Scaling Zero-Shot TTS with Speaker-Agnostic Training

Anonymous ACL submission

Abstract

The goal of language model (LM)-based zero-shot text-to-speech (TTS) is to synthesize speech with voices unseen during training. However, zero-shot TTS requires labeled speaker information for each utterance during training. This information is expensive to acquire, making it difficult to scale systems to large amounts of data. In this paper, we show that these issues can be overcome by simply combining a large dataset without speaker labels and a smaller dataset with speaker labels, before training a TTS model on the mixture. To prevent information mismatch between the two types of data, we introduce new data augmentation techniques to regularize model training: speaker dropout and speaker scrambling. As a result, we achieve relative gains up to 64% better speaker similarity and 80% lower WER, when compared to standard training recipes. We show that our method not only generalizes well to low-resource and cross-lingual settings, but also scales to over 200K hours of training data. We will open-source all code and pre-trained models. Audio samples are available at https://cccmon7.github.io/opus_tts/.

1 Introduction

Auto-regressive language models have recently become a popular formulation for text-to-speech (TTS) systems (Wang et al., 2023; Du et al., 2024; Maiti et al., 2024; Défossez et al., 2024) due to their ability to easily leverage text-only pre-training from Large Language Models (LLMs) (Touvron et al., 2023; Brown et al., 2020). These systems are capable of generating fluent and natural-sounding synthetic speech in a variety of voices while being relatively easy to train, making them a prime target for scaling (Huang et al., 2025).

Increasing amounts of research has focused on this task of multi-speaker TTS, as it allows models to leverage more training data while allowing them to generate more diverse audio. Zero-shot TTS (Wang et al., 2023; Chen et al., 2024a; Casanova et al., 2022, 2024; Wu et al., 2022) is a particularly exciting implementation of this concept, as it allows models to synthesize speech in voices that were unseen during training. These models leverage an enrollment speaker prompt during inference, which contains an audio example of the voice that the model should mimic.

One key limitation of zero-shot TTS models is their dependence on speaker information during training, because learning to clone a voice demands an additional utterance from the same speaker as the target example. Obtaining such speaker labels is a non-trivial cost that requires either manual annotations or complex speaker diarization pipelines (Park et al., 2022), if not outright impossible due to privacy concerns and data access conditions: storing such sensitive biometric data with the corresponding speech is a genuine security issue. Such expenses, combined with the complexity of TTS model architectures (Du et al., 2024; Wang et al., 2023; Wu et al., 2022), only further increase the difficulty in scaling TTS models to more data and larger model sizes.

In this paper, we propose a straightforward two-step method to relax this data constraint: we first 1.) merge **gold**-quality data annotated with speaker information and **silver**-quality data without speaker information, and then 2.) propose new data augmentation techniques to regularize model training on the mismatched data.

We ground our method in the same theoretical mechanisms that motivate classifier-free guidance (Ho and Salimans, 2021), leveraging a generative model's capacity to capture both conditional and unconditional distributions. We exploit this formulation to scale TTS models to large amounts of data: samples with incomplete speaker metadata are modeled unconditionally, while samples with the complete metadata are trained conditionally. Although simply combining the incomplete

data can lead to large gains in generation quality, with relative gains up to 60% better speaker similarity and 60% lower word error rate-based intelligibility (WER), we also show that such an implementation is suboptimal. We propose two straightforward data-augmentation techniques to alleviate the train–test mismatch caused by incomplete data: (1) randomly dropping the speaker prompt and (2) sampling artificial speaker prompts from the target speech. Together, these augmentations boost speaker similarity by 2% and reduce WER by 48%.

Our contributions can be summarized as follows:

1. We present a method that loosens the data constraints of zero-shot TTS, making it far easier to scale models to larger training corpora.

2. Our methods can lead to zero-shot TTS models that are more robust, leading to relative gains up to 64% better speaker similarity and 80% lower WER, when compared to the vanilla TTS training recipe.

3. Our methods reduce the amount of labeled speaker information needed to train TTS systems: our models perform comparably to those trained without the augmentation on twice the amount of data. This method generalizes cross-lingually, enabling the training of zero-shot TTS models without any speaker information in a language.

4. By reducing the labeled information needed for TTS training, we show that our method can reduce the costs of automatic data annotation by as much as 22%, allowing us to train a SOTA TTS model on 200K hours of audio.

2 Related Work

2.1 Zero-Shot TTS

The goal of zero-shot TTS is to synthesize speech with voices unseen during training. Large-scale training for zero-shot TTS has focused on two main branches of work: LM-based (Peng et al., 2024; Wang et al., 2023; Chen et al., 2024a; Du et al., 2024) and diffusion-based (Le et al., 2023; Lipman et al., 2023; Eskimez et al., 2024). Diffusion-based (and by extension flow-matching) approaches center around training a non-autoregressive model on continuous speech representations (Le et al., 2023; Liu et al., 2024). However, their non-autoregressive nature often requires explicit duration modelling or even frame-level speech/text alignments (Le et al., 2023). While these can be addressed with certain training techniques (Eskimez et al., 2024), it comes at the cost of heavier inference time constraints and thus limits the usability of these models.

The LM-based approach typically involves training an auto-regressive language model (Brown et al., 2020; Touvron et al., 2023) on discrete speech tokens (Borsos et al., 2023; Lakhotia et al., 2021; Nguyen et al., 2023) quantized from a speech representation model (Chen et al., 2024b, 2023a; Kumar et al., 2023; Shi et al., 2024b; Chen et al., 2023b). This approach yields several key advantages, namely the ability to leverage text pretraining from LLMs and their highly optimized training frameworks/software (Rasley et al., 2020; Dao et al., 2022; Dao, 2024). Combined with the storage-efficient nature of the discrete speech tokens, this has made the LM-based approach much easier to scale (Huang et al., 2025). Due to these advantages, our work focuses on this formulation.

2.2 Speaker Dependencies in Zero-Shot TTS

Few works have attempted to remove the dependency on speaker labels in zero-shot TTS. The most similar to our work are SPEAR-TTS (Kharitonov et al., 2023) and CosyVoice (Du et al., 2024), which decouple LM-based TTS into two cascaded stages: 1.) text to semantic tokens and 2.) semantic to acoustic tokens. In SPEAR-TTS, the second module is trained on short segments by sampling a prompt and target subsequence from each utterance. While this allows for efficient self-supervised training, it disregards cases where speaker data is in-fact available and prevents the model from learning long-form speaker, phonetic, and prosodic patterns. CosyVoice instead uses an averaged continuous speaker embedding, which removes prosodic and pronunciation information. Furthermore, the embedding is *always* obtained from the same target speech utterance during training, potentially overfitting to a single speaker prompt[1]. Our method can be viewed as a form of semi-supervised training that addresses these issues, being able to scale to more data *and* train on a diverse selection of speaker prompts. Furthermore, our method only requires a single-stage, simplifying the training process and thus reducing the compute requirements, while removing issues caused by potential errors in the cascade during inference.

[1]This limitation also applies to models that instead condition the LM on continuous speaker representations.

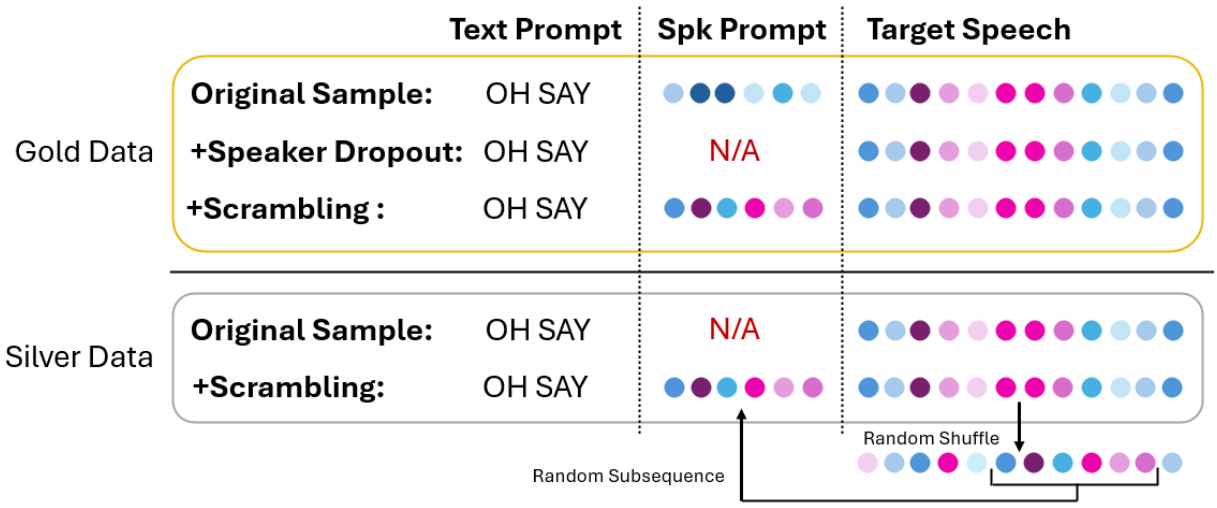

Figure 1: Overview of our proposed data concatenation and augmentation strategies for speaker-agnostic training.

E2 TTS (Eskimez et al., 2024) addresses the speaker dependency in flow matching models, although their goal was simplifying the training pipeline rather than scaling. They accomplish this by framing TTS as conditional masked language modeling, where the model must in-fill audio that corresponds to a masked time span, given the unmasked audio and text prompt. However, the major limitation of this method is that it requires the transcript of the speaker prompt during inference, which limits the cross-lingual capabilities of the model while being expensive to obtain. While this can be ameliorated by force-aligning the speech and text during training, this introduces a significant cost that makes scaling even more difficult. Our method does not introduce any additional preprocessing expenses, making it far more scalable.

3 Method

LM-based TTS auto-regressively models the conditional probability of the t-length target speech tokens $Y = (y_t|t = 1, ..., T)$, given the input text sequence $X = (x_n|n = 1, ..., N)$ and speaker token prompt $S = (s_k|k = 1, ..., K)$:

$$\prod_t P(y_t|y_{1:t-1}, S, X) \quad (1)$$

This formulation places two main constraints on the data that can be used for TTS training: 1.) X must be the text transcript of Y and 2.) Y and S must be obtained from the same speaker. The former constraint is relatively easy to address, since paired speech/text is often found naturally on the internet. The second constraint, which this work focuses on, is more problematic to satisfy. Crawled web data generally does not include per-utterance speaker information (Chen et al., 2021; Galvez et al., 2021; Li et al., 2023; He et al., 2024), and performing speaker identification poses serious privacy concerns. While the former can be addressed by pseudo-labeling (as done by He et al. (2024)), this is very expensive and difficult to scale to large amounts of data, while often introducing inaccuracies from multi-speaker segments (Clifton et al., 2020; He et al., 2024).

3.1 Speaker-Agnostic TTS Data Relaxation

Rather than being constrained by the limitations of the conventional zero-shot TTS paradigm, we propose a training framework that is *agnostic* to speaker labels and facilitates scalable exploration of TTS properties. In other words, the proposed approach enables model training without the need for speaker annotations. We define two types of training data: (1) **gold**-standard data with paired speech/text *and* speaker labels (X, Y, and S), and (2) **silver**-standard data with only paired speech/-text (X and Y). The LM is thus trained to not only perform speaker-conditional generation (Equation 1), but also speaker-unconditional generation:

$$\prod_t P(y_t|y_{1:t-1}, X) \quad (2)$$

In a theoretically ideal setting, this probability is near impossible to model without S since the value of Y is different for every speaker. However, since in practice the LM is trained with teacher-forcing, it instead models the probability of the next speech token y_t given the text sequence X and prior speech tokens $[y_1, ..., y_{t-1}]$. The latter effectively acts as

both prior context and a speaker prompt, making a solution feasible.

3.2 Speaker Dropout

To prevent the model from overfitting its speaker prompting abilities on the **gold**-standard data (as in Equation 1), we randomly drop out the speaker prompt S during training with some probability p. For each utterance in the **gold** data, we uniformly sample a value $\hat{p}$. If $\hat{p} <= p$, the model is only conditioned on the text input X (Equation 2). Otherwise, the model is conditioned on both the text X and speaker prompt S (Equation 1).

We note that the same dropout strategy is standard in classifier-free guidance (Ho et al., 2021), which uses an unconditional model (Eq. 2) to steer a conditional one (Eq. 1) during inference. This dropout allows a single language model to capture both the conditional and unconditional distributions. In our case, however, the dropout is used purely for training regularization to prevent domain-specific overfitting, since our use of the **silver**-quality data (Section 3.1) already enables the training of an unconditional model[2]. To the best of our knowledge, no prior works (Hussain et al., 2025; Darefsky et al., 2024) have studied the impact of this form of `speaker dropout` technique outside the use of classifier-free guidance.

3.3 Speaker Scrambling

Although speaker dropout curbs overfitting in conditional generation, it offers scant benefit on out-of-domain (OOD) inputs relative to the **gold**-standard audio. The underlying issue is a persistent train–test mismatch: the model is never trained to produce speaker-conditioned outputs for the **silver**-standard data. We tackle this limitation with a new straightforward augmentation scheme: `speaker scrambling`. For a target speech sequence Y, we exploit the time-invariance of speaker identity by randomly shuffling its acoustic tokens along the temporal axis. To curb data leakage, we then truncate the shuffled sequence at a random position r, yielding an artificial speaker prompt $\hat{S}$:

$$\hat{S} = \text{Truncate}(\text{Shuffle}(Y), r) \tag{3}$$

If truncation is not performed, the model will be overly biased towards the prompt tokens. During training, we append $\hat{S}$ to each **silver**-standard example—or replace the original prompt S in **gold**-standard data—with probability v. The input text X and the target speech Y remain untouched, resulting in

$$\prod_t P(y_t | y_{1:t-1}, \hat{S}, X) \tag{4}$$

The technique is visualized in Figure 1 and an example Python implementation of this technique is shown in Listing 1.

Listing 1 Example Python code for `speaker dropout` and `speaker scrambling`.

```python
import random
import numpy as np

def preprocess(text, speech, spk, p, v):
    """
    Args:
        text (array): input text
        speech (array): target speech
        spk (array or None): speaker prompt
        p (float): speaker dropout prob
        v (float): scrambling prob
    Returns:
        text (array): input text
        speech (array): target speech
        spk (array or None): speaker prompt
    """
    # speaker dropout
    if random.random() < p:
        spk = None
    # speaker scrambling
    if random.random() < v:
        l = len(speech)
        # randomly shuffle across time
        idx = np.random.permutation(l)
        scrambled = speech[idx]
        # take a random subsequence
        start = random.randint(0, l // 2 - 1)
        spk = scrambled[start: start+l // 4]
    return text, speech, spk
```

4 Experimental Setup

Our models use a delay interleave decoder-only Transformer (Vaswani et al., 2017) architecture (Copet et al., 2024) for multi-stream language modeling. Audio waveforms are converted into discrete tokens with a DAC-style neural codec (Kumar et al., 2023) from ESPnet-Codec[3] (Shi et al., 2024b) (8 streams, vocabulary size= 8192) and XEUS (Chen et al., 2024b) embeddings quantized with K-means (1 stream, vocabulary size= 5000), leading to a total of 9 audio input/output streams. The Transformer decoder and text embeddings are

[2] Since our overall method is a training technique and classifier-free guidance is an inference method, they can be combined and used in a single model. We leave such explorations to future work.

[3] https://huggingface.co/ftshijt/espnet_codec_dac_large_v1.4_360epoch

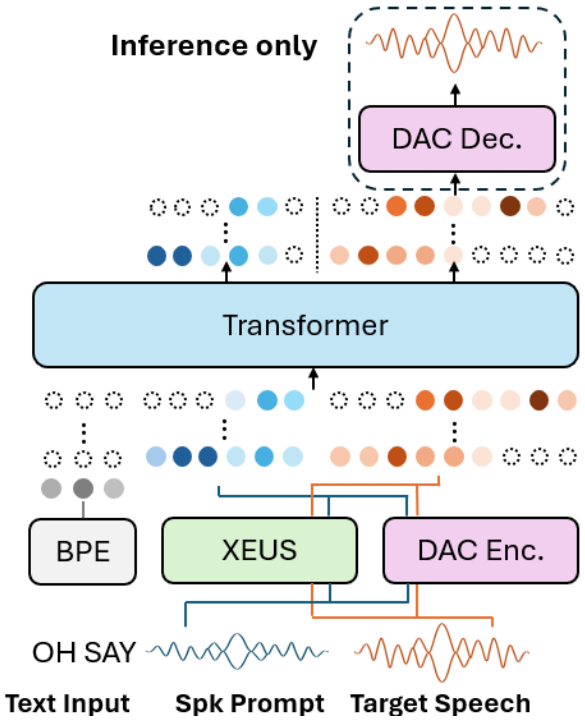

Figure 2: Overview of the multi-stream architecture. The input text tokens are shown in grey, the speaker prompt in blue, and the target speech tokens in orange.

initialized with a pre-trained SmolLM with 360M parameters [4]. We use the same text tokenizer as the pre-trained model, which has a vocabulary size of 49K. Models are trained for 400K steps with pure bfloat16, Deepspeed (Rasley et al., 2020), and the Adam optimizer (Kingma and Ba, 2015). The learning rate is linearly warmed up to 0.0001 for 10K steps, after which it is held constant. We use the final checkpoint for evaluation. Inference is performed with top-k sampling with $k = 30$ and a sampling temperature of 0.7. Sampling is performed 10 times per input, and we report the average result across the 10 generated samples. We set both the speaker dropout p and scrambling rate v to 0.5. Each model is trained on 16 NVIDIA GH200 GPUs for 30 hours. We conduct our experiments using the ESPnet-SLM toolkit (Tian et al., 2025; Watanabe et al., 2018).

4.1 Training Data

Our primary training set consists of 46K hours of English audiobook recordings from LibriVox, obtained by combining LibriSpeech 960 (Panayotov et al., 2015) and the English portion of MLS (Pratap et al.). Importantly, the metadata for these recordings include the ground-truth speaker ID, which allows us to use this dataset as the **gold**-standard TTS data described in Section 3.1.

For the **silver**-standard speech data where the ground-truth speaker information is unknown, we use a cleaned version of YODAS (Li et al., 2023) from Tian et al. (2025), which has 70K hours of English speech mined from YouTube.

Note that this version of YODAS contains speaker pseudo-labels through the use of a pre-trained Pyannote diarization model (Bredin et al., 2020) during the cleaning process, based on He et al. (2024). We generally do not use these pseudo-labels, so that we can better simulate in-the-wild conditions where speaker information is truly unknown. The only exception is in Section 5.5, where we compare the effectiveness of our method against speaker pseudo-labels.

4.2 Evaluation

To evaluate the generalizability of our technique, we use multiple test sets from a wide variety of domains for our evaluation: LibriSpeech test-clean (audiobooks) (Panayotov et al., 2015), VCTK (accented read speech) (Yamagishi et al., 2019), and Genshin [5] (voice dubbing). We consider VCTK as the primary measure of performance, as it is considered OOD for both the **gold** and **silver** quality data while being a standard academic benchmark. Ablative studies (Sections 5.2, 5.3, and 5.5) are performed on the in-domain LibriSpeech test-clean.

Following recent speech synthesis studies (Eskimez et al., 2024; Wang et al., 2023; Chen et al., 2024a; Anastassiou et al., 2024; Huang et al., 2023), we use automatic proxy metrics for objective evaluations. We focus on three such proxy metrics: Word Error Rate (WER) for intelligibility, UTMOS (Baba et al., 2024) for audio quality, and Speaker Embedding Similarity (SPK). WER is obtained from Whisper V3 Turbo (Radford et al., 2023) and SPK is calculated using a pre-trained WavLM+ECAPA-TDNN (Jung et al., 2024; Chen et al., 2022). We use VERSA (Shi et al., 2024a) to calculate each metric.

5 Results

Our main results are shown in Table 1, where we vary the TTS model training along two axes: use of data augmentation (none vs the methods proposed in Section 3) and training data (**Gold** only vs **Gold**+**Silver**). Model A1 is trained on only **gold**-standard data without the proposed data augmenta-

[4] https://huggingface.co/blog/smollm

[5] https://github.com/espnet/espnet/tree/master/egs2/genshin/tts1

Table 1: Cross-domain evaluation of each TTS model, trained with or without the proposed augmentation techniques. Models are evaluated using WER (↓), Proxy MOS (↑), and SPK (↑). Model A1 serves as the baseline system that represents the standard TTS LM training recipe, whereas B2 uses our proposed methods.

ID	Training Data	Augmentation	test-clean			VCTK			Genshin		
			WER	SPK	UTMOS	WER	SPK	UTMOS	WER	SPK	UTMOS
A1	Gold	✗	6.0	0.62	3.96	12.4	0.28	3.68	23.7	0.39	3.34
A2	Gold	✓	5.6	0.62	3.99	2.9	0.39	3.93	20.4	0.39	3.25
B1	Gold+Silver	✗	6.5	0.67	3.91	4.9	0.45	3.80	23.8	0.46	3.08
B2	Gold+Silver	✓	**4.7**	**0.68**	**4.05**	**2.5**	**0.46**	**4.00**	**16.8**	**0.47**	**3.47**

tion, serving as the baseline that represents the standard TTS LM training recipe. While it achieves strong performance on the in-domain evaluation (test-clean), it generalizes poorly to OOD speaker prompts (VCTK and Genshin), with noticeable degradations on the two other test sets. Naively adding the silver-standard data largely alleviates this issue (A1 vs B1), but leads to degradations in in-domain performance. Nevertheless, the largest improvements from scaling to more data are in speaker similarity, with a relative 64% increase on VCTK. The results show that B2, the model trained on the additional silver-quality data with data augmentation, clearly performs the best: model B2 achieves the best scores in all 9 metrics / datasets. The data augmentation not only recovers the degradations on test-clean WER and MOS, but also improves generation quality in all 3 metrics in the OOD test sets. This highlights the importance of using *both* of our proposed techniques in building more robust TTS systems.

5.1 Subjective Evaluations

We also conducted subjective evaluations on A1, B1, and B2 using two tests: Comparative Mean Opinion Score (CMOS) for naturalness and Speaker Similarity Mean Opinion Score (SMOS), following the setups of (Ju et al., 2024; Eskimez et al., 2024; Wang et al., 2023). We generated 20 samples from each system, each from a different speaker. We hired 30 MTurk annotators, leading to 600 total samples per system for each test. For CMOS, annotators compared the generated sample with the ground truth, without knowing which was which. They were asked to rate naturalness on a scale of -3 to 3, where negative values indicate preference for the ground truth and vice versa for positive values. For SMOS, annotators compared the audio prompt with the generated sample or ground truth and rated the speaker similarity on a scale of 1 (not similar at all) to 5 (identical). More details can be found in Appendix 6.

Table 2: Subjective evaluation on VCTK.

ID	Data	Aug.	CMOS↑	SMOS↑
	Ground Truth	-	n/a	$4.11_{\pm 0.05}$
A1	Gold	✗	$-0.016_{\pm 0.084}$	$2.89_{\pm 0.08}$
B1	Gold+Silver	✗	$0.020_{\pm 0.084}$	$3.58_{\pm 0.06}$
B2	Gold+Silver	✓	$-0.085_{\pm 0.086}$	$\mathbf{3.74}_{\pm 0.06}$

We find that all models achieved a degree of naturalness that is indistinguishable from the ground truth (Ju et al., 2024; Anastassiou et al., 2024; Eskimez et al., 2024), with no statistically significant differences across the three evaluated models ($p \approx 0.8$). However, there were clear differences in the SMOS scores, with models also trained on silver-quality data adhering significantly better to the audio prompt (2.89 for A1 vs 3.58 and 3.74 for B1 and B2, respectively). This shows the importance of scaling to more data as well as the effectiveness of our technique for doing so.

5.2 Impact of Data Augmentation

We analyze the effect of changing the speaker prompt dropout/scrambling rate on TTS quality. We train 4 different TTS models on the Gold+Silver mixture with dropout/scrambling rates of $[0.0, 0.1, 0.3, 0.5]$, respectively. Table 3 shows the results of this ablation on test-clean. We found that higher dropout/scrambling rates to be more effective, with values of 0.3 and 0.5 both outperforming the no augmentation baseline in most metrics. Interestingly, using a value that is too low (0.1) leads to significant performance degradations across all 3 metrics.

We also ablate the effects of each data augmentation technique, with results on test-clean shown in Table 4. Applying speaker dropout alone improves WER at the cost of SPK, with no change in MOS, while the addition of speaker scrambling further improves all metrics.

Table 3: Impact of the scrambling/dropout rate on LibriSpeech test-clean.

p/v	WER	SPK	UTMOS
0.0	6.5	0.67	3.91
0.1	9.7	0.65	3.75
0.3	6.3	0.66	3.93
0.5	**4.7**	**0.68**	**4.05**

Table 4: Ablating the effect of each data augmentation on LibriSpeech test-clean, with $p = 0.5$.

Dropout	Scrambling	WER	SPK	UTMOS
✗	✗	6.5	0.67	3.91
✓	✗	6.3	0.66	3.91
✓	✓	**4.7**	**0.68**	**4.05**

5.3 Impact of Speaker Information Amount

In this section, we analyze the effect of decreasing the amount of ground-truth speaker labels in our data, with and without our proposed data augmentation techniques. We train 4 TTS models on different subsets of the Gold+Silver data mixture by decreasing the amount of gold-standard data from 46K hours to 960 hours, 400 hours, and 200 hours respectively. This process is conducted once with models trained using the proposed augmentation techniques, and once without any augmentation.

The change in WER, SPK, and MOS on LibriSpeech test-clean as the amount of gold-standard data is decreased is shown in Figure 3. Models trained with the data augmentation generally perform better across all 3 metrics at each resource level, showing the generalizability of our method. One can also observe that our proposed methods scale well to the low-resource regime, showing that our technique can enable stronger TTS systems with minimal amounts of labeled speaker data.

5.4 Cross-lingual Generalization

The lack of speaker labels is an even larger limitation to non-English TTS, as speaker information is even more scarce. Diarization models are also more inconsistent in multilingual settings (Kalluri et al., 2024), making pseudo-labels weak. We thus ask the following research question: can we still create a zero-shot TTS system without any speaker information in a language? We analyze the cross-lingual generalizability of our method, where the gold-standard with speaker information is in one language, and the silver-standard data is in another.

In this experiment, we use the 2000 hour German subset of MLS (Pratap et al.) for the silver-standard data, and compare it to a model trained only on German MLS, but with the ground truth speaker labels. We evaluate models with the same 3 metrics, although we note that UTMOS is not trained on German, hence lower overall UTMOS scores when compared to previous sections. In Table 5, our method without any German speaker labels achieves only 0.8 worse WER, while achieving higher SPK and UTMOS. This shows the potential of our method in aiding the development of large-scale TTS systems in non-English languages.

Table 5: Cross-lingual benchmark on MLS German.

Spk Label	WER	SPK	UTMOS
Ground Truth	**20.1**	0.40	2.53
Augmentation	20.9	**0.57**	**2.70**

Table 6: Comparison on test-clean of different speaker pseudo-labeling methods for silver-quality data.

Spk Label	WER	SPK	UTMOS
None	6.5	0.67	3.91
Diarization	8.3	0.65	3.89
Scrambled (proposed)	**4.7**	**0.68**	**4.05**

5.5 Comparing with Speaker Pseudo-labels

An alternative (albeit expensive) method to our proposed technique is to generate pseudo speaker labels for the training data by using a speaker diarization or clustering model. This is the process used to create large-scale open TTS datasets, such as Emilia (He et al., 2024). However, diarization is costly, and may represent up to 22% of the compute used for pseudo-labeling (Figure 4).

Table 6 shows the compares models trained on the Gold+Silver data mixture with different speaker pseudo-labeling strategies on the silver-quality data: no speaker pseudo-labels, diarization-based pseudo-labels, and our speaker scrambling-based pseudo-labels. We find that the use of the diarization-based pseudo-labels can actually *degrade* performance, likely due to mis-identifying different speakers as a single person and thus lead to training on bad prompts. On the other hand, our proposed method shows clear improvements over the no labeling baseline.

5.6 Scaling Properties

We combine the speaker-agnostic training lessons learned in the previous sections to develop a new

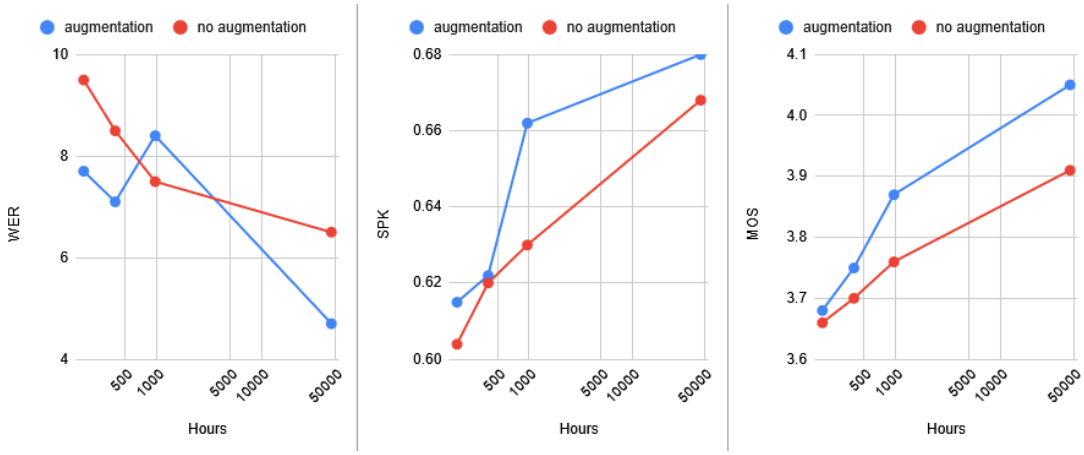

Figure 3: Change in evaluation metrics as the amount of gold-standard data increases.

Figure 4: Comparison of data cleaning costs with and without speaker diarization.

SOTA TTS model trained on over 200K hours of English audio. We collect additional data from the 100,00 Podcasts Dataset (Clifton et al., 2020) and the Emilia (He et al., 2024). 100,00 Podcasts consists of 60K hours of English podcast audio from Spotify, whereas Emilia contains 45K hours of cleaned web-crawled English. Due to the noise in the original meta-data, we clean and re-segment the Spotify audio by performing voice activity detection and ASR, intentionally omitting diarization to show the usefulness of our technique in real-world settings. By omitting the diarization step, we are able to make data processing 22% faster, reducing the average processing time from 212 seconds per clip to 166 seconds (Figure 4).

We compare our scaled model (which we refer to as OpusTTS) to 5 SOTA open-source models in Table 7 on LibriSpeech test-clean: CosyVoice (Du et al., 2024), reproduced versions for VallE-X and VallE-2 (Wang et al., 2023; Chen et al., 2024a), ESPnet-SLM (Tian et al., 2025), and Sesame CSM (Schalkwyk et al., 2025). OpusTTS achieves the best SPK, second-best WER, and third-best UTMOS, despite being relatively lightweight in parameter count and only using transparent data.

Table 7: Comparison with SOTA zero-shot TTS systems on test-clean. ⋆ indicates reproduction. The best model is **bold**, while the second best is underlined.

Model	Params.	WER	SPK	UTMOS
CosyVoice	300M	5.0	0.51	**4.15**
VallE-X ⋆	300M	27.3	0.35	3.38
VallE 2 ⋆	800M	27.8	0.46	3.65
ESPnet-SLM	440M	**3.1**	0.55	4.03
CSM	1B	20.3	0.65	4.11
OpusTTS (ours)	440M	4.5	**0.71**	4.07

6 Conclusion

In this paper, we present a novel yet simple technique to relax the data constraints of language model-based zero-shot TTS systems, allowing them to be more easily scaled to over hundreds of thousands of hours of audio. Our method exploits the teacher-forcing paradigm used to train Transformer-based language models, bypassing the dependency on speaker information. We showcase the effectiveness of this formulation by training models on a mixture of gold-quality TTS with speaker labels and silver-quality data without speaker labels. In doing so, we show the importance of incorporating certain data augmentation techniques in making this formulation work, such as speaker dropout and our newly introduced speaker scrambling. Our technique outperforms the use of diarization-based pseudo-labeling and can generalize to cross-lingual transfer learning. Finally, we showcase the production capabilities of our method by scaling TTS training to 200K hours, leading to a new SOTA open-source model. In the future, we hope to scale to even more data and enhance our model's capabilities with post-training.

Limitations

While our method allows for LM-based TTS systems to be more easily scaled to large amounts of training data, the effectiveness of this technique is still bound by the quality of the data collected. Carelessly including large amounts of low-quality data, such inaccurate speech/text pairs or noisy audio, can instead harm performance. The performance of LM-based TTS methods are also bound by the quality of their speech tokenizer, whose reconstruction accuracy may bottleneck the quality of the generated speech. Finally, our training and evaluation is performed primarily on English. While we expect that our technique is sufficiently general and simple enough to generalize to many languages, more in-depth studies would be beneficial. We note that these techniques are known weaknesses of LM-based TTS models in general, and are not unique to our proposed method.

Ethical Considerations

Speaker identity and attributes are sensitive biometric information. While our proposed technique can prevent the use of performing speaker identification on audio data, it does not completely eliminate the need for labeled speaker information in zero-shot TTS training. We also acknowledge the dangers of voice cloning technologies like zero-shot TTS, such as impersonation or fraud. However, there is also a reproducibility crisis in large-scale TTS research, as many SOTA models are not released nor transparent on the training data used. As such, we choose to release our pre-trained models under non-commercial licenses (CC-BY-NC 4.0) following the access conditions of the training data, designated solely for research purposes, and explicitly forbid their use for malicious activities. While one may argue that bad actors may still use the models regardless of the license, we note that there are already several similarly performant models released for non-research use (Nari Labs, 2025; Schalkwyk et al., 2025).

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP*, pages 7124–7128.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proc. NeurIPS*, volume 33, pages 1877–1901.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

Guoguo Chen and 1 others. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*.

William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. 2023a. Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute. In *Interspeech 2023*.

William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. 2023b. Joint prediction and denoising for large-scale multilingual self-supervised learning. In *ASRU 2023*.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024b. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jade Copet and 1 others. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS 2022*.

Jordan Darefsky, Ge Zhu, and Zhiyao Duan. 2024. Parakeet.

Alexandre Défossez and 1 others. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Zhihao Du and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689.

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A large-scale diverse English speech recognition dataset for commercial usage. In *NeurIPS 2021*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *arXiv preprint arXiv:2407.05361*.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. 2023. The singing voice conversion challenge 2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Mikyas T Desta, Roy Fejgin, Rafael Valle, and Jason Li. 2025. Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance. *arXiv preprint arXiv:2502.05236*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Alex Gichamba, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe. 2024. Espnet-spk: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models. In *Interspeech 2024*, pages 4278–4282.

Shareef Babu Kalluri, Prachi Singh, Pratik Roy Chowdhuri, Apoorva Kulkarni, Shikha Baghel, Pradyoth Hegde, Swapnil Sontakke, Deepak K T, S.R. Mahadeva Prasanna, Deepu Vijayasenan, and Sriram Ganapathy. 2024. The second displace challenge: Diarization of speaker and language in conversational environments. In *Interspeech 2024*, pages 1630–1634.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR 2015*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

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

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. YODAS: Youtube-oriented dataset for audio and speech. In *ASRU 2023*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Alexander H. Liu, Matthew Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. 2024. Generative pre-training for speech with flow matching. In *The Twelfth International Conference on Learning Representations*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP*, pages 13326–13330. IEEE.

Nari Labs. 2025. Dia. `https://github.com/nari-labs/dia`.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and 1 others. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Vassil Panayotov and 1 others. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech and Language*, 72:101317.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12442–12462, Bangkok, Thailand. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML 2023*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Johan Schalkwyk, Ankit Kumar, Dan Lyth, Sefik Emre Eskimez, Zack Hodari, Cinjon Resnick, Ramon Sanabria, and Raven Jiang. 2025. Crossing the Uncanny Valley of Conversational Voice.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, and 1 others. 2024a. Versa: A versatile evaluation toolkit for speech, audio, and music. *arXiv preprint arXiv:2412.17667*.

Jiatong Shi and 1 others. 2024b. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. *arXiv preprint arXiv:2409.15897*.

Jinchuan Tian, Jiatong Shi, William Chen, Siddhant Arora, Yoshiki Masuyama, Takashi Maekaku, Yihan Wu, Junyi Peng, Shikhar Bharadwaj, Yiwen Zhao, and 1 others. 2025. Espnet-speechlm: An open speech language model toolkit. *arXiv preprint arXiv:2502.15218*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arxiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS 2017*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Interspeech 2018*.

Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. 2022. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. In *Interspeech 2022*, pages 2568–2572.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, and 1 others. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pages 271–350.

Appendix – Subjective Evaluation

We conducted subjective evaluations using two tests: Comparative Mean Opinion Score (CMOS) for naturalness and Speaker Similarity Mean Opinion Score (SMOS), following the setups of (Ju et al., 2024; Eskimez et al., 2024; Wang et al., 2023) after review board approval. Participants were informed that their work would be used to evaluate speech processing systems. We generated 20 samples from each system, each from a different speaker. 30 MTurk annotators scored each system, leading to 600 total samples per system for each test. Annotators were paid upon completion of all annotations, at a rate of $0.11 USD per minute.

For CMOS, annotators were shown the generated sample next to the ground truth sample and were asked "*Which recording sounds more natural? Please rate on a scale of* -3 *to* 3." They were presented with a 7 point scale, with -3 indicating that the ground truth was "*clearly more natural*" and 3 being the generated sample was "*clearly more natural*". A score of 0 indicated that "Both samples are equally natural". We shuffled the order in which the recordings were presented, both within and between pairs. Annotators were not told which recording was the ground truth.

For SMOS, annotators were shown the speaker prompt paired with either the ground truth or a generated sample. They were asked "How similar are the two speakers in the recording? Please rate on a scale of 1 to 5." A score of 1 meant that the speakers in the two recordings were "Not similar at all", while a score of 5 indicated that the speakers were "identical." A score of 3 meant that the two speakers were "moderately similar." We shuffled the order in which the recordings were presented, both within and between pairs. Annotators were not told which recording was the speaker prompt.