EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control

emoknob.cs.columbia.edu

Haozhe Chen
Columbia University
hc3295@columbia.edu

Run Chen
Columbia University
runchen@cs.columbia.edu

Julia Hirschberg
Columbia University
julia@cs.columbia.edu

Abstract

While recent advances in Text-to-Speech (TTS) technology produce natural and expressive speech, they lack the option for users to select emotion and control intensity. We propose EmoKnob, a framework that allows fine-grained emotion control in speech synthesis with few-shot demonstrative samples of arbitrary emotion. Our framework leverages the expressive speaker representation space made possible by recent advances in foundation voice cloning models. Based on the few-shot capability of our emotion control framework, we propose two methods to apply emotion control on emotions described by open-ended text, enabling an intuitive interface for controlling a diverse array of nuanced emotions. To facilitate a more systematic emotional speech synthesis field, we introduce a set of evaluation metrics designed to rigorously assess the faithfulness and recognizability of emotion control frameworks. Through objective and subjective evaluations, we show that our emotion control framework effectively embeds emotions into speech and surpasses emotion expressiveness of commercial TTS services.[1]

1 Introduction

The complexity of human communication extends far beyond mere verbal exchange. Vocal inflections and emotional undertones play pivotal roles in conveying meaning. While text alone can be ambiguous in meaning (Jenkins, 2020), different emotions in voices can articulate different messages in the same piece of text (Nygaard and Lunders, 2002). Consider Shakespeare's iconic phrase, *To be or not to be*. This line can express despair, contemplation, defiance, or resignation, depending on the speaker's emotional delivery, illustrating the profound impact of vocal emotions in communication.

The ultimate objective in the field of conversational systems is to develop intelligent agents capable of comprehending, deciding, and synthesizing speech with nuanced emotional undertones. While recent advances in Text-to-Speech (TTS) technology have achieved remarkable naturalness and expressiveness in synthesized voices(ElevenLabs; OpenAI, 2024b; Microsoft), these systems lack the capability for users to select and control the emotional tone and intensity. The emotion conveyed in the generated speech is solely determined by the text, without allowing for variability or intensity control.

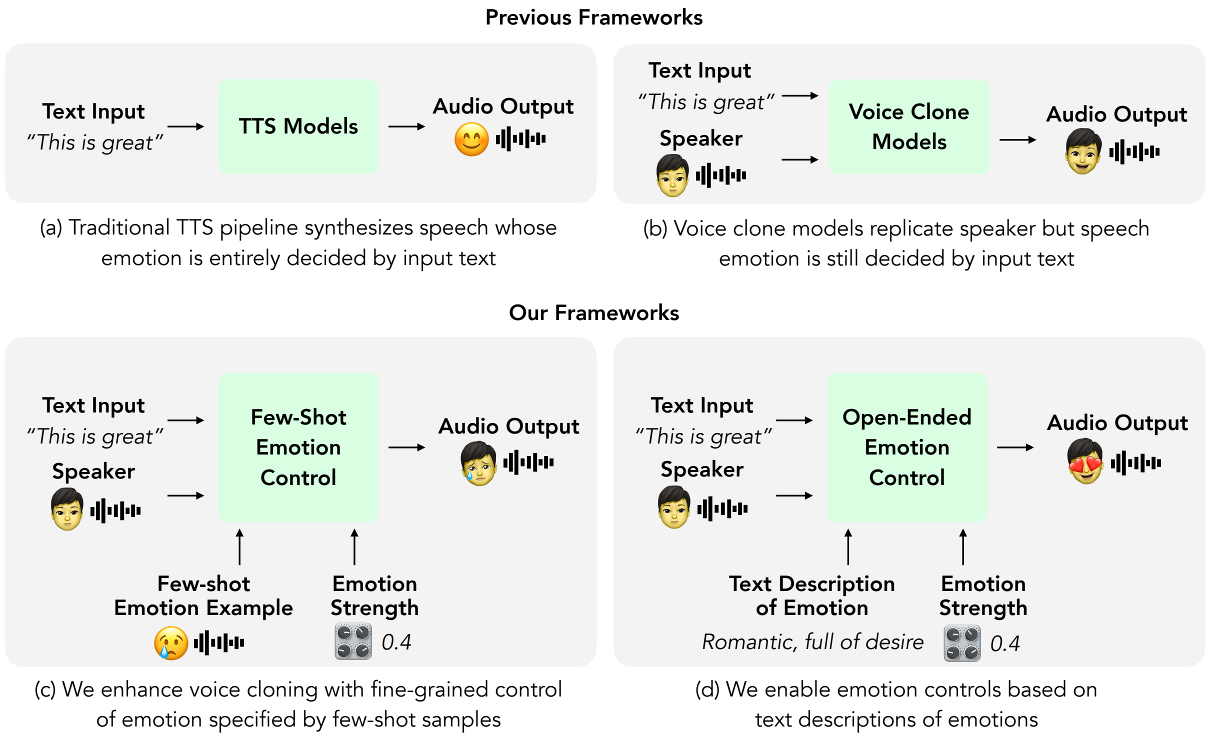
Previous works on emotion control in speech synthesis primarily focus on a few simple emotion categories (Lei et al., 2022; Lorenzo-Trueba et al., 2018; Kang et al., 2023; Qin et al., 2024). These methods do not allow control of a more diverse array of emotions. Synthesis for more complex and heterogeneous emotions like charisma (Yang et al., 2020) and empathy (Chen et al., 2024) is not well studied.

Our work leverages recent breakthroughs in foundation models for voice cloning (MetaVoice; Anastassiou et al., 2024; suno.ai, 2023; Casanova et al., 2024; Shen et al., 2018). By exploring the rich expressiveness in these models' latent embedding spaces, we develop methods to extract a representation for any emotion with just a few demonstrative samples. These representations are inherently synergistic with the speech generation capabilities of rapidly advancing voice cloning/TTS models, enabling us to generate high quality speech while applying fine-grained emotion controls. This approach proves effective for both simple and complex emotions and includes mechanisms to adjust emotional intensity with a scalar knob.

Our framework's capability of applying fine-grained emotion control for any emotion with a few demonstrative examples enables us to propose two methods for applying emotion control based

[1] See audio samples, code, and live demo at emoknob.cs.columbia.edu.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8170–8180
November 12-16, 2024 ©2024 Association for Computational Linguistics



(a) Traditional TTS pipeline synthesizes speech whose emotion is entirely decided by input text

(b) Voice clone models replicate speaker but speech emotion is still decided by input text

(c) We enhance voice cloning with fine-grained control of emotion specified by few-shot samples

(d) We enable emotion controls based on text descriptions of emotions

Figure 1: Fine-grained emotion control with EmoKnob. While existing TTS and voice cloning frameworks lack the option for users to control emotions in speech, our framework allows users to embed arbitrary emotion with a specified intensity in speech with few-shot samples. This framework allows us to propose two methods for controlling emotions based on open-ended text descriptions of emotions.

on arbitrary text descriptions of emotions. We use a synthetic-data-based and a retrieval-based method to leverage recent advances in Large Language Models (LLMs) and text embedding models (OpenAI; Meng et al., 2024), in conjunction with our few-shot emotion control framework, to address a lack of open-ended captioned emotional speech dataset.

We recognize that emotion control in speech synthesis is still at its early stage, and traditional evaluation metrics for TTS systems cannot comprehensively evaluate emotion control frameworks. We therefore introduce a set of rigorous evaluation metrics designed to systematically measure the effectiveness of an emotion control framework at faithfully conveying recognizable emotions.

With a set of subjective and objective evaluations, we show that our framework produces faithful and recognizable emotion control on speech. We find that 83% of the participants consider that speech with emotion enhancement by our framework surpasses leading commercial TTS services at conveying these emotions.

	Expressive Emotion Control	Few-Shot Emotion Control	Open-Ended Emotion Control	Synergetic with TTS Model Advances
Classifier-Based Style Transfer[1]	✓	✗	✗	✗
Domain Adversarial Training [2]	✓	✓	✗	✗
Voice Text Descriptions[3]	✗	✗	✓	✗
Ours	✓	✓	✓	✓

Table 1: Comparison between our framework and prior works on emotion control in speech synthesis. Our framework allows few-shot emotion control of arbitrary emotions and is synergetic with rapidly advancing text-to-speech models. We also propose two frameworks that allow users to control emotions with open-ended text emotion description. [1]Lei et al. (2022); Lorenzo-Trueba et al. (2018); Kang et al. (2023); Qin et al. (2024). [2] Jo et al. (2023). [3]Guo et al. (2022); Yang et al. (2023); Lacombe et al. (2024); Lyth and King (2024).

2 Related Work

2.1 Foundational Model for TTS and Voice Cloning

Large foundational models have become the basis of many machine learning fields such as text (OpenAI et al., 2024) and images (Radford et al., 2021). These large foundational models are trained in an unsupervised manner with massive datasets and learn high quality representations of data, which are commonly used directly or through fine-tuning for downstream tasks.

The TTS domain also sees a rising trend in large, foundational models. These end-to-end models trained on large corpora provide natural speech rendering from text. MetaVoice trains a 1.2B parameter model with 100K hours of speech for TTS; Lajszczak et al. (2024) trains a 1B parameter model on 100K open-domain speech data. Many of these models are capable of replicating a speaker's voice in zero-shot or few-shots (MetaVoice; Anastassiou et al., 2024; suno.ai, 2023; Casanova et al., 2024; Shen et al., 2018). Our work explores how to leverage the high quality speaker representation learned by these foundational models to enhance voice cloning with few-shot fine-grained emotion control. In particular, we focus on manipulating the latent speaker embedding in MetaVoice.

2.2 Emotion and Style Control in Speech Synthesis

While models discussed in Section 2.1 and existing commercial services (OpenAI, 2024b; Microsoft; ElevenLabs) produce natural sounding speech, their speech output's emotions are primarily decided by input text, and the emotion strength cannot be controlled. Users thus cannot select arbitrary emotions for a piece of text. However, emotions expressed through acoustic-prosody serve an important additional channel for conveying information (Gobl and Chasaide, 2003; Patel et al., 2011; Laukkanen et al., 1997).

Previous work trains latent speech style space on small corpora and cannot generalize to style transfer beyond the training corpus (Zhang et al., 2019). In addition, existing labeled emotional speech datasets (Martinez-Lucas et al.; Poria et al., 2019) are limited to a few categories of basic emotions. Previous work thus commonly bases emotion controls on categorical emotion label inputs and is limited in types of emotions that can be controlled (Lei et al., 2022; Lorenzo-Trueba et al., 2018; Kang et al., 2023; Qin et al., 2024). Extending these methods to control new emotions require extensive retraining of models, preventing expressive emotion control over many emotions. These methods' requirement on large labeled datasets also prevents emotional control on more complex, nuanced emotions represented by more specialized, heterogeneous datasets such as charisma (Yang et al., 2020) and empathy (Chen et al., 2024).

While Jo et al. (2023) uses domain adversarial training to achieve few-shot emotion transfer, their method requires training a style encoder built from scratch and is not compatible with existing and future large foundational models. Thus, it is unable to improve naturalness and expressiveness in current and future TTS model developments. Our work provides a training-free framework that leverages a foundation model's TTS capability for single/few-shot emotion control and is inherently synergetic with growing foundation speech models.

2.3 Open-ended Text Prompt Control on Voice

A recent strand of works use text description to control voices. Guo et al. (2022); Yang et al. (2023); Lacombe et al. (2024); Lyth and King (2024) allow users to describe qualities such as tone, pitch, gender, and emotions of a voice before synthesizing speech with the described voice. While existing speech datasets lack text descriptions of voices, this work bypasses this obstacle by creating synthetic text captions based on acoustic-prosodic features and speaker metadata. These methods do not generalize well to text descriptions beyond the format and the scope of the synthetic captions. The emotion control with these methods is limited to the categorical emotion labels in speaker metadata. These methods also do not allow voice cloning and emotion variation on an unseen speaker. Based on our method's capability to enhance voice cloning with single/few samples, we propose retrieval and synthetic data based frameworks for synthesizing expressive emotions with open-ended text descriptions.

3 Methods

We apply fine-grained emotion controls by manipulating the speaker embedding space of pre-trained foundation voice cloning models. This framework allows us to apply emotion control with few-shot emotional speech samples. The few-shot capability enables us to design two frameworks for applying control with emotion specified by arbitrary text descriptions.

3.1 Preliminaries: Pre-Trained Foundational Voice Cloning Model

Existing voice cloning models (MetaVoice; Anastassiou et al., 2024; suno.ai, 2023; Casanova et al., 2024; Shen et al., 2018) can be abstracted into a two-stage architecture with a speaker encoder E

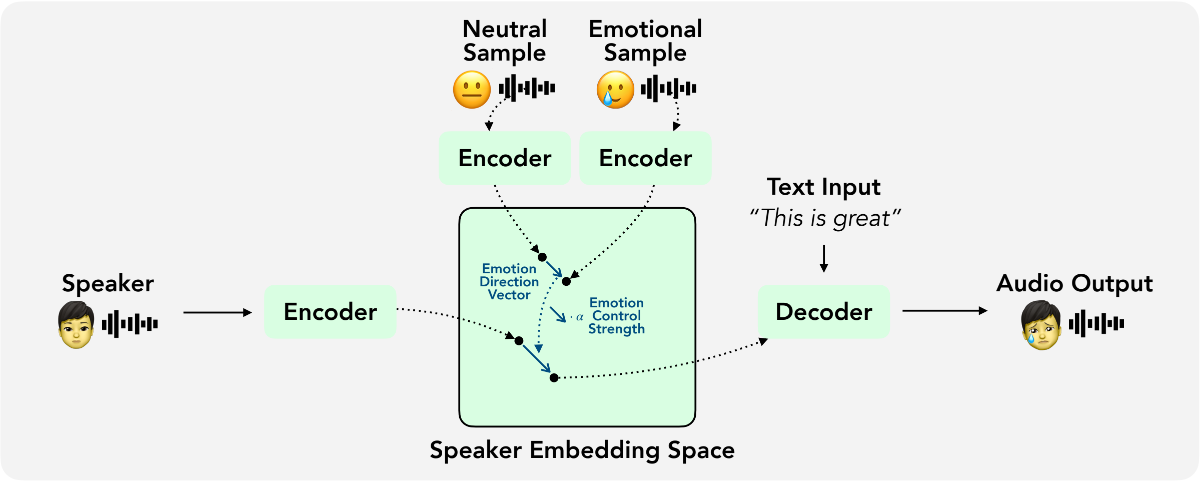


Figure 2: EmoKnob's few-shot emotion control pipeline. EmoKnob first extracts an emotion direction vector in speaker embedding space of pre-trained foundation voice cloning models with a pair of neutral and emotional sample. Then, EmoKnob manipulates the reference speaker's embedding with the obtained emotion direction vector and a specified emotion strength to embed the emotion into speech.

that takes in a speaker reference clip x_s and outputs a speaker embedding u_s. A conditional text-to-speech decoder D then takes in input text I to output speech audio $y_{s,I} = D(u_s, I)$ that utters I replicating speaker's voice. We will manipulate the speaker embedding space (output space of E and conditional input space of D) to obtain an emotion representation and obtain few-shot emotion control.

3.2 Few-Shot Fine-Grained Emotion Control

We hypothesize that a pre-trained foundation voice cloning model's speaker embedding provides expressive representations for acoustic-prosodic qualities. Our framework disentangles how an speaker embedding represents speaker-specific qualities and speaker-independent emotions. We then use the speaker-independent emotions obtained to apply fine-grained emotion control on an arbitrary speaker representation. We show this process for few-shot fine-grained emotion control in Figure 2.

We disentangle speaker-specific qualities and speaker-independent emotion representations by using paired samples of emotional speech x_e^i and neutral speech x_n^i from the same speaker. We encode representations u_e^i, u_n^i for these i-th pairs of samples in a speaker embedding space with the pre-trained speaker encoder E: $u_e^i = E(x_e^i), u_n^i = E(x_n^i)$.

We hypothesize that taking their difference results in a speaker-independent emotion direction vector v_e^i. In addition, we normalize v_e^i for convenient fine-grained emotion strength control later: $v_e^i = \frac{u_e^i - u_n^i}{||u_e^i - u_n^i||}$

We can obtain the emotion direction vector by averaging over many pairs of samples. We will show in experiments that single-shot ($N = 1$) suffices to produce high-quality emotion control in many cases:

$$v_e = \frac{1}{N} \sum_{i=1}^{N} \frac{u_e^i - u_n^i}{||u_e^i - u_n^i||}$$

Given a new speaker reference sample x_s, we hope to replicate the speaker's voice qualities while controlling emotions in an utterance. We first obtain the reference speaker's speaker embedding with $u_s = E(x_s)$. Then, we apply emotion control with

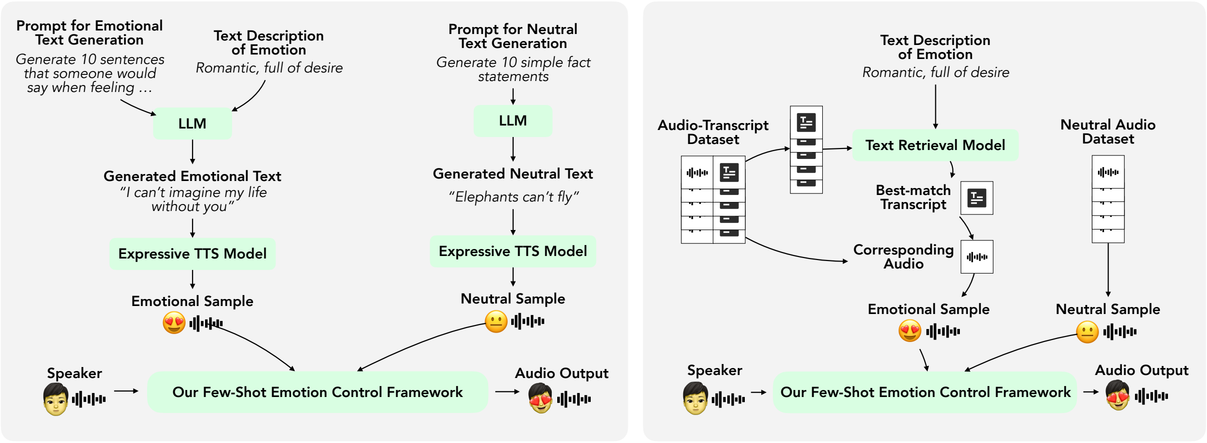
$$u_{s,e} = u_s + \alpha \cdot v_e$$

where emotion control strength α is a scalar that enables fine-grained control of emotion intensity. We hypothesize that larger α values lead to more intense emotions in the speech produced.

Finally, we use pre-trained decoder D to synthesize $y_{s,I,e}$, a speech utterance of text I replicating speaker s's voice while conveying emotion e.

3.3 Towards Open-Ended Text Prompted Emotion Control

Our framework's ability to apply emotion controls with the few-shot demonstration allows us to design two frameworks that take in open-ended text description of an emotion and apply fine-grained control on output speech for the specified emotion. These frameworks allow synthesis of speech with emotions such as *Romantic, full of desire* and *Grateful, appreciative, thankful, indebted,*



(a) Synthetic-data based open-ended emotion control

(b) Retrieval based open-ended emotion control

Figure 3: EmoKnob enables emotion control with open-ended text descriptions of emotion. Based on recent advances in LLMs and EmoKnob's capability of applying emotion control with few-shot samples, we propose two methods that bypass the data insuffiency problem in emotional speech and embed emotions described by open-ended text descriptions into speech.

blessed that are nuanced in details and lack existing datasets. Both frameworks take advantage of recent development in LLMs to overcome the lack of a labeled emotional speech dataset.

3.3.1 Synthetic Data-Based Method

While existing TTS models and services do not allow emotion control, they produce expressive and accurate emotions for texts that obviously convey the emotions (OpenAI, 2024b; Microsoft; ElevenLabs). We leverage this quality to generate synthetic emotional samples that can be used for emotion control with our framework. We show this process in Figure 3(a).

Given a text description T of an emotion e, we prompt an LLM to generate N text samples $I_e^{1,\cdots,N}$, that obviously convey the emotion: $I_e^{1,\cdots,N} = \text{LLM}(T)$. Prompted with prompts such as *Generate 10 sentences that someone would say when feeling [emotion]*, LLM generates emotional texts that conveys emotion e. Then, we use expressive commercial TTS services to obtain an emotional speech sample $x_e^{1\cdots i}$: $x_e^i = \text{TTS}(x_e^i)$.

We can obtain neutral audio samples with the same procedure by first prompting LLM with prompts such as *Generate 10 simple fact statements* to generate neutral texts. Then, we can obtain the neutral audio samples x_n^i with the TTS services.

We can then use the emotional speech samples obtained x_e^i and x_n^i with our few-shot emotion control framework to apply fine-grained emotion control on new speakers.

3.3.2 Transcript Retrieval-Based Method

While a synthetic-data-based method enables open-ended emotion control while bypassing the lack of captioned emotion datasets, the high cost of expressive TTS services limits the framework's wide usage. In this section, we hypothesize that in existing datasets with speech-transcript pairs, transcripts that obviously convey an emotion are matched with audio clips that convey the emotion. We leverage recent developments of text embedding models and document retrieval pipeline to find emotional audio samples that we can use for few-shot emotion control. We show this pipeline in Figure 3(b).

Given a text description T of an emotion e and a text embedding model M, we retrieve transcript-audio pairs (I_e^j, x_e^j) in a dataset such that the transcript I_e^j best matches the emotion description: $j = \arg\max_j M(I_e^j)^T M(T)$.

We can find neutral samples x_n either with the same retrieval pipeline or neutral labels in the dataset, which are more widely available than diverse emotion labels.

4 Experiments

4.1 Evaluation Metrics

4.1.1 Subjective Evaluations

Given the novelty of fine-grained emotion control in text-to-speech synthesis, there is not an established paradigm for examining this capability. To rigorously test the objective of providing fine-grained, faithful emotion control, we proposed the following subjective evaluation metrics:

Emotion Selection Accuracy (ESA): Participants compare audio samples with and without control generated from emotion-neutral text, selecting which better conveys the emotion. ESA measures the percentage choosing the controlled audio and tests the system's ability to embed any emotions to any text.

Emotion Enhancement Accuracy (EEA): Participants compare audio samples with and without control generated from emotion-matching text, selecting which better conveys the emotion. EEA measures the percentage choosing the controlled audio and tests the method's ability to amplify text's emotions.

Emotion Discrimination Test (EDT): Participants compare two audio samples generated from the same neutral text and controlled with different emotions, selecting the one matching a given emotion. EDT evaluates the distinguishability and faithfulness of emotion control.

Emotion Identification Test (EIT): Participants identify the emotion in a controlled audio sample from neutral text, choosing between two emotion labels. EIT measures the accuracy of emotion identification and verifies the recognizability of emotions resulted from emotion control.

Emotion Selection Comparison (ESC): Participants compare our emotion-controlled audio to commercial TTS audio with neutral text, selecting which conveys more specified emotion. ESC measures percentage of selecting our controlled audio and evaluates system advantage over existing TTS services to embed any emotion into any text.

Emotion Enhancement Comparison (EEC): Similar to ESC, but with emotion-matching text. EEC evaluates emotion expressiveness after control compared to commercial TTS without emotion control functionality.

Emotion Strength Test (EST): Participants compare two audio samples controlled with the same emotion but different emotion strengths α, selecting which conveys more emotion. EST measures correct response percentage and evaluates our framework's effectiveness at fine-grained control over emotion intensity.

Since all metrics are calculated from binary choice questions, 50% serves as the random guess baseline to all metrics. We asked one question for each emotion and each metric to 23 university student volunteers from our lab and recruited on campus. Participants are told that responses are used to evaluate a new emotional text-to-speech framework. This study is approved by IRB. We anonymized the participant response. We provided the full subjective evaluation survey we used at `https://frolicking-baklava-af4770.netlify.app/`. For EEC and ESC, we compared speech generated from our framework with speech generated with ElevenLabs (ElevenLabs).

4.1.2 Objective Evaluation

Since our goal is to preserve source speaker identity and maintain accurate text-to-speech synthesis while conducting emotion control, we follow previous voice cloning work (Anastassiou et al., 2024; Shah et al., 2023) on measuring word error rate (WER) and speaker similarity (SIM). We use 100 texts from Common Voice dataset (Ardila et al., 2020) to calculate WER and SIM.

For WER, we first transcribe the generated clips with Whisper-large-v3 (Radford et al., 2022) and calculate WER with jiwer library (nikvaessen). We use the WER of audio generated without any emotion control (original voice cloning model) as a baseline of comparison. Similar WER between emotion-controlled audio and baseline suggests that our framework preserves the high quality TTS in base voice cloning models.

For SIM, we used spkrec-ecapa-voxceleb (Ravanelli et al., 2021) to measure the similarity between generated audio and a reference speaker clip. We use SIM between audio generated without any emotion control and a reference speaker clip as baseline. Similar SIM between the baseline and using emotion-controlled audio suggests our framework's faithful replication of reference speaker while applying emotion control.

4.2 Experiment Details

We use MetaVoice-1B (MetaVoice) as the base voice cloning model, while our framework can be easily extended to any embedding-conditioned voice cloning model. We conduct speech generation on a single NVIDIA L40 GPU. We use an additional NVIDIA L40 GPU for text retrieval in text-retrieval based open-ended emotion control. We provide an audio sample page at `emoknob.cs.columbia.edu`.

4.3 Single-Shot Control of Simple Emotions

We first show our framework's effectiveness on fine-grained emotion control with simple emotion categories: *Happy, Surprise, Angry, Sad, Disgust, Contempt*. We obtain an emotion direction vector in single-shot (one pair of same-speaker emotional and neutral speech clips) from the MSP Podcast dataset (Lotfian and Busso, 2019). We fix emotion strength α to be 0.4 for all samples in evaluation.

We report the subjective evaluation results in Table 2 and the objective evaluate results with a standard deviation in Table 3. High ESA and ESC values shows that our emotion control framework is capable of embedding arbitrary emotion in any text, surpassing commercial TTS services without emotion control option. High EEA and EEC values show that our framework enhances emotions into emotion-matching text, surpassing emotion utterances of commercial TTS services. High EDT and EIT values show that our framework produces recognizable emotions in speech. High EST values show that the emotion strength α option in our framework faithfully produces different strengths of emotions specified by corresponding values.

Speech produced from emotion control shows similar WER within uncertainty as a baseline of no emotion control and thus preserves high quality TTS of the base model. Emotion-controlled speech also shows similar SIM within uncertainty as the baseline, showing that our framework preserves speaker identity well while conducting emotion control.

	ESA↑	EEA↑	EDT↑	EIT↑	ESC↑	EEC↑	EST↑
Happy	100%	100%	100%	100%	100%	100%	83%
Surprise	100%	100%	91%	44%	100%	61%	91%
Angry	82%	100%	82%	74%	100%	100%	100%
Sad	100%	83%	91%	100%	100%	83%	74%
Disgust	74%	91%	91%	74%	61%	83%	91%
Contempt	61%	83%	13%	52%	74%	74%	74%
Averages	86%	93%	78%	74%	89%	83%	86%
Baseline	50%	50%	50%	50%	50%	50%	50%

Table 2: Subjective evaluation results for emotion controls with simple emotions.

	WER ↓	SIM ↑
Happy	0.143 ± 0.349	0.662 ± 0.087
Surprise	0.061 ± 0.107	0.703 ± 0.076
Angry	0.082 ± 0.211	0.712 ± 0.060
Sad	0.113 ± 0.297	0.719 ± 0.059
Disgust	0.05 ± 0.139	0.719 ± 0.063
Contempt	0.053 ± 0.098	0.712 ± 0.069
Average	0.085 ± 0.208	0.705 ± 0.077
w/o Emotion Control	0.079± 0.160	0.719 ± 0.071

Table 3: Objective evaluation results for controls with simple emotions.

4.4 Two-Shot Control of Complex Emotion

Our framework allows a few-shot transfer of emotion onto new speakers and bases such transfer on expressive representation of foundation voice cloning models. We show that these features enable previously not studied controls on more complex, composite, and nuanced emotions. Our experiments focus on two emotions with corresponding datasets: (1) *charisma* defined as conveying the personality of leadership and persuasiveness (Yang et al., 2020); and (2) *compassionate empathy* defined as understanding another's pain as if we are having it ourselves and taking action to mitigate problems producing it (Chen et al., 2024). For each emotion, we use two pairs of emotional and neutral speech from two speakers. We fix emotion strength $\alpha = 0.4$ for all samples.

We report the subjective and objective evaluation results in 4. Subjective evaluation results show that our framework produces recognizable, faithful emotion selection and enhancement, surpassing commercial TTS on uttering specified emotions. Speech produced from emotion control shows similar WER and SIM within uncertainty as the baseline of no emotion control, showing that our framework preserves accurate TTS of the base model and speaker identity while conducting emotion control.

	ESA↑	EEA↑	ESC↑	EEC↑	WER↓	SIM↑
Empathy	74%	83%	100%	22%	0.074± 0.07	0.712± 0.06
Charisma	83%	91%	74%	74%	0.031± 0.08	0.680 ± 0.07
Baseline	50%	50%	50%	50%	0.079 ± 0.16	0.719 ± 0.07

Table 4: Subjective and objective evaluation results for controls with complex emotions

4.5 Synthetic Data-Based Open-Ended Emotion Control

	ESA↑	EEA↑	ESC↑	EEC↑	WER↓	SIM↑
Desire	61%	61%	61%	83%	0.066± 0.13	0.713 ± 0.07
Envy	83%	74%	61%	74%	0.085± 0.13	0.704± 0.07
Romance	61%	91%	52%	91%	0.076± 0.12	0.713± 0.06
Sarcasm	61%	61%	74%	74%	0.120± 0.20	0.717± 0.07
Baseline	50%	50%	50%	50%	0.079 ± 0.16	0.719± 0.07

Table 5: Subjective and objective evaluation results for open-ended controls with emotion text descriptions through a synthetic data-based method.

We experiment with our synthetic-data based framework for emotion control on arbitrary text emotion description with emotions that do not have previously collected labeled datasets for emotional speech synthesis: *Desire, enviousness, romance, sarcasm*. We use GPT4-o (OpenAI) to generate emotional and neutral speech texts. We use ElevenLabs (ElevenLabs) to generate 10 pairs of samples

(10 speakers) for each emotion. We fix emotion strength of $\alpha = 0.4$ for all samples.

We report the subjective and objective evaluation results in Table 5. Subjective evaluations indicate our recognizable and faithful emotion control in speech, outperforming commercial TTS in expressing specific emotions. Additionally, speech from our emotion control maintains similar WER and SIM to the baseline, confirming that our framework effectively preserves the base model's accuracy and speaker identity while controlling emotions.

4.6 Retrieval-Based Open-Ended Emotion Control

Since a text retrieval model works best with descriptive, detailed texts, we focus on longer emotion descriptions of three emotions that lack established labeled datasets shown in Table 6. We prefix the emotion descriptions with the retrieval prompt of *Given a description, retrieve relevant transcript lines whose overall style/emotions matches the description* to enable retrieval models focused on the overall emotion of the transcript and avoid keyword matching. We use SFR-Embedding-Mistral (Meng et al., 2024) as the text embedding model. We use 10 pairs of emotional and neutral samples for each emotion. We fix emotion strength $\alpha = 0.5$ for all samples.

We report the subjective evaluation results and objective evaluate results with a standard deviation in Table 6. The evaluation results show that our framework produces recognizable, faithful emotion selection and enhancement while preserving base model accuracy and reference speaker identity.

	ESA↑	EEA↑	ESC↑	EEC↑	WER↓	SIM↑
Grateful ··· [1]	83%	83%	83%	61%	0.146± 0.38	0.650± 0.07
Curious, ··· [2]	61%	100%	61%	22%	0.124± 0.29	0.655± 0.06
Blaming	65%	69%	74%	74%	0.112± 0.21	0.630± 0.06
Desire ··· [3]	78%	100%	100%	91%	0.062± 0.14	0.664± 0.09
Baseline	50%	50%	50%	50%	0.079 ± 0.16	0.719± 0.07

Table 6: Subjective and objective evaluation results for open-ended controls with emotion text descriptions through retrieval-based methods. [1] Grateful, appreciative, thankful, indebted, blessed. [2] Curious intrigued. [2] Desire and excitement.

5 Ablation Studies

In this section, we conduct ablation studies that vary the shot (sample) number when obtaining the emotion direction vector and the emotion control strength α when applying emotion control. These ablation studies help users decide how to select these hyper-parameters. We report in Table 4 SIM and WER with audio generated with 100 Common Voice texts while applying emotion control of simple emotions with varying shot numbers and emotion strengths. We observe that both SIM and WER are insensitive to shot number and degrades as emotion control strength increases. Users thus need to trade off between generating more emotional clip with higher emotion strength and accurate TTS and voice clone. However, a larger number of samples make the method more robust in larger emotion control strength. Users thus could employ a larger number of samples to compensate the TTS quality decrease while obtaining more emotional speech.

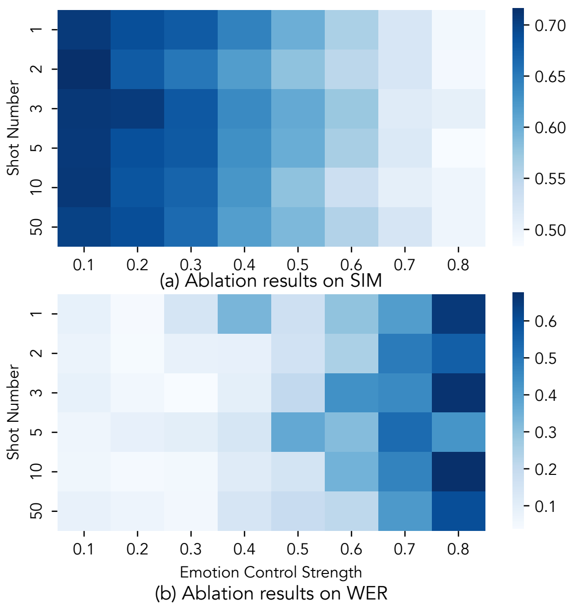


Figure 4: Ablation results measuring SIM and WER with varying shot number and emotion strength.

6 Conclusion and Future Works

We proposed EmoKnob, a framework that enables fine-grained emotion control in voice cloning with few-shot samples. We also propose a synthetic-data-based and a retrieval-based method to embed emotions described by open-ended text into speech synthesis. Given novelty of the emotion control domain, we proposed a set of metrics to rigorously evaluate faithfulness and recognizability of emotion control. Our method establishes a new way of extracting emotion representation in foundation speech models thus bypassing data limitations. Future works can further explore emotion control paradigms such as synthesizing emotions in conversation turns based on these representations.

Limitations

Naturalness and expressiveness of speech created by our framework is constrained by base voice cloning model. However, since we are seeing rapid advances in foundation speech models, and our method is inherently synergetic with these advances, speech produced by EmoKnob will naturally improve as voice cloning models scale up and improve.

Potential Risks

Risks in speech identity theft in voice cloning apply to our work. Practices such as voice cloning detection (Malik, 2019) and phasing out voice-based authentication systems (OpenAI, 2024a) help mitigate risks of our works.

Acknowledgements

This research is being developed with funding from the Columbia Center of AI Technology (CAIT) research award and the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *Preprint*, arXiv:2406.04904.

Run Chen, Haozhe Chen, Anushka Kulkarni, Eleanor Lin, Linda Pang, Divya Tadimeti, Jun Shin, and Julia Hirschberg. 2024. Detecting empathy in speech. In *Proc. INTERSPEECH 2024*.

ElevenLabs. Text to Speech & AI Voice Generator — elevenlabs.io. `https://elevenlabs.io`. [Accessed 14-06-2024].

Christer Gobl and A. N. Chasaide. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun.*, 40:189–212.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2022. Promptts: Controllable text-to-speech with text descriptions. *Preprint*, arXiv:2211.12171.

Jeffrey Jenkins. 2020. Detecting emotional ambiguity in text. *MOJ Applied Bionics and Biomechanics*.

Suhee Jo, Younggun Lee, Yookyung Shin, Yeongtae Hwang, and Taesu Kim. 2023. Cross-speaker emotion transfer by manipulating speech style latents. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Minki Kang, Wooseok Han, Sung Ju Hwang, and Eunho Yang. 2023. Zet-speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models. *Preprint*, arXiv:2305.13831.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024. Parler-tts. `https://github.com/huggingface/parler-tts`.

Mateusz Lajszczak, Guillermo Cambara Ruiz, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín Cortinas, Ammar Abbas, Adam Michalski, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Haohan Guo, Bartosz Putrycz, Soledad López Gambino, Kayeon Yoo, Elena Sokolova, and Thomas Drugman. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv*.

A. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen. 1997. On the perception of emotions in speech: the role of voice quality. *Logopedics Phoniatrics Vocology*, 22:157–168.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *CoRR*, abs/2201.06460.

Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai. 2018. Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. *Speech Communication*, 99:135–143.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *Preprint*, arXiv:2402.01912.

Hafiz Malik. 2019. Securing voice-driven interfaces against fake (cloned) audio attacks. *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 512–517.

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. The msp-conversation corpus. *Interspeech 2020*.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning — blog.salesforceairesearch.com. https://blog.salesforceairesearch.com/sfr-embedded-mistral/. [Accessed 15-06-2024].

MetaVoice. MetaVoice - Text to Speech & AI Voice Changer.

Microsoft. Text to Speech – Realistic AI Voice Generator | Microsoft Azure — azure.microsoft.com. https://azure.microsoft.com/en-us/products/ai-services/text-to-speech#features. [Accessed 14-06-2024].

nikvaessen. GitHub - jitsi/jiwer: Evaluate your speech-to-text system with similarity measures such as word error rate (WER) — github.com. https://github.com/jitsi/jiwer. [Accessed 15-06-2024].

L. Nygaard and Erin R. Lunders. 2002. Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, 30:583–593.

OpenAI. Chatgpt. https://chatgpt.com/. [Accessed 15-06-2024].

OpenAI. 2024a. Navigating the challenges and opportunities of synthetic voices. https://openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/. [Accessed 15-06-2024].

OpenAI. 2024b. Text to speech. https://platform.openai.com/docs/guides/text-to-speech. [Accessed 13-06-2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sona Patel, K. Scherer, E. Björkner, and J. Sundberg. 2011. Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, 87:93–98.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Preprint*, arXiv:1810.02508.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2024. Openvoice: Versatile instant voice cloning. *Preprint*, arXiv:2312.01479.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Neil Shah, Saiteja Kosgi, Vishal Tambrahalli, Neha Sahipjohn, Niranjan Pedanekar, and Vineet Gandhi. 2023. Parrottts: Text-to-speech synthesis by exploiting self-supervised representations. *Preprint*, arXiv:2303.01261.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *Preprint*, arXiv:1712.05884.

suno.ai. 2023. Bark. `https://github.com/suno-ai/bark`.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2023. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *Preprint*, arXiv:2301.13662.

Zixiaofan Yang, Jessica Huynh, Riku Tabata, Nishmar Cestero, Tomer Aharoni, and Julia Hirschberg. 2020. What makes a speaker charismatic? producing and perceiving charismatic speech. pages 685–689.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. *Preprint*, arXiv:1812.04342.