

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

Ye Jia* Yu Zhang* Ron J. Weiss* Quan Wang Jonathan Shen Fei Ren
 Zhifeng Chen Patrick Nguyen Ruoming Pang Ignacio Lopez Moreno Yonghui Wu
 Google Inc.
 {jiaye,ngyuzh,ronw}@google.com

Abstract

We describe a neural network-based system for **text-to-speech (TTS)** synthesis that is able to generate speech audio in the voice of different speakers, including those unseen during training. Our system consists of **three independently trained components**: (1) a *speaker encoder network*, trained on a speaker verification task using an independent dataset of noisy speech without transcripts from thousands of speakers, to generate a fixed-dimensional embedding vector from only seconds of reference speech from a target speaker; (2) a *sequence-to-sequence synthesis network* based on Tacotron 2 that generates a mel spectrogram from text, conditioned on the speaker embedding; (3) an *auto-regressive WaveNet-based vocoder network* that converts the mel spectrogram into time domain waveform samples. We demonstrate that the proposed model is able to transfer the knowledge of speaker variability learned by the discriminatively-trained speaker encoder to the multispeaker TTS task, and is able to synthesize natural speech from speakers unseen during training. We quantify the importance of training the speaker encoder on a large and diverse speaker set in order to obtain the best generalization performance. Finally, we show that randomly sampled speaker embeddings can be used to synthesize speech in the voice of novel speakers dissimilar from those used in training, indicating that the model has learned a high quality speaker representation.

1 Introduction

The goal of this work is to build a TTS system which can generate natural speech for a variety of speakers in a data efficient manner. We specifically address a zero-shot learning setting, where a few seconds of untranscribed reference audio from a target speaker is used to synthesize new speech in that speaker’s voice, without updating any model parameters. Such systems have accessibility applications, such as restoring the ability to communicate naturally to users who have lost their voice and are therefore unable to provide many new training examples. They could also enable new applications, such as transferring a voice across languages for more natural speech-to-speech translation, or generating realistic speech from text in low resource settings. However, it is also important to note the potential for misuse of this technology, for example impersonating someone’s voice without their consent. In order to address safety concerns consistent with principles such as [1], we verify that voices generated by the proposed model can easily be distinguished from real voices.

Synthesizing natural speech requires training on a large number of high quality speech-transcript pairs, and supporting many speakers usually uses tens of minutes of training data per speaker [8]. Recording a large amount of high quality data for many speakers is impractical. Our approach is to decouple speaker modeling from speech synthesis by independently training a speaker-discriminative embedding network that captures the space of speaker characteristics and training a high quality TTS

*Equal contribution.

model on a smaller dataset conditioned on the representation learned by the first network. Decoupling the networks enables them to be trained on independent data, which reduces the need to obtain high quality multispeaker training data. We train the speaker embedding network on a speaker verification task to determine if two different utterances were spoken by the same speaker. In contrast to the subsequent TTS model, this network is trained on untranscribed speech containing reverberation and background noise from a large number of speakers.

We demonstrate that the speaker encoder and synthesis networks can be trained on unbalanced and disjoint sets of speakers and still generalize well. We train the synthesis network on 1.2K speakers and show that training the encoder on a much larger set of 18K speakers improves adaptation quality, and further enables synthesis of completely novel speakers by sampling from the embedding prior.

There has been significant interest in end-to-end training of TTS models, which are trained directly from text-audio pairs, without depending on hand crafted intermediate representations [17, 23]. Tacotron 2 [15] used WaveNet [19] as a vocoder to invert spectrograms generated by an encoder-decoder architecture with attention [3], obtaining naturalness approaching that of human speech by combining Tacotron’s [23] prosody with WaveNet’s audio quality. It only supported a single speaker.

Gibiansky et al. [8] introduced a multispeaker variation of Tacotron which learned low-dimensional speaker embedding for each training speaker. Deep Voice 3 [13] proposed a fully convolutional encoder-decoder architecture which scaled up to support over 2,400 speakers from LibriSpeech [12].

These systems learn a fixed set of speaker embeddings and therefore only support synthesis of voices seen during training. In contrast, VoiceLoop [18] proposed a novel architecture based on a fixed size memory buffer which can generate speech from voices unseen during training. Obtaining good results required tens of minutes of enrollment speech and transcripts for a new speaker.

Recent extensions have enabled few-shot speaker adaptation where only a few seconds of speech per speaker (without transcripts) can be used to generate new speech in that speaker’s voice. [2] extends Deep Voice 3, comparing a *speaker adaptation* method similar to [18] where the model parameters (including speaker embedding) are fine-tuned on a small amount of adaptation data to a *speaker encoding* method which uses a neural network to predict speaker embedding directly from a spectrogram. The latter approach is significantly more data efficient, obtaining higher naturalness using small amounts of adaptation data, in as few as one or two utterances. It is also significantly more computationally efficient since it does not require hundreds of backpropagation iterations.

Nachmani et al. [10] similarly extended VoiceLoop to utilize a target speaker encoding network to predict a speaker embedding. This network is trained jointly with the synthesis network using a contrastive triplet loss to ensure that embeddings predicted from utterances by the same speaker are closer than embeddings computed from different speakers. In addition, a cycle-consistency loss is used to ensure that the synthesized speech encodes to a similar embedding as the adaptation utterance.

A similar spectrogram encoder network, trained without a triplet loss, was shown to work for transferring target prosody to synthesized speech [16]. In this paper we demonstrate that training a similar encoder to discriminate between speakers leads to reliable transfer of speaker characteristics. Our work is most similar to the speaker encoding models in [2, 10], except that we utilize a network independently-trained for a speaker verification task on a large dataset of untranscribed audio from tens of thousands of speakers, using a state-of-the-art generalized end-to-end loss [22]. [10] incorporated a similar speaker-discriminative representation into their model, however all components were trained jointly. In contrast, we explore transfer learning from a pre-trained speaker verification model.

Doddipatla et al. [7] used a similar transfer learning configuration where a speaker embedding computed from a pre-trained speaker classifier was used to condition a TTS system. In this paper we utilize an end-to-end synthesis network which does not rely on intermediate linguistic features, and a substantially different speaker embedding network which is not limited to a closed set of speakers. Furthermore, we analyze how quality varies with the number of speakers in the training set, and find that zero-shot transfer requires training on thousands of speakers, many more than were used in [7].

2 Multispeaker speech synthesis model

Our system is composed of three independently trained neural networks, illustrated in Figure 1: (1) a recurrent *speaker encoder*, based on [22], which computes a fixed dimensional vector from a speech

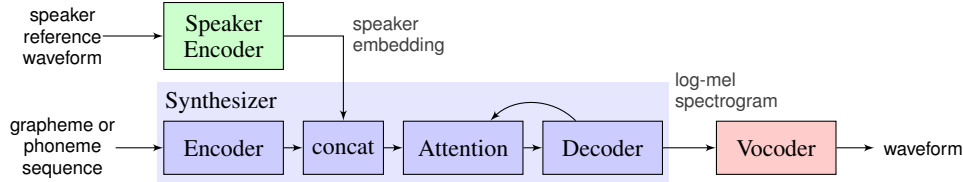


Figure 1: Model overview. Each of the three components are trained independently.

signal, (2) a sequence-to-sequence *synthesizer*, based on [15], which predicts a mel spectrogram from a sequence of grapheme or phoneme inputs, conditioned on the speaker embedding vector, and (3) an autoregressive WaveNet [19] *vocoder*, which converts the spectrogram into time domain waveforms.¹

2.1 Speaker encoder

The speaker encoder is used to condition the synthesis network on a reference speech signal from the desired target speaker. Critical to good generalization is the use of a representation which captures the characteristics of different speakers, and the ability to identify these characteristics using only a short adaptation signal, independent of its phonetic content and background noise. These requirements are satisfied using a speaker-discriminative model trained on a text-independent speaker verification task.

We follow [22], which proposed a highly scalable and accurate neural network framework for speaker verification. The network maps a sequence of log-mel spectrogram frames computed from a speech utterance of arbitrary length, to a fixed-dimensional embedding vector, known as *d-vector* [20, 9]. The network is trained to optimize a generalized end-to-end speaker verification loss, so that embeddings of utterances from the same speaker have high cosine similarity, while those of utterances from different speakers are far apart in the embedding space. The training dataset consists of speech audio examples segmented into 1.6 seconds and associated speaker identity labels; no transcripts are used.

Input 40-channel log-mel spectrograms are passed to a network consisting of a stack of 3 LSTM layers of 768 cells, each followed by a projection to 256 dimensions. The final embedding is created by L_2 -normalizing the output of the top layer at the final frame. During inference, an arbitrary length utterance is broken into 800ms windows, overlapped by 50%. The network is run independently on each window, and the outputs are averaged and normalized to create the final utterance embedding.

Although the network is not optimized directly to learn a representation which captures speaker characteristics relevant to synthesis, we find that training on a speaker discrimination task leads to an embedding which is directly suitable for conditioning the synthesis network on speaker identity.

2.2 Synthesizer

We extend the recurrent sequence-to-sequence with attention Tacotron 2 architecture [15] to support multiple speakers following a scheme similar to [8]. An embedding vector for the target speaker is concatenated with the synthesizer encoder output at each time step. In contrast to [8], we find that simply passing embeddings to the attention layer, as in Figure 1, converges across different speakers.

We compare two variants of this model, one which computes the embedding using the speaker encoder, and a baseline which optimizes a fixed embedding for each speaker in the training set, essentially learning a lookup table of speaker embeddings similar to [8, 13].

The synthesizer is trained on pairs of text transcript and target audio. At the input, we map the text to a sequence of phonemes, which leads to faster convergence and improved pronunciation of rare words and proper nouns. The network is trained in a transfer learning configuration, using a pretrained speaker encoder (whose parameters are frozen) to extract a speaker embedding from the target audio, i.e. the speaker reference signal is the same as the target speech during training. No explicit speaker identifier labels are used during training.

Target spectrogram features are computed from 50ms windows computed with a 12.5ms step, passed through an 80-channel mel-scale filterbank followed by log dynamic range compression. We extend [15] by augmenting the L_2 loss on the predicted spectrogram with an additional L_1 loss. In practice,

¹See https://google.github.io/tacotron/publications/speaker_adaptation for samples.

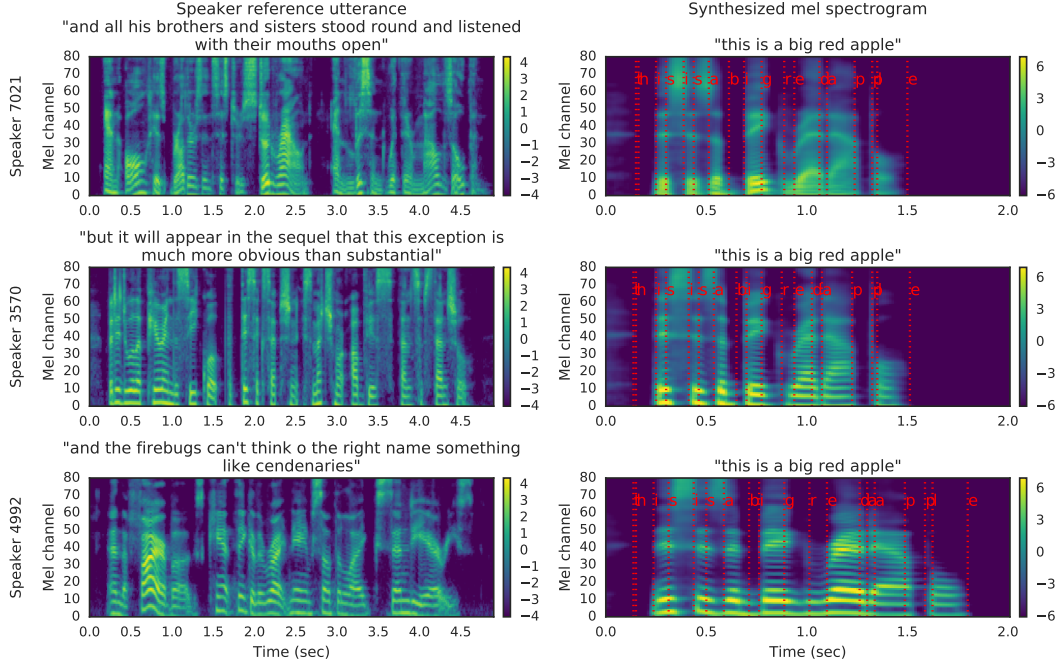


Figure 2: Example synthesis of a sentence in different voices using the proposed system. Mel spectrograms are visualized for reference utterances used to generate speaker embeddings (left), and the corresponding synthesizer outputs (right). The text-to-spectrogram alignment is shown in red. Three speakers held out of the train sets are used: one male (top) and two female (center and bottom).

we found this combined loss to be more robust on noisy training data. In contrast to [10], we don't introduce additional loss terms based on the speaker embedding.

2.3 Neural vocoder

We use the sample-by-sample autoregressive WaveNet [19] as a vocoder to invert synthesized mel spectrograms emitted by the synthesis network into time-domain waveforms. The architecture is the same as that described in [15], composed of 30 dilated convolution layers. The network is not directly conditioned on the output of the speaker encoder. The mel spectrogram predicted by the synthesizer network captures all of the relevant detail needed for high quality synthesis of a variety of voices, allowing a multispeaker vocoder to be constructed by simply training on data from many speakers.

2.4 Inference and zero-shot speaker adaptation

During inference the model is conditioned using arbitrary untranscribed speech audio, which does not need to match the text to be synthesized. Since the speaker characteristics to use for synthesis are inferred from audio, it can be conditioned on audio from speakers that are outside the training set. In practice we find that using a single audio clip of a few seconds duration is sufficient to synthesize new speech with the corresponding speaker characteristics, representing zero-shot adaptation to novel speakers. In Section 3 we evaluate how well this process generalizes to previously unseen speakers.

An example of the inference process is visualized in Figure 2, which shows spectrograms synthesized using several different 5 second speaker reference utterances. Compared to those of the female (center and bottom) speakers, the synthesized male (top) speaker spectrogram has noticeably lower fundamental frequency, visible in the denser harmonic spacing (horizontal stripes) in low frequencies, as well as formants, visible in the mid-frequency peaks present during vowel sounds such as the 'i' at 0.3 seconds – the top male F_2 is in mel channel 35, whereas the F_2 of the middle speaker appears closer to channel 40. Similar differences are also visible in sibilant sounds, e.g. the 's' at 0.4 seconds contains more energy in lower frequencies in the male voice than in the female voices. Finally, the characteristic speaking rate is also captured to some extent by the speaker embedding, as can be seen

Table 1: Speech naturalness Mean Opinion Score (MOS) with 95% confidence intervals.

System	VCTK Seen	VCTK Unseen	LibriSpeech Seen	LibriSpeech Unseen
Ground truth	4.43 ± 0.05	4.49 ± 0.05	4.49 ± 0.05	4.42 ± 0.07
Embedding table	4.12 ± 0.06	N/A	3.90 ± 0.06	N/A
Proposed model	4.07 ± 0.06	4.20 ± 0.06	3.89 ± 0.06	4.12 ± 0.05

by the longer signal duration in the bottom row compared to the top two. Similar observations can be made about the corresponding reference utterance spectrograms in the right column.

3 Experiments

We used two public datasets for training the speech synthesis and vocoder networks. VCTK [21] contains 44 hours of clean speech from 109 speakers, the majority of which have British accents. We downsampled the audio to 24 kHz, trimmed leading and trailing silence (reducing the median duration from 3.3 seconds to 1.8 seconds), and split into three subsets: train, validation (containing the same speakers as the train set) and test (containing 11 speakers held out from the train and validation sets).

LibriSpeech [12] consists of the union of the two “clean” training sets, comprising 436 hours of speech from 1,172 speakers, sampled at 16 kHz. The majority of speech is US English, however since it is sourced from audio books, the tone and style of speech can differ significantly between utterances from the same speaker. We resegmented the data into shorter utterances by force aligning the audio to the transcript using an ASR model and breaking segments on silence, reducing the median duration from 14 to 5 seconds. As in the original dataset, there is no punctuation in transcripts. The speaker sets are completely disjoint among the train, validation, and test sets.

Many recordings in the LibriSpeech clean corpus contain noticeable environmental and stationary background noise. We preprocessed the target spectrogram using a simple spectral subtraction [4] denoising procedure, where the background noise spectrum of an utterance was estimated as the 10th percentile of the energy in each frequency band across the full signal. This process was only used on the synthesis target; the original noisy speech was passed to the speaker encoder.

We trained separate synthesis and vocoder networks for each of these two corpora. Throughout this section, we used synthesis networks trained on phoneme inputs, in order to control for pronunciation in subjective evaluations. For the VCTK dataset, whose audio is quite clean, we found that the vocoder trained on ground truth mel spectrograms worked well. However for LibriSpeech, which is noisier, we found it necessary to train the vocoder on spectrograms predicted by the synthesizer network. No denoising was performed on the target waveform for vocoder training.

The speaker encoder was trained on a proprietary voice search corpus containing 36M utterances with median duration of 3.9 seconds from 18K English speakers in the United States. This dataset is not transcribed, but contains anonymized speaker identities. It is never used to train synthesis networks.

We primarily rely on crowdsourced Mean Opinion Score (MOS) evaluations based on subjective listening tests. All our MOS evaluations are aligned to the *Absolute Category Rating* scale [14], with rating scores from 1 to 5 in 0.5 point increments. We use this framework to evaluate synthesized speech along two dimensions: its naturalness and similarity to real speech from the target speaker.

3.1 Speech naturalness

We compared the naturalness of synthesized speech using synthesizers and vocoders trained on VCTK and LibriSpeech. We constructed an evaluation set of 100 phrases which do not appear in any training sets, and evaluated two sets of speakers for each model: one composed of speakers included in the train set (Seen), and another composed of those that were held out (Unseen). We used 11 seen and unseen speakers for VCTK and 10 seen and unseen speakers for LibriSpeech (Appendix D). For each speaker, we randomly chose one utterance with duration of about 5 seconds to use to compute the speaker embedding (see Appendix C). Each phrase was synthesized for each speaker, for a total of about 1,000 synthesized utterances per evaluation. Each sample was rated by a single rater, and each evaluation was conducted independently: the outputs of different models were not compared directly.

Table 2: Speaker similarity Mean Opinion Score (MOS) with 95% confidence intervals.

System	Speaker Set	VCTK	LibriSpeech
Ground truth	Same speaker	4.67 ± 0.04	4.33 ± 0.08
Ground truth	Same gender	2.25 ± 0.07	1.83 ± 0.07
Ground truth	Different gender	1.15 ± 0.04	1.04 ± 0.03
Embedding table	Seen	4.17 ± 0.06	3.70 ± 0.08
Proposed model	Seen	4.22 ± 0.06	3.28 ± 0.08
Proposed model	Unseen	3.28 ± 0.07	3.03 ± 0.09

Results are shown in Table 1, comparing the proposed model to baseline multispeaker models that utilize a lookup table of speaker embeddings similar to [8, 13], but otherwise have identical architectures to the proposed synthesizer network. The proposed model achieved about 4.0 MOS in all datasets, with the VCTK model obtaining a MOS about 0.2 points higher than the LibriSpeech model when evaluated on seen speakers. This is the consequence of two drawbacks of the LibriSpeech dataset: (i) the lack of punctuation in transcripts, which makes it difficult for the model to learn to pause naturally, and (ii) the higher level of background noise compared to VCTK, some of which the synthesizer has learned to reproduce, despite denoising the training targets as described above.

Most importantly, the audio generated by our model for unseen speakers is deemed to be at least as natural as that generated for seen speakers. Surprisingly, the MOS on unseen speakers is higher than that of seen speakers, by as much as 0.2 points on LibriSpeech. This is a consequence of the randomly selected reference utterance for each speaker, which sometimes contains uneven and non-neutral prosody. In informal listening tests we found that the prosody of the synthesized speech sometimes mimics that of the reference, similar to [16]. This effect is larger on LibriSpeech, which contains more varied prosody. This suggests that additional care must be taken to disentangle speaker identity from prosody within the synthesis network, perhaps by integrating a prosody encoder as in [16, 24], or by training on randomly paired reference and target utterances from the same speaker.

3.2 Speaker similarity

To evaluate how well the synthesized speech matches that from the target speaker, we paired each synthesized utterance with a randomly selected ground truth utterance from the same speaker. Each pair is rated by one rater with the following instructions: “You should not judge the content, grammar, or audio quality of the sentences; instead, just focus on the similarity of the speakers to one another.”

Results are shown in Table 2. The scores for the VCTK model tend to be higher than those for LibriSpeech, reflecting the cleaner nature of the dataset. This is also evident in the higher ground truth baselines on VCTK. For seen speakers on VCTK, the proposed model performs about as well as the baseline which uses an embedding lookup table for speaker conditioning. However, on LibriSpeech, the proposed model obtained a lower similarity MOS than the baseline, which is likely due to the wider degree of within-speaker variation (Appendix B), and background noise level in the dataset.

On unseen speakers, the proposed model obtains lower similarity between ground truth and synthesized speech. On VCTK, the similarity score of 3.28 is between “moderately similar” and “very similar” on the evaluation scale. Informally, it is clear that the proposed model is able to transfer the broad strokes of the speaker characteristics for unseen speakers, clearly reflecting the correct gender, pitch, and formant ranges (as also visualized in Figure 2). But the significantly reduced similarity scores on unseen speakers suggests that some nuances, e.g. related to characteristic prosody, are lost.

The speaker encoder is trained only on North American accented speech. As a result, accent mismatch constrains our performance on speaker similarity on VCTK since the rater instructions did not specify how to judge accents, so raters may consider a pair to be from different speakers if the accents do not match. Indeed, examination of rater comments shows that our model sometimes produced a different accent than the ground truth, which led to lower scores. However, a few raters commented that the tone and inflection of the voices sounded very similar despite differences in accent.

As an initial evaluation of the ability to generalize to out of domain speakers, we used synthesizers trained on VCTK and LibriSpeech to synthesize speakers from the other dataset. We only varied the train set of the synthesizer and vocoder networks; both models used an identical speaker encoder. As

Table 3: Cross-dataset evaluation on naturalness and speaker similarity for unseen speakers.

Synthesizer Training Set	Testing Set	Naturalness	Similarity
VCTK	LibriSpeech	4.28 ± 0.05	1.82 ± 0.08
LibriSpeech	VCTK	4.01 ± 0.06	2.77 ± 0.08

Table 4: Speaker verification EERs of different synthesizers on unseen speakers.

Synthesizer Training Set	Training Speakers	SV-EER on VCTK	SV-EER on LibriSpeech
Ground truth	—	1.53%	0.93%
VCTK	98	10.46%	29.19%
LibriSpeech	1.2K	6.26%	5.08%

shown in Table 3, the models were able to generate speech with the same degree of naturalness as on unseen, but in-domain, speakers shown in Table 1. However, the LibriSpeech model synthesized VCTK speakers with significantly higher speaker similarity than the VCTK model is able to synthesize LibriSpeech speakers. The better generalization of the LibriSpeech model suggests that training the synthesizer on only 100 speakers is insufficient to enable high quality speaker transfer.

3.3 Speaker verification

As an objective metric of the degree of speaker similarity between synthesized and ground truth audio for unseen speakers, we evaluated the ability of a limited speaker verification system to distinguish synthetic from real speech. We trained a new *eval-only* speaker encoder with the same network topology as Section 2.1, but using a different training set of 28M utterances from 113K speakers. Using a different model for evaluation ensured that metrics were not only valid on a specific speaker embedding space. We enroll the voices of 21 real speakers: 11 speakers from VCTK, and 10 from LibriSpeech, and score synthesized waveforms against the set of enrolled speakers. All enrollment and verification speakers are unseen during synthesizer training. Speaker verification equal error rates (SV-EERs) are estimated by pairing each test utterance with each enrollment speaker. We synthesized 100 test utterances for each speaker, so 21,000 or 23,100 trials were performed for each evaluation.

As shown in Table 4, as long as the synthesizer was trained on a sufficiently large set of speakers, i.e. on LibriSpeech, the synthesized speech is typically most similar to the ground truth voices. The LibriSpeech synthesizer obtains similar EERs of 5-6% using reference speakers from both datasets, whereas the one trained on VCTK performs much worse, especially on out-of-domain LibriSpeech speakers. These results are consistent with the subjective evaluation in Table 3.

To measure the difficulty of discriminating between real and synthetic speech for the same speaker, we performed an additional evaluation with an expanded set of enrolled speakers including 10 synthetic versions of the 10 real LibriSpeech speakers. On this 20 voice discrimination task we obtain an EER of 2.86%, demonstrating that, while the synthetic speech tends to be close to the target speaker (cosine similarity > 0.6 , and as in Table 4), it is nearly always even closer to other synthetic utterances for the same speaker (similarity > 0.7). From this we can conclude that the proposed model can generate speech that resembles the target speaker, but not well enough to be confusable with a real speaker.

3.4 Speaker embedding space

Visualizing the speaker embedding space further contextualizes the quantitative results described in Section 3.2 and 3.3. As shown in Figure 3, different speakers are well separated from each other in the speaker embedding space. The PCA visualization (left) shows that synthesized utterances tend to lie very close to real speech from the same speaker in the embedding space. However, synthetic utterances are still easily distinguishable from the real human speech as demonstrated by the t-SNE visualization (right) where utterances from each synthetic speaker form a distinct cluster adjacent to a cluster of real utterances from the corresponding speaker.

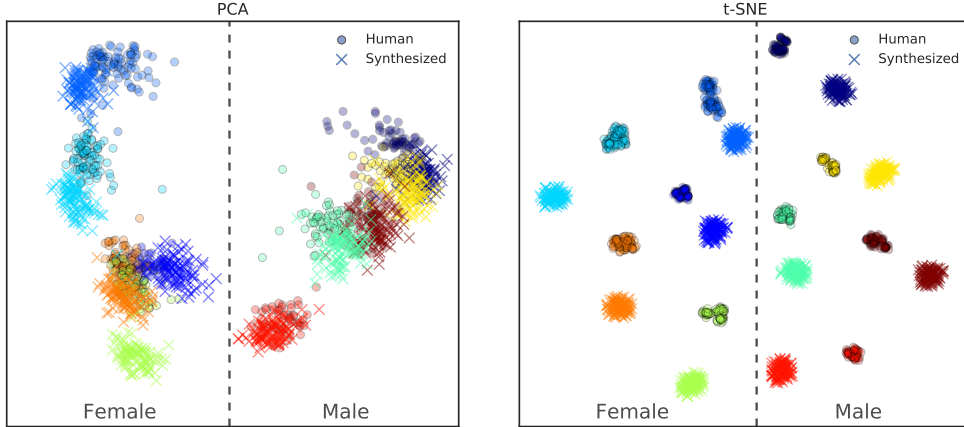


Figure 3: Visualization of speaker embeddings extracted from LibriSpeech utterances. Each color corresponds to a different speaker. Real and synthetic utterances appear nearby when they are from the same speaker, however real and synthetic utterances consistently form distinct clusters.

Table 5: Performance using speaker encoders (SEs) trained on different datasets. Synthesizers are all trained on LibriSpeech Clean and evaluated on held out speakers. LS: LibriSpeech, VC: VoxCeleb.

SE Training Set	Speakers	Embedding Dim	Naturalness	Similarity	SV-EER
LS-Clean	1.2K	64	3.73 ± 0.06	2.23 ± 0.08	16.60%
LS-Other	1.2K	64	3.60 ± 0.06	2.27 ± 0.09	15.32%
LS-Other + VC	2.4K	256	3.83 ± 0.06	2.43 ± 0.09	11.95%
LS-Other + VC + VC2	8.4K	256	3.82 ± 0.06	2.54 ± 0.09	10.14%
Internal	18K	256	4.12 ± 0.05	3.03 ± 0.09	5.08%

Speakers appear to be well separated by gender in both the PCA and t-SNE visualizations, with all female speakers appearing on the left, and all male speakers appearing on the right. This is an indication that the speaker encoder has learned a reasonable representation of speaker space.

3.5 Number of speaker encoder training speakers

It is likely that the ability of the proposed model to generalize well across a wide variety of speakers is based on the quality of the representation learned by the speaker encoder. We therefore explored the effect of the speaker encoder training set on synthesis quality. We made use of three additional training sets: (1) LibriSpeech Other, which contains 461 hours of speech from a set of 1,166 speakers disjoint from those in the clean subsets, (2) VoxCeleb [11], and (3) VoxCeleb2 [6] which contain 139K utterances from 1,211 speakers, and 1.09M utterances from 5,994 speakers, respectively.

Table 5 compares the performance of the proposed model as a function of the number of speakers used to train the speaker encoder. This measures the importance of speaker diversity when training the speaker encoder. To avoid overfitting, the speaker encoders trained on small datasets (top two rows) use a smaller network architecture (256-dim LSTM cells with 64-dim projections) and output 64 dimensional speaker embeddings.

We first evaluate the speaker encoder trained on LibriSpeech Clean and Other sets, each of which contain a similar number of speakers. In Clean, the speaker encoder and synthesizer are trained on the same data, a baseline similar to the non-fine tuned speaker encoder from [2], except that it is trained discriminatively as in [10]. This matched condition gives a slightly better naturalness and a similar similarity score. As the number of training speakers increases, both naturalness and similarity improve significantly. The objective EER results also improve alongside the subjective evaluations.

These results have an important implication for multispeaker TTS training. The data requirement for the speaker encoder is much cheaper than full TTS training since no transcripts are necessary, and the audio quality can be lower than for TTS training. We have shown that it is possible to synthesize very

Table 6: Speech from fictitious speakers compared to their nearest neighbors in the train sets. Synthesizer was trained on LS Clean. Speaker Encoder was trained on LS-Other + VC + VC2.

Nearest neighbors in	Cosine similarity	SV-EER	Naturalness MOS
Synthesizer train set	0.222	56.77%	3.65 ± 0.06
Speaker Encoder train set	0.245	38.54%	

natural TTS by combining a speaker encoder network trained on large amounts of untranscribed data with a TTS network trained on a smaller set of high quality data.

3.6 Fictitious speakers

Bypassing the speaker encoder network and conditioning the synthesizer on random points in the speaker embedding space results in speech from fictitious speakers which are not present in the train or test sets of either the synthesizer or the speaker encoder. This is demonstrated in Table 6, which compares 10 such speakers, generated from uniformly sampled points on the surface of the unit hypersphere, to their nearest neighbors in the training sets of the component networks. SV-EERs are computed using the same setup as Section 3.3 after enrolling voices of the 10 nearest neighbors. Even though these speakers are totally fictitious, the synthesizer and the vocoder are able to generate audio as natural as for seen or unseen real speakers. The low cosine similarity to the nearest neighbor training utterances and very high EER indicate that they are indeed distinct from the training speakers.

4 Conclusion

We present a neural network-based system for **multispeaker TTS synthesis**. The system combines an independently trained **speaker encoder network** with a **sequence-to-sequence TTS synthesis network** and **neural vocoder** based on Tacotron 2. By leveraging the knowledge learned by the discriminative speaker encoder, the synthesizer is able to generate high quality speech not only for speakers seen during training, but also for speakers never seen before. Through evaluations based on a speaker verification system as well as subjective listening tests, we demonstrated that the synthesized speech is reasonably similar to real speech from the target speakers, even on such unseen speakers.

We ran experiments to analyze the impact of the amount of data used to train the different components, and found that, given sufficient speaker diversity in the synthesizer training set, speaker transfer quality could be significantly improved by increasing the amount of speaker encoder training data.

Transfer learning is critical to achieving these results. By separating the training of the speaker encoder and the synthesizer, the system significantly lowers the requirements for multispeaker TTS training data. It requires neither speaker identity labels for the synthesizer training data, nor high quality clean speech or transcripts for the speaker encoder training data. In addition, training the components independently significantly simplifies the training configuration of the synthesizer network compared to [10] since it does not require additional triplet or contrastive losses. However, modeling speaker variation using a low dimensional vector limits the ability to leverage large amounts of reference speech. Improving speaker similarity given more than a few seconds of reference speech requires a model adaptation approach as in [2], and more recently in [5].

Finally, we demonstrate that the model is able to generate realistic speech from fictitious speakers that are dissimilar from the training set, implying that the model has learned to utilize a realistic representation of the space of speaker variation.

The proposed model does not attain human-level naturalness, despite the use of a WaveNet vocoder (along with its very high inference cost), in contrast to the single speaker results from [15]. This is a consequence of the additional difficulty of generating speech for a variety of speakers given significantly less data per speaker, as well as the use of datasets with lower data quality. An additional limitation lies in the model’s inability to transfer accents. Given sufficient training data, this could be addressed by conditioning the synthesizer on independent speaker and accent embeddings. Finally, we note that the model is also not able to completely isolate the speaker voice from the prosody of the reference audio, a similar trend to that observed in [16].

Acknowledgements

The authors thank Heiga Zen, Yuxuan Wang, Samy Bengio, the Google AI Perception team, and the Google TTS and DeepMind Research teams for their helpful discussions and feedback.

References

- [1] Artificial Intelligence at Google – Our Principles. <https://ai.google/principles/>, 2018.
- [2] Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. *arXiv preprint arXiv:1802.06006*, 2018.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*, 2015.
- [4] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.
- [5] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al. Sample efficient adaptive text-to-speech. *arXiv preprint arXiv:1809.10460*, 2018.
- [6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, pages 1086–1090, 2018.
- [7] Rama Doddipatla, Norbert Braunschweiler, and Rannieri Maia. Speaker adaptation in dnn-based speech synthesis using d-vectors. In *Proc. Interspeech*, pages 3404–3408, 2017.
- [8] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep Voice 2: Multi-speaker neural text-to-speech. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2962–2970. Curran Associates, Inc., 2017.
- [9] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5115–5119. IEEE, 2016.
- [10] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf. Fitting new speakers based on a short untranscribed sample. *arXiv preprint arXiv:1802.06984*, 2018.
- [11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [12] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.
- [13] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: 2000-speaker neural text-to-speech. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.
- [14] ITUT Rec. P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 1996.
- [15] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyriannakis, and Yonghui. Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.
- [16] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. *arXiv preprint arXiv:1803.09047*, 2018.

- [17] Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2Wav: End-to-end speech synthesis. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [18] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. VoiceLoop: Voice fitting and synthesis via a phonological loop. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.
- [19] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.
- [20] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4052–4056. IEEE, 2014.
- [21] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.
- [22] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.
- [23] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010, August 2017.
- [24] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.

Appendix A Additional joint training baselines

Table 7: Speech naturalness and speaker similarity Mean Opinion Score (MOS) with 95% confidence intervals of baseline models where the speaker encoder and synthesizer networks are trained jointly (top two rows). Included for comparison are the separately trained baseline from Table 5 (middle row) as well as the embedding lookup table baseline and proposed model from Tables 1 and 2 (bottom two rows). All but the bottom row, are trained entirely on LibriSpeech. The bottom row uses a speaker encoder trained on a separate speaker corpus. All evaluations are on LibriSpeech.

System	Embedding Dim	Naturalness MOS		Similarity MOS	
		Seen	Unseen	Seen	Unseen
Joint training	64	3.72 ± 0.06	3.59 ± 0.07	2.47 ± 0.08	2.44 ± 0.09
Joint training + speaker loss	64	3.71 ± 0.06	3.71 ± 0.06	2.82 ± 0.08	2.12 ± 0.08
Separate training (Table 5)	64	3.88 ± 0.06	3.73 ± 0.06	2.64 ± 0.08	2.23 ± 0.08
Embedding table (Tables 1,2)	64	3.90 ± 0.06	N/A	3.70 ± 0.08	N/A
Proposed model (Tables 1,2,5)	256	3.89 ± 0.06	4.12 ± 0.05	3.28 ± 0.08	3.03 ± 0.09

Although separate training of the speaker encoder and synthesizer networks is necessary if the speaker encoder is trained on a larger corpus of untranscribed speech, as described in Section 3.5, in this section we evaluate the effectiveness of joint training of the speaker encoder and synthesizer networks as a baseline, similar to [10].

We train on the Clean subset of LibriSpeech, containing 1.2K speakers, and use a speaker embedding dimension of 64 following Section 3.5. We compare two baseline jointly-trained systems: one without any constraints on the output of the speaker encoder, analogous to [16], and another with an additional speaker discrimination loss formed by passing the 64 dimension speaker embedding through a linear projection to form the logits for a softmax speaker classifier, optimizing a corresponding cross-entropy loss.

Naturalness and speaker similarity MOS results are shown in Table 7, comparing these jointly trained baselines to results reported in previous sections. We find that both jointly trained models obtain similar naturalness MOS on Seen speakers, with the variant incorporating a discriminative speaker loss performing better on Unseen speakers. In terms of both naturalness and similarity on Unseen speakers, the model which includes the speaker loss has nearly the same performance as the baseline from Table 5, which uses a separately trained speaker encoder that is also optimized to discriminate between speakers. Finally, we note that the proposed model, which uses a speaker encoder trained separately on a corpus of 18K speakers, significantly outperforms all baselines, once again highlighting the effectiveness of transfer learning for this task.

Appendix B Speaker variation

The tone and style of LibriSpeech utterances varies significantly between utterances even from the same speaker. In some examples, the speaker even tries to mimic a voice in a different gender. As a result, comparing the speaker similarity between different utterances from a same speaker (i.e. self-similarity) can sometimes be relatively low, and varies significantly speaker by speaker. Because of the noise level in LibriSpeech recordings, some speakers have significantly lower naturalness scores. This again varies significantly speaker by speaker. This can be seen in Table 8. In contrast, VCTK is more consistent in terms of both naturalness and self-similarity.

Table 4 shows the variance in naturalness MOS across different speakers on synthesized audio. It compares the MOS of different speakers for both ground truth and synthesized on VCTK, revealing that the performance of our proposed model on VCTK is also very speaker dependant. For example, speaker “p240” obtained a MOS of 4.48, which is very close to the MOS of the ground truth (4.57), but speaker “p260” is a full 0.5 points behind its ground truth.

Table 8: Ground truth MOS evaluations breakdown on unseen speakers. Similarity evaluations compare two utterances by the same speaker.

(a) VCTK				(b) LibriSpeech			
Speaker	Gender	Naturalness	Similarity	Speaker	Gender	Naturalness	Similarity
p230	F	4.22	4.65	1320	M	4.64	4.43
p240	F	4.57	4.67	2300	M	4.67	4.22
p250	F	4.31	4.72	3570	F	4.31	4.38
p260	M	4.56	4.31	3575	F	4.59	4.36
p270	M	4.29	4.77	4970	F	3.77	4.16
p280	F	4.41	4.71	4992	F	4.40	3.81
p300	F	4.60	4.87	6829	F	4.24	4.39
p310	F	4.56	4.52	7021	M	4.71	4.55
p330	F	4.34	4.77	7729	M	4.55	4.48
p340	F	4.44	4.71	8230	M	4.65	4.70
p360	M	4.36	4.63				

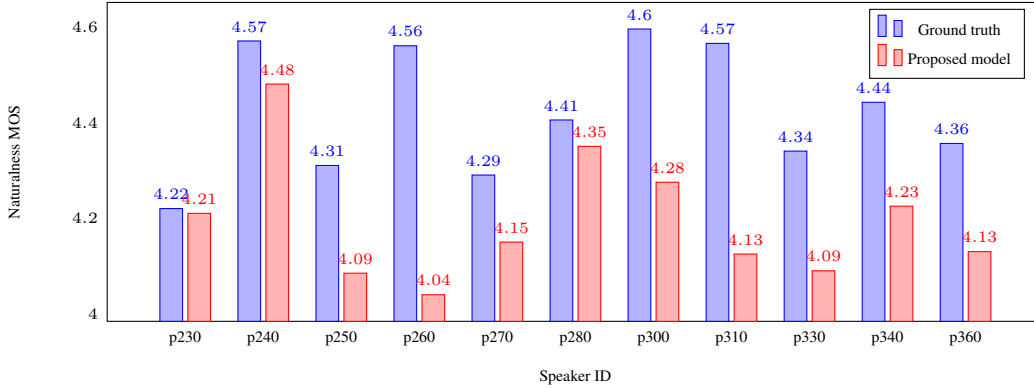


Figure 4: Per-speaker naturalness MOS of ground truth and synthesized speech on unseen VCTK speakers.

Appendix C Impact of reference speech duration

Table 9: Impact of duration of reference speech utterance. Evaluated on VCTK.

	1 sec	2 sec	3 sec	5 sec	10 sec
Naturalness (MOS)	4.28 ± 0.05	4.26 ± 0.05	4.18 ± 0.06	4.20 ± 0.06	4.16 ± 0.06
Similarity (MOS)	2.85 ± 0.07	3.17 ± 0.07	3.31 ± 0.07	3.28 ± 0.07	3.18 ± 0.07
SV-EER	17.28%	11.30%	10.80%	10.46%	11.50%

The proposed model depends on a reference speech signal fed into the speaker encoder. As shown in Table 9, increasing the length of the reference speech significantly improved the similarity, because we can compute more precise speaker embedding with it. Quality saturates at about 5 seconds on VCTK. Shorter reference utterances give slightly better naturalness, because they better match the durations of reference utterances used to train the synthesizer, whose median duration is 1.8 seconds. The proposed model achieves close to the best performance using only 2 seconds of reference audio. The performance saturation using only 5 seconds of speech highlights a limitation of the proposed model, which is constrained by the small capacity of the speaker embedding. Similar scaling was found in [2], where adapting a speaker embedding alone was shown to be effective given limited adaptation data, however fine tuning the full model was required to improve performance if more data was available. This pattern was also confirmed in more recent work [5].

Appendix D Evaluation speaker sets

Table 10: Speaker sets used for evaluation.

(a) VCTK											
Seen											
Speaker	p231	p241	p251	p261	p271	p281	p301	p311	p341	p351	p361
Gender	F	M	M	F	M	M	F	M	F	F	F
Unseen											
Speaker	p230	p240	p250	p260	p270	p280	p300	p310	p330	p340	p360
Gender	F	F	F	M	M	F	F	F	F	F	M

(b) LibriSpeech											
Seen											
Speaker	446	1246	2136	4813	4830	6836	7517	7800	8238	8123	
Gender	M	F	M	M	M	M	F	F	F	F	
Unseen											
Speaker	1320	2300	3570	3575	4970	4992	6829	7021	7729	8230	
Gender	M	M	F	F	F	F	F	M	M	M	

Appendix E Fictitious speakers

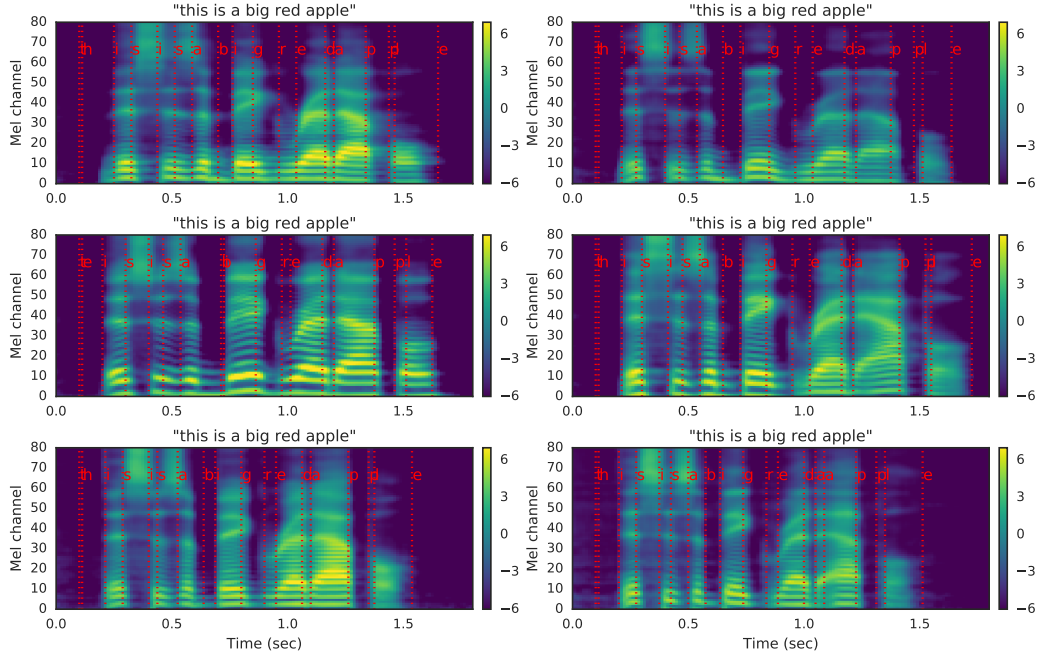


Figure 5: Example synthesis of a sentence conditioned on several random speaker embeddings sampled from the unit hypersphere. All samples contain consistent phonetic content, but there is clear variation in fundamental frequency and speaking rate. Audio files corresponding to these utterances are included in the demo page (https://google.github.io/tacotron/publications/speaker_adaptation).

Appendix F Speaker similarity MOS evaluation interface

Instructions

In this task, your job is to evaluate if the two speech audio samples are from the same speaker. Please release this task if any of the following are true:

- You think you do not have good listening ability.
- There is considerable background noise (street noise, loud fan/air-conditioner, open TV/radio, people talking, etc).
- For any reason, you can't hear the audio samples.

Task



How are you listening to the speech samples?

- ☒ **Headphones, with no noise in the background.** I am listening to the speech sample using headphones and there is **no noise** around me (people talking, music playing, air-conditioners, fans, etc.).
- ☐ **Headphones, with some low-level noise in the background.** I am listening to the speech sample using headphones and there is some **low-level** noise around me (people talking, music playing, air-conditioners, fans, etc.).
- ☐ **Audio speakers or other.**

Speaker Similarity Between Two Speech Samples

Please listen to the two speech samples below (Sample A and Sample B) and rate how similar they are. Your rating should reflect your evaluation of how close the voices of the two speakers sound. You **should not judge the content, grammar, or audio quality** of the sentences; instead, just focus on the similarity of the speakers to one another.

Speech samples (please listen at least **two times each**)

Sample A	
Sample B	

Please rate the similarity of the two speech samples

A horizontal slider scale for rating similarity. The scale has seven tick marks labeled from left to right: "N/A", "Not at all similar", "Slightly similar", "Moderately similar", "Very similar", and "Extremely similar". A green vertical bar is positioned at the "Very similar" mark.

Comment (optional)

Figure 6: Interface of MOS evaluation for speaker similarity.