Audiobox: Unified Audio Generation with Natural Language Prompts

Audiobox Team

Apoorv Vyas[*], Bowen Shi[*], Matthew Le[*], Andros Tjandra[*], Yi-Chiao Wu*,
Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang,
Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster
Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane
Mary Williamson[†], Wei-Ning Hsu[†]

Fundamental AI Research (FAIR), Meta

Abstract

Audio is an essential part of our life, but creating it often requires expertise and is time-consuming. Research communities have made great progress over the past year advancing the performance of large scale audio generative models for a single modality (speech, sound, or music) through adopting more powerful generative models and scaling data. However, these models lack controllability in several aspects: speech generation models cannot synthesize novel styles based on text description and are limited on domain coverage such as outdoor environments; sound generation models only provide coarse-grained control based on descriptions like "a person speaking" and would only generate mumbling human voices. This paper presents AUDIOBOX, a unified model based on flow-matching that is capable of generating various audio modalities. We design description-based and example-based prompting to enhance controllability and unify speech and sound generation paradigms. We allow transcript, vocal, and other audio styles to be controlled independently when generating speech. To improve model generalization with limited labels, we adapt a self-supervised infilling objective to pre-train on large quantities of unlabeled audio. AUDIOBOX sets new benchmarks on speech and sound generation (0.745 similarity on Librispeech for zero-shot TTS; 0.77 FAD on AudioCaps for text-to-sound) and unlocks new methods for generating audio with novel vocal and acoustic styles. We further integrate Bespoke Solvers, which speeds up generation by over 25 times compared to the default ODE solver for flow-matching, without loss of performance on several tasks. Our demo is available at `https://audiobox.metademolab.com/`.

1 Introduction

Why building audio generative models: Audio is a key component in creating many forms of content, such as movies, podcasts, audiobooks, and Ads. However, audio creation is time-consuming and requires various expertise, such as voice acting, music composing and performing, Foley sound effect creation, and sound engineering. This imposes a great barrier to entry for the general public,

[*] Research team, equal contribution

[†] Research and engineering leadership, equal contribution

Corresponding authors: Apoorv Vyas, Wei-Ning Hsu (`{vyasapoorv,wnhsu}@meta.com`)

Preprint. Under review.

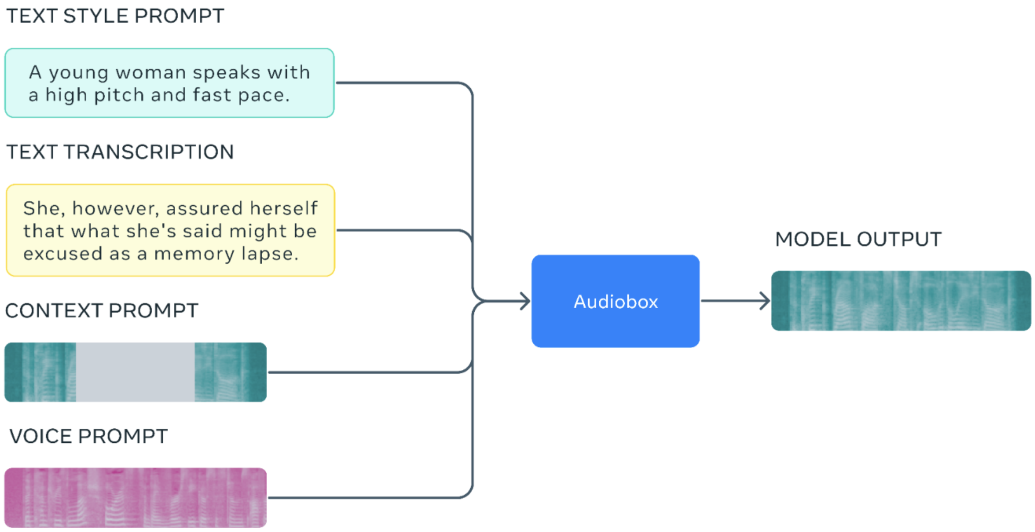

TEXT STYLE PROMPT
A young woman speaks with a high pitch and fast pace.
TEXT TRANSCRIPTION
She, however, assured herself that what she's said might be excused as a memory lapse.
CONTEXT PROMPT
VOICE PROMPT
Audiobox
MODEL OUTPUT

Figure 1: Audiobox model diagram

making it hard for people to become audio creators. Even for professionals, performing these tasks can still take a lot of time and resources, limiting their productivity. Developing audio generative models that are generalizable, controllable, and high quality can bring transformative changes to the audio creation process, improving the efficiency of the professionals as well as unleashing the creativity for everyone.

Progress of audio generative models: Recently, researchers have made significant progress advancing audio generative models. Speech generative models can mimic any vocal style using audio prompts that are as short as three seconds [Wang et al., 2023a, Shen et al., 2023, Le et al., 2023, Kharitonov et al., 2023], infill a portion of speech to remove transient noise or edit words for any speaker [Le et al., 2023, Shen et al., 2023], synthesize foreign languages in anyone's voice [Zhang et al., 2023, Le et al., 2023], and create dialogues [Borsos et al., 2023]. Music generative models can create music in various styles using a short text description [Schneider et al., 2023, Huang et al., 2023a, Agostinelli et al., 2023, Copet et al., 2023] and infill a portion of music [Li et al., 2023]. Sound effect generative models follows a similar paradigm. They are capable of creating and infilling complex acoustic scenes like *"birds chirping and water dripping with some banging in the background"* given a text description [Yang et al., 2023c, Kreuk et al., 2022, Huang et al., 2023b, Ghosal et al., 2023, Liu et al., 2023b,c]. Recent models also extends to more general editing, such as removal or addition of sound events with natural language instructions [Wang et al., 2023b, Liu et al., 2023d].

Limitation of existing models: Existing audio generative models are still limited in controllability and generalizability. First, the real world audio content often contain a mix of speech, music, and sound effects. However, existing audio generative models are mostly modality-specific, which only generate either speech, music, or sound effects. In particular, existing large scale speech generative models [Wang et al., 2023a, Le et al., 2023, Shen et al., 2023] are trained mostly on audiobooks [Zen et al., 2019, Kahn et al., 2019, Pratap et al., 2020], which lacks diversity compared to truly in-the-wild data such as AudioSet [Gemmeke et al., 2017] in terms of expressivity (e.g., non-verbal sounds like coughing, screaming, laughing) and acoustic conditions (e.g., urban, rural, public indoor, stadiums). These models can only generate audio of limited styles and do not capture the correlation between different audio modalities.

On the other hand, there is a discrepancy between speech and sound/speech generation paradigm. Recent speech generation models mostly use example-based control, where an audio sample of the target style is provided and the style control is more precise; in contrast, description-based control is adopted for music and sound generation, where the model can create novel styles based on natural language prompts. Both approaches have their strengths and weaknesses, but such a discrepancy prevents development of unified models that enjoy the best of both worlds.

Last but not least, existing sound generation models only provide coarse control such as *"a man is speaking"* when generating speech. Existing datasets do not offer finer-grained captions that characterizes vocal styles in greater details, such as *"A middle aged woman from the American South is speaking over the phone in a passionate voice. She speaks in at a fast pace with a high pitch."* Neither do these models enable transcript input to controlling the textual content. Hence, these models can only generate mumbling speech.

Due to a lack of consideration in the language-guided generation of speech within a natural setting, designing proper objective evaluation metrics for such universal models remains an open question that has not been fully addressed by prior works. In objective evaluation, previous speech-oriented studies Guo et al. [2023], Leng et al. [2023], Yang et al. [2023a] often adopt ad-hoc evaluation metrics (e.g., accuracy of pre-defined attributes), making it challenging to generalize to free-form instructions. The joint audio-text embedding network (e.g., CLAP Wu et al. [2023]), widely utilized in text-to-audio generation, is tailored to sound events and frequently falls short in capturing intricate attributes such as accents in speech (see Section 7.1.1).

Goals and overview of our model: To tackle these problems, there are three key objectives of this work. First, we aim to build a unified model for sound and speech in order to generate a wider variety of real-world audio, which is often a mix of both. Second, we want to improve controllability for creating novel styles through enabling multiple input methods, using either reference audio, text description, or a combination of both. Last but not least, to improve model generalization, we want to scale training data and utilize data with different level of supervision.

To that end, we present the Audiobox framework. Audiobox is built upon Voicebox [Le et al., 2023] and SpeechFlow [Liu et al., 2023a], which are flow-matching based models for transcript-guided speech generation and self-supervised speech pre-training, respectively. To facilitate data scaling and development of downstream models, we first adopt the SpeechFlow pre-training method and pre-train a unified model using large quantities of unlabeled speech, music, and sound effects, referred to as AUDIOBOX SSL (Section 4). To validate the effectiveness of the unified pre-trained model, we fine-tune AUDIOBOX SSL for transcript-guided speech generation (AUDIOBOX SPEECH, Section 5) and description-guided sound generation (AUDIOBOX SOUND, Section 6), showing significant improvements from prior studies.

Combining the best of both worlds, we present AUDIOBOX, the unified model for sound and speech generation in Section 7. It bridges the gap between sound and speech generation by enabling natural language prompts for holistic style control, and furthers disentangled speech control with voice prompts. Our joint model achieves unprecedented controllability for universal audio generation and superior versatility with additional capabilities on top of what Voicebox offers. AUDIOBOX outperforms existing domain specific models on multiple tasks and is close to AUDIOBOX SPEECH and AUDIOBOX SOUND on their corresponding benchmark tasks.

To facilitate the evaluation of Audiobox and advance research in text-guided universal audio generative models, we propose Joint-CLAP, trained on both sound and speech description data. In comparison to CLAP Wu et al. [2023], Joint-CLAP significantly outperforms CLAP in retrieving description-based speech, and the text-to-audio similarity exhibits a stronger correlation with human judgment.

Orthogonally, to improve performance-efficiency trade-off, we integrate Bespoke Solver, a novel post-training inference optimization methods for flow-matching models. With Bespoke Solver, our models are able speed up by 25x compared to using the adaptive step size dopri5 solver without loss of performance.

As generative models become more powerful and essential parts of everyone's life, it is more important than ever to conduct research responsibly and mitigate potential risks. We conducted a series of study demonstrating the fairness is achieved through better representing voices of different demographic groups with data scaling. We also validate the effectiveness of a recent watermarking system [Seamless Communication, 2023], showing the verification is highly effective and robust to adversarial perturbation.

2 Related Work

This paper is related to a large body of work on large scale generative modeling for audio. As the focus of this work is on universality and controllability, we first discuss controllable generation for modality specific models and then compare with recent studies on universal models that can perform multiple tasks or generate audio in multiple modalities and domains. For the rest of the paper, we will refer to speech, sound, music as different *audio modalities*, and within modality style variation, such as read speech, spontaneous speech, conversational speech, as different *domains*.

Large scale in-context text-to-speech generative models: Over the past few months, there has been significant progress in developing large scale speech generative models [Wang et al., 2023a, Shen et al., 2023, Kharitonov et al., 2023, Le et al., 2023, Yang et al., 2023b, Borsos et al., 2023] that are trained on in-the-wild data at the scale of close to 100K hours [Kahn et al., 2019, Pratap et al., 2020] with minimal supervision, which leads to much better generalization for synthesizing unseen speech styles in a zero-shot fashion. These models are in sharp contrast to conventional regression-based models such as Ren et al. [2021], Shen et al. [2017], Łańcucki [2021], which are trained on highly curated datasets [Yamagishi et al., 2019] containing clean audio, limited style variation, and extensive labels (e.g., speaker and emotion labels).

The key to successful data scaling in recent work is the adoption of powerful generative models that can capture highly stochastic input-output relationships. For example, VALL-E [Wang et al., 2023a] adopt the token-based autoregressive language modeling approach, which converts speech into discrete tokens with a neural codec model [Défossez et al., 2022] and formulate text-to-speech (TTS) as a conditional language modeling problem given a transcript and an audio prompt (the first few seconds of the target speech). NaturalSpeech2 [Shen et al., 2023] and Voicebox [Le et al., 2023] adopt non-autoregressive diffusion [Ho et al., 2020] and conditional flow-matching models [Lipman et al., 2023]. Given a transcript and an audio context (the audio surrounding the target speech), these models iteratively transform a noise sampled from a simple prior to speech, represented as learned latent features or mel spectrograms.

At the high level, VALL-E performs transcript-guided speech continuation while NaturalSpeech2 and Voicebox perform transcript-guided speech infilling. These models are trained with only transcript supervision, which facilitates data scaling. The *style* of the generated audio is controlled through the audio prompt or audio context. Note that the style refers to not only voice, but everything other than transcript, including prosody, emotion, acoustic environment, channel, noise, etc. This can be understood as a form of *in-context learning*: because the audio style tends to be coherent within an utterance, these models learn to infer the style of the target based on its context. In turn, it enables generalization to unseen style, such that speech of any style can be generated by conditioning on an audio prompt/context of the desired style.

While the in-context style transfer paradigm is powerful, it also possesses several limitations in terms of controllability. First, audio prompt is the only input mechanism of controlling the audio style. Users cannot provide a descriptive text, such as “a young man speaking with a happy tone in an auditorium” to create diverse speech matching the description, whereas this feature is commonly supported and widely enjoyed for image [Ramesh et al., 2022, Rombach et al., 2022], music [Agostinelli et al., 2023], and sound [Kreuk et al., 2022] generation. Second, disentangled style control is not enabled with the paradigm, where voice and other attributes, such as emotion and acoustic condition, can be controlled independently. This feature is often desired as exemplified in earlier work where emotion and voice can be controlled independently [Hsu et al., 2019, Kulkarni et al., 2021, Nguyen et al., 2023].

Natural language style prompting for controllable speech generation: Studies on controllable speech generation aims to develop models which can generate speech of many different domains and provide input methods for disentangled, flexible, and accurate control. Earlier models often enable control over only a small number of attributes (e.g., speaker and emotion) with a fixed number of options (e.g., happy/sad/neutral for emotion) through one-hot vectors [Nguyen et al., 2023]. Such methods are difficult to generalize as it is difficult to represent many speech attributes, such as audio quality, acoustic environment, with one-hot vectors. Nor could information such as “a speaker starts with a slow pace and speeds up” be accurately represented. In-context TTS [Wang et al., 2023a] models greatly improves domain coverage, but has the limitation on flexibility and disentangled control described above.

To address the limitation, several recent studies also propose to control speech style through natural language prompts. InstructTTS [Yang et al., 2023a] and PromptTTS [Guo et al., 2023] are the two earliest works. They are trained on small scale data with mainly emotion variation and limited number of speakers (7 for InstructTTS and 2 for PromptTTS synthetic setup). In particular, InstructTTS collects human descriptions for 44 hours of speech focusing on only the emotion and a separate speaker ID input is used as model input. Therefore, the natural language prompt is only used for controlling the emotion. PromptTTS recruits human annotators to write descriptions to given four to five attribute labels (emotion, gender, volume, speed, and pitch; emotion label is not available for the real data), and trains models on 2-voice synthetic data as well as LibriTTS [Zen et al., 2019]. Because the descriptions of PromptTTS are created based on attribute labels instead of speech samples, these descriptions do not contain additional information compared to the labels and theoretically does not enable finer grained attribute control.

PromptTTS2 [Leng et al., 2023] is a concurrent work which improves upon PromptTTS in two aspects. First, it proposes a automatic description creation pipeline based on speech attribute labeler and large language models, which enables scaling to training on 44K hours of audiobook data. Second, PromptTTS2 adopts a diffusion model to capture the one-to-many relationship given input (transcript and description), whereas PromptTTS adopts a regression model assuming deterministic mapping. Nevertheless, similar to PromptTTS, all the descriptions PromptTTS2 create are derived from four categorical attributes with two to three options each (total 54 combinations). Hence, PromptTTS2 does not provide finer grained control than PromptTTS and has limited coverage on the attributes it can control via natural language prompt.

Large scale general-domain models for sound and music generation: Text-to-sound [Kreuk et al., 2022] and text-to-music [Schneider et al., 2023] are the emerging paradigms for general-domain sound and music generation, in contrast to earlier studies that generate finite sound effects [Donahue et al., 2018] or instruments [Huang et al., 2018]. The text here refers to a holistic description of the target audio, such as *"A child shouts while an emergency vehicle siren sounds with the horn blowing."* [Kim et al., 2019] and *"The low quality recording features a ballad song that contains sustained strings... It sounds sad and soulful, like something you would hear at Sunday services."* for music [Agostinelli et al., 2023].

Similar to speech generation, the recent progress can be largely attributed to the advancement in generative models for continuous data [Ho et al., 2020, Huang et al., 2023a, Liu et al., 2023b] and audio tokenizers [Zeghidour et al., 2022, Défossez et al., 2022, Kreuk et al., 2022, Copet et al., 2023, Agostinelli et al., 2023], which enables modeling methods capable of capturing highly stochastic conditional distributions of audio given descriptions for general domain sound/music data.

A key limitation of these models is the ability to control transcript and generate intelligible speech or vocals. These models only take a description as input, which does not specify the transcript when speech is presented. Hence, generating samples with prompts like "a person speaking" often results in speech-like mumbling sound with unintelligible content [Liu et al., 2023b]. In other words, these models does not offer an input for users to control transcript, and have not learned language models that allow it to construct and synthesize meaningful sentences given only the description.

Unified model for audio generation: With the great progress made in developing general-domain models for each audio modality, researchers also start exploring unified model that can generate audio beyond a single modality and perform multiple generative tasks. Such a model could potentially learn from different sources of supervision and benefit from knowledge transfer across tasks. There are three concurrent studies that are related to this work.

UniAudio [Yang et al., 2023b] focuses on building a single model that can perform multiple tasks, including text-to-music, text-to-sound, and in-context TTS and natural language style prompted TTS. It follows the VALL-E [Wang et al., 2023a] framework, which tokenizes audio and serializes conditioning input and output audio tokens for training a conditional token-based language model. It is trained on the same speech descriptions collected by PromptTTS, which inherits the same limitations in terms what attributes and how granular they can be controlled through natural language prompts as discussed earlier.

VoiceLDM [Lee et al., 2023] is the most related work. It introduces a transcript input to AudioLDM [Liu et al., 2023b] and controls style through text description embedded with a frozen Contrastive Language-Audio Pre-training (CLAP) model [Wu et al., 2023]. During training, CLAP embedding from audio is used for conditioning. VoiceLDM is trained on datasets with rich acoustic variation, and hence is capable of generating speech in diverse acoustic environments. However, the performance in terms of controllability is bounded by the pre-trained CLAP model. Since the CLAP model are trained on audio-caption pairs focus on sound events, the embedding only encodes very coarse information regarding speech attributes. Furthermore, VoiceLDM also follows the sound generation paradigm which always generate audio clips of a fixed size (10 seconds), which is not ideal for speech generation that have variable length in general. Finally, despite that the model can generate non-speech sounds when conditioned on empty transcripts, the performance of sound generation lags behind state-of-the-art models by a large margin.

AudioLDM 2 [Liu et al., 2023c] presents a two-stage model that is applicable to speech, sound, and music generation. It is comprised of a deterministic auto-regressive model that maps conditioning input (e.g., CLAP-embedded audio, description, transcript, image) to semantic features sequence, and a diffusion model which mapping semantic to acoustic features. The structure is similar to SPEAR-TTS [Kharitonov et al., 2023] but with different modeling methods and representations for each stage. Hence, similarly it can leverage unlabeled audio for training the second stage model. While AudioLDM 2 presents a unified framework, empirically separate models for speech and sound/music generation are trained, as the authors noted that different model architecture hyperparameters are required for different modalities.

3 Background

This work is heavily built upon the training objective and model architecture of Voicebox [Le et al., 2023], and the self-supervised objective of SpeechFlow [Liu et al., 2023a]. Both studies adopt conditional flow-matching [Lipman et al., 2023] as the modeling backbone, which is a powerful non-autoregressive generative model for continuous data. We provide a technical overview here.

Conditional flow-matching: Conditional flow-matching (FM) [Lipman et al., 2023] is a novel generative modeling method derived from the continuous normalizing flow [Chen et al., 2018] framework. It models the paths that transform samples from a simple prior distribution p_0 to the corresponding samples from the complex data distribution p_1 in a continuous manner. We use *flow step* t to describe the progress of transformation, where the prior is at $t = 0$ and the data is at $t = 1$.

The training objective of FM resembles the objective diffusion models [Ho et al., 2020]: during training, given a sample x_1 drawn from the data distribution, a random flow step $t \sim \mathcal{U}[0, 1]$ is sampled, and a noisy version of the data x_t as well as its derivative $v_t = dx_t/dt$ for the chosen condition path are computed. A FM model u is trained to predict the derivative v_t given t and x_t. During inference, to draw a sample x_1 from the learned data distribution, a sample x_0 is first drawn from the prior distribution, and then the ordinary differential equation (ODE) solver is used to estimate x_1 given x_0 and the derivative parameterized by the FM model through integration. Trade-off between accuracy of x_1 estimation and speed can be flexibly selected by configuring the ODE solver.

At a high level, FM subsumes diffusion models, which correspond to specific paths of the transformation. The authors of Lipman et al. [2023] presented an alternative called optimal transport (OT), which are conditional paths with constant directions and speeds. It is arguably easier to learn and can be more accurately estimated by the ODE solver with fewer steps. The OT path results in better training and inference efficiency as empirically verified in Lipman et al. [2023] and Le et al. [2023].

Given a sample x_1 and a flow-step t, with the OT conditional path we have $x_t = (1-(1-\sigma_{min})t)x_0 + tx_1$ and $v_t = x_1 - (1 - \sigma_{min})x_0$, where x_0 is drawn from the prior distribution $N(0, I)$ and σ_{min} is a small value (10^{-5}). The FM model u minimizes:

$$\mathbb{E}_{t,x_1,x_0}||u(x_t, t) - v_t||^2. \tag{1}$$

Voicebox: Voicebox [Le et al., 2023] is a conditional generative model based on FM which additionally conditions on frame-aligned phonetic transcript and masked audio for audio prediction, and conditions on phonetic transcript and masked duration sequence for phone duration prediction. Audio is represented as 80-dimensional Mel spectrograms and are converted to waveform using a HiFi-GAN vocoder [Kong et al., 2020]. Duration sequence denotes the number of frames for each phoneme in the transcript.

Voicebox adopts the Transformer [Vaswani et al., 2017] model with U-Net [Ronneberger et al., 2015] connections. Masked spectrogram (or masked duration), frame-aligned phone embeddings (or phone embeddings), and noisy audio x_t (or noisy duration) are concatenated along the channel dimension and projected to the Transformer feature dimension. The flow step sinusoidal embedding is then concatenated with the project features along the time dimension, passed as input to the Transformer model. The Transformer output is then projected to 80 dimensions (or 1 dimension for duration) and predicts the derivative v_t.

It is a supervised model trained on 60K hours of audiobooks and achieves state-of-the-art performance on in-context text-to-speech synthesis that can mimic the audio style given a three second audio prompt. It is also high versatile due to the generality of transcript-guided infilling, where the model can perform transient noise removal, diverse style generation, speech editing, cross-lingual style transfer by simply forming transcript and audio inputs differently.

SpeechFlow: SpeechFlow [Liu et al., 2023a] is a self-supervised framework based on FM with learns to infill speech given the audio context. This is equivalent to Voicebox without conditioning on transcripts. The self-supervised objective tackles label scarcity issues and enables the model to learn from large quantities of unlabeled speech the distribution of speech as well as the correlation between temporal segments within an utterance.

Fine-tuning SpeechFlow with the same transcript-guided infilling objective as Voicebox shows superior performance and sample efficiency, matching style similarity of VALL-E [Wang et al., 2023a] with only 10 hours of labeled data. The pre-trained model also demonstrates promising improvements on other speech generation tasks, including source separation and speech enhancement. It also enables parameter efficient fine-tuning like LoRA [Hu et al., 2021] and fine-tuning with a much lower batch size, demonstrating the efficiency and reusability of self-supervised pre-train models

4 AUDIOBOX SSL: Self-supervised Generative Audio Pre-training

Our first step is to develop AUDIOBOX SSL, a foundation model that can be fine-tuned for any downstream audio generation tasks. Because labeled data are not always available or of high quality, and data scaling is the key to generalization, our strategy is to train this foundation model using audio without any supervision, such as transcripts, captions, or attribute labels, which can be found in larger quantities.

4.1 Method

We adapt AUDIOBOX SSL from SpeechFlow, which was originally designed for generative speech pre-training. The same learning objective is also meaningful for general audio: through learning to infill, the model can also capture the temporal relationship of audio events (e.g., clock ticking sound at fixed time interval, approaching train producing sounds with increasing volume), and learns the distribution of general audio. Therefore, during supervised fine-tuning, a model does not need to learn what a natural audio sample sounds like, but only needs to learn aligning the label with the corresponding mode of distribution.

The original SpeechFlow model is trained to predict spectrograms and uses a HiFi-GAN model to generate waveform given spectrogram. However, HiFi-GAN does not generalize well to non-speech audio such as sound or music [Lee et al., 2022]. To tackle that, we train the model to predict latent features learned by an autoencoder. In particular, we use the dense Encodec [Défossez et al., 2022] features which are extracted prior to the residual quantization layer, which demonstrates good resynthesis quality in various audio modalities and has been adopted for sound and music generation [Kreuk et al., 2022, Copet et al., 2023]. This is similar to the latent diffusion framework [Rombach et al., 2022] that is also adopted in NaturalSpeech2 [Shen et al., 2023].

During training, the model is conditioned on fully masked features with probability p_{cond}. With probability $1 - p_{\text{cond}}$, a subset (n_{mask}) of frames are masked with minimum span length l_{mask}. The FM loss is computed only on masked frames. When a frame is masked, its features are set to 0.

4.2 Experimental Setup

Training data: We collect an large scale audio dataset that greatly increases the domain coverage, modality coverage, and quantities compared to previous large scale audio generative model studies [Yang et al., 2023b, Borsos et al., 2023, Wang et al., 2023a, Liu et al., 2023c], which leverage datasets ranging between 10K to 100K hours containing mostly speech from a single domain (e.g., audiobooks).

Specifically, our dataset includes over 160K hours of speech (primarily English), 20K hours of music and 6K hours of sound samples. The speech portion covers audiobooks, podcasts, read sentences, talks, conversations, and in-the-wild recordings including various acoustic conditions and non-verbal voices. To ensure fairness and a good representation for people from various groups, it includes speakers from over 150 countries speaking over 200 different primary languages. We refer to this set as "Mix-185K."

Model and training: We train a 24 layer Transformer Vaswani et al. [2017] with convolutional position embeddings Baevski et al. [2020] and symmetric bi-directional ALiBi self-attention bias Press et al. [2021]. The model has 16 attention heads, 1024/4096 embedding/feed-forward network (FFN) dimension, and 330M parameters. We add UNet-style skip connections, where states are concatenated channel-wise and then combined using a linear layer.

The model is trained for 1 million updates with an effective batch size of 480K frames. For efficiency, samples are randomly chunked if they exceed 1,600 frames. We set $p_{\text{cond}} = 0.1$, $n_{\text{mask}} \sim \mathcal{U}[70\%, 100\%]$, and $l_{\text{mask}} = 10$. We use the Adam Kingma and Ba [2014] optimizer with learning rate 1e-4, linearly warmed up for 5k steps and linearly decayed over the rest of training. For stability, we use gradient norm clipping with a norm threshold of 0.2.

5 AUDIOBOX SPEECH: Scaling In-context Text-to-speech Synthesis

In this section, we study the effectiveness of pre-training and fine-tuning data scaling for speech generation. We present AUDIOBOX SPEECH, which fine-tunes AUDIOBOX SSL with the same transcript-guided speech infilling objective as Voicebox using transcribed speech. The resulting model can be applied to multiple downstream tasks just like Voicebox.

5.1 Method

To incorporate the frame-aligned transcript z, we follow Liu et al. [2023a]. Specifically, given the noisy Encodec features x_t at the flow-step t, masked Encodec features x_{ctx}, we first concatenate x_t and x_{ctx} channel-wise and apply a linear project to get x_h. We then apply another linear layer to the frame-aligned transcript embeddings z_{emb}, and add this to the hidden state x_h. The resulting features are concatenated with the flow step sinusoidal embedding along the time dimension and fed to the Transformer as input. The Transformer output is projected and predicts the derivative v_t.

There are two different approaches to fine-tuning the model. The first one is low-rank adaptation (LoRA) Hu et al. [2021], where we add LoRA adapters to the linear input projection of each self-attention layer. With this approach, only the transcript embedding, projection parameters, along with the LoRA adapter parameters are optimized. The second approach is full fine-tuning, where all parameters are optimized together. Liu et al. [2023a] showed that LoRA achieves better performance when fine-tuning SpeechFlow on 960 hours of speech, but we suspect that full fine-tuning may prevail when we scale fine-tuning data.

In addition, many prior studies [Le et al., 2023, Wang et al., 2023a] represent transcripts as phoneme sequences and using the off-the-shelf Montreal Forced Aligner [McAuliffe et al., 2017] for aligning the training data. Instead, we represent transcript with raw characters, including punctuation and with true cases, and utilize the SEAMLESSM4T V2 multilingual char-to-unit forced aligner presented in Seamless Communication [2023] adapted from RAD-TTS [Shih et al., 2021]. This aligner is trained on large quantities of multilingual data and can align raw text with speech. There are several benefits with the replacement. First, it circumvents the need of phonemizers and avoids error propagation due to incorrect phonemization. Second, raw text preserves more information than phonemized text, such as casing (e.g., all caps for emphasis) and punctuation. Third, the SEAMLESSM4T V2 aligner is much more robust than MFA and can handle multilingual/code-switching text, which enables easier extension to multilingual TTS systems and is more suitable for aligning challenging speech such as conversational and noisy samples.

Following Le et al. [2023], we train a flow-matching duration model only with labeled data. It was shown in Le et al. [2023] that FM duration model has better diversity compared to regression duration models. However, it is less stable and sometimes produces unnatural prosody. To alleviate the issue, we propose to average over a small number of duration sequences for stabilization, which empirically shows better trade-off between diversity and quality. The averaging operation is reasonable as duration distributions are relatively unimodal. When averaging more samples, it approaches the mean, which is the estimation produced by regression models.

5.2 Task and Evaluation

We consider the in-context TTS (also known as zero-shot TTS) task. In-context TTS aims to synthesize speech that resembles the audio style of the given an audio example which may be unseen during training. The audio style refers to not only voice, but everything other than transcript, such as prosody and acoustic condition. To perform the task, input raw/frame-level transcript is the concatenation of the raw/frame-level transcript of the audio example and the target raw/frame-level transcript, while the masked audio/duration is the concatenation of the example audio/duration and a mask for the speech/duration to be generated. We first sample duration sequence for the target raw transcript to create frame-level target transcript using the duration model, and then sample audio with the audio model.

The performance is measured in terms of style similarity, content correctness, and quality. A proxy automatic metric for style similarity is the cosine similarity between the audio prompt and the generated audio in some embedding space that reflects the audio style. WavLM-TDCNN [Chen et al., 2022b] is commonly used for embedding [Wang et al., 2023a, Kharitonov et al., 2023, Le et al., 2023]. Le et al. [2023] advocates for reporting both similarity with respect to raw audio (SIM-orig) and to audio resynthesized from the same vocoder (SIM-resyn) for comparability across studies (SIM-orig). Content correctness can be approximated with the word error rate (WER) from some speech recognition model; however, WER can result from both synthesis error and recognition error, and hence is less reliable when numbers are close or when the target style is more difficult to recognize (e.g., accented speech, conversational speech, noisy speech). In this paper we use Whisper `large-v2` instead of HuBERT-L Hsu et al. [2021] used in prior studies [Wang et al., 2023a, Le et al., 2023] because the latter is less robust and has higher WER on real data for non audiobook domains. Subjective evaluations are often used for assessing style similarity and audio quality, measured by mean opinion scores (MOS).

5.3 Experimental Setup

Training data: We train AUDIOBOX SPEECH on a transcribed English subset of the speech data used for pre-training. The subset contains 100K hours of speech covering similar domains as the full set, which we refer to as "SP-multi-100K." We create the transcribed subset with the following pre-processing methods:

For unsegmented multi-speaker conversational datasets information, we first segment our dataset using PyAnnote diarization toolkit [Plaquet and Bredin, 2023, Bredin, 2023] to create single speaker speech segments. For untranscribed speech, we transcribe data using two speech recognition models, Whisper Radford et al. [2022] `large-v2` and `medium.en`. For each audio with unknown language, we additional use the Whisper `large-v2` model for language identification (LID). We then remove the utterances where the probability being English is lower than 50% or the the word error rate (WER) between the transcriptions from the two models is greater than 50%.

To create a similar text distributions across multiple datasets, we apply inverse text normalization to create true-cased and punctuated transcript for any dataset with normalized transcript using Whisper-punctuation library.[3] It performs the task through constrained search where the produced transcript needs to match the original transcript after normalization.

Model and training: We adopt the full fine-tuning method and train the audio model for 200K steps with an effective batch size of 240K frames. Samples are randomly chunked if they exceed 1,600 frames. Character embeddings are 128 dimensions. For each batch, audio is entire masked with probability 0.3; otherwise a contiguous chunk is masked where the chunk size 70% to 100% of the frames. The same optimizer, learning rate, scheduler, and gradient clipping as AUDIOBOX SSL are used.

The duration model has 8 heads, 768/2048 embedding/FFN dimensions, 10 layers, with 40 dimension character embeddings. It is trained for 600K updates with an effective batch size of 120K frames. For each batch, duration is entirely masked with probability 0.2 and otherwise a chunk of 10% to 100%

[3] `https://github.com/jumon/whisper-punctuator`

of the sequence length is masked. The rest of the optimization parameters are the same as the audio model.

Evaluation data and configuration: For in-context TTS, three second prompts are used following Wang et al. [2023a]. Voicebox uses the last three seconds of the reference as the prompt, which often contains a considerable amount of trailing silence. We instead use the last three seconds after removing the trailing silences based on the forced alignment for all experiments in this paper. Duration is estimated by averaging over five samples and following [Le et al., 2023] predicted silence at both ends are trimmed to 0.1 second max.

The `torchdiffeq` [Chen, 2018] package is used. By default, we use the midpoint solver with a step size of 0.0625, which invokes the derivatives being evaluated 32 times. When using classifier free guidance the model does 2 forward passes per evaluation, leading to a total of 64 calls to the model. A guidance weight for classifier-free guidance [Ho and Salimans, 2022] of 0.7 is applied.

Models are evaluated on five datasets representing different domains. (1) Librispeech test-clean (LS) [Panayotov et al., 2015]: audiobook recordings that are scripted and relatively clean. Following Wang et al. [2023a], we keep only samples between 4 to 10 seconds for evaluation to compare with prior studies. (2) CommonVoice v13.0 English test set (CV) [Ardila et al., 2019]: sentences read by volunteers worldwide. It covers broader accents and are noisier compared to Librispeech. (3) Switchboard (SWBD) [Godfrey et al., 1992]: a conversational speech corpus. We evaluate on a subset of 611 samples from 8 speakers. (4) Expresso [Nguyen et al., 2023] (Expr) is a multispeaker expressive speech dataset covering 7 different speaking styles, which we evaluate on a subset of 999 samples. (5) An internal expressive and accented dataset (Accent): read sentences with speakers covering a wider range of accents and 10 emotions. We create a subset of 500 samples for evaluation.

5.4 Main Results

We compare AUDIOBOX SPEECH with several state-of-the-art in-context speech generation models. Voicebox, VALL-E, NaturalSpeech 2 (NS2), and YourTTS are trained on 60K, 60K, 44K, 600 hours of audiobooks respectively. UniAudio is trained on about 100K hours of audio, where speech accounts for 81K hours and are mostly audiobooks. Results are shown in Table 1 and 2.

AUDIOBOX SPEECH achieves a new best on style similarity (0.745 vs. 0.710 from UniAudio) on the audiobook domain test set (LS). More importantly, AUDIOBOX SPEECH drastically improves Voicebox on all other domains, with similarity improvement ranging from 0.096 to 0.156. The results suggest that AUDIOBOX SPEECH generalizes much better thanks to scaling data to cover more domains. The subjective evaluations presented in Table 2 again confirms that AUDIOBOX SPEECH transfers styles significantly better than the baselines, and generate audio with better quality.

Table 1: In-context TTS style similarity and content correctness. We cite Yang et al. [2023b] for the NS2 results which are not in the original paper[Shen et al., 2023]. WER with * are computed using HuBERT-L ASR that is not comparable with the other numbers.

	Sim-r ↑	Sim-o ↑						Word error rate (%) ↓					
	LS	LS	CV	SWBD	Expr	Accent	Avg	LS	CV	SWBD	Expr	Accent	Avg
VALL-E	0.580	-	-	-	-	-	-	5.9*	-	-	-	-	-
NS2	0.620	-	-	-	-	-	-	2.3*	-	-	-	-	-
UniAudio	0.710	-	-	-	-	-	-	2.0*	-	-	-	-	-
YourTTS	-	0.455	0.312	0.291	0.290	0.366	0.343	6.8	10.4	11.8	9.5	4.0	8.5
Voicebox	0.696	0.674	0.477	0.452	0.487	0.563	0.531	**2.6**	7.9	10.6	7.2	2.1	6.1
AUDIOBOX SPEECH	**0.745**	**0.734**	**0.607**	**0.608**	**0.603**	**0.659**	**0.642**	3.2	**3.7**	**9.1**	**3.2**	**0.9**	**4.0**

5.5 Ablation Study

We present ablation studies in Table 3. To understand the effect of data scaling, we create a subset containing 60K hours of audiobook speech referred to as "SP-book-60K", which is a subset of the 100K hour multi-domain speech we have (SP-multi-100K).

We first compare the top two rows, which differ in the pre-training data and are both fine-tuned with LoRA. Results suggest that while WER remains similar, scaling pre-training data greatly improves style similarity, especially on domains not covered in the fine-tuning data (CV, SWBD, Expr, Accent). On the other hand, scaling fine-tuning data from SP-book-60K to SP-multi-100K does not appear to

Table 2: In-context TTS style similarity and quality subjective evaluation

	Style similarity MOS ↑				
	LS	CV	SWBD	Expr	Accent
YourTTS	1.67 ± 0.09	1.61 ± 0.09	1.55 ± 0.08	1.41 ± 0.07	1.46 ± 0.07
Voicebox	2.85 ± 0.12	2.66 ± 0.13	2.89 ± 0.13	2.42 ± 0.13	2.51 ± 0.11
Audiobox Speech	**3.88** ± **0.11**	**3.77** ± **0.11**	**3.63** ± **0.12**	**3.85** ± **0.11**	**3.77** ± **0.11**
	Quality MOS ↑				
	LS	CV	SWBD	Expr	Accent
YourTTS	1.89 ± 0.10	2.19 ± 0.12	1.57 ± 0.08	1.74 ± 0.09	1.92 ± 0.10
Voicebox	3.70 ± 0.11	3.06 ± 0.12	2.94 ± 0.12	2.76 ± 0.12	3.38 ± 0.12
Audiobox Speech	**4.11** ± **0.08**	**4.00** ± **0.09**	**3.74** ± **0.09**	**4.00** ± **0.09**	**4.22** ± **0.07**

improve much on similarity. This potentially results from the fact that pre-training data is a superset of fine-tuning data, and hence fine-tuning has little to learn on style transfer and focuses on aligning transcript with speech.

Comparing the third and the fourth row, we see that by fine-tuning the whole model, style similarity improves slightly and WER improves greatly on most of the domains (23% to 43% relative WER reduction). The only exception is on SWBD, which are 8kHz narrowband recordings that are likely less represented in the fine-tuning data. Finally, we compare the last two rows and confirm that using audio prompts without silence leads to drastic improvements on similarity on datasets which tend to have long trailing silences (CV, Accent), while overall maintaining the WER. This is because the silence is not informative for inferring the target style.

Table 3: Ablation study for in-context TTS. PT and FT data denote the data used for pre-training and fine-tuning repsectively. FT method denotes whether LoRA or full fine-tuning (full) is adopted. "has sil" denote whether the conditioned audio prompt contains silence.

						Sim-o ↑		
PT data	FT data	FT method	has sil	LS	CV	SWBD	Expr	Accent
SP-book-60K	SP-book-60K	LoRA	Y	0.708	0.461	0.530	0.552	0.529
Mix-185K	SP-book-60K	LoRA	Y	0.718	0.505	0.592	0.571	0.584
Mix-185K	SP-multi-100K	LoRA	Y	0.714	0.502	0.583	0.559	0.590
Mix-185K	SP-multi-100K	full	Y	0.720	0.508	0.556	0.603	0.596
Mix-185K	SP-multi-100K	full	N	0.734	0.607	0.608	0.603	0.659
						WER (%) ↓		
PT data	FT data	FT method	has sil	LS	CV	SWBD	Expr	Accent
SP-book-60K	SP-book-60K	LoRA	Y	4.4	4.4	8.7	4.2	1.5
Mix-185K	SP-book-60K	LoRA	Y	3.8	4.7	8.9	3.9	1.4
Mix-185K	SP-multi-100K	LoRA	Y	3.8	6.0	9.0	4.0	1.4
Mix-185K	SP-multi-100K	full	Y	2.5	3.6	10.1	3.1	0.8
Mix-185K	SP-multi-100K	full	N	3.2	3.7	9.1	3.2	0.9

6 Audiobox Sound: Simple Text-to-sound Generation and Infilling

In this section, we present Audiobox Sound, a model for text-guided generation of general sound. The task is also referred to as text-to-audio generation (TTA) in many prior works[Liu et al., 2023b, Huang et al., 2023b, Kreuk et al., 2022]. It aims to generate general audios given a holistic text description. In contrast to text-to-speech synthesis, the text cannot be frame-wise aligned to audio. Furthermore, sound data only constitutes a small portion of the whole training data. Thus we investigate whether general audio pre-training is able to bring gains to generation of audios of specific domain, which we take sound generation as an example. While we focus on generation of sound events, the technique can similarly apply to other areas (e.g., music).

Most prior works Liu et al. [2023b], Ghosal et al. [2023], Liu et al. [2023c], Huang et al. [2023b], Yang et al. [2023c] build the diffusion models upon a constrained latent space, commonly learned through autoencoding. Such strategy has shown to improve the data efficiency Rombach et al. [2021]. In this work, we adopt a different approach, which directly builds the flow matching network on auto-encoding based latent representation of *raw waveforms*. Such methodology has been largely explored in the language model space Kreuk et al. [2022], Copet et al. [2023], Agostinelli et al. [2023], which typically requires to build a billion-scale model to achieve comparable performance to the alternatives aforementioned. Here we show that by leveraging such simple strategy the flow matching models can achieve SOTA performance while being highly efficient (e.g., $>$ 2x smaller than Kreuk et al. [2022]).

6.1 Method

Similar to speech generation, we model the text-conditional sound distribution with flow matching. In contrast to learning phoneme encoding from scratch, we employ a pre-trained text encoder to map audio captions into word embeddings. Due to the lack of alignment between audio and text embedding, a cross-attention layer is applied in each transformer layer to allow the model attend to the whole text sequence in modeling the gradient distribution, similar to Ghosal et al. [2023], Liu et al. [2023b,c], Kreuk et al. [2022].

Different from prior works in TTA such as AudioLDM [Liu et al., 2023b], AudioLDM2 [Liu et al., 2023c], Tango [Ghosal et al., 2023], we do not rely on an off-the-shelf variational auto-encoder [Kingma and Welling, 2014] to map the low-level audio representation (mel spectrogram) into a latent space and model the distribution in the original embedding space directly. This streamlines the model architecture and reduces the necessity of introducing excessive trainable parameters during fine-tuning, thus bridging the gap between pre-training and fine-tuning.

Except for the cross-attention layers, all the remaining parameters are initialized based on the pre-trained model introduced in Section 4. Similar to text-to-speech synthesis, parameter-efficient fine-tuning strategy like LoRA Hu et al. [2021] can be applied in text-to-audio generation. In practice, we observed fine-tuning the whole model leads to significantly better performance and thus choose to fine-tune the whole model by default (see Section 6.5).

Multi-stage fine-tuning: Compared to transcripts for text-to-speech synthesis, high-quality audio captioning data are much more scarce. Typically, public audio captioning datasets include fewer than 1000 hours of audios, which is orders of magnitude smaller than the speech datasets. On the other hand, the larger-scale sound data often contain noisy category labels and has distributional shift in the audio category [Kim et al., 2019]. To mitigate this issue, we divide the fine-tuning process into two stages, which is based on low-quality (e.g., tags) and high-quality (e.g., human written captions) audio descriptions respectively. Weights of the first model are used to initialize the subsequent model. We argue the labeled data used in first stage, despite its noisy nature, is helpful for learning the text conditional distribution (see Section 6.5).

6.2 Tasks and Evaluation

We consider the following two sound generation tasks: text-to-sound (TTA) generation and text-guided audio infilling (TAI). We use AudioCaps test set [Kim et al., 2019], a standard benchmark for sound generation [Kreuk et al., 2022, Liu et al., 2023b,c, Yang et al., 2023b, Lee et al., 2023, Ghosal et al., 2023], to evaluate all models. For TTA, the model is evaluated standard Frechet Audio Distance (FAD) [Kilgour et al., 2019], Frechet Distance (FD) and KL divergence (KLD) based on the pre-trained audio event tagger PANN [Kong et al., 2019], and Inception score (IS) [Salimans et al., 2016]. FAD and FD measure distribution-level similarity between reference samples and generated samples. KLD is an instance level metric computing the divergence of the acoustic event posterior between the reference and the generated sample for a given description. IS measures specificity and coverage for a set of samples without requiring references, which assigns a higher score if instance posteriors have low entropy and marginal posterior has high entropy. The metrics are implemented following the `audioldm_eval` toolkit.[4]. In addition, we calculate the similarity between generated audio and text description using the CLAP model Wu et al. [2023] [5].

[4] `https://github.com/haoheliu/audioldm_eval`

[5] We use the *630k-best* checkpoint of `https://github.com/LAION-AI/CLAP`

In TAI, the model is conditioned on $p\%$ of the ground-truth audio as context to infill the remaining $(100 - p)\%$, in addition to the text description of the whole audio. In particular, p is set to be 30 and the middle 70% are the region to fill in. In addition to the metrics for TTA, we further measure the similarity to the reference audio (*CLAP-aa*), which is the cosine similarity between CLAP embeddings of the generated and reference audio.

In addition to the objective metrics aforementioned, we also conduct subjective evaluation to evaluate two main aspects of the generated audio: overall naturalness (OVL) and relevance to text input (REL), similar to Kreuk et al. [2022], Liu et al. [2023b]. For these two metrics, raters were asked to rate the perceptual quality and the match between audio and text of the audio samples in a range between 1 and 5 similar to MOS. Based on the evaluation protocol Kreuk et al. [2022], the subjective evaluation is done on 100 randomly sampled files from AudioCaps test set. Each sample is evaluated by 5 annotators from professional annotation service. We list the annotation interface in Appendix D.

6.3 Experimental Setup

Data: To train AUDIOBOX SOUND, we use about 6K hours of audio data, among which ~ 150 hours are captioned audios (SD-cap-150) and the remaining ones only consist of audio tags (SD-tag-6K). During the first-stage fine-tuning, the whole dataset is used while only the captioning data are used in the second stage. To tackle the ontology of audio tags, we concatenate the tags of different levels as the pseudo-caption of the audio. See Table 4 for example audio description in these two sources.

Table 4: Examples of audio descriptions in tag-based and caption-based datasets (Note: the two columns of each row are unaligned.)

Tag-based description	Caption-based description
Animal	A woman talks nearby as water pours
Drill	Multiple clanging and clanking sounds
Fill, Liquid	The sizzling of food while a dish is clanking
Bell, Hall, Room, Inside, Large	a motorboat cruises along, and a man talks
Wolves, Domestic, Animal, Canidae, Dogs, Pets	The wind is blowing, insects are
Bark, Bow-wow, Animals, Growling	singing, and rustling occurs

Implementation Details: We use T5-base [Raffel et al., 2020] to map the text description into embeddings. Each cross-attention layer has 16 heads and its implementation remains same as the self-attention layers except that keys and values are text embeddings. The time-step embedding is added to the T5 embedding before being attended to. In the first stage, we fine-tune the model for 200K updates with an effective batch size of 720K frames. During the second stage, we further fine-tune the model for 100K updates with an effective batch size 240K frames. For both stages, the learning rate and gradient clipping are set to 0.0002 and 0.2 respectively. For inference, we use `dopri5` solver with absolute and relative tolerance of 10^{-5} as the default option. The classifier-free guidance weight is tuned between 0 and 5 and we found setting it to 1 leads to the best result. For each text prompt, we generate 32 random samples and select the one with the highest CLAP similarity to the text prompt. For audio infilling, the masked audio is always kept for conditioning and only the text description is optionally dropped for classifier free guidance.

Baselines: We compare Audiobox Sound against models from the faimily of AudioLDM2 Liu et al. [2023c] and TANGO Ghosal et al. [2023], which stand as current SOTA approaches for general audio generation Liu et al. [2023c].

6.4 Main Results

Text-To-Audio: Table 5 compares our model to prior audio audio generation models in TTA. AUDIOBOX SOUND consistently outperforms all prior works in both objective and subjective evaluation by a large margin, though it is significantly more parameter efficient. It is also worth noting compared to many approaches listed in Table 5, the sound training data we used is also fewer. This further reveals the effect of general domain pre-training for sound generation.

Text-To-Audio Infilling: Table 6 shows the the performance of AUDIOBOX SOUND on TAI, as well as its comparison to prior works. Our model outperforms prior works by a large margin as well on this task. Compared to TAI, we noticed a mixing result according to different metrics. Noticably, the trend on FAD and KLD is not consistently, as in the comparison between TTA and TAI. This can be related to the sensitivity of metrics. On the other hand, the similarity between the generation and reference is greatly increased (CLAP-aa: 0.61→0.77) when the context is fed into the model, which suggests the improvement of coherence to the original audio when context is employed.

Table 5: Text-to-audio generation results on AudioCaps evaluation set. Baselines are evaluated based on the respective official repos. Subjective scores are computed based on 95% confidence interval.

	objective					subjective	
	FAD ↓	FD ↓	KLD ↓	IS ↑	CLAP ↑	OVL ↑	REL ↑
Ground-truth	-	-	-	13.28	0.49	3.36±0.18	3.86±0.18
AudioLDM-L-Full Liu et al. [2023b]	3.37	28.76	1.66	8.72	0.43	2.48±0.14	3.20±0.18
AudioLDM 2-Full Liu et al. [2023c]	1.76	32.12	1.71	8.56	0.43	2.90±0.16	2.98±0.19
AudioLDM 2-Full-Large Liu et al. [2023c]	1.89	33.28	1.60	8.55	0.45	2.90±0.16	3.13±0.17
TANGO Ghosal et al. [2023]	1.57	23.78	1.37	8.30	0.51	3.10±0.14	3.51±0.16
TANGO-full-FT Ghosal et al. [2023]	2.19	18.47	1.20	8.80	0.56	3.04±0.13	3.78±0.15
AUDIOBOX SOUND	**0.77**	**8.30**	**1.15**	**12.70**	**0.71**	**3.43±0.15**	**4.09±0.15**

Table 6: Text-to-audio infilling results on AudioCaps evaluation set. Baselines are evaluated based on the respective official repos. Subjective scores are computed based on 95% confidence interval.

	objective						subjective	
	FAD ↓	FD ↓	KLD ↓	IS ↑	CLAP ↑	CLAP-aa ↑	OVL ↑	REL ↑
Ground-truth	-	-	-	13.28	0.49	-	3.13±0.13	4.21±0.15
AudioLDM-L-Full Liu et al. [2023b]	2.65	21.27	0.84	8.27	0.51	0.76	2.58±0.12	3.58±0.17
TANGO Ghosal et al. [2023]	**1.25**	18.02	0.78	8.53	0.53	**0.78**	2.75±0.12	3.94±0.15
TANGO-full-FT Ghosal et al. [2023]	1.86	15.00	0.71	8.95	0.56	**0.78**	2.79±0.12	4.07±0.14
AUDIOBOX SOUND	1.29	**7.19**	**0.65**	**12.05**	**0.63**	0.77	**2.95±0.12**	**4.20±0.12**

Inference efficiency: In addition to quality metrics, we further show the quality-speed trade-off at inference time in Figure 2. Specifically, we vary the number of inference steps, which correspond to the step size in the ODE solver for our model and the number of DDIM steps in TANGO and AudioLDM2. AUDIOBOX SOUND achieves consistently higher quality (lower FAD) with the same number of inference steps compared to AudioLDM2 and Tango. This implies the better efficiency of the flow-matching approach Audiobox is based on, as is similarly demonstrated in Le et al. [2023].

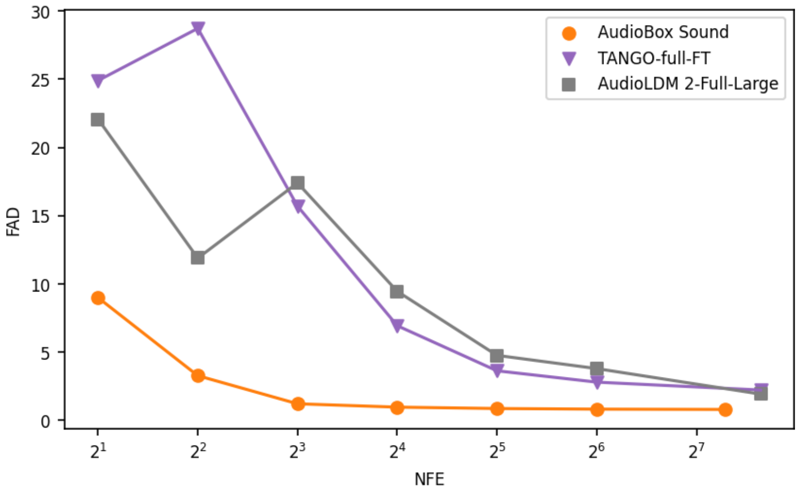


Figure 2: Quality-speed trade-off of AUDIOBOX SOUND, Tango and AudioLDM2. NFE: Number of function evaluations.

6.5 Analysis and Ablation Study

Ablation Study: Here we conduct an ablation study showing the effect of different components of AUDIOBOX SOUND. Specifically, we vary the following training strategies: training with SD-cap-150 only, training with SD-tag-6K and SD-cap-150, training with the whole speech, music and sound datasets.

As is shown in Table 7, using a general pre-trained model boosts the performance by $\sim 20\%$ in FAD. Despite the discrepancy in task and data domain, generation of universal audios is a beneficial pretext task for text-to-sound generation. As music and speech constitutes a significant portion of our evaluation set, increasing the scale of these two modalities in pre-training provides additional benefits. Furthermore, the two-stage fine-tuning also consistently outperforms fine-tuning with SD-cap-150 only regardless of using a pre-trained model or not. The gain is mostly attributed to scaling up in-domain training data (i.e., sound only). Despite the labels being different, simply using audio tags can still enhance learning the mapping between the description of events and the actual audio. Finally, comparing the last two rows of Table 7 suggests reranking with CLAP model is an effective approach to improving the overall performance in both the audio quality (FAD: $0.91 \rightarrow 0.78$) and text-audio relatedness (CLAP score: $0.60 \rightarrow 0.71$).

Fine-tuning strategy We compare the two different fine-tuning strategies: LoRA vs. full model fine-tuning. For LoRA, we add LoRA adaptors described in Section 5 to self-attention layers. In contrast to full-tuning where the whole model is fine-tuned, only the adaptors and cross-attention layers will be updated during fine-tuning and all the remaining parts are frozen. LoRA fine-tuning is on average 15% to 30% worse (relative) than its full fine-tuning counterpart. The incorporation of cross-attention layers induces large architectural change to the model, which increases the necessity of fine-tuning the whole model.

Table 7: Ablation for sound generation on AudioCaps evaluation set. Tag: audio tagging data, Cap: captioning data. Note the results of this table are based on the midpoint solver with a step size of $1/32$ (equivalent to 64 NFE) for the purpose of inference speed-up.

PT (SSL)	FT-1	FT-2	w/ rerank	FAD ↓	FD ↓	KLD ↓	IS ↑	CLAP ↑
✗	-	SD-cap-150	✓	1.17	9.88	1.17	11.43	**0.71**
✗	SD-tag-6K + SD-cap-150	-	✓	1.61	13.16	1.34	10.17	0.67
✗	SD-tag-6K + SD-cap-150	SD-cap-150	✓	0.97	8.70	1.17	12.19	**0.71**
✓	-	SD-cap-150	✓	0.95	8.70	1.15	12.21	0.70
✓	SD-tag-6K + SD-cap-150	SD-cap-150	✗	0.91	8.95	1.33	12.41	0.60
✓	SD-tag-6K + SD-cap-150	SD-cap-150	✓	**0.78**	**8.31**	**1.14**	**12.62**	**0.71**

7 AUDIOBOX: Toward Universal and Controllable Audio Generation

In previous sections, we discussed speech and sound generation independently. This section presents AUDIOBOX, a single model that can produce both speech and audio conditioned on text description or audio example. Fine-tuning our pre-trained model for this joint task enables natural language instruction to control the output speech attributes like perceived age, gender, quality on top of example-based control (ZS-TTS). Furthermore, training on wide variety of data enables simulating voices in different environments and accompanied by acoustic events such as birds chirping, applause. We further envision a scenario where the user would like to *restyle* the given audio example with natural language instruction. For example, change the audio style to make it sound like it is recorded in a cathedral. This requires disentangled vocal style control using an additional utterance from the same speaker called voice prompt.

We design AUDIOBOX to enable speech and sound generation capabilities previously discussed in Sections 5 and 6. Furthermore through voice prompt and description we also envision vocal style transfer to more complex acoustic scenes enabled through joint training. Below we discuss in details speech caption and voice prompt modeling, data creation, and experiments.

7.1 Data Creation

7.1.1 Speech Captions

We aim to bridge the gap between speech and sound datasets by supporting description-based control for speech generation. We consider both human annotations and automatically created captions

Automatic captions: Given the lack of any dataset with fine-grained description for speech, we generate speech captions using a large language model (LLM) with speech attribute tags extracted either using existing metadata or use pseudo labels using classifiers. We extract the following attributes: (1) age: 4 classes (2) gender: 2 classes (3) audio quality: 3 classes (4) pitch: 3 classes (5) speaking rate: 3 classes (6) accent: open-vocabulary (7) emotion: open-vocabulary (8) environment: open-vocabulary More details can be found in Appendix A

Given the above attributes, we use the LLAMA2 7B model Touvron et al. [2023] to convert them into captions. To capture different writing styles, we prompt the model a style bank mimicking different characters with example writing samples. A few of them are listed below:

- A young male adult voice, conveys anger and frustration. The audio, of normal quality, is recorded inside a small space. The person speaks with South Asia accent and a normal speaking pace.
- This young bloke's ticked off, audio's all good. He's in some small space and has a South Asian accent. Talks normal speed.
- Got this young dude who's mad, audio's decent. He's in a tight spot, has that South Asian accent, and talks at a chill pace.
- Young man is angry. Audio is okay, small place. Accent from South Asia. Speaks normal.

To further improve coverage over different environment and background sounds, for each utterance, we apply a random augmentation by convolving with a random room impulse responses (RIR) from a set of known environments and optionally add add a background noise from a set with known tags.

We also generate the corresponding caption with updated environment and background noises using the LLAMA2 7B model. When adding any background noise to the utterance, we update the quality to "low". For utterances applied only RIR we update the quality to be "normal" if the original quality was "studio". We do not apply utterances with low audio quality since those may not be suited for RIR augmentations.

Human annotations: We create human-based annotation to gather more fine-grained description and better alignment towards human hearing perception. We select a 500 hour subset of SP-multi-100K described in Section 5.5.

In the annotation guidelines, we ask the annotator to describe the perceived attribute such as: gender, age, accent, emotion, environment, tonal variation, speaking pace, pitch, emotion, audio quality, vocal style and any miscellaneous details from the speech utterances. In addition to this we also collect categories for the attributes. To ensure we get high quality description, we filter annotators in two stages. First, we keep annotators who successfully labeled pre-selected gold samples with high accuracy. We additionally use an LLM to automatically rate the quality annotations to ensure high quality detailed captions to complement our automatic caption above. More details on quality can be found in Appendix B. Here are some captions example curated by our human annotator:

1. A young woman with an American accent speaks in a higher pitched voice. She speaks at a normal pace with a bit of a muffled voice. She is outside in an urban area and cars can be heard passing by in the background. She has a happy and excited tone that is slightly melodious. The audio is of poor quality and dog barking can be heard at the end.
2. A middle aged man with a mildly masculine voice seems to be outside in a rural or natural environment with a moderately background noise of birds singing. He seems to be in a neutral mood when show casing a house to some people. His voice is hoarse/rough speaking at a slow pace with an average voice pitch.

7.1.2 Voice Prompts

Natural language description alone allows user to control styles through describing attributes such as age, accent, emotion, pitch, and environment. However, a user maybe interested in synthesizing a specific vocal style and while changing other attributes such as quality, emotion, background. This requires disentangled control between the input voice sample and natural language text prompt.

For each target utterance, we sample an additional utterance from the same speaker to serve as voice prompt during training. The voice prompt is selected such that it differs from the target utterance on one or more attribute such as emotion, environment, and speaking rate. This is to de-correlate the target and prompt on everything but vocal similarity. We additionally apply a random room impulse response and background noise augmentation to the voice prompt to increase robustness as well as further de-correlation.

Note that this is different from passing the audio as audio context (zero-shot TTS) where we expect the model to copy over emotion, environment and other background details as well. Here we would want the model to transfer only the vocal style from prompt and use the description for other details such as environment and emotions.

7.2 Method

AUDIOBOX (Fig. 1) conditions on both transcript and masked audio features (same as AUDIOBOX SPEECH) and captions (same as AUDIOBOX SOUND) for description conditional generation. To unify training, for sound inputs without transcript, we create a pseudo-transcript that contains "<sound>" tokens each of length 1 second filling the length of audio. We additionally condition on the another utterance from the same speaker (voice prompt). As described in Section 7.1.2, the voice prompt is selected in a adversarial fashion to enable disentangled control. For audios with missing prompts, we feed a pseudo voice prompt of length 0.1s filled with zeros. The voice prompt is embedded by a lightweight Transformer. We then concatenate the output with the caption description embedding for cross-attention. We randomly initialize the parameters for the cross-attention, description projection, and character embedding weights. All other parameters are initialized based on AUDIOBOX SSL in Section 4. Similar to the sound model training in Section 6, we use multi-stage fine-tuning as described next.

Multi-stage fine-tuning: Except for the high quality 500 hours of speech captions that we collect, the rest of our speech captions are generated using attribute tags and an LLM. Furthermore most of the datasets do not provide any meta-data further limiting the quality of the captions. To mitigate this issue we train our model in two stages. In the first stage we use all the captions for speech and audios. To avoid under-fitting on the audio events generation, we upsample the audio data such that the ratio of total speech and audio data in hours is about $3 : 1$. In the second stage, we initialize the model from first stage weights and only train on the high quality data that comprises 500 hour of annotated speech captions and a few other datasets with emotion and accent metadata for rich LLM captions. We again upsample the audio data such that the ratio of total speech and audio data is about $2.6 : 1$.

7.3 Task and Evaluation

In our unified AUDIOBOX model, the model is capable of new generation tasks such as description-guided TTS (transcript + description) and description-guided TTS with extra voice conditioning generation (transcript + description + voice prompt). Additionally, AUDIOBOX also maintains generation capability from all prior section including: diverse speech sampling (transcript only), zero-shot TTS (transcript + context prompt) (see Sec.5.2), text-to-sound (TTA) generation (description only) and text-guided infilling (TAI, description + context prompt) (see Sec.6.2). In Appendix C, describe the tasks and inputs in detail.

For all speech generation tasks, we measure the WER and similarity of vocal style if context or voice prompt is provided. In addition, for any generation task with description conditioning, we measure the similarity between description and generated audio with cosine similarity between CLAP text and audio embedding. For the description-guided TTS, in addition to objective metric, we also conduct subjective evaluation to assess the QMOS and REL. Below, we provide details on the CLAP model used for speech evaluation.

7.3.1 Joint-CLAP similarity

In terms of tasks, generating speech conditioned on text descriptions is similar to description-guided sound generation (TTA). As is common in TTA, we also employ the text-to-audio similarity to measure how well the generated audio matches the description. However, unlike TTA scenario, joint text-audio embedding models such as CLAP Wu et al. [2023] cannot be straightforwardly applied to the speech domain. Existing CLAP models are trained with coarse description about speech, such as "a person speaking". The model is unable to distinguish fine-grained speaking styles like accent or emotion. Although there exist public CLAP models which are trained with speech data, most of them are trained with (speech, transcript) pairs which is orthogonal to the text description. Thus, for the purpose of evaluating description-conditioned speech generative models, we propose *Joint-CLAP* model, which is designed for both description-based speech and audio evaluation.

Training Similar to CLAP Wu et al. [2023], Joint-CLAP consists of an audio and text branch, each responsible for encoding audio waveforms and the natural language sentences respectively. Given a speech-text pair (x^a, x^t), the audio and text branch f_a and f_t encodes it into the embedding pair (e^a, e^t): $e^a = f_a(x^a)$, $e^t = f_t(x^t)$. We use the same contrastive loss for model training following Wu et al. [2023], Radford et al. [2021], where τ is a learnable parameter.

$$L = \frac{1}{2N}\sum_{i=1}^{N}(\log\frac{\exp{(e^a_i \cdot e^t_i/\tau)}}{\sum_{j=1}^{N}\exp{(e^a_i \cdot e^t_j/\tau)}} + \log\frac{\exp{(e^t_i \cdot e^a_i/\tau)}}{\sum_{j=1}^{N}\exp{(e^t_i \cdot e^a_j/\tau)}}) \tag{2}$$

In practice, we use pre-trained RoBERTa Liu et al. [2019] as the text encoder f_t. In contrast to CLAP, which uses pretrained audio taggers (e.g., HSTAT Chen et al. [2022a]) for audio encoding, here we use WavLM Chen et al. [2022b] as the backbone for encoding. Self-supervised speech models can better capture detailed information (e.g., speaking style) than general audio classifiers. Both RoBERTa and WavLM encoders are fine-tuned in model training.

Data The training data of Speech-CLAP consists of SD-tag-6K, SD-cap-150, and 2K hours of speech datasets including both human and automatic captions. The training set includes both speech and non-speech data in order to equip the model the discriminative capabilities for speaking with environmental sound use cases (e.g., *a man speaks as birds chirp and dogs bark*). The speech portion is a subset of the captioned speech described in Section 7.1.1, which are selected to balance the ratio of human annotated and LLM-augmented captions. The model is evaluated on the evaluation sets of the sound and speech subset respectively.

Implementation Details For audio and text encoder, we use WavLM-base+ and RoBERTa base respectively. Using alternative speech encoders within the same family such as WavLM-large brings similar results. The audio and text embeddings are normalized before calculating the loss (equation 2). The model is trained using Adam optimizer Kingma and Ba [2014] with a learning rate of $5e-5$. We use 64 volta32 GPUs with a batch size of 75 per GPU for 200K updates. For training stability, the gradient is clipped to 10 by norm and raw floating point precision is used without any quantization. We track the recall (A2T@10) on the validation set at the end of each epoch and select the model checkpoint with the highest value.

Retrieval Performance We compare Joint-CLAP to the original CLAPs proposed by Wu et al. [2023], measuring the text-to-audio and audio-to-text retrieval performance. Specifically, we take two public CLAP models trained general audios: *CLAP (general audio)* [6], and general audios plus speech: *CLAP (w/ speech)* [7] Per retrieval task, we report the recall under three thresholds: 1, 5 and 10. As is shown in Table 8, public CLAPs, regardless of whether speech data are utilized or not, achieves significantly lower performance on speech retrieval based text descriptions, with $\sim$ 30x performance degradation compared to the sound benchmark. This might be due to the naturally larger ambiguity in the task, where description of speech may exhibit higher variance. For instance, different people may have varying opinions on what constitutes fast speaking versus slow speaking. In spite of such ambiguity, Joint-CLAP still significantly improves the retrieval performance under the same setting (T2A@10 on speech: 2.29 → 22.01), while maintaining the performance for general audios (T2A@10 on sound: 63.64 → 67.64). The gain is attributed to fine-tuning with speech-specific

[6] https://huggingface.co/lukewys/laion_clap/blob/main/630k-best.pt

[7] https://huggingface.co/lukewys/laion_clap/blob/main/music_speech_audioset_epoch_15_esc_89.98.pt

datasets and using a high-performing speech encoder. To further ablate this effect, we trained a CLAP model without altering the model architecture using in-domain speech data. The retrieval performance is considerably lower than the WavLM-based Joint-CLAP (e.g., T2A@10 on speech: 12.01 vs. 22.01).

Table 8: Comparison between Speech-CLAP and public CLAP models on retrieval performance in sound and speech.

	Speech					
	Text→Audio			Audio→Text		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP (general audio) Wu et al. [2023]	0.36	1.29	2.29	0.64	2.26	3.55
CLAP (w/ speech) Wu et al. [2023]	0.82	2.42	3.37	0.51	1.90	2.60
Speech-CLAP	**7.10**	**16.30**	**22.01**	**5.96**	**16.07**	**22.34**

	Sound					
	Text→Audio			Audio→Text		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP (general audio) Wu et al. [2023]	11.03	45.33	63.64	9.45	44.36	61.70
CLAP (w/ speech) Wu et al. [2023]	11.15	42.42	60.36	9.70	43.15	59.03
Speech-CLAP	**13.33**	**51.88**	**67.64**	**11.27**	**47.27**	**64.48**

Correlation between Joint-CLAP scores and human opionion scores In practice, we also notice the Joint-CLAP model is more closely correlated to human-perceived text-audio similarity, as opposed to the public CLAP model (see Figure 3). Specifically, we take six Audiobox models of varying performance and run subjective evaluation with these models on the four evaluation sets. As is shown in Figure 3, the Pearson correlation coefficient between the text-audio similarity and REL score is increased from 0.028 to 0.727 with a joint CLAP model, suggesting that its text-audio similarity score is a reliable metric for evaluating description-controlled speech generation.

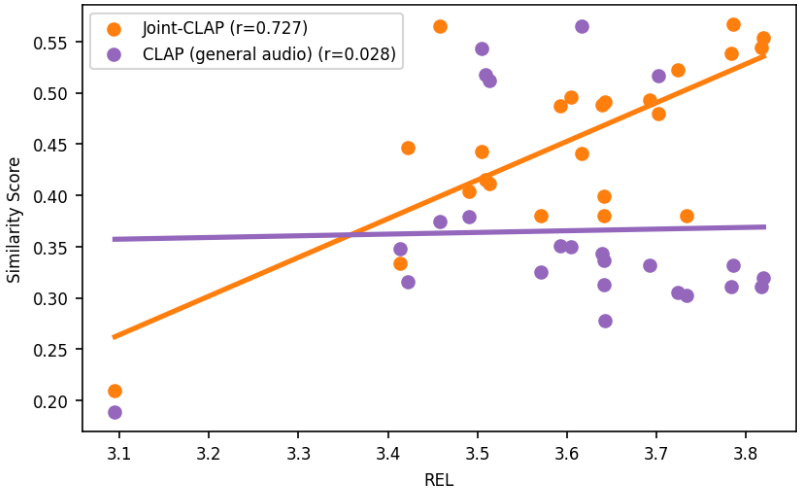


Figure 3: Correlation between text-audio similarity and REL score in different CLAP models. r: Pearson correlation coefficient.

7.4 Experimental Setup

Training data: We train unified Audiobox with a combination of (1) English speech dataset (SP-Multi-100K, see Sec.5.3) with additional text description and voice prompt for each corresponding utterances and (2) sound dataset with text description or tags (SD-TAG-6K and SD-CAP-150, see Sec.6.3). In both cases, each description is either generated from an LLM, or annotated by humans.

We employ two-stage fine-tuning to improve our model fidelity and quality. In the first stage fine-tuning, we incorporate all speech (SP-Multi-100K) and sound (SD-TAG-6K and SD-CAP-150) datasets into our training dataset. In the second stage fine-tuning, we use a subset of our first-stage fine-tuning dataset comprised of higher quality dataset with total about 2,310 hours.

Implementation details: Unified AUDIOBOX model takes four different inputs: 1) frame-aligned transcript, 2) description, 3) voice prompts, and 4) context prompt (masked audio features). First, we first embed the input character sequence in frame-aligned transcript to 128 dimension features. The embedded sequence is then projected using a linear layer and added to the projected masked audio features as input to the Transformer. Next, we use T5-base to extract 512-dimension continuous embedding from the description. The parameters of T5-base are kept frozen during training. We add a trainable linear layer to project the output from 512-dimensions to match the Transformer embedding dimensions (1024). For the voice prompts, we first extract dense features using the same Encodec model described in Section 4. These features are then input to a 3-layered Transformer model with 1024 embedding dimensions, 16 attention heads, and a feed-forward dimension of 4096. We then concatenate the time-step embedding, voice prompt encoder output, and description embedding which form the input for cross-attention.

During training, we randomly drop voice prompt, captions, and context with the probabilities specified in Table 9:

Table 9: Drop-out probabilities for context (ctx), voice prompt (vp), and caption (cap). “F” (false) / “T” (true) refers whether the input is used.

Hyper-parameters			
P(vp=F)	p(ctx=F | vp=T)	P(ctx=F | vp=F)	p(cap=F)
0.5	0.7	0.5	0.3

These probabilities are designed with specific use cases discussed previously. Note that zero-shot TTS requires the model to copy each and every attribute from the audio prompt while restylization requires model to maintain high similarity of vocal style while discarding emotion, environment and other attributes. This requires us to distinguish the context from the voice prompt.

Setting the dropout probabilities as defined in Table 9 lead to the joint probabilities presented in Table 10. The joint probabilities correspond to each of the use case that the model can support. Note that the generative pre-training already tunes model for *ZS-TTS* and *diverse speech sampling* applications. Therefore, we select the hyper-parameters to bias the model towards *description-guided TTS* with and without vocal conditioning.

Table 10: Derived joint probabilities for context, voice prompt, and caption for different use cases.

Hyper-parameters			
ZS-TTS	Description-TTS w/ vocal	Description-TTS	Sampling
P(ctx=T, vp=F, cap=F)	P(ctx=F, vp=T, cap=T)	P(ctx=F, vp=F, cap=T)	P(ctx=F, vp=F, cap=F)
0.075	0.245	0.175	0.075

In the first stage fine-tuning, we fine-tune all parameters for a maximum of 600K updates with 32 A100-80GB GPUs. We stopped training after 350K steps as we didnot find any gains in model performance beyond this. In the second stage, we further fine-tune our model parameter with LoRA fine-tuning on the self-attention parameters with $r = 64$ and cross attention input projection layers for 100K updates with 16 A100-80GB GPUs.

For the unified AUDIOBOX duration model, we use both transcript and the description text as the input. We use 12 Transformers decoder layer with 8 heads, 768/2048 embedding/FFN dimensions self-attention and cross-attention layer to attend the description embedding. We use 40 dimension for the character embedding. During training, we set the description embedding drop probability 0.3. The model trained with 600K updates with flow-matching loss with 8 A100-80GB GPUS. For evaluation, we use the checkpoint at 200K steps.

Evaluation data: We measure the effectiveness of description-guided TTS and description-guided TTS with vocal prompts on the following test sets. First, we annotate a set of 1,946 recordings sampled from diverse sources, including LibriTTS [Zen et al., 2019], Common Voice [Ardila et al., 2019], Switchboard [Godfrey et al., 1992], Fisher [Cieri, Christopher, et al. , 2004,2005a,,], Spotify [Clifton et al., 2020], AudioSet [Gemmeke et al., 2017], Expresso [Nguyen et al., 2023] in order to evaluate the ability to generalize. This set is denoted as SpCap (SC). The second set is AC-filtered (AC-filt) [Lee et al., 2023] with 825 utterances. It constructed from AudioCaps test set by transcribing and keeping samples with reliable ASR transcriptions.

The third one is the Expresso test set (Expr) with 999 utterances. Finally, the fourth one contains utterances from the internal Accent set. We apply randomly sampled RIR and noise augmentation to construct this set and denote it as "Accent+" (500 utterances). Expr and Accent+ use speech captions derived from LLM using the available attributes. For Accent+, we additionally pass the environment and background noises tags to the LLM to incorporate the information into generated captions. Together these sets cover a wide variety of acoustic events, emotions, accents, environments, and vocal styles.

To evaluate description-based TTS with vocal prompt, we use Expr and Accent+ datasets and select another utterance from the same speaker. The prompt is selected such that is different from the target utterance on either emotion or speaking style (enunciated, whisper, etc). Furthermore, we also compare against AUDIOBOX SOUND and AUDIOBOX SPEECH on speech and sound applications using the evaluation sets described in Sections 5 and 6 respectively.

Inference: We use duration model described in this section with averaging over 5 samples. For description-guided TTS (with or without voice prompt), we additionally sample a silence duration of between 0 and 3 seconds and pad it to both ends. We find this generates audios that are coherent with the description particularly when they also mention acoustic events. For example: a man speaks and car passes by while a dog is barking. However, this can cause model to hallucinate sounds when there are no acoustic events described. To cover all scenarios involving description-guided TTS, we generate $N = 8$ samples with stochastic silence padding and then output the best sample based on clap re-ranking using the joint model. We use a guidance weight of 0.75 for the description-guided TTS (with/without voice prompt) applications.

For sound only generation, we always generate 10s long audios with pseudo-transcripts using a guidance weight of 1.33. We use clap reranking with $N = 16$ samples using the sound clap model. For zero-shot in-context TTS applications, we trim the end-silences similar to the AUDIOBOX SPEECH model and use a guidance weight of 1.0. Given that this application doesn't involve any descriptions, we do not use clap re-ranking. Unless specified, both acoustic and duration AUDIOBOX models use the midpoint solver with a step size of $1/32$, which invokes the function being integrated 64 times. When using classifier free guidance the model does 2 forward passes, leading to a total of 128 calls to the model forward pass.

7.5 Main Results

In this section, we investigate the effectiveness of the unified AUDIOBOX model on a number of use cases. We first compare the description-guided TTS with and without voice prompt in Tables 11 and 12 respectively. For this task, we compare with VoiceLDM Lee et al. [2023] and AudioLDM2 Liu et al. [2023c] models as baselines. Next, in Table 13 we evaluate how well AUDIOBOX performs speech tasks as compared to non-description speech only model, AUDIOBOX SPEECH. Finally, Table 14, we compare against the sound-only AUDIOBOX SOUND model on the TTA task.

7.5.1 Description-based control for speech generation

Table 11 compares AUDIOBOX with VoiceLDM Lee et al. [2023] and AudioLDM2 Liu et al. [2023c] models on description-guided TTS and description-guided TTS with voice prompt (voice restylization) tasks. We find that AUDIOBOX outperforms both baselines on all datasets and metrics. In particular, AUDIOBOX is able to consistently generate audios for rich descriptions in SC, background events (AC-filt), expressive audios (Expr), and accented audios with diverse backgrounds (Accent+).

Table 11: Description-based control for speech generation. Audiobox outperforms both AudioLDM2 and VoiceLDM on all datasets and metrics. VoiceLDM and AudioLDM2 models struggle in particular of Expr and Accent+ datasets with expressive audios.

	JointCLAP ↑				WER (%) ↓			
	SC	AC-filt	Expr	Accent+	SC	AC-filt	Expr	Accent+
ground truth	0.403	0.479	0.548	0.561	8.4	23.5	5.8	13.5
VoiceLDM	0.245	0.449	0.060	0.235	8.0	6.8	5.3	4.4
AudioLDM2-SP	0.241	0.225	0.066	0.110	32.5	26.3	33.8	23.9
Audiobox	**0.430**	**0.489**	**0.387**	**0.596**	**7.2**	**5.2**	**4.5**	**2.6**

	QMOS ↑				REL ↑			
	SC	AC-filt	Expr	Accent+	SC	AC-filt	Expr	Accent+
ground truth	3.60± 0.11	3.25± 0.14	4.00± 0.09	3.24± 0.13	3.66± 0.10	3.86± 0.12	4.01± 0.10	3.51± 0.11
VoiceLDM	3.01± 0.10	2.95± 0.13	2.92± 0.12	2.87± 0.12	2.90± 0.10	3.08± 0.14	2.78± 0.11	3.2± 0.11
AudioLDM2-SP	2.19± 0.11	2.17± 0.12	2.47± 0.11	2.25± 0.10	2.37± 0.11	2.11± 0.12	2.48± 0.11	2.22± 0.10
Audiobox	3.58± 0.10	3.38± 0.12	3.82± 0.09	3.54± 0.12	3.74± 0.09	3.61± 0.12	3.94± 0.11	3.61± 0.10

We also note that AudioLDM2 and VoiceLDM struggle in particular on expressive datasets (Expr and Accent+). In particular, we find that utterances generated by AudioLDM2 and VoiceLDM models are significantly worse than the ground truth especially in complicated scenarios involving description of both speech, environment (cathedral), and background sounds. This results in worse scores on the Accent+ dataset. Furthermore, Expr test set contains voices exploring expressive styles like enunciation, whispering, non-binary gender which is where AudioLDM2 and VoiceLDM struggle. We hypothesize this could be because they are out-of-distribution cases w.r.t training. Both VoiceLDM and AudioLDM2 model tend to struggle on such utterances leading to low scores on objective metrics.

Our subjective evaluations also align with the objective metrics where we find the the Audiobox model significantly outperforms the baselines in particular to similarity to the description. The worse scores on Accent+ and Expr dataset for AudioLDM2 and VoiceLDM model further confirms our own observations.

In Table 12, we present the results for description-guided TTS with voice prompt. VoiceLDM and AudioLDM2 model do not simultaneously support conditioning based on vocal and text descriptions for a transcript. Towards our best effort comparison, we combine the CLAP embedding for the audio vocal prompt and the textual description by averaging them and use it as a conditioning input. We find that Audiobox outperforms both baselines. We also notice that in the absence of voice-prompt, the speaker similarity of Audiobox is greatly reduced as the description cannot capture all aspects of voice. The subjective evaluations aligns with the objective metrics both for description and generated audio similarity and speaker similarity. We find that the voice prompt greatly improves the speaker similarity while matching the descriptions.

7.5.2 Comparison to Audiobox Speech and Audiobox Sound

Table 13 compares the unified Audiobox and speech only Audiobox Speech models for zero-shot TTS on 5 different datasets. We use the same duration model for both acoustic models for this task. We find that the unified Audiobox model gives higher speaker similarity but performs marginally worse on the word error rate. This is also confirmed by subjective evaluations where we find only minor differences between the Audiobox and Audiobox Speech models.

In Table 14, we present the results comparing the unified Audiobox to the Audiobox Sound, VoiceLDM, and AudioLDM2 models on the task of TTA task as described in Section 6.2. We find that Audiobox significantly outperforms all baselines achieving the state-of-the-art performance for joint models and even outperforms sound only models such as TANGO. The Audiobox performs worse only to the Audiobox Sound model which specializes in sound generation. The subjective evaluations further confirm that both our Audiobox and Audiobox Sound outperform all other baselines by a significant margin.

Table 12: Description-based control with extra voice conditioning for speech generation

Comparing on objective metrics.

Model	Voice cond.	JointCLAP ↑		Sim-o ↑		WER ↓	
		Expr	Accent+	Expr	Accent+	Expr	Accent+
ground truth	n/a	0.548	0.561	0.395	0.526	5.8	13.5
VoiceLDM	avg. CLAP	0.093	0.204	0.115	0.076	4.8	3.9
AudioLDM2-SP	avg. CLAP	0.067	0.118	0.045	0.089	34.6	30.2
AUDIOBOX	No	0.387	**0.596**	0.181	0.141	**4.5**	**2.6**
AUDIOBOX	Yes	**0.480**	0.593	**0.377**	**0.344**	7.7	2.8

Comparing on subjective metrics for Speaker similarity, quality and description aspects

Model	Voice cond.	QMOS ↑		REL ↑		Speaker Similarity MOS ↑	
		Expr	Accent+	Expr	Accent+	Expr	Accent+
ground truth	n/a	4.0± 0.09	3.24± 0.13	4.01± 0.1	3.51± 0.11	3.38± 0.11	3.27± 0.10
AUDIOBOX	No	3.82± 0.09	3.54± 0.12	3.94± 0.11	3.61± 0.1	3.02± 0.12	3.03± 0.10
AUDIOBOX	Yes	3.86± 0.09	3.58± 0.12	3.99± 0.11	3.57± 0.11	**3.36± 0.11**	**3.24± 0.11**

Table 13: Comparing AUDIOBOX and AUDIOBOX SPEECH model for In-context TTS application. Both model use the same regression based duration model

Style similarity and content correctness using objective metrics

	Sim-o ↑						Word error rate (%) ↓					
	LS	CV	SWBD	Expr	Accent	Avg	LS	CV	SWBD	Expr	Accent	Avg
AUDIOBOX SPEECH	**0.734**	0.607	0.608	0.603	0.659	0.642	**3.2**	3.7	**9.1**	3.2	0.9	**4.0**
AUDIOBOX	0.732	**0.624**	**0.610**	**0.643**	**0.674**	**0.656**	4.8	**3.0**	12.6	**2.7**	0.9	4.8

Style similarity MOS subjective evaluation ↑

	LS	CV	SWBD	Expr	Accent
AUDIOBOX SPEECH	**3.88 ± 0.11**	3.77 ± 0.11	3.63 ± 0.12	3.85 ± 0.11	3.77 ± 0.11
AUDIOBOX	3.72 ± 0.11	**4.03 ± 0.11**	**3.72 ± 0.12**	**4.01 ± 0.10**	**3.88 ± 0.11**

Quality MOS subjective evaluation ↑

	LS	CV	SWBD	Expr	Accent
AUDIOBOX SPEECH	**4.11 ± 0.08**	**4.00 ± 0.09**	3.74 ± 0.09	**4.00 ± 0.09**	**4.22 ± 0.08**
AUDIOBOX	3.95 ± 0.08	3.97 ± 0.09	**3.88 ± 0.08**	3.93 ± 0.09	4.17 ± 0.07

8 Inference Optimization with Bespoke Solver

To generate samples from a flow-matching model, an ODE solver is used at inference time to approximate the integration. There are many solvers that one can choose from, such as adaptive step-size `dopri5` solver or fixed step-size `midpoint` solver. These solvers can be configured to operate at different speed-accuracy trade-off (accuracy in computing the integral). While flow-matching with OT path produces higher quality samples compared to diffusion models [Lipman et al., 2023, Le et al., 2023] for the same number of ODE steps and achieves better trade-off, very aggressive settings like `midpoint` with only 4 steps may still dramatically decrease the sample quality.

Inference efficiency is quantified by the number of function evaluation (NFE), which denotes the number of time an ODE solver evaluates the derivative. To improve the inference speed at the extreme low NFE regime (i.e., 4), we adopt Bespoke Solvers Shaul et al. [2023] to recover similar sample quality as the original model with a much lower NFE.

Assume the initial noise sample $x(0) = x_0 \sim p(x_0)$. Bespoke solver learns extra parameters $\theta \in \mathbb{R}^p$ where p is very small and minimize the *global truncation error* (sum of *local truncation error*) between approximate sample x_n^θ and ground truth data point $x(1)$ in the following formula: $\mathbb{E}_{x_0 \sim p(x_0)} \|x(1) - x_n^\theta\|$, where x_n^θ is the output of the solver step$^\theta$.

At a high level, Bespoke solvers aims to learn transformation for paths such that transformed can be more accurately estimated with the desired number of ODE steps. Bespoke Solver work by transforming the sample trajectory $x(t)$ using two components $t_r : [0, 1] \to [0, 1]$ as time reparameterization and invertible function $\varphi : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$, where those functions are

Table 14: Comparing unified AUDIOBOX for Text-to-audio generation results on AudioCaps evaluation set. We find that AUDIOBOX outperforms all baselines except the sound only AUDIOBOX SOUND. Most notably it even outperforms TANGO-full-FT model on most metrics by significant margin.

	objective					subjective	
	FAD ↓	FD ↓	KLD ↓	IS ↑	CLAP ↑	OVL ↑	REL ↑
ground truth	-	-	-	**13.28**	0.49	3.36± 0.18	3.86± 0.18
Unified Models							
VoiceLDM Lee et al. [2023]	10.28	49.48	2.95	4.79	0.37	2.07± 0.16	2.62± 0.22
UniAudio Yang et al. [2023b]	3.12	-	2.60	-	-	-	-
AUDIOBOX (ours)	**1.10**	**10.14**	**1.19**	**11.90**	**0.70**	**3.19**± **0.14**	**3.94**± **0.14**
Sound-only models							
TANGO-full-FT Ghosal et al. [2023]	2.19	18.47	1.20	8.80	0.56	3.04± 0.13	3.78± 0.15
AUDIOBOX SOUND (ours)	**0.77**	**8.30**	**1.15**	**12.70**	**0.71**	**3.43**± **0.15**	**4.09**± **0.15**

parameterized by extra parameters θ. Let the parametric solver be $\text{step}^{\theta}(t, x; u_t)$. First we transform input (t, x) into $(r, \bar{x}) = (r_t, \varphi_{r_t}(x))$. Next, we perform a step in the transformed space as $(r_{next}, \bar{x}_{next}) = \text{step}(r, \bar{x}; \bar{u}_r)$, using the chosen base solver (e.g., `midpoint`), where $\bar{u}_r$ is vector field on transformed trajectory. To transform back to original space, we compute $(t_{next}, x_{next}) = step^{\theta}(x, t; u_t) = (t_{r_{next}}, \varphi^{-1}_{r_{next}}(\bar{x}_{next}))$.

To train the Bespoke solver, we generate the ground-truth path $x(t)$ at times t_i where $i \in [N]$ using standard ODE solver, and we calculate the *local truncation error* $d_i^{\theta} = \|x(t_i) - step_x^{\theta}(t_{i-1}, x(t_{i-1}); u)\|$ between ground truth and predicted sample from parameterized solver θ, and finally we minimize the Bespoke loss $\mathcal{L}(\theta) = \mathbb{E}_{x_0 \sim p(x_0)} \sum_{i=1}^{n} d_i^{\theta}$.

In this paper, we generate ground truth paths for training Bespoke Solvers for speech generation using `dopri5` ODE solver to estimate $N = 200$ steps with guidance weight (GW) of 0.7. Table 15 top half shows the evaluation result on zero-shot TTS with matched guidance weight (0.7) comparing two standard ODE solvers: `midpoint` and `dopri5` with the Bespoke Solver. As we can see, by using bespoke solver, we could reduce ODE steps down to 4 and still retain similar performance in term of style similarity and WER.

In addition, we also study if a Bespoke Solver trained for a specific guidance weight generalizes to a different guidance weight, and present comparison between the default `midpoint` solver with the bespoke solver using GW=0.0. Results suggest that it can generalize to different guidance setups.

Table 15: Comparison between the standard ODE solver using midpoint, `dopri5` and parameterized Bespoke solver in term of NFE, speaker similarity and WER.

Solver	NFE	GW	Sim-o ↑					Word error rate (%) ↓				
			LS	CV	SWBD	Expr	Accent	LS	CV	SWBD	Expr	Accent
`dopri5`	∼280	0.7	0.733	0.607	0.605	0.602	0.657	3.0	3.6	9.5	2.8	0.9
`midpoint`, 16 steps	32		0.734	0.607	0.608	0.603	0.659	3.2	3.7	9.1	3.2	0.9
Bespoke, 4 steps	8		0.735	0.607	0.606	0.606	0.658	3.0	3.5	8.3	3.0	0.7
`midpoint`, 16 steps	32	0.0	0.671	0.546	0.578	0.541	0.601	3.6	5.1	12.1	3.1	1.3
Bespoke, 4 steps	8		0.672	0.548	0.576	0.544	0.604	3.6	5.1	12.1	3.0	1.3

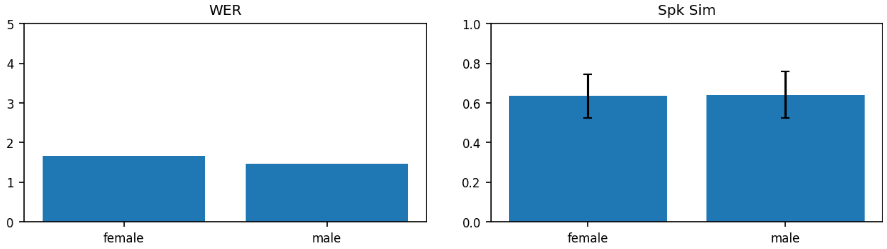
9 Responsible AI

In order to build a system responsibly, we conduct evaluations to gauge the fairness aspect and studies methods to defend misuse. In this section, we first analyze if our model produces similar performance on different groups like genders and accents. Second, we also perform watermarking experiments to evaluate if a recently proposed watermarking system generalizes to our models such that watermarked samples from our models can be reliably detected.

9.1 Fairness across groups

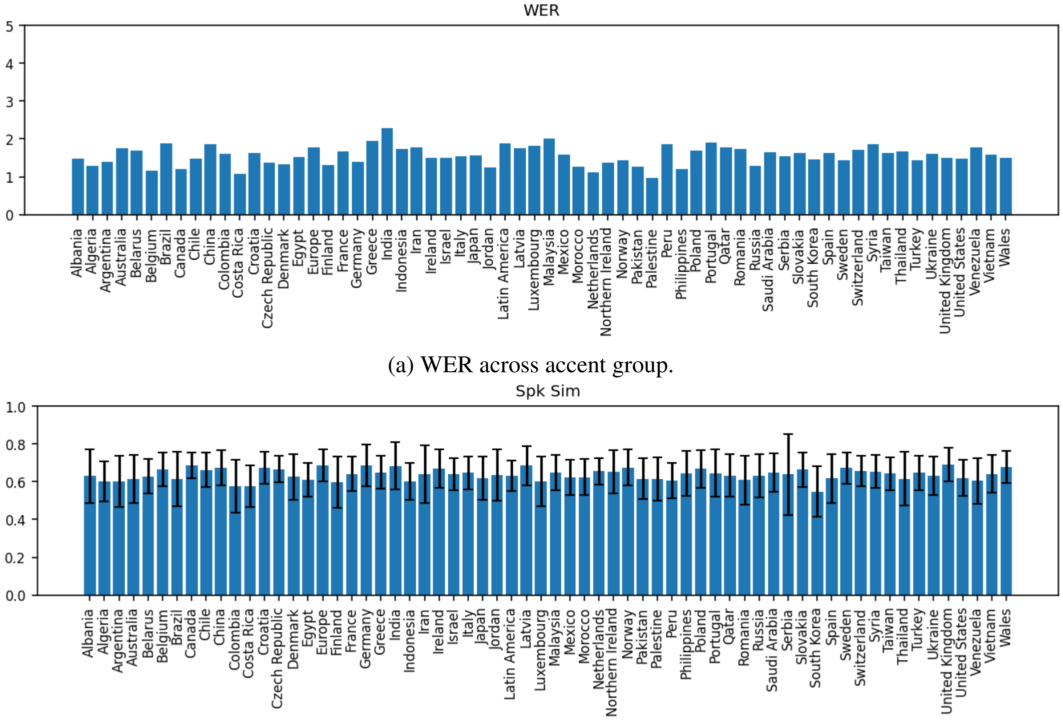
We train our model on large quantities of data from various sources. We believe through scaling training data, our model can perform well across many different groups. We assess this aspects by evaluating model performance by genders and by accents. In particular, we consider gender bias or accent bias are observed if there is a groups that has significantly worse performance in term of content correctness (measured by WER) and style similarity (measured by cosine similarity between style embeddings) compared to those of the entire population.

To conduct our experiment, we consider the zero-shot TTS task conditioned on a context prompt. We use a dataset with country and gender labels for this experiment. For the TTS transcript, we sample 20 transcripts from the test set. For the TTS prompts, we evaluate on accents of which there are at least 5 unique speakers in the dataset, which leave us to 64 accents. Then, we sample 20 random utterances (10 for male, 10 for female) from each accent groups. In total, we have 400 (20 transcripts × 20 prompts) for each accent groups and 12800 (20 transcripts × 10 prompts × 64 accents) for each gender groups.



(a) WER across gender group.

(b) Speaker similarity across gender group (mean ± 1 stddev).



(a) WER across accent group.

(b) Speaker similarity across accent group (mean ± 1 stddev).

Figure 4a shows average WER and 4b shows average speaker similarity across different gender group. We observed that the number are very similar and the speaker similary mean fall between ± 1 standard deviation. Figure 5a shows average WER and 5b shows average speaker similarity across different accent group. Similar with the gender groups, WER over all accents remain similar and each group speaker similarity falls within ± 1 standard deviation. Across gender and accent, WER remains very low around 1.5% which means 1 mistake for every 66 words in the transcript. We come to the conclusion that our model has no significant performance difference given different group of gender and accents.

9.2 Watermarking for Generated Audio Detection

Recent advancement on quality and fidelity in audio generative model has empower novel applications and use case on the model. However, at the same time, there are many people has their raising concerns about the risks of misused. Therefore, the ability to recognize which audio is generated or real is crucial to prevent the misused of the technology and enable certain platform to comply with their policy Fernandez et al. [2023].

In this section, we use Seamless Watermark [Seamless Communication, 2023] to see we can reliably put and detect an imperceptible watermark on top of our model generated audio. The watermarking model has similar building block as Encodec Défossez et al. [2022]. The training objectives are based on weighted combination of two losses: 1) perceptual loss to ensure the watermark is imperceptible (Si-SNR and L1 loss), 2) localization loss based on binary cross entropy to ensure accurate localized detection on watermark in frame level.

Here, we use the output generated from most scenarios such as zero-shot TTS, description-based TTS, voice+description-based TTS, and sound generation and apply various data augmentation on top of them. We measure the performance of watermark detection by their false positive rate (FPR) and false negative rate (FNR).

Augmentation	FPR	FNR
No augmentation	0.001	0
Bandpass filter	0.001	0
Boost audio	0.001	0
Duck audio	0.001	0
Echo	0.001	0.001
Highpass filter	0.001	0
Lowpass filter	0.001	0
Pink noise	0.001	0
Random noise	0	0
Speed slower	0	0.003
Smoothing	0	0.001
Up-down resampling	0.001	0

Table 16: List of audio augmentation technique applied on top of watermarked audio with their detection performance respectively averaged on all scenarios.

Table 16 shows the average FPR and FNR over all tasks for each data augmentations. We observed very low FPR and FNR, close to 0%, which means the watermark works very robustly against various type of generated audio and speech and data augmentations. Simultaneously, the watermarked audio also have very low scale-invariant signal-to-noise ratio (SI-SNR) -20.6db, which means the watermarks residual is in-perceivable from human perspective.

10 Discussion

10.1 Limitations

Fine-grained Control: With the recent advances in generative models, the performance in terms of controllability is mainly determined by the domain coverage and the quantity of the training data. We have demonstrated that for in-context TTS (example-based control), style similarity can be significantly improved by scaling the data used for self-supervised pre-training, which learns to infill audio given the audio context.

In contrast, description-based control requires a higher level of supervision, using paired audio and description to align concepts described in text with variations observed in audio. Hence, it is harder to generalize description-based control due to the scarcity of labeled data covering various concepts and concepts of different granularity.

To give a concrete examples, our training data may contain both instances of chihuahua barking and those of labrador barking; however, all those instances are likely captioned as "a dog barking." Hence, when prompted with "a chihuahua barking," the best the model can do is generating a dog barking audio clip if the text embedding of "chihuahua" and "dog" are close to each other, but it would not be able to generate the correct chihuahua barking sound if such supervision was not provided during training. The same idea also applies to speech attributes such as accents, where regional accents cannot be accurately generated if the training dataset does not include those paired examples.

Data creation: Given the coverage and the quantity of paired data is the key to improve description-based control, it is natural to consider strategies to create such data. However, it is in fact very challenging to create fine-grained descriptions given audio. While it is easy for annotators to differentiate cat meowing and dog barking, labeling which dog species solely based on audio is difficult task for most of the people. Similar challenges exist as well regarding labeling speech attributes such as accents. Moreover, annotators can often disagree on attributes such as emotion, perceived age and quality of audio. Hence, it is difficult to create large scale fine-grained description datasets for audio. The lack of large such datasets also leads to difficulty in developing attribute taggers and captioning models that can automate description creation and be used for evaluation.

10.2 Broader Impact

This work greatly advances controllability for speech generation and improves coverage of styles. The ability to generate speech with desired vocal and acoustic styles using natural language descriptions unlocks a wealth of applications. For example, it can be used to create new voices for characters in immersive audiobooks, Ads, and movie scripts, where creators have in mind what style of voice the characters should have. Compared to exampled-based control (in-context TTS), description-based control can create novel voice of the desired without having to clone from an existing individual and saves the time creators spends on searching the reference voice.

The ability to generate speech in diverse acoustic conditions is especially crucial for applications such as film making and immersive audiobook creation, where characters may be presented in different environment such as caves and it is essential to create audio reflecting the acoustic properties of those scenes. The ability to preserve the voice while changing emotion and acoustic scenes is also crucial for generating long form audio content such as stories. Overall, Audiobox makes it much easier for creators to generate content with higher quality compared to prior models.

While Audiobox can help spark everyone's creativity and bring many positive social impacts, similar to other powerful generative models, it also carries risks of being misused and causing unintended harm. In particular, speech synthesis may be used for spreading misinformation and impersonation. We presented studies on watermarking to effectively mitigate this risk in a robust fashion. On other hand, we also demonstrated that model perform similarly well across variations demographic groups, ensuring bias is reduced through data scaling.

11 Conclusion

This paper presents AUDIOBOX, a unified model for audio generation with unprecedented versatility, controllability, and quality. AUDIOBOX is capable of generating both speech and sound from text description, audio example, or a combination of vocal style reference and description. In particular, for speech generation, AUDIOBOX is able to control very fine-grained vocal styles such as accent, emotion, timbre and create speech simulating more diverse environment compared to previous models. Asides from showing novel capabilities, AUDIOBOX outperforms all prior in-context speech generation and sound generation models on well-studied benchmarks evaluating existing capabilities.

More importantly, we believe this work pioneers in building universal audio generative models with unified controls and sheds light for future research on audio generative modeling. In essence, we demonstrate that with large quantities of data, it is possible to build a unified model that outperforms modality specific ones. This points toward a path similar to the evolution of language generation models, where a large scale model trained with a simple objective on large quantities of data eventually surpasses task or language specific models with significantly better generalization ability and emerging capabilities.

Acknowledgement

The authors would like to thank Ricky Chen, Hady Elsahar, Ilia Kulikov, Hirofumi Inaguma, Jing Xu, and Yossi Adi, Alexander H. Liu, Chung-Ming Chien, Qing He, Thilo Koehler, Fuchun Peng, Xiaohui Zhang, Vimal Manohar, Po-Wei Chou, Kaustubh Kalgaonkar, Anurag Kumar, Yangyang Shi, Zhaoheng Ni, Gael Le Lan, and Varun Nagaraja for their helpful discussion on research, thank Juan Pino, Ian Stewart, Alexander Miller, and Joelle Pineau for organizational support, thank Adina Williams, Christophe Ropers, Chloe Bakalar, Imanol Arrieta Ibarra, and Esteban Arcaute for discussion on responsible AI, thank Wei Zhu, Yichen Wang, Jiun-Ren Lin, Chao Zhou, Peter Weng, Stephen Fink, Ivan Evtimov, David Renardy, Sonia Kim for responsible AI and safety implementation, thank Neil Seejoor, Somya Jain, Chandan Avdhut, Chris Henry, and KC Braunschweig for support on infrastructure, thank Carolyn Krol, Ernest Hammond, Mo Metanat, David Soofian, Ndidi Elue, Mallika Malhotra, Kelechi Ebi Kamanu, Maeve Ryan, Harrison Rudolph, Jennifer Okafor for their support on research and grant review, thank Ana Paula Kirschner Mofarrej, Lydia Baillergeau, Steph Miles, Raghu Nayani, Michelle Restrepo, Tamara Piksa, Chris Wiltz, Orialis Valentin, Aiman Farooq, Gopika Jhala and Ashley Gabriel on cross-functional support

Contribution

Apoorv Vyas proposed and implemented LLM caption, audio augmentation and annotation quality control pipelines, and implemented voice prompting

Bowen Shi led Audiobox-Sound experiments, implemented and conducted experiments for Joint-CLAP, proposed two-stage fine-tuning and led studies on evaluation

Matthew Le implemented and conducted experiments for Audiobox-SSL, Audiobox-Speech, and Bespoke Solver, led model integration to demo

Andros Tjandra implemented speech attribute labelers and responsible AI studies

Yi-Chiao Wu created Audiobox baseline results and implemented audio infilling for baselines

Liang Tan explore speech representation and conducted preliminary experiments on forced aligners

Bowen Shi and Wei-Ning Hsu prepared sound data and implemented, proposed Joint-CLAP and conducted experiments for Audiobox-Sound

Andros Tjandra and Apoorv Vyas implemented Audiobox

Andros Tjandra and Matthew Le conducted experiments for duration models

Apoorv Vyas, Andros Tjandra and Bowen Shi iterated on LLM prompting for text-to-speech and sound training

Apoorv Vyas, Andros Tjandra, Matthew Le, Bowen Shi, Liang Tan and Wei-Ning Hsu prepared speech data

Apoorv Vyas, Andros Tjandra, Matthew Le and Bowen Shi conducted Audiobox experiments

Wei-Ning Hsu, Bowen Shi, Apoorv Vyas, Andros Tjandra, Matthew Le wrote the paper

Baishan Guo, Apoorv Vyas and Andros Tjandra implemented the human annotation pipeline

Baishan Guo ran human annotation and subjective evaluation, and analyzed annotation and evaluation results

Bapi Akula explored audio pre-processing and transformation, assisted in developing data pipeline

Carleigh Wood coordinated and facilitated data annotation

Jiemin Zhang led the demo development, designed and implemented demo infra, model integration, early demo mitigation and capacity testing

Xinyue Zhang designed and implemented demo backend, data logging, mitigation verification and toxic content filtering.

Robbie Adkins designed and implemented demo frontend and supported backend implementation.

Akinniyi Akinyemi conducted demo deployment, demo and mitigation infra set up.

Joshua Lane implemented early UI structure.

William Ngan designed the demo experience and implemented front-end demo interfaces.

Brian Ellis prototyped demo concepts and created audio for demos

Alice Rakotoarison, Chris Summers conducted demo user experience research

Yael Yungster provided design management support

Jeff Wang provided product management support for the team, contributed to overall research vision, strategy, project milestones and execution.

Ivan Cruz provided technical program management support, coordinated responsible AI study, red teaming, and cross-functional support

Rashel Moritz provided program management support, contributed to early project planning, mitigation planning, review, and cross-functional support

Mary Williamson provided management support for the team and co-led the project, contributed to research vision, and oversaw demo

Wei-Ning Hsu designed and led the project, advised Apoorv, Bowen, Matthew, Andros, Yi-Chiao, and Liang on the research, and coordinated research, demo, and data streams.

References

A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *International Conference on Language Resources and Evaluation*, 2019.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.

Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

H. Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022a.

R. T. Q. Chen. torchdiffeq, 2018. URL `https://github.com/rtqichen/torchdiffeq`.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Neural Information Processing Systems*, 2018.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022b.

Cieri, Christopher, et al. . Fisher English training speech parts 1 and 2 LDC200{4,5}S13. *Web Download. Linguistic Data Consortium, Philadelphia*, 2004,2005a.

Cieri, Christopher, et al. . Fisher English training speech parts 1 and 2 transcripts LDC200{4,5}T19. *Web Download. Linguistic Data Consortium, Philadelphia*, 2004,2005b.

A. Clifton, A. Pappu, S. Reddy, Y. Yu, J. Karlgren, B. Carterette, and R. Jones. The spotify podcast dataset. *arXiv preprint arXiv:2004.04270*, 2020.

J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez. Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022.

C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22466–22477, October 2023.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

D. Ghosal, N. Majumder, A. Mehrish, and S. Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan. PromptTTS: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, et al. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*, 2019.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023a.

R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023b.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazar'e, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. rahman Mohamed, and E. Dupoux. Libri-Light: A benchmark for asr with limited or no supervision. *International Conference on Acoustics, Speech and Signal Processing*, 2019.

E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision, 2023.

K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

C. D. Kim, B. Kim, H. Lee, and G. Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

J. Kong, J. Kim, and J. Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2019.

F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

A. Kulkarni, V. Colotte, and D. Jouvet. Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 31–35. IEEE, 2021.

A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096680.

A. Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.

M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023.

S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

Y. Lee, I. Yeon, J. Nam, and J. S. Chung. Voiceldm: Text-to-speech with environmental context, 2023.

Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, et al. PromptTTS 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*, 2023.

P. Li, B. Chen, Y. Yao, Y. Wang, A. Wang, and A. Wang. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. *arXiv preprint arXiv:2308.04729*, 2023.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu. Generative pre-training for speech with flow matching, 2023a.

H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023b.

H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023c.

X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang. Separate anything you describe. *arXiv preprint arXiv:2308.05037*, 2023d.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. URL `https://api.semanticscholar.org/CorpusID:198953378`.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, 2017.

T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*, 2023.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. *International Conference on Acoustics, Speech and Signal Processing*, 2015.

A. Plaquet and H. Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020.

O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *ArXiv*, abs/2108.12409, 2021.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL `https://api.semanticscholar.org/CorpusID:231591445`.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021.

A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016. URL `https://api.semanticscholar.org/CorpusID:1687220`.

F. Schneider, Z. Jin, and B. Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

Seamless Communication. Seamless: Multilingual expressive and streaming speech translation. 2023.

N. Shaul, J. Perez, R. T. Chen, A. Thabet, A. Pumarola, and Y. Lipman. Bespoke solvers for generative flow models. *arXiv preprint arXiv:2310.19075*, 2023.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *International Conference on Acoustics, Speech and Signal Processing*, 2017.

K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

K. J. Shih, R. Valle, R. Badlani, A. Lancucki, W. Ping, and B. Catanzaro. Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

C. Wang, S. Chen, Y. Wu, Z.-H. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv*, abs/2301.02111, 2023a.

Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao. Audit: Audio editing by following instructions with latent diffusion models. *arXiv preprint arXiv:2304.00830*, 2023b.

Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

J. Yamagishi, C. Veaux, and K. MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.

D. Yang, S. Liu, R. Huang, G. Lei, C. Weng, H. Meng, and D. Yu. InstructTTS: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*, 2023a.

D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023b.

D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023c.

N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2022.

H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023.

A Speech Attributes for Speech Caption Creation

As described in Section 7.1.1, we extract attributes for creating speech captions. We obtain speech attributes from the associated metadata or by pseudo-labeling for a subset of attributes which can be labeled more reliably. Details for each attribute are listed below

- Age: We first bin the age into 4 different categories namely less than twenty (<20), young adults (20-35), middle age (40-60), and elders (>60). We then fine-tune our dataset from pre-trained WavLM-base checkpoint with 3200 hours speeech and age metadata from our training set (consisted of conversational and reading speech with various quality).
- Gender: We fine-tune on top of WavLM-base checkpoint with 4300 hours speech and gender metadata from our training set (consisted of conversational and reading speech with various quality).
- Audio Quality: We use TorchAudio-Squim Kumar et al. [2023] library and extract Perceptual Evaluation of Speech Quality (PESQ) Rix et al. [2001] score. We then bin the score into three categories: Low quality (0-2.39), Normal quality (2.39-3.8) and Studio Quality (>3.8).
- Pitch: We use PyWorld vocoder [8] to extract fundamental frequency (f0) and then calculate the geometric mean across all voiced region. We use gender dependent threshold for binning the pitch into three different categories: low, normal, high. For gender masculine, we set low pitch (0-40 percentile), normal pitch (40-90 percentile) and high pitch (>90 percentile). For gender feminine, we set low pitch (percentile 0-10), normal pitch (10-60 percentile) and high pitch (>60 percentile). The logic behind asymmetric threshold is because in general people will perceive most of masculine voice have lower pitch and most of feminine voice have higher pitch.
- Speaking rate: Given the transcript and audio, we first apply VAD to remove the silence segments. We then calculate character per seconds (CPS) and bin them into 3 categories: slow (<9.2 CPS), high (>20.8 CPS) and normal (9.2 <= x <= 20.8 CPS).
- Accent: We use the accent from the metadata whenever available in the metadata, otherwise leave it blank.
- Emotion: We use the emotion labels whenever available in the metadata, otherwise we leave it as blank.
- Environment: We use the environment tags such as inside a room, outside whenever available from the datasets.

B Automatic Captions: Quality

To ensure we get high quality descriptions, we deploy a two-stage approach to filter the annotator candidate. First, we keep only annotators that successfully labeled pre-selected gold samples with high accuracy (> 73%). Later, we score their submitted captions using LLM and keep the annotator if their averaged score is above certain threshold. More specifically, for a speech segment, we first use a LLM to generate a caption based on annotated audio attributes. We then run a second stage where we ask the another LLM to compare the LLM-generated caption with human-written caption and rate the quality of human-written captions from 1 to 5. We prompt this LLM to give low score to human-written captions where no interesting audio events were added in additional to the annotated audio attributes or some important audio attributes are missing. Annotators with an averaged caption score less than 3 were removed. This resulted in high quality and detailed captions that complement our pseudo-labeled captions above. Here are some captions example curated by our human annotator:

C Unified Audiobox Task Description

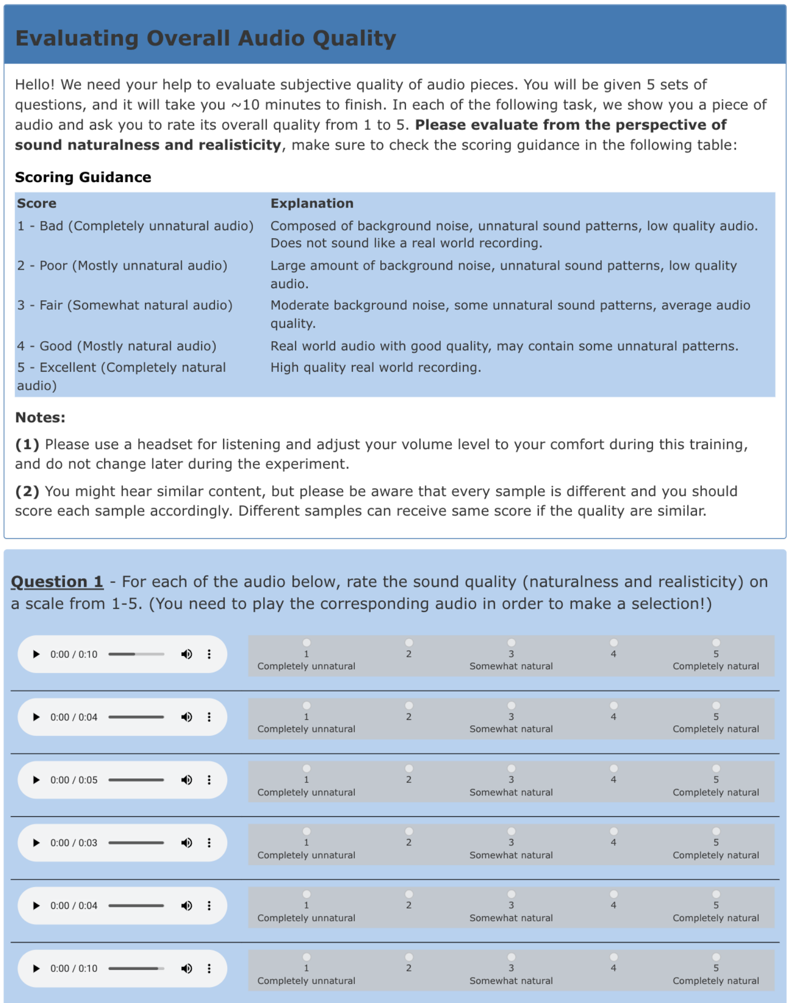
Below we describe different tasks that unified AUDIOBOX model can solve along with the inputs required.

[8] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

- Zero-shot TTS (in-context TTS): model takes as input a transcript and an audio example and generates speech that resembles the example's audio style (described in Section 5.2). Inputs: (context, transcript).
- Description-TTS/TTA: model takes as input a transcript/pseudo-transcript and a text description and generates speech / audio matching the description. Inputs: (description, transcript/pseudo-transcript)
- Voice Restylization: model receives a transcript, a voice prompt and a description. The generated output needs to match speaker's vocal style and the description. Note that the description could contain attributes different from the voice prompt. For the voice prompt could have been recorded in a small room with neutral emotion and the description could specify happy emotion in a church with bells ringing. Inputs: (voice, description, transcript)
- Sampling: The model receives a transcript as input and samples diverse voices. Inputs: (transcript)
- Speech Infilling/Editing: model takes as input an masked speech with accompanying transcript and an optional description and infills the masked portion. Inputs: (context, transcript, optional description)
- Audio Infilling/Editing: model takes as input an masked audio with pseudo transcript and description to infill the masked portion with matching description. Inputs: (context, pseudo transcript, description)

D Subjective Evaluation Interface

We show human annotation interfaces for sound in Figure D1 D2, for speech in Figure D3 D5 D4.



Evaluating Overall Audio Quality

Hello! We need your help to evaluate subjective quality of audio pieces. You will be given 5 sets of questions, and it will take you ~10 minutes to finish. In each of the following task, we show you a piece of audio and ask you to rate its overall quality from 1 to 5. **Please evaluate from the perspective of sound naturalness and realisticity**, make sure to check the scoring guidance in the following table:

Scoring Guidance

Score	Explanation
1 - Bad (Completely unnatural audio)	Composed of background noise, unnatural sound patterns, low quality audio. Does not sound like a real world recording.
2 - Poor (Mostly unnatural audio)	Large amount of background noise, unnatural sound patterns, low quality audio.
3 - Fair (Somewhat natural audio)	Moderate background noise, some unnatural sound patterns, average audio quality.
4 - Good (Mostly natural audio)	Real world audio with good quality, may contain some unnatural patterns.
5 - Excellent (Completely natural audio)	High quality real world recording.

Notes:

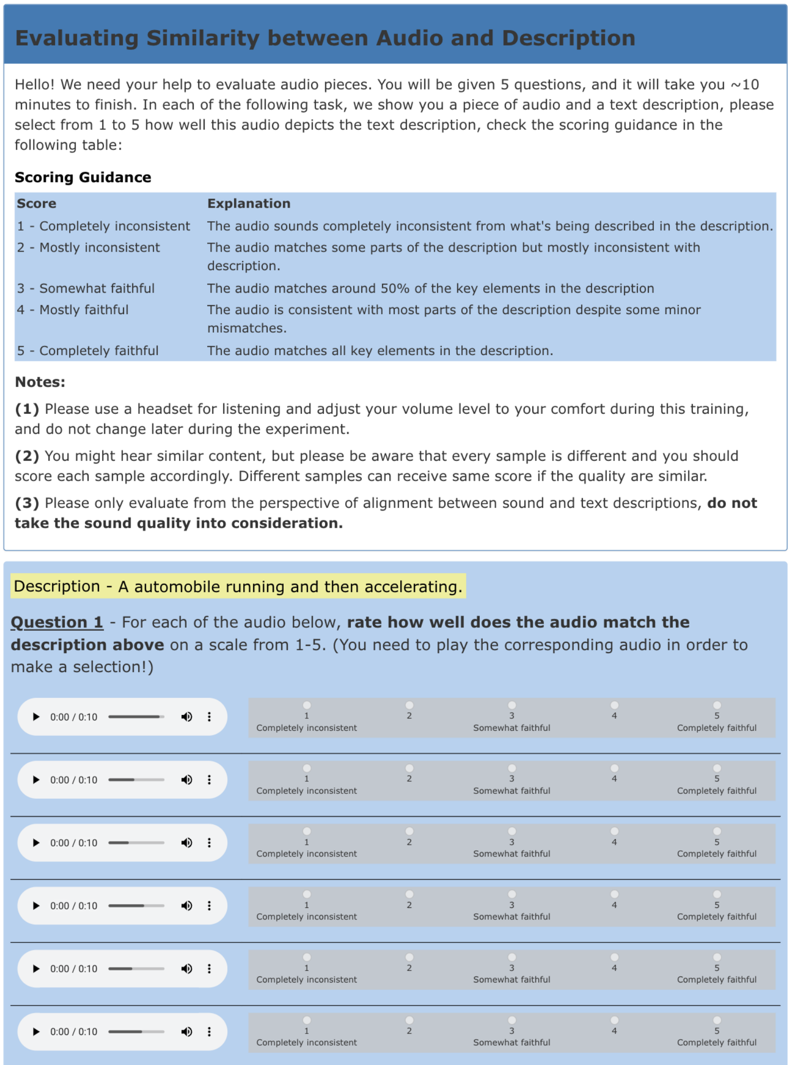
(1) Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

(2) You might hear similar content, but please be aware that every sample is different and you should score each sample accordingly. Different samples can receive same score if the quality are similar.

Question 1 - For each of the audio below, rate the sound quality (naturalness and realisticity) on a scale from 1-5. (You need to play the corresponding audio in order to make a selection!)

▶ 0:00 / 0:10	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:04	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:05	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:03	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:04	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:10	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural

Figure D1: OVL evaluation for sound



Evaluating Similarity between Audio and Description

Hello! We need your help to evaluate audio pieces. You will be given 5 questions, and it will take you ~10 minutes to finish. In each of the following task, we show you a piece of audio and a text description, please select from 1 to 5 how well this audio depicts the text description, check the scoring guidance in the following table:

Scoring Guidance

Score	Explanation
1 - Completely inconsistent	The audio sounds completely inconsistent from what's being described in the description.
2 - Mostly inconsistent	The audio matches some parts of the description but mostly inconsistent with description.
3 - Somewhat faithful	The audio matches around 50% of the key elements in the description
4 - Mostly faithful	The audio is consistent with most parts of the description despite some minor mismatches.
5 - Completely faithful	The audio matches all key elements in the description.

Notes:

(1) Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

(2) You might hear similar content, but please be aware that every sample is different and you should score each sample accordingly. Different samples can receive same score if the quality are similar.

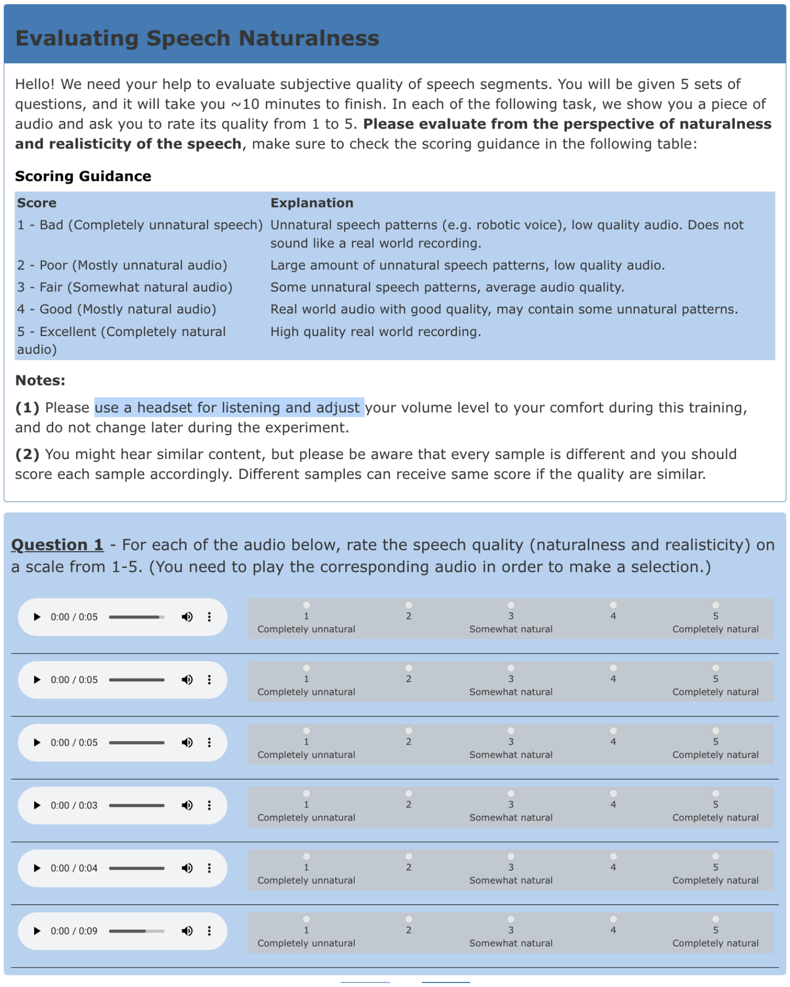
(3) Please only evaluate from the perspective of alignment between sound and text descriptions, **do not take the sound quality into consideration.**

Description - A automobile running and then accelerating.

Question 1 - For each of the audio below, **rate how well does the audio match the description above** on a scale from 1-5. (You need to play the corresponding audio in order to make a selection!)

▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful
▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful
▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful
▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful
▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful
▶ 0:00 / 0:10	○ 1 Completely inconsistent	○ 2	○ 3 Somewhat faithful	○ 4	○ 5 Completely faithful

Figure D2: REL evaluation for sound



Evaluating Speech Naturalness

Hello! We need your help to evaluate subjective quality of speech segments. You will be given 5 sets of questions, and it will take you ~10 minutes to finish. In each of the following task, we show you a piece of audio and ask you to rate its quality from 1 to 5. **Please evaluate from the perspective of naturalness and realisticity of the speech**, make sure to check the scoring guidance in the following table:

Scoring Guidance

Score	Explanation
1 - Bad (Completely unnatural speech)	Unnatural speech patterns (e.g. robotic voice), low quality audio. Does not sound like a real world recording.
2 - Poor (Mostly unnatural audio)	Large amount of unnatural speech patterns, low quality audio.
3 - Fair (Somewhat natural audio)	Some unnatural speech patterns, average audio quality.
4 - Good (Mostly natural audio)	Real world audio with good quality, may contain some unnatural patterns.
5 - Excellent (Completely natural audio)	High quality real world recording.

Notes:

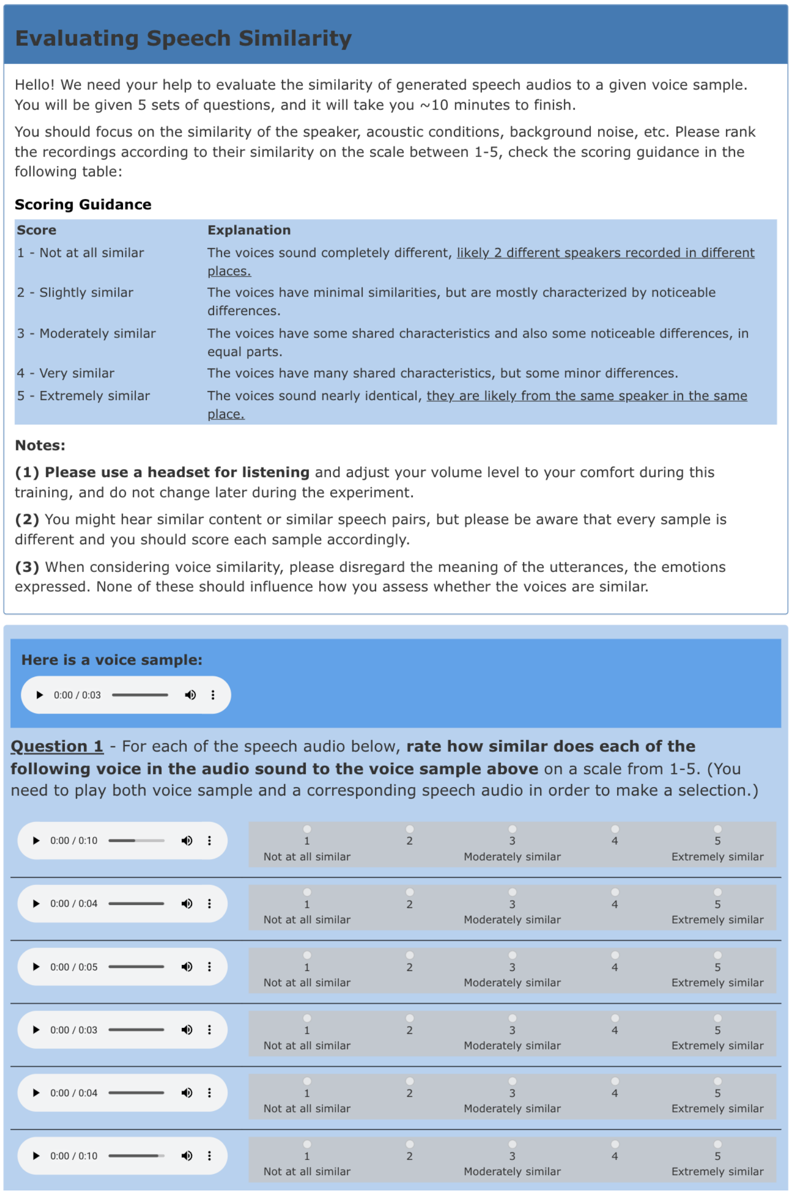
(1) Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

(2) You might hear similar content, but please be aware that every sample is different and you should score each sample accordingly. Different samples can receive same score if the quality are similar.

Question 1 - For each of the audio below, rate the speech quality (naturalness and realisticity) on a scale from 1-5. (You need to play the corresponding audio in order to make a selection.)

▶ 0:00 / 0:05	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:05	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:05	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:03	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:04	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural
▶ 0:00 / 0:09	○ 1 Completely unnatural	○ 2	○ 3 Somewhat natural	○ 4	○ 5 Completely natural

Figure D3: Quality MOS evaluation for speech



Evaluating Speech Similarity

Hello! We need your help to evaluate the similarity of generated speech audios to a given voice sample. You will be given 5 sets of questions, and it will take you ~10 minutes to finish.

You should focus on the similarity of the speaker, acoustic conditions, background noise, etc. Please rank the recordings according to their similarity on the scale between 1-5, check the scoring guidance in the following table:

Scoring Guidance

Score	Explanation
1 - Not at all similar	The voices sound completely different, likely 2 different speakers recorded in different places.
2 - Slightly similar	The voices have minimal similarities, but are mostly characterized by noticeable differences.
3 - Moderately similar	The voices have some shared characteristics and also some noticeable differences, in equal parts.
4 - Very similar	The voices have many shared characteristics, but some minor differences.
5 - Extremely similar	The voices sound nearly identical, they are likely from the same speaker in the same place.

Notes:

(1) Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

(2) You might hear similar content or similar speech pairs, but please be aware that every sample is different and you should score each sample accordingly.

(3) When considering voice similarity, please disregard the meaning of the utterances, the emotions expressed. None of these should influence how you assess whether the voices are similar.

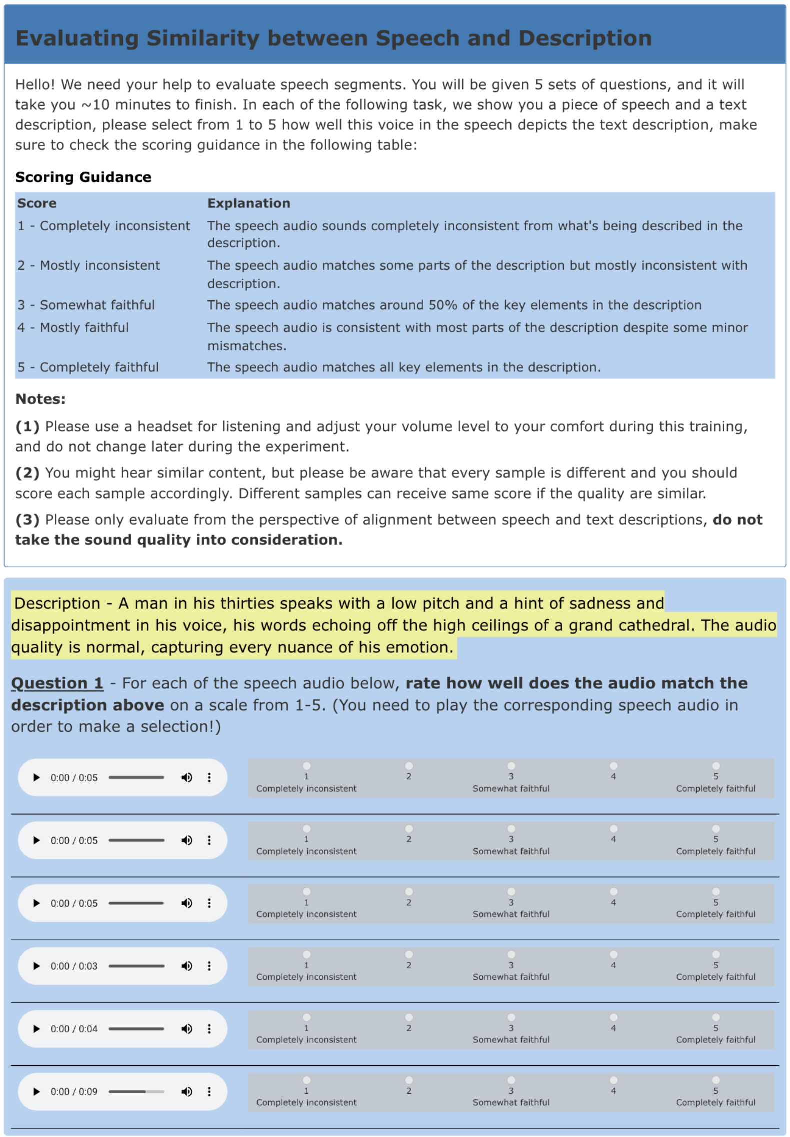
Here is a voice sample:

▶ 0:00 / 0:03

Question 1 - For each of the speech audio below, **rate how similar does each of the following voice in the audio sound to the voice sample above** on a scale from 1-5. (You need to play both voice sample and a corresponding speech audio in order to make a selection.)

▶ 0:00 / 0:10	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar
▶ 0:00 / 0:04	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar
▶ 0:00 / 0:05	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar
▶ 0:00 / 0:03	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar
▶ 0:00 / 0:04	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar
▶ 0:00 / 0:10	○ 1 Not at all similar	○ 2	○ 3 Moderately similar	○ 4	○ 5 Extremely similar

Figure D4: Similarity MOS evaluation for speech



Evaluating Similarity between Speech and Description

Hello! We need your help to evaluate speech segments. You will be given 5 sets of questions, and it will take you ~10 minutes to finish. In each of the following task, we show you a piece of speech and a text description, please select from 1 to 5 how well this voice in the speech depicts the text description, make sure to check the scoring guidance in the following table:

Scoring Guidance

Score	Explanation
1 - Completely inconsistent	The speech audio sounds completely inconsistent from what's being described in the description.
2 - Mostly inconsistent	The speech audio matches some parts of the description but mostly inconsistent with description.
3 - Somewhat faithful	The speech audio matches around 50% of the key elements in the description
4 - Mostly faithful	The speech audio is consistent with most parts of the description despite some minor mismatches.
5 - Completely faithful	The speech audio matches all key elements in the description.

Notes:

(1) Please use a headset for listening and adjust your volume level to your comfort during this training, and do not change later during the experiment.

(2) You might hear similar content, but please be aware that every sample is different and you should score each sample accordingly. Different samples can receive same score if the quality are similar.

(3) Please only evaluate from the perspective of alignment between speech and text descriptions, **do not take the sound quality into consideration.**

Description - A man in his thirties speaks with a low pitch and a hint of sadness and disappointment in his voice, his words echoing off the high ceilings of a grand cathedral. The audio quality is normal, capturing every nuance of his emotion.

Question 1 - For each of the speech audio below, **rate how well does the audio match the description above** on a scale from 1-5. (You need to play the corresponding speech audio in order to make a selection!)

▶ 0:00 / 0:05 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

▶ 0:00 / 0:05 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

▶ 0:00 / 0:05 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

▶ 0:00 / 0:03 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

▶ 0:00 / 0:04 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

▶ 0:00 / 0:09 — ○ 1 Completely inconsistent ○ 2 ○ 3 Somewhat faithful ○ 4 ○ 5 Completely faithful

Figure D5: REL evaluation for speech