# Neural Voice Cloning with a Few Samples

**Sercan Ö. Arık** [*1] **Jitong Chen** [*1] **Kainan Peng** [*1] **Wei Ping** [*1] **Yanqi Zhou** [1]

## Abstract

Voice cloning is a highly desired feature for personalized speech interfaces. Neural network based speech synthesis has been shown to generate high quality speech for a large number of speakers. In this paper, we introduce a neural voice cloning system that takes a few audio samples as input. We study two approaches: speaker adaptation and speaker encoding. Speaker adaptation is based on fine-tuning a multi-speaker generative model with a few cloning samples. Speaker encoding is based on training a separate model to directly infer a new speaker embedding from cloning audios and to be used with a multi-speaker generative model. In terms of naturalness of the speech and its similarity to original speaker, both approaches can achieve good performance, even with very few cloning audios. [1] While speaker adaptation can achieve better naturalness and similarity, the cloning time or required memory for the speaker encoding approach is significantly less, making it favorable for low-resource deployment.

## 1. Introduction

Generative models based on deep learning have been successfully applied to many domains such as image synthesis (van den Oord et al., 2016; Karras et al., 2017), audio synthesis (Wang et al., 2017; Engel et al., 2017; Arik et al., 2017a), and language modeling (Jozefowicz et al., 2016; Merity et al., 2017). Deep neural networks are capable of modeling complex data distributions and scale well to large training data. They can be further conditioned on external inputs to control high-level behaviors, such as dictating the content and style of generated samples.

For speech synthesis, generative models can be conditioned on text (Wang et al., 2017) and speaker identity (Arik et al., 2017b; Ping et al., 2017). While text carries linguistic information and controls the content of the generated speech, speaker representation captures speaker characteristics such as pitch range, speech rate and accent. One approach for multi-speaker speech synthesis is to jointly train a generative model and speaker embeddings on triplets of (text, audio, speaker identity) (Arik et al., 2017b; Ping et al., 2017). Embeddings for all speakers are randomly initialized and trained with a generative loss. The idea is to encode the speaker-dependent information in low-dimensional embeddings, while sharing the majority of the model parameters for all speakers. One limitation of such a model is that it can only generate speech for speakers observed during training. A more interesting task is to learn the voice of an unseen speaker from a few speech samples, or *voice cloning*. Voice cloning can be used in many speech-enabled applications to provide personalized user experience.

In this work, we focus on voice cloning with limited speech samples from an unseen speaker, which can also be considered in the context of few-shot generative modeling of speech. With a large number of samples, a generative model can be trained from scratch for any target speaker. Yet, few-shot generative modeling is challenging besides being appealing. The generative model needs to learn the speaker characteristics from limited information provided by a few audio samples and generalize to unseen texts. We explore voice cloning methods with the recently proposed end-to-end neural speech synthesis approaches (Wang et al., 2017; Ping et al., 2017), which apply sequence-to-sequence modeling with attention mechanism. In neural speech synthesis, an encoder converts text (character or phoneme sequences) to hidden representations, and a decoder estimates the time-frequency representation of speech in an autoregressive way. Compared to traditional unit-select speech synthesis (Sagisaka et al., 1992) and statistical parametric speech synthesis (Zen et al., 2009), neural speech synthesis has a simpler pipeline and produces more natural speech (Shen et al., 2017b).

An end-to-end multi-speaker speech synthesis model is typically parameterized by the weights of generative model and a speaker embedding look-up table, where the latter is supposed to carry speaker characteristics. In this work, we investigate two questions. First, how well can speaker em-

[*]Equal contribution [1]Baidu Research. Correspondence to: Sercan Ö. Arık <sercanarik@baidu.com>, Jitong Chen <chenjitong01@baidu.com>, Kainan Peng <pengkainan@baidu.com>, Wei Ping <pingwei01@baidu.com>.

[1]Cloned audio samples can be found in `https://audiodemos.github.io`

beddings capture the differences between speakers? Second, how well can speaker embeddings be learned for an unseen speaker with only a few samples? We compare these two voice cloning approaches: (i) speaker adaptation and (ii) speaker encoding, in terms of speech naturalness, speaker similarity, cloning/inference time and model footprint.

## 2. Voice Cloning

We consider a multi-speaker generative model, $f(\mathbf{t}_{i,j}, s_i; W, \mathbf{e}_{s_i})$, which takes a text $\mathbf{t}_{i,j}$ and a speaker identity $s_i$. The model is parameterized by $W$, the trainable parameters in encoder and decoder, and $\mathbf{e}_{s_i}$, the trainable speaker embedding corresponding to $s_i$. Both $W$ and $\mathbf{e}_{s_i}$ are optimized by minimizing a loss function $L$ that penalizes the difference between generated and ground truth audios (for example, a L1 or L2 loss on spectrogram):

$$\min_{W,\mathbf{e}} \mathbb{E}_{\substack{s_i \sim \mathcal{S}, \\ (\mathbf{t}_{i,j}, \mathbf{a}_{i,j}) \sim \mathcal{T}_{s_i}}} \left\{ L\left( f(\mathbf{t}_{i,j}, s_i; W, \mathbf{e}_{s_i}), \mathbf{a}_{i,j} \right) \right\} \quad (1)$$

where $\mathcal{S}$ is a set of speakers, $\mathcal{T}_{s_i}$ is a training set of text-audio pairs for speaker $s_i$, and $\mathbf{a}_{i,j}$ is the ground-truth audio for $\mathbf{t}_{i,j}$ of speaker $s_i$. The expectation is estimated over text-audio pairs of all training speakers. In practice, $\mathbb{E}$ operator for the loss function is approximated by minibatch. We use $\widehat{W}$ and $\widehat{e}$ to denote the trained parameters and embeddings.

Speaker embeddings have been shown to effectively capture speaker differences for multi-speaker speech synthesis. They are low-dimension continuous representations of speaker characteristics (Arik et al., 2017b; Ping et al., 2017). Despite being trained with a purely generative loss, discriminative properties (e.g. gender or accent) can indeed be observed in embedding space (Arik et al., 2017b).

In voice cloning, we aim to extract the speaker characteristics for an unseen speaker $s_k$ (that is not in $\mathcal{S}$) from a set of cloning audios $\mathcal{A}_{s_k}$, and generate a different audio conditioned on a given text for that speaker. The two performance metrics for the generated audio are (i) how natural it is and (ii) whether it sounds like it is pronounced by the same speaker.

The two approaches for neural voice cloning are summarized in Fig. 1 and explained in the following sections.

### 2.1. Speaker adaptation

The idea of speaker adaptation is to fine-tune a trained multi-speaker model for an unseen speaker using a few audios and corresponding texts by applying gradient descent. Fine-tuning can be applied to either the speaker embedding (Taigman et al., 2017) or the whole model.

For embedding-only adaptation, we have the following objective:

$$\min_{\mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j}, \mathbf{a}_{k,j}) \sim \mathcal{T}_{s_k}} \left\{ L\left( f(\mathbf{t}_{k,j}, s_k; \widehat{W}, \mathbf{e}_{s_k}), \mathbf{a}_{k,j} \right) \right\} \quad (2)$$

where $\mathcal{T}_{s_k}$ is a set of text-audio pairs for the target speaker $s_k$. For whole model adaptation, we have the following objective:

$$\min_{W,\mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j}, \mathbf{a}_{k,j}) \sim \mathcal{T}_{s_k}} \left\{ L\left( f(\mathbf{t}_{k,j}, s_k; W, \mathbf{e}_{s_k}), \mathbf{a}_{k,j} \right) \right\} \quad (3)$$

Although the entire model provides more degrees of freedom for speaker adaptation, its optimization is challenging especially for a small number of cloning samples. While running the optimization, careful choice of the number of iterations is crucial for avoiding underfitting or overfitting.

### 2.2. Speaker encoding

We propose a speaker encoding method to directly estimate the speaker embedding from audio samples of an unseen speaker. Such a model does not require fine-tuning during voice cloning and therefore the same model can be used for all unseen speakers.

Specifically, the speaker encoding function, $g(\mathcal{A}_{s_k}; \Theta)$, takes a set of cloning audio samples $\mathcal{A}_{s_k}$ and estimates $\mathbf{e}_{s_k}$. The function is parametrized by $\Theta$. Ideally, the speaker encoder can be jointly trained with multi-speaker generative model from scratch, with a loss function defined for generated audio quality:

$$\min_{W,\mathbf{\Theta}} \mathbb{E}_{\substack{s_i \sim \mathcal{S}, \\ (\mathbf{t}_{i,j}, \mathbf{a}_{i,j}) \sim \mathcal{T}_{s_i}}} \left\{ L\left( f(\mathbf{t}_{i,j}, s_i; W, g(\mathcal{A}_{s_i}; \Theta)), \mathbf{a}_{i,j} \right) \right\}. \quad (4)$$

Note that speaker encoder is trained with existing speakers. During training, a set of cloning audio samples $\mathcal{A}_{s_i}$ are randomly sampled for training speaker $s_i$. During inference, $\mathcal{A}_{s_k}$, audio samples from the target speaker $s_k$, is used to compute $g(\mathcal{A}_{s_k}; \Theta)$. However, we have observed optimization challenges when the training process formulated in Eq. 4 is started from scratch. One potential problem is that the model fits an average voice to minimize the overall generative loss, commonly referred as *mode collapse* in generative modeling literature. One idea to address mode collapse is to introduce discriminative loss functions for intermediate embeddings, [2] or generated audios. [3] In our case, however, such approaches only slightly improve speaker difference.

Instead, we propose training speaker encoder separately. Speaker embeddings $\widehat{\mathbf{e}}_{\mathbf{s_i}}$ are extracted from a trained multi-speaker generative model $f(\mathbf{t}_{i,j}, s_i; W, \mathbf{e}_{s_i})$. Then, the

[2]We have experimented classification accuracy by mapping the embeddings to speaker class labels via a softmax layer.

[3]We have experimented integrating a pre-trained speaker classifier to promote speaker difference of generated audios.

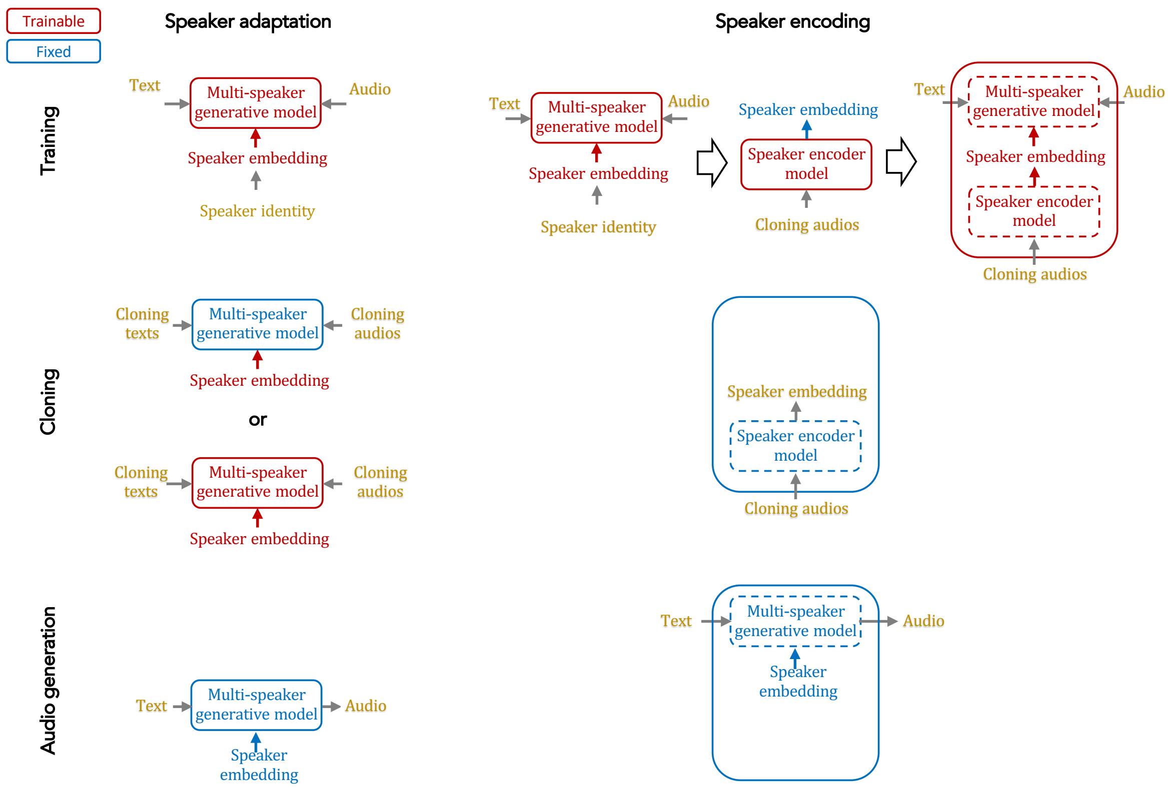


*Figure 1.* Speaker adaptation and speaker encoding approaches for training, cloning and audio generation.

speaker encoder model $g(\mathcal{A}_{s_k}; \Theta)$ is trained to predict the embeddings from sampled cloning audios. There can be several objective functions for the corresponding regression problem. We obtain the best results by simply using an L1 loss between the estimated and target embeddings:

$$\min_{\Theta} \mathbb{E}_{s_i \sim \mathcal{S}} \left\{ |g(\mathcal{A}_{s_i}; \Theta) - \widehat{\mathbf{e}}_{\mathbf{s_i}})| \right\}, \tag{5}$$

Eventually, the entire model can be jointly fine-tuned based on the objective function Eq. 4, using pre-trained $\widehat{W}$ and pre-trained $\widehat{\Theta}$ as initial points. Fine-tuning helps the generative model learn how to compensate the errors of embedding estimation, and we observe less attention problems. However, generative loss still dominates learning and speaker differences in generated audios may be slightly reduced (see Appendix A for details).

For speaker encoder model $g(\mathcal{A}_{s_k}; \Theta)$, we propose a neural network architecture comprising the following three parts (show in Fig. 2):

(i) *Spectral processing*: We compute mel-spectrograms for cloning audio samples and pass them to PreNet, which contains fully-connected (FC) layers with exponential linear unit (ELU) for feature transformation.

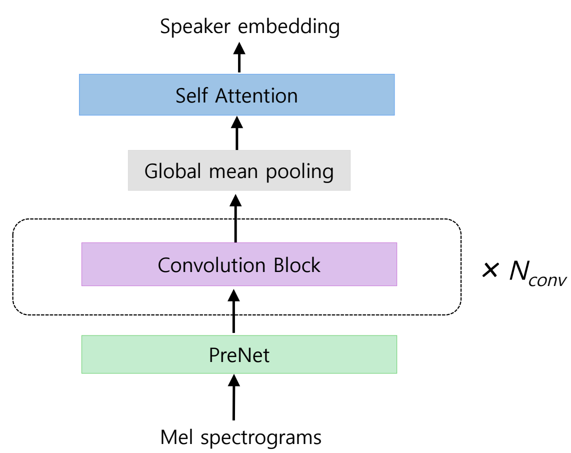


*Figure 2.* Speaker encoder architecture. See Appendix A for details.

(ii) *Temporal processing*: We incorporate temporal contexts using several convolutional layers with gated linear unit and residual connections. Then, average pooling is applied to summarize the whole utterance.

(iii) *Cloning sample attention*: Considering that different cloning audios contain different amount of speaker information, we use a multi-head self-attention mechanism (Vaswani et al., 2017) to compute the weights for different audios and get aggregated embeddings.

### 2.3. Discriminative models for evaluation

Voice cloning performance metrics can be based on human evaluations through crowdsourcing platforms, but they are slow and expensive during model development. In this section, we propose two evaluation methods using discriminative models.

#### 2.3.1. Speaker classification

Speaker classifier determines which speaker an audio sample belongs to. For voice cloning evaluation, a speaker classifier can be trained on the set of target speakers used for cloning. High-quality voice cloning would result in high speaker classification accuracy. We use a speaker classifier with similar spectral and temporal processing layers shown in Fig. 7 and an additional embedding layer before the softmax function.

#### 2.3.2. Speaker verification

Speaker verification is the task of authenticating the claimed identity of a speaker, based on a test audio and enrolled audios from the speaker. In particular, it performs binary classification to identify whether the test audio and enrolled audios are from the same speaker (e.g., Snyder et al., 2016). In this work, we consider the end-to-end text-independent speaker verification framework (Snyder et al., 2016) (see Appendix C for more details of model architecture). One can train a speaker verification model on a multi-speaker dataset, then directly test whether the cloned audio and the ground truth audio are from the same speaker. Unlike the speaker classification approach, speaker verification model does not require training with the audios from the target speaker for cloning, hence it can be used for unseen speakers with a few samples. As the quantitative performance metric, the equal error-rate (EER) [4] from speaker verification model can be used to measure how close the cloned audios are to the ground truth audios.

## 3. Experiments

We compare two approaches for voice cloning. For speaker adaptation approach, we train a multi-speaker generative model and adapt it to a target speaker by fine-tuning the embedding or the whole model. For speaker encoding approach, we train a multi-speaker generative model and a speaker encoder. The estimated speaker embedding is then fed to multi-speaker generative model to generate audios for a target speaker.

### 3.1. Datasets

Multi-speaker generative model and speaker encoder model are trained using LibriSpeech dataset (Panayotov et al., 2015), which contains audios for 2484 speakers sampled at 16 KHz, totalling 820 hours. LibriSpeech is a dataset for automatic speech recognition, and its audio quality is lower compared to speech synthesis datasets.[5]

Voice cloning is performed using VCTK dataset (Veaux et al., 2017). VCTK consists of audios for 108 native speakers of English with various accents. To be consistent with LibriSpeech dataset, VCTK audio samples are downsampled to 16 KHz. For a chosen speaker, a few cloning audios are sampled randomly for each experiment. The sentences presented in Appendix B are used to generate audios for evaluation.

### 3.2. Specifications

#### 3.2.1. Multi-speaker generative model

Our multi-speaker generative model is based on the convolutional sequence-to-sequence architecture proposed in (Ping et al., 2017), with similar hyperparameters and Griffin-Lim vocoder. To get better performance, we increase the time-resolution by reducing the hop length and window size parameters to 300 and 1200, and add a quadratic loss term to penalize larger amplitude components superlinearly. For speaker adaptation experiments, we reduce the embedding dimensionality to 128, as it yields less overfitting problems. Overall, the baseline multi-speaker generative model has around 25M trainable parameters.

#### 3.2.2. Speaker adaptation

For speaker adaptation approach, either the entire multi-speaker generative model parameters or only its speaker embeddings are fine-tuned. For both cases, optimization is separately applied to each of the 108 speakers from VCTK dataset.

#### 3.2.3. Speaker encoder model

For speaker encoding approach, we train speaker encoders for different number of cloning audios separately, to obtain the minimum validation loss. We convert cloning audios to log-mel spectrograms with 80 frequency bands, a hop length of 400, a window size of 1600. Log-mel spectrograms are fed to spectral processing layers, which are composed of 2-layer prenet of size 128. Then, temporal processing is

[4] One may change decision threshold to trade-off between false acceptance rate and false rejection rate. When the two rates are equal, it is referred to as the equal error rate.

[5] We designed a segmentation and denoising pipeline to process LibriSpeech, as described in (Ping et al., 2017)

applied with two 1-dimension convolutional layers with a filter width of 12. Finally, multi-head attention is applied with 2 heads and a unit size of 128 for keys, queries and values. The final embedding size is 512. To construct validation set, 25 speakers are held out from training set. A batch size of 64 is used while training, with an initial learning rate of 0.0006 with annealing rate of 0.6 applied every 8000 iterations. Mean absolute error for the validation set is shown in Fig. 11 in Appendix D. An increasing number of cloning audios leads to more accurate speaker embedding estimation, especially with introduction of the attention mechanism proposed (see Appendix D for more details about the learned attention coefficients).

#### 3.2.4. Speaker classification model

We train a speaker classifier on VCTK dataset to classify which of the 108 speakers an audio sample belongs to. Speaker classifier has a fully-connected layer of size 256, 6 convolutional layers with 256 filters of width 4, and a final embedding layer of size 32. The model achieves 100% accuracy for validation set of size 512.

#### 3.2.5. Speaker verification model

We train a speaker verification model on the LibriSpeech dataset to measure the quality of cloned audios compared to ground truth audios, to be used for any unseen speaker. We hold out 50 speakers from Librispeech as a validation set for unseen speakers. The equal-error-rates (EERs) are estimated by randomly pairing up utterances from the same or different speakers (50% for each case) in test set. We perform 40960 trials for each test set. We describe the details of speaker verification model in Appendix C.

### 3.3. Voice cloning performance

For speaker adaptation approach, we pick the number of iterations using speaker classification accuracy. For whole model adaptation, we pick the number of iterations.[6] For speaker embedding adaptation, we fix the number of iterations as 100K for all cases. For speaker encoding, we consider voice cloning with and without joint fine-tuning of the speaker encoder and multi-speaker generative model.[7]

Table 1 summarizes the approaches and lists the requirements for training, data, cloning time and footprint size.

#### 3.3.1. Evaluations by discriminative models

For speaker adaptation approaches, Fig. 3 show the speaker classification accuracy vs. the number of iterations. Both

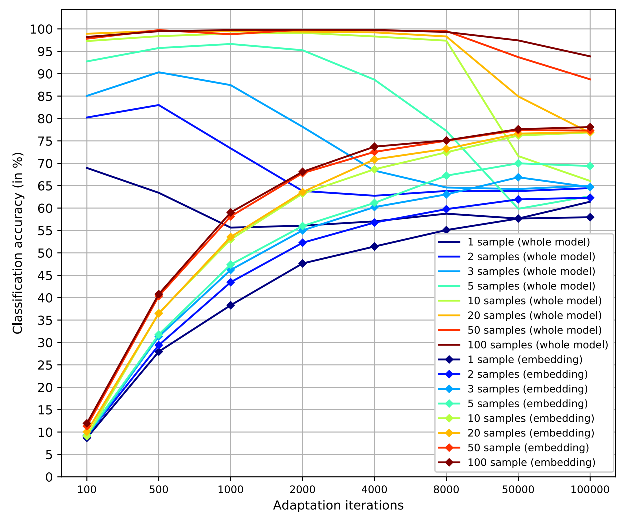


*Figure 3.* Performance of whole model adaptation and speaker embedding adaptation for voice cloning in terms of speaker classification accuracy. Different numbers of cloning samples and fine-tuning iterations are evaluated.

adaptation methods benefit from more cloning samples when the sample count is low. After 10 samples, increasing sample count does not significantly improve speaker classification accuracy. In the low sample count regime, adapting the speaker embedding is less likely to overfit the samples than adapting the whole model. The two methods also require different numbers of iterations to converge. Compared to whole model adaptation, which converges around 1000 iterations for even 100 cloning audio samples, embedding adaptation takes significantly more iterations to converge.

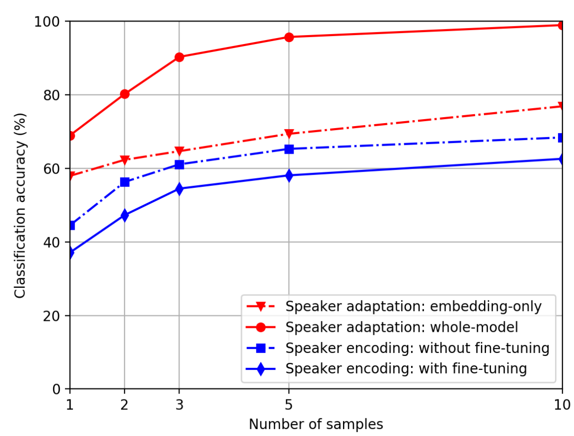


*Figure 4.* Comparison of speaker adaptation and speaker encoding approaches in term of speaker classification accuracy with different numbers of cloning samples.

[6] 100 for 1, 2 and 3 cloning audio samples, 1000 for 5 and 10 cloning audio samples

[7] The learning rate and annealing parameters are optimized for joint fine-tuning.

| | Speaker adaptation | | Speaker encoding | |
|---|---|---|---|---|
| **Approaches** | Embedding-only | Whole-model | Without fine-tuning | With fine-tuning |
| **Pre-training** | Multi-speaker generative model | | | |
| **Data** | Text and audio | | Audio | |
| **Cloning time** | $\sim 8$ hours | $\sim 0.5 - 5$ mins | $\sim 1.5 - 3.5$ secs | $\sim 1.5 - 3.5$ secs |
| **Inference time** | $\sim 0.4 - 0.6$ secs | | | |
| **Parameters per speaker** | 128 | $\sim 25$ million | 512 | 512 |

*Table 1.* Comparison of requirements for speaker adaptation and speaker encoding. Cloning time interval assumes 1-10 cloning audios. Inference time is for an average sentence. All assume implementation on a TitanX GPU.

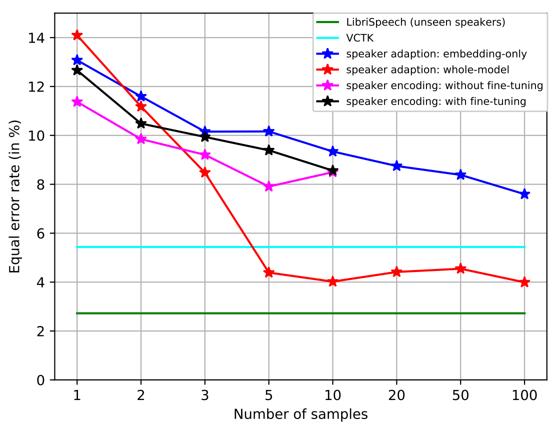


*Figure 5.* Speaker verification EER (using 5 enrollment audio) for different numbers of cloning samples.

Figs. 4 and 5 show the classification accuracy and equal-error-rate (EER) for both speaker adaptation and encoding approaches, obtained by speaker classification and speaker verification models. Both speaker adaptation and speaker encoding approaches benefit from more cloning audios. When the number of cloning audio samples exceed 5, whole model adaptation outperforms the other techniques in both metrics. Speaker encoding approaches yield a lower classification accuracy compared to embedding adaptation, but they achieve similar speaker verification performance.

#### 3.3.2. Human evaluations

Besides evaluations by discriminative models, we also conduct subject tests on Amazon Mechanical Turk framework. For assessment of the naturalness of the generated audios, we use 5-scale mean opinion score (MOS). For assessment of how similar the generated audios are to the ground truth audios from target speakers, we use 4-scale similarity score with the same question and categories in (Wester et al., 2016). We conduct each evaluation independently, so the cloned audios of two different models are not directly compared during rating. Multiple votes on the same sample are aggregated by a majority voting rule.

Tables 2 and 3 show the results of human evaluations. Both speaker adaptation approaches benefit from higher number of cloning audios. The improvement with high sample count is more significant for whole model adaptation as expected, since there are more degrees of freedom provided to be tuned for that particular speaker. There is a very small difference in naturalness for speaker encoding approaches when the number of cloning audios is increased. Most importantly, speaker encoding does not degrade the naturalness of the baseline multi-speaker generative model. Fine-tuning improves the naturalness of speaker encoding as expected, since it allows the generative model to learn how to compensate the errors of the speaker encoder while training. Similarity scores slightly improve with higher sample counts for speaker encoding, and match the scores for speaker embedding adaptation. The gap of similarity with ground truth is mostly attributed to the limited naturalness of the outputs, as they are trained with LibriSpeech dataset.

### 3.4. Speaker embedding space and manipulation

As shown in Fig. 6 and Appendix E, speaker encoder models map speakers into a meaningful latent space. Inspired by word embedding manipulation (e.g. via simple algebraic operations as *king* - *queen* = *male* - *female*), we consider applying algebraic operations to the inferred embeddings to transform their speech characteristics.

To transform gender, we get the averaged speaker embeddings for female and male, and add their difference to a particular speaker. For example,

$$\begin{aligned} \mathit{BritishMale} \; &+ \mathit{AveragedFemale} \\ &- \mathit{AveragedMale} \end{aligned}$$

can yield a British female speaker. Similarly, we can consider region of accent transformation by

$$\begin{aligned} \mathit{BritishMale} \; &+ \mathit{AveragedAmerican} \\ &- \mathit{AveragedBritish} \end{aligned}$$

to obtain an American male speaker. Our results (`https://audiodemos.github.io/`) demonstrate high qual-

| Approach | Sample count | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 5 | 10 |
| Ground-truth (at 16 KHz) | 4.66±0.06 | | | | |
| Multi-speaker generative model | 2.61±0.10 | | | | |
| Speaker adaptation: embedding-only | 2.27±0.10 | 2.38±0.10 | 2.43±0.10 | 2.46±0.09 | 2.67±0.10 |
| Speaker adaptation: whole-model | 2.32±0.10 | 2.87±0.09 | 2.98±0.11 | 2.67±0.11 | 3.16±0.09 |
| Speaker encoding: without fine-tuning | 2.76±0.10 | 2.76±0.09 | 2.78±0.10 | 2.75±0.10 | 2.79±0.10 |
| Speaker encoding: with fine-tuning | 2.93±0.10 | 3.02±0.11 | 2.97±0.1 | 2.93±0.10 | 2.99±0.12 |

*Table 2.* Mean Opinion Score (MOS) evaluations for naturalness with 95% confidence intervals.

| Approach | Sample count | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 5 | 10 |
| Ground-truth: same speaker | 3.91±0.03 | | | | |
| Ground-truth: different speakers | 1.52±0.09 | | | | |
| Speaker adaptation: embedding-only | 2.66±0.09 | 2.64±0.09 | 2.71±0.09 | 2.78±0.10 | 2.95±0.09 |
| Speaker adaptation: whole-model | 2.59±0.09 | 2.95±0.09 | 3.01±0.10 | 3.07±0.08 | 3.16±0.08 |
| Speaker encoding: without fine-tuning | 2.48±0.10 | 2.73±0.10 | 2.70±0.11 | 2.81±0.10 | 2.85±0.10 |
| Speaker encoding: with fine-tuning | 2.59±0.12 | 2.67±0.12 | 2.73±0.13 | 2.77±0.12 | 2.77±0.11 |

*Table 3.* Similarity score evaluations with 95% confidence intervals.

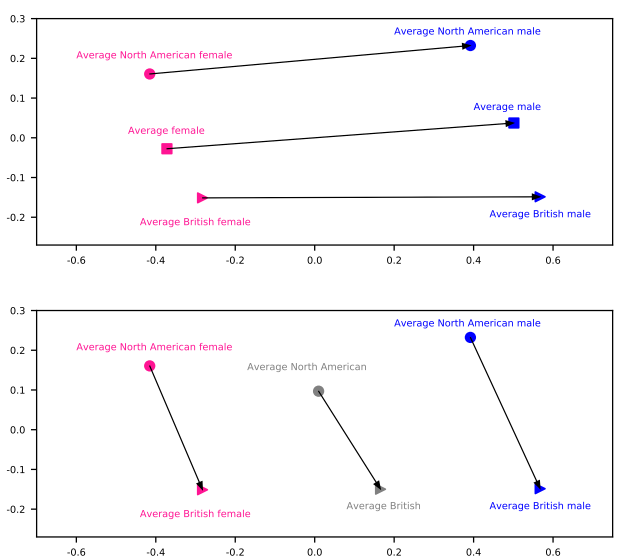


*Figure 6.* Visualization of estimated speaker embeddings by speaker encoder. The first two principal components of the average speaker embeddings for the speaker encoder with 5 sample count. Only British and North American regional accents are shown as they constitute the majority of the labeled speakers in the VCTK dataset. Please see Appendix E for more detailed analysis.

ity audios with specific gender and accent characteristics obtained in this way.

# 4. Related Work

## 4.1. Few-shot generative models

Humans can learn most new generative tasks from only a few examples, and it has motivated research on few-shot generative models.

Early studies on few-shot generative modeling mostly focus on Bayesian models. In (Lake et al., 2013) and (Lake et al., 2015), hierarchical Bayesian models are used to exploit compositionality and causality for few-shot generation of characters. In (Lake et al., 2014), a similar idea is modified to acoustic modeling task, with the goal of generating new words in a different language.

Recently, deep learning approaches are adapted to few-shot generative modeling particularly for image generation applications. In (Reed et al., 2017), few-shot distribution estimation is considered using an attention mechanism and meta-learning procedure, for conditional image generation. In (Azadi et al., 2017), few-shot learning is applied to font style transfer, by modeling the glyph style from a few observed letters, and synthesizing the whole alphabet conditioned on the estimated style. The technique is based on multi-content generative adversarial networks, penalizing the unrealistic synthesized letters compared to the ground truth. In (Rezende et al., 2016), sequential generative modeling is applied for one-shot generalization in image generation, using a spatial attentional mechanism.

### 4.2. Speaker embeddings in speech processing

Speaker embedding is a well-established approach to encode discriminative information in speakers. It has been used in many speech processing tasks such as speaker recognition/verification (Li et al., 2017), speaker diarization (Rouvier et al., 2015), automatic speech recognition (Doddipatla, 2016) and speech synthesis (Arik et al., 2017b). In some of these, the model explicitly learns to output embeddings with a discriminative task such as speaker classification. In others, embeddings are randomly initialized and implicitly learned from an objective function that is not directly related to speaker discrimination. For example, in (Arik et al., 2017b), a multi-speaker generative model is trained to generate audio from text, where speaker embeddings are implicitly learned from a generative loss function. This approach encourages most parts of the model to be speaker-independent and shared, while pushing speaker-related information into embeddings.

### 4.3. Voice conversion

The goal of voice conversion is to modify an utterance from source speaker to make it sound like the target speaker, while keeping the linguistic contents unchanged.

One common approach is dynamic frequency warping, to align spectra of different speakers. (Agiomyrgiannakis & Roupakia, 2016) proposes a dynamic programming algorithm that simultaneously estimates the optimal frequency warping and weighting transform while matching source and target speakers using a matching-minimization algorithm. Wu et al. (2016) uses a spectral conversion approach integrated with the locally linear embeddings for manifold learning. There are also approaches to model spectral conversion using neural networks such as (Desai et al., 2010), (Chen et al., 2014), (Hwang et al., 2015), that are typically trained with a large amount of audio pairs of target and source speakers.

## 5. Conclusions

In this paper, we have demonstrated two approaches for neural voice cloning: speaker adaptation and speaker encoding. We demonstrate that both approaches can achieve reasonable cloning quality even with only a few cloning audios.

For naturalness, we demonstrate that both speaker adaptation and speaker encoding can achieve an MOS for naturalness similar to baseline multi-speaker generative model. Thus, the proposed techniques can potentially be used with better multi-speaker models in the future (such as replacing Griffin-Lim with WaveNet vocoder as in (Shen et al., 2017a)).

For similarity, we demonstrate that both approaches benefit from a larger number of cloning audios. The performance gap between whole-model and embedding-only adaptation indicates that some discriminative speaker information still exists in the generative model besides speaker embeddings. The benefit of compact representation via embeddings is fast cloning and small footprint size per user. Especially for the applications with resource constraints, these practical considerations should clearly favor for the use of speaker encoding approach. Methods to fully embed the speaker information into the embeddings would be an important research direction to improve performance of voice cloning.

Training multi-speaker generative model in this paper is done using a speech recognition dataset with low-quality audios and limited diversity in representation of universal set of speakers. Improvements in the quality of dataset will result in higher naturalness and similarity of generated samples. Also, increasing the amount and diversity of speakers should enable a more meaningful speaker embedding space, which can improve the similarity obtained by both approaches. We expect our techniques to benefit significantly from a large-scale and high-quality multi-speaker speech dataset.

We believe that there are many promising horizons for improvement in voice cloning. Advances in meta-learning, i.e. systematic approach of learning-to-learn while training, should be promising to improve voice cloning, e.g. by integrating speaker adaptation or encoding into training, or by inferring the model weights in a more flexible way than the speaker embeddings are being used.

## References


Agiomyrgiannakis, Yannis and Roupakia, Zoi. Voice morphing that improves tts quality using an optimal dynamic frequency warping-and-weighting transform. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5650–5654, 2016.

Amodei, Dario, Ananthanarayanan, Sundaram, Anubhai, Rishita, Bai, Jingliang, Battenberg, Eric, Case, Carl, Casper, Jared, Catanzaro, Bryan, Cheng, Qiang, Chen, Guoliang, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pp. 173–182, 2016.

Arik, Sercan Ömer, Chrzanowski, Mike, Coates, Adam, Diamos, Greg, Gibiansky, Andrew, Kang, Yongguo, Li, Xian, Miller, John, Raiman, Jonathan, Sengupta, Shubho, and Shoeybi, Mohammad. Deep voice: Real-time neural text-to-speech. *CoRR*, abs/1702.07825, 2017a.

Arik, Sercan Ömer, Diamos, Gregory F., Gibiansky, Andrew, Miller, John, Peng, Kainan, Ping, Wei, Raiman, Jonathan, and Zhou, Yanqi. Deep voice 2: Multi-speaker neural text-to-speech. *CoRR*, abs/1705.08947, 2017b.

Azadi, Samaneh, Fisher, Matthew, Kim, Vladimir, Wang, Zhaowen, Shechtman, Eli, and Darrell, Trevor. Multi-content gan for few-shot font style transfer. *CoRR*, abs/1708.02182, 2017.

Chen, L. H., Ling, Z. H., Liu, L. J., and Dai, L. R. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1859–1872, Dec 2014. ISSN 2329-9290. doi: 10.1109/TASLP.2014.2353991.

Desai, S., Black, A. W., Yegnanarayana, B., and Prahallad, K. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):954–964, July 2010. ISSN 1558-7916. doi: 10.1109/TASL.2010.2047683.

Doddipatla, R. Speaker adaptive training in deep neural networks using speaker dependent bottleneck features. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5290–5294, March 2016. doi: 10.1109/ICASSP.2016.7472687.

Engel, Jesse, Resnick, Cinjon, Roberts, Adam, Dieleman, Sander, Eck, Douglas, Simonyan, Karen, and Norouzi, Mohammad. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

Hwang, H. T., Tsao, Y., Wang, H. M., Wang, Y. R., and Chen, S. H. A probabilistic interpretation for artificial neural network-based voice conversion. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 552–558, Dec 2015. doi: 10.1109/APSIPA.2015.7415330.

Jozefowicz, Rafal, Vinyals, Oriol, Schuster, Mike, Shazeer, Noam, and Wu, Yonghui. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Karras, Tero, Aila, Timo, Laine, Samuli, and Lehtinen, Jaakko. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

Lake, Brenden M., Salakhutdinov, Ruslan, and Tenenbaum, Joshua B. One-shot learning by inverting a compositional causal process. In *NIPS*, 2013.

Lake, Brenden M., ying Lee, Chia, Glass, James R., and Tenenbaum, Joshua B. One-shot learning of generative speech concepts. In *CogSci*, 2014.

Lake, Brenden M., Salakhutdinov, Ruslan, and Tenenbaum, Joshua B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050.

Li, Chao, Ma, Xiaokong, Jiang, Bing, Li, Xiangang, Zhang, Xuewei, Liu, Xiao, Cao, Ying, Kannan, Ajay, and Zhu, Zhenyao. Deep speaker: an end-to-end neural speaker embedding system. *CoRR*, abs/1705.02304, 2017.

Merity, Stephen, Keskar, Nitish Shirish, and Socher, Richard. Regularizing and optimizing LSTM language models. *CoRR*, abs/1708.02182, 2017.

Panayotov, Vassil, Chen, Guoguo, Povey, Daniel, and Khudanpur, Sanjeev. Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210. IEEE, 2015.

Ping, Wei, Peng, Kainan, Gibiansky, Andrew, Arik, Sercan Ömer, Kannan, Ajay, Narang, Sharan, Raiman, Jonathan, and Miller, John. Deep voice 3: 2000-speaker neural text-to-speech. *CoRR*, abs/1710.07654, 2017.

Prince, Simon JD and Elder, James H. Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE, 2007.

Reed, Scott E., Chen, Yutian, Paine, Thomas, van den Oord, Aäron, Eslami, S. M. Ali, Rezende, Danilo Jimenez, Vinyals, Oriol, and de Freitas, Nando. Few-shot autoregressive density estimation: Towards learning to learn distributions. *CoRR*, abs/1710.10304, 2017.

Rezende, Danilo, Shakir, Danihelka, Ivo, Gregor, Karol, and Wierstra, Daan. One-shot generalization in deep generative models. In Balcan, Maria Florina and Weinberger, Kilian Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1521–1529, New York, New York, USA, 20–22 Jun 2016. PMLR.

Rouvier, M., Bousquet, P. M., and Favre, B. Speaker diarization through speaker embeddings. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2082–2086, Aug 2015. doi: 10.1109/EUSIPCO.2015.7362751.

Sagisaka, Yoshinori, Kaiki, Nobuyoshi, Iwahashi, Naoto, and Mimura, Katsuhiko. Atr $\mu$-talk speech synthesis system. In *Second International Conference on Spoken Language Processing*, 1992.

Shen, Jonathan, Pang, Ruoming, J. Weiss, Ron, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerry-Ryan, RJ, A. Saurous, Rif, Agiomyrgiannakis, Yannis, and Wu, Yonghui. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. 12 2017a.

Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerry-Ryan, RJ, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *arXiv preprint arXiv:1712.05884*, 2017b.

Snyder, David, Ghahremani, Pegah, Povey, Daniel, Garcia-Romero, Daniel, Carmiel, Yishay, and Khudanpur, Sanjeev. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170, 2016.

Taigman, Yaniv, Wolf, Lior, Polyak, Adam, and Nachmani, Eliya. Voice synthesis for in-the-wild speakers via a phonological loop. *CoRR*, abs/1707.06588, 2017.

van den Oord, Aaron, Kalchbrenner, Nal, Espeholt, Lasse, Vinyals, Oriol, Graves, Alex, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pp. 4790–4798, 2016.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Ł ukasz, and Polosukhin, Illia. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6000–6010. Curran Associates, Inc., 2017.

Veaux, Christophe, Yamagishi, Junichi, and MacDonald, Kirsten et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit, 2017.

Wang, Yuxuan, Skerry-Ryan, R. J., Stanton, Daisy, Wu, Yonghui, Weiss, Ron J., Jaitly, Navdeep, Yang, Zongheng, Xiao, Ying, Chen, Zhifeng, Bengio, Samy, Le, Quoc V., Agiomyrgiannakis, Yannis, Clark, Rob, and Saurous, Rif A. Tacotron: A fully end-to-end text-to-speech synthesis model. *CoRR*, abs/1703.10135, 2017.

Wester, Mirjam, Wu, Zhizheng, and Yamagishi, Junichi. Analysis of the voice conversion challenge 2016 evaluation results. In *Interspeech*, pp. 1637–1641, 09 2016.

Wu, Yi-Chiao, Hwang, Hsin-Te, Hsu, Chin-Cheng, Tsao, Yu, and Wang, Hsin-min. Locally linear embedding for exemplar-based spectral conversion. In *Interspeech*, pp. 1652–1656, 09 2016.

Zen, Heiga, Tokuda, Keiichi, and Black, Alan W. Statistical parametric speech synthesis. *Speech Communication*, 51 (11):1039–1064, 2009.

# Appendices

## A. Detailed speaker encoder architecture

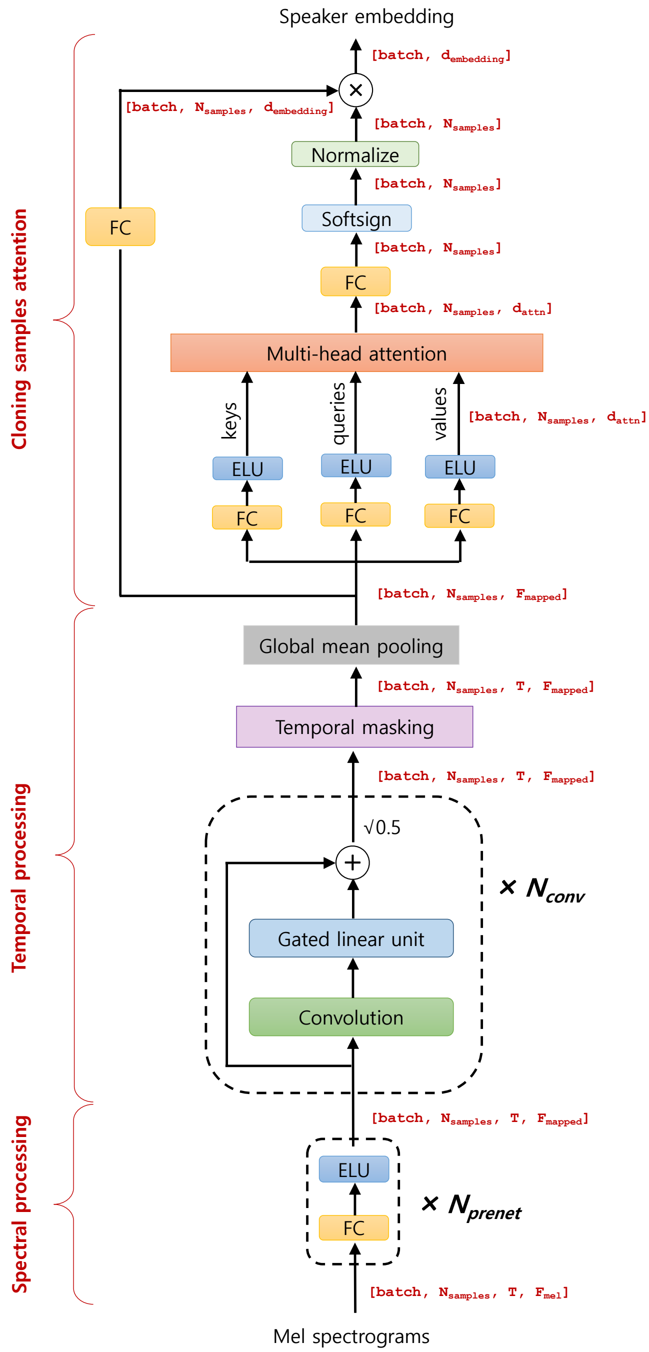


*Figure 7.* Speaker encoder architecture with intermediate state dimensions. ($batch$: batch size, $N_{samples}$: number of cloning audio samples $|A_{s_k}|$, $T$: number of mel spectrograms timeframes, $F_{mel}$: number of mel frequency channels, $F_{mapped}$: number of frequency channels after prenet, $d_{embedding}$: speaker embedding dimension). Multiplication operation at the last layer represents inner product along the dimension of cloning samples.

## B. Voice cloning test sentences

```
Prosecutors have opened a massive investigation/into allegations of/fixing games/and illegal betting%.
Different telescope designs/perform differently%and have different strengths/and weaknesses%.
We can continue to strengthen the education of good lawyers%.
Feedback must be timely/and accurate/throughout the project%.
Humans also judge distance/by using the relative sizes of objects%.
Churches should not encourage it%or make it look harmless%.
Learn about/setting up/wireless network configuration%.
You can eat them fresh cooked%or fermented%.
If this is true%then those/who tend to think creatively%really are somehow different%.
She will likely jump for joy%and want to skip straight to the honeymoon%.
The sugar syrup/should create very fine strands of sugar%that drape over the handles%.
But really in the grand scheme of things%this information is insignificant%.
I let the positive/overrule the negative%.
He wiped his brow/with his forearm%.
Instead of fixing it%they give it a nickname%.
About half the people%who are infected%also lose weight%.
The second half of the book%focuses on argument/and essay writing%.
We have the means/to help ourselves%.
The large items/are put into containers/for disposal%.
He loves to/watch me/drink this stuff%.
Still%it is an odd fashion choice%.
Funding is always an issue/after the fact%.
Let us/encourage each other%.
```

*Figure 8.* The sentences used to generate test samples for the voice cloning models. The white space characters / and % follow the same definition as in (Ping et al., 2017).

## C. Speaker verification model

Given a set of (e.g., 1∼5) enrollment audios [8] and a test audio, speaker verification model performs a binary classification and tells whether the enrollment and test audios are from the same speaker. Although using other speaker verification models (e.g., Snyder et al., 2016) would also suffice, we choose to create our own speaker verification models using convolutional-recurrent architecture (Amodei et al., 2016). We note that our equal-error-rate results on test set of unseen speakers are on par with the state-of-the-art speaker verification models. The architecture of our model is illustrated in Figure 9. We compute mel-scaled spectrogram of enrollment audios and test audio after resampling the input to a constant sampling frequency. Then, we apply two-dimensional convolutional layers convolving over both time and frequency bands, with batch normalization and ReLU non-linearity after each convolution layer. The output of last convolution layer is feed into a recurrent layer (GRU). We then mean-pool over time (and enrollment audios if there are many), then apply a fully connected layer to obtain the speaker encodings for both enrollment audios and test audio. We use the probabilistic linear discriminant analysis (PLDA) for scoring the similarity between the two encodings (Prince & Elder, 2007; Snyder et al., 2016). The PLDA score (Snyder et al., 2016) is defined as,

$$s(\boldsymbol{x}, \boldsymbol{y}) = w \cdot \boldsymbol{x}^\top \boldsymbol{y} - \boldsymbol{x}^\top S \boldsymbol{x} - \boldsymbol{y}^\top S \boldsymbol{y} + b \tag{6}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are speaker encodings of enrollment and test audios respectively after fully-connected layer, $w$ and $b$ are scalar parameters, and $S$ is a symmetric matrix. Then, $s(\boldsymbol{x}, \boldsymbol{y})$ is feed into a sigmoid unit to obtain the probability that they are from the same speaker. The model is trained using cross-entropy loss. Table 4 lists hyperparameters of speaker verification model for Librispeech dataset.

[8] Enrollment audios are from the same speaker.

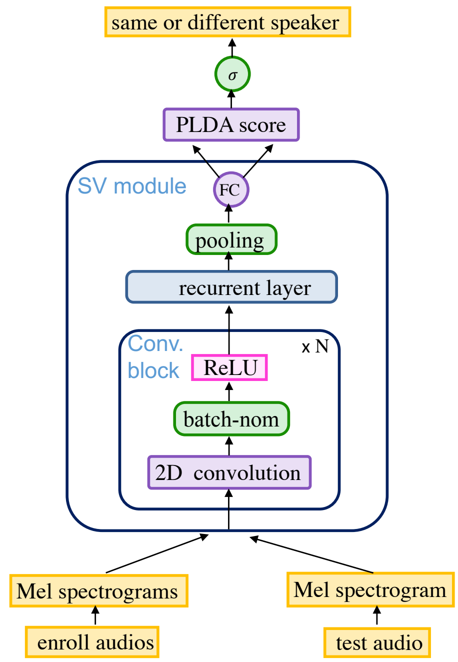


*Figure 9.* Architecture of speaker verification model.

| Parameter | |
|---:|:---:|
| Audio resampling freq. | 16 KHz |
| Bands of Mel-spectrogram | 80 |
| Hop length | 400 |
| Convolution layers, channels, filter, strides | 1, 64, $20 \times 5$, $8 \times 2$ |
| Recurrent layer size | 128 |
| Fully connected size | 128 |
| Dropout probability | 0.9 |
| Learning Rate | $10^{-3}$ |
| Max gradient norm | 100 |
| Gradient clipping max. value | 5 |

*Table 4.* Hyperparameters of speaker verification model for LibriSpeech dataset.

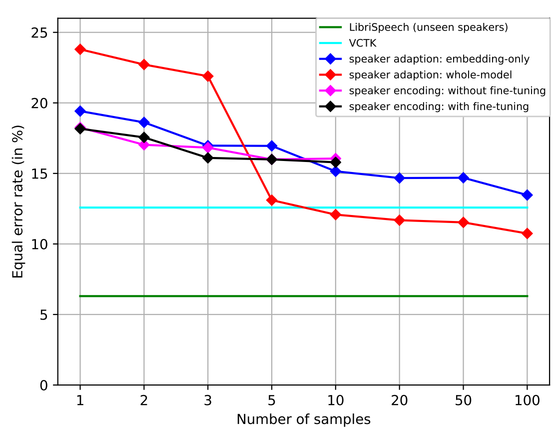


*Figure 10.* Speaker verification EER (using 1 enrollment audio) vs. number of cloning audio samples.

## D. Implications of attention

For a trained speaker encoder model, Fig. 12 exemplifies attention distributions for different audio lengths. The attention mechanism can yield highly non-uniformly distributed coefficients while combining the information in different cloning samples, and especially assigns higher coefficients to longer audios, as intuitively expected due to the potential more information content in them.

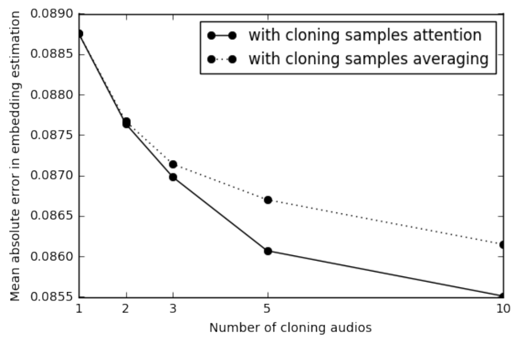


*Figure 11.* Mean absolute error in embedding estimation vs. the number of cloning audios for a validation set of 25 speakers, shown with the attention mechanism and without attention mechanism (by simply averaging).

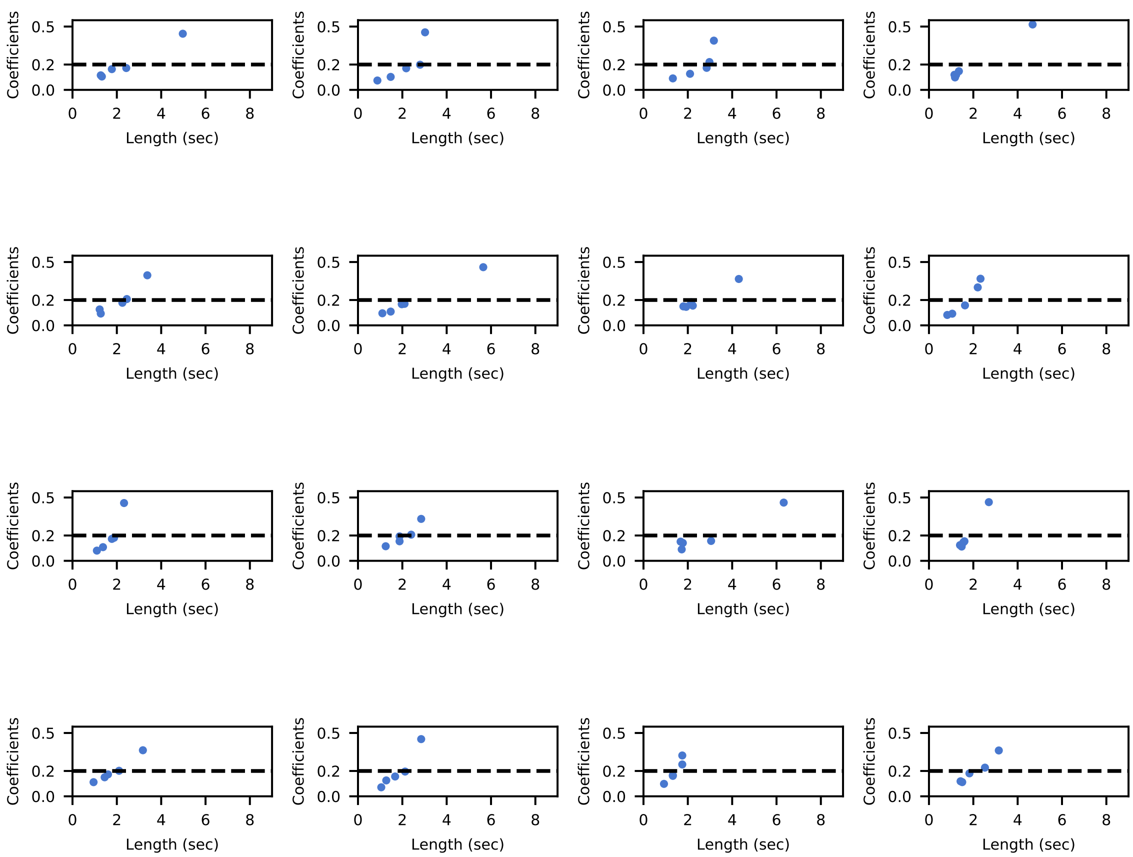


*Figure 12.* Inferred attention coefficients for the speaker encoder model with $N_{samples}$ = 5 vs. lengths of the cloning audio samples. The dashed line corresponds to the case of averaging all cloning audio samples.

## E. Speaker embedding space learned by the encoder

To analyze the speaker embedding space learned by the trained speaker encoders, we apply principal component analysis to the space of inferred embeddings and consider their ground truth labels for gender and region of accent from the VCTK dataset. Fig. 13 shows visualization of the first two principal components. We observe that speaker encoder maps the cloning audios to a latent space with highly meaningful discriminative patterns. In particular for gender, a one dimensional linear transformation from the learned speaker embeddings can achieve a very high discriminative accuracy - although the models never see the ground truth gender label while training.

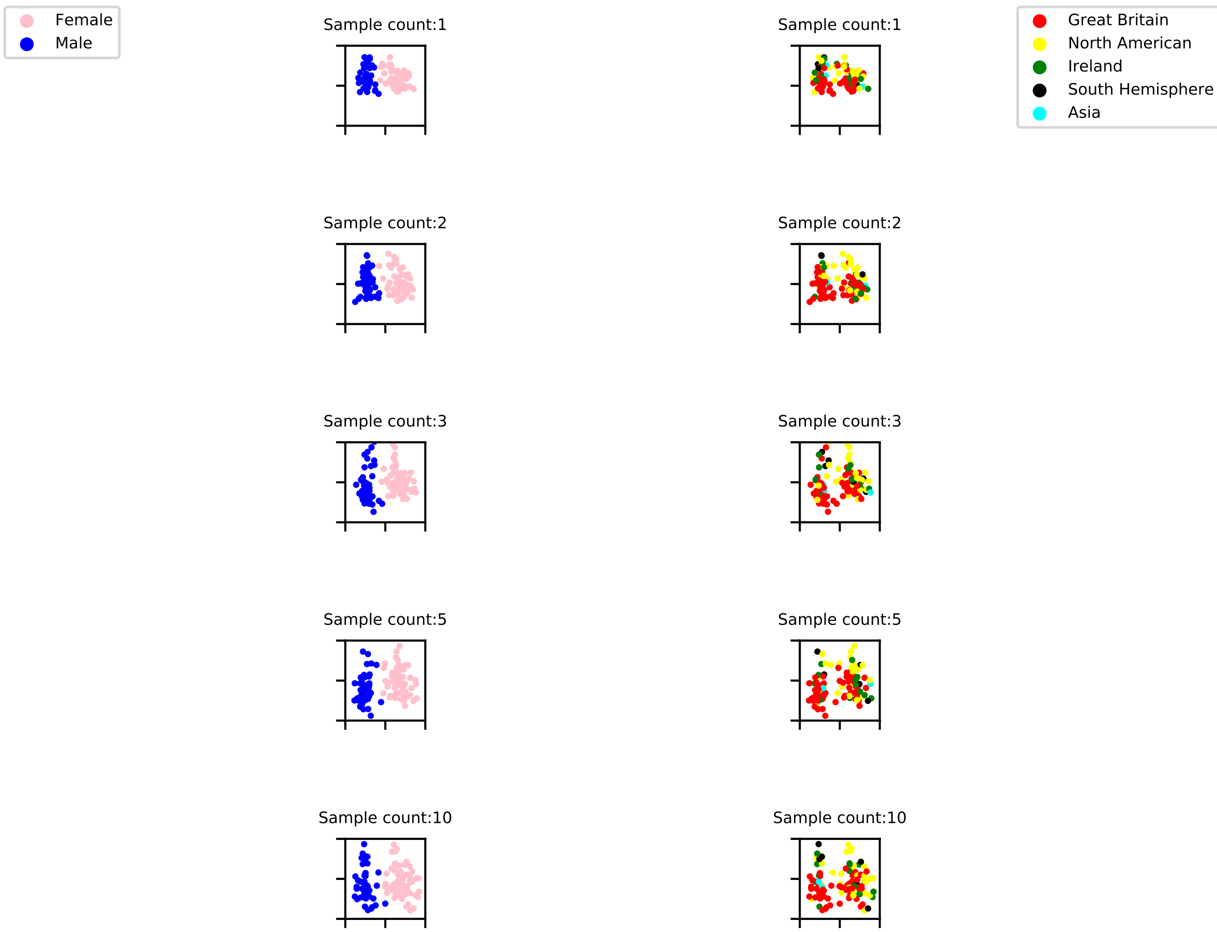


*Figure 13.* First two principal components of the inferred embeddings, with the ground truth labels for gender and region of accent for the VCTK speakers as in (Arik et al., 2017b).

## F. Similarity scores

For the result in Table 3, Fig. 14 shows the distribution of the scores given by MTurk users as in (Wester et al., 2016). For 10 sample count, the ratio of evaluations with the 'same speaker' rating exceeds 70 % for all models.

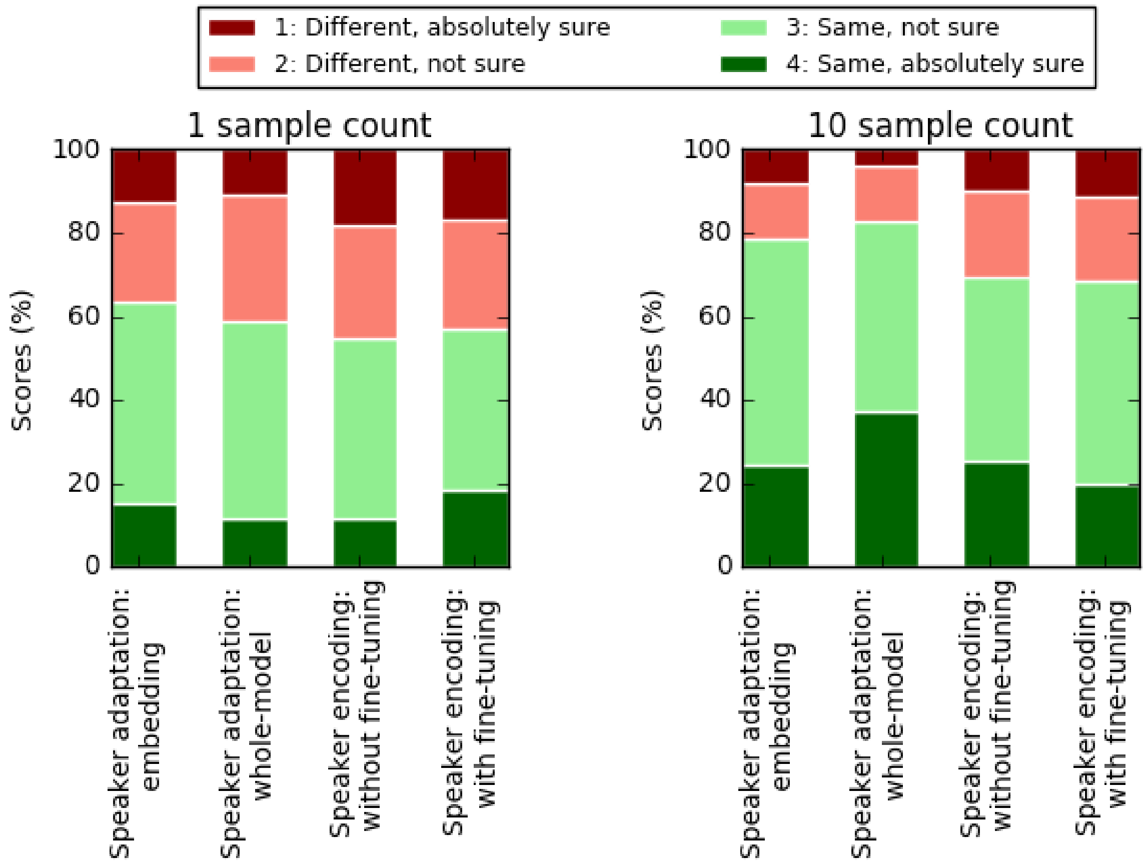


*Figure 14.* Distribution of similarity scores for 1 and 10 sample counts.