [19]: Firstly, emotions can be categorized into different emotion classes, such as happiness or sadness, reflecting distinct emotional states. Secondly, emotions can be described using two attributes, arousal, and valence, sometimes with the third attribute of dominance. Arousal refers to the level of intensity or activation of the emotion, ranging from calm to highly stimulated. Valence refers to how pleasant or unpleasant an emotion is, ranging from very positive to very negative. Dominance relates to how much control one feels over the situation.

Finding an effective emotion embedding is crucial for our framework. In our preliminary experiment, we used the eight emotion categories determined in [20] where each emotion category was represented by a learnable embedding, which is then used as e^{spk}. However, we found the TTS model struggled to generate emotionally expressive speech with this approach. We also explored using the prosody encoder of the FACodec [21]. However, we found that the output of the prosody encoder contains phonetic information, resulting in the speech generation following the contents of the emotion prompt audio rather than the text prompt.

Ultimately, we identify a promising representation: arousal and valence values predicted by a pre-trained arousal-valence-dominance extractor [22][2]. This extractor is initialized with a wav2vec 2 model [23] and fine-tuned on MSP-PODCAST data [20] to predict arousal, valence, and dominance values. Chunk-wise arousal-valence values ($D^{emo} = 2$) are extracted using a sliding window with a window size of 0.5 seconds and a hop size of 0.25 seconds. Because the extractor outputs each value in the range of 0.0 to 1.0, we subtract 0.5 from the estimated value to adjust the range from -0.5 to 0.5. We align the length of the extracted values with the phoneme embedding through linear interpolation. This representation allows for capturing more nuanced emotional variations within each utterance. Note that our preliminary investigations revealed that the additional use of the dominance value hurt the audio quality; therefore, we omitted the dominance value.

3.4. Collecting large-scale emotional data with pseudo-labeling

The quantity and quality of training data is a crucial factor for achieving high-quality TTS. However, either the recording of emotional speech or manual annotation of the recording is costly, making it difficult to scale the data size to more than 100 hours.

In this work, we curate 27k hours of highly emotional data, referred to as In-house Emotion Data (IH-EMO) in this paper, from 200k hours of in-house unlabeled anonymized English audio [24]. The data curation procedure is as follows. We first employ the emotion2vec model [25][3] to obtain predicted emotion confidence scores. We retain the samples if the predicted emotion is {angry, disgusted, fearful, sad, surprised} or the predicted emotion is {neutral, happy} with a confidence score of 1.0.[4] We further apply DNSMOS [26] and retain only samples whose OVLR score is greater than 3.0. Finally, we also apply an in-house speaker change detection model and discard the sample whenever a speaker change is detected. As a result, 27k hours of emotional audio are collected. We use an off-the-shelf speech recognition model[5] to obtain the transcription.

4. EXPERIMENTS

4.1. Data

4.1.1. Training data

We used three training datasets: Libri-light, LAUGH, and IH-EMO. Libri-light was used for pre-training the audio model without NV and emotion embeddings. On the other hand, LAUGH and IH-EMO were used for fine-tuning the audio model with NV and emotion embeddings. During fine-tuning, we also used the Libri-light data with a certain probability as suggested in [14]. The overview of each dataset is as follows.

Libri-light [27]: 60k hours of untranscribed English audiobooks from over 7,000 speakers. We transcribed the data by using a pre-trained Kaldi ASR model[5], which was trained on the 960-hour LibriSpeech dataset [28].

LAUGH: 460 hours of laughing speech collected from the AMI meeting corpus [29], Switchboard corpus [30], and Fisher corpus [31]. We gather all the utterances marked with laughter from the transcriptions of each corpus. Note that the dataset still contains a substantial amount of neutral speech, as laughter tends to occur at certain parts of the speech.

---

[2]https://github.com/audeering/w2v2-how-to
[3]https://github.com/ddlBoJack/emotion2vec
[4]The dominant predictions were neutral or happy. To balance the emotion category, we decided to exclude the samples with low confidence scores for these two emotion classes. Even after this procedure, these two emotion classes were still the top-2 categories in the collected data.
[5]https://kaldi-asr.org/models/m13

Table 2. Summary of evaluation datasets. S2ST: Speech-to-speech translation.

Dataset	Source Language	Utterances	Type	Task
JVNV S2ST	Japanese (ja)	1615	Staged	S2ST (ja→en) with various emotions
EMO-change	English (en)	84	Simulated	Zero-shot TTS with emotion change
Laughter-test	Chinese (zh)	154	Real	S2ST (zh→en) with laughter
Crying-test	Chinese (zh)	33	Real	S2ST (zh→en) with crying

IH-EMO: 27k hours of collected emotional speech as described in Section 3.4.

4.1.2. Evaluation data

We used four evaluation datasets as presented in Table 2.

JVNV speech-to-speech translation (S2ST): To evaluate the emotion transferability of zero-shot TTS models, we established an experimental setting based on the Japanese-to-English S2ST scenario. In the evaluation, we used a Japanese speech from the JVNV corpus [32]. The JVNV corpus includes speeches from four speakers (two males and two females) with six emotions: anger, disgust, fear, happiness, sadness, and surprise with various NVs. With its intense emotional expressions and various NVs, the JVNV dataset offers comprehensive coverage of expressive emotions, making it ideal for our emotion transfer testing. In the evaluation process, we first applied speech-to-text translation to the Japanese speech to obtain the English translations. We then used a zero-shot TTS model, using the English translations as a text prompt and the Japanese speech as an audio prompt, from which we extracted both NV and emotion embeddings. The generated speech is expected to be English-translated speech with the original speaker's voice and emotional characteristics. We assessed the similarity of the speaker and emotion between the source Japanese speech and the translated English speech using various metrics described in the next section.

EMO-change: To test the model's capacity for fine-grained emotional speech generation, we created the EMO-change dataset based on the RAVDESS dataset [33], which contains English emotional speech data with emotions such as calm, happy, sad, angry, fearful, surprised, and disgusted. The RAVDESS dataset includes two transcriptions: "kids are talking by the door" and "dogs are sitting by the door" with intense emotional expressions. To generate the EMO-change dataset, we randomly select two utterances with different emotions with the transcription "kids are talking by the door," remove the silence from each, and concatenate them to create the audio prompts. During testing, the concatenated emotion change sample serves as the audio prompt, and a repeated sentence "dogs are sitting by the door dogs are sitting by the door" serves as the text prompt for the zero-shot TTS model. This setup evaluates the model's ability to generate speech that mimics the emotional transitions in the audio prompt while following the given text prompt, and speaker characteristics.

Laughter-test: To evaluate the zero-shot TTS capability in generating laughing speech, we employed a Chinese-to-English S2ST experiment by following the protocol presented in [14]. Specifically, we used 154 Chinese utterances containing laughter[6] from the evaluation subset of the DiariST-AliMeeting test set [34]. We applied zero-shot TTS, where we used the ground-truth English transcription as the text prompt and the Chinese speech as the audio prompt to extract both NV and emotion embeddings. The generated speech is expected to be English speech with the original speaker's voice and laughter characteristics.

---

[6]Data can be found at https://aka.ms/elate.

Crying-test: To evaluate the zero-shot TTS capability in generating crying speech, we further conducted a Chinese-to-English S2ST experiment using 33 Chinese crying speech samples collected from publicly available data sources. We followed the same procedure with the Laughter-test, and the generated speech is expected to be English speech with the original speaker's voice and crying characteristics.

4.2. Evaluation metrics

4.2.1. Objective evaluation metrics

We utilized the following objective metrics. Among them, AutoPCP, EMO SIM, and Aro-Val SIM are closely related to the emotion controllability.[7]

Word error rate (WER): To evaluate the intelligibility of the generated audio, we applied a Whisper-Large [35] to the generated audio and computed the WER. In all our tables, we express the WER in percentage.

Speaker SIM-o: To evaluate the speaker similarity between the generated audio and the audio prompt, we computed a cosine similarity between speaker embeddings of the two audios. We used a WavLM-large-based speaker verification model [36][8], following previous works [13, 37].

AutoPCP: AutoPCP [38] is an utterance-level estimator to quantify the prosody similarity between two speech samples. We computed the score between the generated audio and the audio prompt. We leveraged AutoPCP_multilingual_v2[9].

Emo SIM: To evaluate the similarity of time-varying emotion states, we applied the emotion2vec model [25][3] to extract the emotion embeddings. We performed interpolation to ensure the embeddings of the audio prompt and the generated audio have the same length. We then computed the cosine similarity between these two embedding sequences for each frame and took the average to obtain the EMO SIM score.

Aro-Val SIM: As another metric for the similarity of time-varying emotion states, we computed the arousal-valence values based on [22][2] using a sliding window with a window size of 0.5 sec and a hop size of 0.25 sec. Similar to EMO SIM, we computed the cosine similarity between the audio prompt and the generated audio for every frame and took the average as the Aro-Val SIM.

4.2.2. Subjective evaluation metrics

We used the following subjective evaluation metrics.

SMOS: Speaker similarity mean opinion score, which is the similarity between the speaker prompt and the generated speech from 1 (not at all similar) to 5 (extremely similar).

NMOS: Naturalness MOS, which is the naturalness of the generated speech from 1 (bad) to 5 (excellent).

EMOS: Emotion MOS, which measures the similarity of emotion between the audio prompt and the generated speech from 1 (not at all similar) to 5 (extremely similar).

4.3. Model configuration

The architecture of the EmoCtrl-TTS audio model closely followed the configurations of the Voicebox [13]. Specifically, we used a Transformer [39] with 24 layers, featuring 16 attention heads, a 1024-dimensional embedding, and a feed-forward layer with a dimension of 4096.

---

[7]https://github.com/hbwu-ntu/EmoCtrlTTS-Eval
[8]https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification
[9]https://github.com/facebookresearch/seamless_communication

Table 3. Objective evaluation results for various models on JVNV S2ST and EMO-change test sets. A model with [(+)] was fine-tuned with 200k steps with more exposure to the IH-EMO. LL: Libri-light.

ID	Model	Init.	h	e	Training Data (hours)	SIM-o↑	WER (%)↓	AutoPCP↑	Emo SIM↑	Aro-Val SIM↑
					JVNV S2ST					
(B1)	SeamlessExpressive [38]	-	-	-	-	0.268	**1.2**	2.91	0.653	0.494
(B2)	Voicebox (reproduction) [13]	-	-	-	LL (60k)	0.347	2.1	2.96	0.655	0.443
(B3)	ELaTE [14]	B2	✓	-	LL (60k) + LAUGH (460)	0.441	3.8	3.36	0.671	0.548
(B4)	Voicebox (fine-tuned)	B2	-	-	LL (60k) + LAUGH (460)	0.410	2.5	3.07	0.645	0.438
(B5)	Voicebox (fine-tuned)	B2	-	-	LL (60k) + IH-EMO (27k)	0.479	2.2	2.96	0.641	0.397
(B6)	Voicebox (fine-tuned)	B2	-	-	LL (60k) + IH-EMO (27k) + LAUGH (460)	0.455	3.0	3.17	0.659	0.470
(P1)	EmoCtrl-TTS	B2	✓	✓	LL (60k) + IH-EMO (27k) + LAUGH (460)	0.448	4.4	3.38	0.693	**0.647**
(P2)	EmoCtrl-TTS[(+)]	B2	✓	✓	LL (60k) + IH-EMO (27k) + LAUGH (460)	**0.497**	3.2	**3.50**	**0.697**	0.643
					EMO-change					
(B2)	VoiceBox (reproduction) [13]	-	-	-	LL (60k)	0.600	1.2	3.31	0.685	0.663
(B3)	ELaTE [14]	B2	✓	-	LL (60k) + LAUGH (460)	0.643	0.2	**3.52**	**0.700**	0.761
(B6)	Voicebox (fine-tuned)	B2	-	-	LL (60k) + IH-EMO (27k) + LAUGH (460)	0.622	1.1	3.31	0.678	0.655
(P1)	EmoCtrl-TTS	B2	✓	✓	LL (60k) + IH-EMO (27k) + LAUGH (460)	0.671	**0.0**	3.45	0.685	**0.822**
(P2)	EmoCtrl-TTS[(+)]	B2	✓	✓	LL (60k) + IH-EMO (27k) + LAUGH (460)	**0.684**	0.9	3.44	0.679	0.811

Table 4. Subjective evaluation results on the JVNV S2ST test set are presented. The group with the top score (scores within the 95% confidence interval of the highest score) is displayed in **bold** font.

ID	Model	SMOS	NMOS	EMOS
		JVNV S2ST		
(B1)	SeamlessExpressive [38]	$2.40_{\pm0.18}$	$2.83_{\pm0.16}$	$3.46_{\pm0.17}$
(B2)	Voicebox (repro.) [13][†]	$3.12_{\pm0.21}$	$3.28_{\pm0.15}$	$3.51_{\pm0.16}$
(B3)	ELaTE [14]	$\mathbf{3.62}_{\pm0.20}$	$\mathbf{3.43}_{\pm0.14}$	$\mathbf{3.68}_{\pm0.15}$
(B6)	Voicebox (fine-tuned)	$\mathbf{3.76}_{\pm0.18}$	$3.29_{\pm0.14}$	$\mathbf{3.67}_{\pm0.16}$
(P2)	EmoCtrl-TTS	$\mathbf{3.72}_{\pm0.18}$	$\mathbf{3.53}_{\pm0.13}$	$\mathbf{3.66}_{\pm0.15}$
		EMO-change		
(B3)	ELaTE [14]	$4.37_{\pm0.09}$	$\mathbf{3.56}_{\pm0.12}$	$\mathbf{3.65}_{\pm0.13}$
(B6)	Voicebox (fine-tuned)	$4.36_{\pm0.10}$	$\mathbf{3.53}_{\pm0.13}$	$2.95_{\pm0.13}$
(P2)	EmoCtrl-TTS	$\mathbf{4.49}_{\pm0.08}$	$\mathbf{3.55}_{\pm0.12}$	$3.54_{\pm0.14}$

The model was pre-trained using Libri-light data, without NV or emotion embedding. This resulted in the reproduction of the Voicebox model (B2 in Table 3). The model was trained for 390K steps with an effective mini-batch size of 307,200 audio frames. A linear-decay learning rate scheduler with a peak learning rate at 7.5e-5 was used, along with 20K steps of linear warmup. After the pre-training, we further fine-tuned the model by combining Libri-light, LAUGH, and IH-EMO. During fine-tuning, the effective mini-batch size was set to 307,200 audio frames. A linear-decay learning rate scheduler was used with a peak learning rate of 7.5e-5. Unless otherwise stated, we fine-tuned the model with 40k steps.

During the inference, we used classifier-free guidance with a guidance strength of 1.0, and the number of function evaluations was set to 32. A MelGAN-based vocoder [40] was used to convert the mel spectrogram into waveforms.

4.4. S2ST pipeline

For the JVNV S2ST evaluation data, we leveraged the Whisper large-v3 model [35] to transcribe the Japanese utterances with time stamps. We then employed GPT-4 [41] to translate the time-stamped Japanese text into English by keeping the time stamps. We then used a total-duration-aware (TDA) duration model [42] to obtain a frame-wise phoneme alignment. For the Laughter-test and Crying-test, we used the ground-truth English text translation, and the same TDA duration model to obtain the frame-wise phoneme alignment.

4.5. Results and discussion

4.5.1. Objective evaluation

Table 3 presents the objective evaluation results on the JVNV and EMO-change datasets. During the evaluation, we generated speech with three different random seeds, and the average of the scores was taken to report.

Baseline analysis: For the JVNV S2ST test set, we first observed that the ELaTE (B3), trained with LL and LAUGH data, achieved superior performance over the SeamlessExpressive model [38][10] (B1) and the reproduced VoiceBox model (B2) across all the metrics, except for WER. The comparison between B2 and B4 highlights the effect of LAUGH data, showing that the latter improved performance in SIM-o and AutoPCP while maintaining WER, EMO SIM, and Aro-Val SIM. By replacing LAUGH with the IH-EMO data (B4 vs. B5), SIM-o and WER were improved from 0.410 to 0.479, and from 2.5% to 2.2%, respectively. However, this came at the cost of a slight degradation of AutoPCP (3.07→2.96), Emo SIM (0.645→0.641), and Aro-Val SIM (0.438→0.397). Combining all training data (B6) gave us a balanced result. However, even after these trials, no Voicebox model achieved better AutoPCP, EmoSIM, and Aro-Val SIM compared to ELaTE. The same trends were also observed for the EMO-change test set. These results suggest that simply adding emotional data does not improve the emotion transferability of the zero-shot TTS models.

Results of EmoCtrl-TTS: From JVNV S2ST results, we first observed (P1) showed a significant improvement in AutoPCP, Emo SIM, and Aro-Val SIM compared to the baselines. By executing a longer fine-tuning (P2), EmoCtrl-TTS achieved further better SIM-o, WER, AutoPCP, and Emo SIM while almost preserving the high Aro-Val SIM. For the EMO-change test set, EmoCtro-TTS achieved the best SIM-o and Aro-Val SIM and the second-best AutoPCP. It is also noteworthy that SIM-o score was significantly improved from (B6) to (P1), which suggests the effectiveness of the use of NV and emotion embeddings not only for the emotion-related metrics but also for the speaker similarity. Note that we have trained more variants of EmoCtrl-TTS to analyze the impact of each data and each embedding, which will be discussed in Section 4.5.3 with Table 5.

---

[10]Our experiment is based on SeamlessExpressive, supported by the Seamless Licensing Agreement. Copyright © Meta Platforms, Inc. All Rights Reserved.

Table 5. Impact of training configurations on the EmoCtrl-TTS on JVNV S2ST test set.

Training Data (hours)	Data Ratio	Steps	IH-EMO h	IH-EMO e	LAUGH h	LAUGH e	SIM-o↑	WER (%)↓	AutoPCP↑	Emo SIM↑	Aro-Val SIM↑
LL (60k) + LAUGH (460)	0.5:0.5	40k					0.410	2.5	3.07	0.645	0.438
LL (60k) + LAUGH (460) (= ELaTE)	0.5:0.5	40k			✓		0.441	3.8	3.36	0.671	0.548
LL (60k) + LAUGH (460)	0.5:0.5	40k				✓	0.332	4.7	3.22	0.680	0.632
LL (60k) + LAUGH (460)	0.5:0.5	40k			✓	✓	0.391	4.4	3.39	0.702	**0.663**
LL (60k) + IH-EMO (27k)	0.5:0.5	40k					0.479	**2.2**	2.96	0.641	0.397
LL (60k) + IH-EMO (27k)	0.5:0.5	40k	✓				0.403	6.7	3.14	0.687	0.590
LL (60k) + IH-EMO (27k)	0.5:0.5	40k		✓			0.487	2.5	3.36	0.682	0.608
LL (60k) + IH-EMO (27k)	0.5:0.5	40k	✓	✓			0.438	7.7	3.29	**0.707**	0.637
LL (60k) + IH-EMO (27k) + LAUGH (460)	0.5:0.25:0.25	40k					0.455	3.0	3.17	0.659	0.470
LL (60k) + IH-EMO (27k) + LAUGH (460)	0.5:0.25:0.25	40k		✓	✓		0.448	4.4	3.38	0.693	0.647
LL (60k) + IH-EMO (27k) + LAUGH (460)	0.5:0.4:0.1	40k		✓	✓		0.487	3.4	3.44	0.690	0.626
LL (60k) + IH-EMO (27k) + LAUGH (460)	0.5:0.4:0.1	200k		✓	✓		**0.497**	3.2	**3.50**	0.697	0.643

Table 6. Results of Laughter-test dataset.

ID	Model	SIM-o↑	WER↓	AutoPCP↑	Emo SIM↑	Aro-Val SIM↑
(B1)	SeamlessExpressive [38]	0.210	11.4	2.31	0.587	0.248
(B2)	Voicebox (repro.) [13][†]	0.328	7.8	2.46	0.634	0.410
(B3)	ELaTE [14]	**0.399**	7.8	**3.44**	0.806	0.700
(B6)	Voicebox (fine-tuned)	0.383	**6.6**	2.47	0.689	0.410
(P2)	EmoCtrl-TTS	0.392	9.9	3.38	**0.848**	**0.795**

Table 7. Results of Crying-test dataset.

ID	Model	SIM-o↑	WER↓	AutoPCP↑	Emo SIM↑	Aro-Val SIM↑
(B1)	SeamlessExpressive [38]	0.258	8.9	2.77	0.576	0.378
(B2)	Voicebox (repro.) [13][†]	0.294	5.5	2.71	0.569	0.367
(B3)	ELaTE [14]	0.384	5.6	3.14	0.642	0.471
(B6)	Voicebox (fine-tuned)	0.383	**5.1**	2.81	0.589	0.413
(P2)	EmoCtrl-TTS	**0.408**	7.5	**3.21**	**0.662**	**0.597**

4.5.2. Subjective evaluation

We performed the subjective evaluation for selected models. We randomly picked 24 samples from JVNV S2ST data and 28 samples from EMO-change data, and asked native English testers to rate SMOS, NMOS and EMOS. For JVNV S2ST test set, each sample was judged by 9, 11, and 10 testers for SMOS, NMOS, and EMOS, respectively. For EMO-change test set, each sample was judged by 12 testers for all metrics.

The results are presented in Table 4. From the JVNV S2ST result, the Voicebox model (B6), leveraging the IH-EMO data, demonstrated significant improvements of SMOS and EMOS from the vanilla Voicebox model (B2), which illustrates the importance of the training data. Meanwhile, ELaTE (B3) excelled in NMOS, suggesting that the integration of NV features significantly boosts the naturalness of the audio by generating an appropriate NV. Finally, EmoCtrl-TTS (P2) achieved even better NMOS while keeping the high SMOS and EMOS scores.

In the EMO-change dataset, Voicebox (B6) showed significantly worse EMOS compared to ELaTE (B3) and EmoCtlr-TTS (P2). This result demonstrated that the conventional zero-shot TTS cannot mimic the time-varying emotional states presented in the audio prompt. The proposed EmoCtrl-TTS (P2) achieved best SMOS and comparable NMOS and EMOS to other top baselines, again showcasing the efficacy of the proposed approach.

4.5.3. Impact of training configurations

Table 5 presents the results of the JVNV S2ST test set with various training data configurations.

The first four rows compare the performance of models trained by Libri-light and LAUGH data with a 0.5:0.5 ratio. By comparing the first two rows, we observe that the NV embedding h significantly improved SIM-o, AutoPCP, Emo SIM, and Aro-Val SIM with the expense of degradation of WER. By introducing emotion embedding e,

AutoPCP, EmoSIM, and Aro-Val SIM were further improved. However, it came with a cost of further degradation of WER as well as a significant drop in SIM-o.

In the 5th to 9th rows, we trained models using the Libri-light and IH-EMO data. Similar to the case with LAUGH data, we observed that the emotion embedding e improved all the emotion-related metrics while keeping SIM-o and WER nearly intact. However, in contrast to the results with the LAUGH data, we observed a severe degradation in WER when the NV embedding h was introduced. Upon examination, we found that the NV embedding often resulted in unwanted NV generation, such as laughter from multiple speakers.

Both observations suggest that including both NV and emotion embeddings in model training is not always beneficial. Based on these findings, we opted to use only emotion embeddings for IH-EMO and only NV embeddings for LAUGH when combining all training data.

The last four rows in Table 5 show the impact of data ratio and training steps on model performance with the combination of all training data. We first observed that the data ratio of 0.5:0.4:0.1 for Libri-light, IH-EMO, and LAUGH provided us with a better SIM-o and WER than the ratio of 0.5:0.25:0.25 while keeping the emotion-related metrics similar. By fine-tuning the model longer, further marginal improvement was obtained for all evaluation metrics.

4.5.4. Results on real laughter and crying data

Table 6 and 7 present a comparison between the EmoCtrl-TTS model and selected baselines on the Laughter-test and Crying-test datasets, both of which are composed of real data. Compared to the baselines, the EmoCtrl-TTS consistently achieved the best performance for Emo SIM and Aro-Val SIM while achieving either the best or the second-best Auto PCP and SIM-o. It demonstrated EmoCtrl-TTS's robustness for the real data. On the other hand, we observe a moderate degradation of WERs.

5. CONCLUSIONS

In this work, we presented EmoCtrl-TTS, an emotion-controllable zero-shot TTS model that can generate highly emotional speech with NVs for any speaker. EmoCtrl-TTS leveraged arousal and valence values as well as the laughter embeddings to control the time-varying characteristics of emotional speech including NVs. Our comprehensive experiments demonstrated that EmoCtrl-TTS can closely mimic the voice characteristics and nuances of the source audio prompts by generating emotional speech with NVs.

6. REFERENCES

[1] Wei Zhao and Zheng Yang, "An emotion speech synthesis method based on VITS," *Applied Sciences*, vol. 13, no. 4, pp. 2225, 2023.

[2] Haobin Tang, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang, "ED-TTS: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis," *arXiv preprint arXiv:2401.08166*, 2024.

[3] Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu, "EmoDiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[4] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao, "EmoMix: Emotion mixing via diffusion models for emotional speech synthesis," *arXiv preprint arXiv:2306.00648*, 2023.

[5] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, 2022.

[6] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao, "QI-TTS: Questioning intonation control for emotional speech synthesis," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[7] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie, "MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[8] Yookyung Shin, Younggun Lee, Suhee Jo, Yeongtae Hwang, and Taesu Kim, "Text-driven emotional style control and cross-speaker style transfer in neural tts," *arXiv preprint arXiv:2207.06000*, 2022.

[9] Younggun Lee, Azam Rabiee, and Soo-Young Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.

[10] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *ISCSLP 2021*. IEEE, 2021, pp. 1–5.

[11] Xiong Cai, Dongyang Dai, Zhiyong Wu, Xiang Li, Jingbei Li, and Helen Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021*. IEEE, 2021, pp. 5734–5738.

[12] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee, "EmoSphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech," *arXiv e-prints*, pp. arXiv–2406, 2024.

[13] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al., "VoiceBox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[14] Naoyuki Kanda, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Canrun Li, Steven Tsai, Zhen Xiao, et al., "Making flow-matching-based zero-shot text-to-speech laugh as you like," *arXiv preprint arXiv:2402.07383*, 2024.

[15] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le, "Flow matching for generative modeling," in *ICLR 2023*, 2023.

[16] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[17] Kimiko Ryokai, Elena Durań López, Noura Howell, Jon Gillick, and David Bamman, "Capturing, representing, and interacting with laughter," in *CHI 2018*, 2018, pp. 1–12.

[18] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman, "Robust laughter detection in noisy environments.," in *Interspeech*, 2021, vol. 8, p. 2021.

[19] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[20] Reza Lotfian and Carlos 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.

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

[22] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[23] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[24] Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Yufei Xia, Jinzhu Li, Sheng Zhao, Jinyu Li, et al., "An investigation of noise robustness for flow-matching-based zero-shot tts," in *Interspeech*, 2024.

[25] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[26] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022*. IEEE, 2022, pp. 886–890.

[27] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-Light: A benchmark for ASR with limited or no supervision," in *ICASSP 2020*. IEEE, 2020, pp. 7669–7673.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP 2015*. IEEE, 2015, pp. 5206–5210.

[29] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *ICMI 2005*. Springer, 2005, pp. 28–39.

[30] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP 1992*. IEEE Computer Society, 1992, vol. 1, pp. 517–520.

[31] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher corpus: A resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[32] Detai Xin, Junfeng Jiang, Shinnosuke Takamichi, Yuki Saito, Akiko Aizawa, and Hiroshi Saruwatari, "JVNV: A corpus of Japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, 2024.

[33] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[34] Mu Yang, Naoyuki Kanda, Xiaofei Wang, Junkun Chen, Peidong Wang, Jian Xue, Jinyu Li, and Takuya Yoshioka, "DiariST: Streaming speech translation with speaker diarization," in *ICASSP 2024*. IEEE, 2024, pp. 10866–10870.

[35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML 2023*. PMLR, 2023, pp. 28492–28518.

[36] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

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

[38] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al., "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

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

[40] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[41] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[42] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Jinyu Li, et al., "Total-duration-aware duration modeling for text-to-speech systems," *arXiv e-prints*, pp. arXiv–2406, 2024.