

UNIVERSITÉ DE LIÈGE

FACULTÉ DES SCIENCES APPLIQUÉES

---

# Automatic Multispeaker Voice Cloning

---

*Author:*

Corentin JEMINE

*Supervisor:*

Prof. Gilles LOUPPE

Academic year 2018 - 2019



*Graduation studies conducted for obtaining the Master's degree  
in Data Science by Corentin Jemine*

# Abstract

Recent advances in deep learning have shown impressive results in the domain of text-to-speech. To this end, a deep neural network is usually trained using a corpus of several hours of professionally recorded speech from a single speaker. Giving a new voice to such a model is highly expensive, as it requires recording a new dataset and retraining the model. A recent research introduced a three-stage pipeline that allows to clone a voice unseen during training from only a few seconds of reference speech, and without retraining the model. The authors share remarkably natural-sounding results, but provide no implementation. We reproduce this framework and open-source the first public implementation of it. We adapt the framework with a newer vocoder model, so as to make it run in real-time.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>A review of text-to-speech methods in machine learning</b>	<b>5</b>
2.1	Statistical parametric speech synthesis . . . . .	5
2.2	Evolution of the state of the art in text-to-speech . . . . .	6
<b>3</b>	<b>Related voice cloning methods</b>	<b>9</b>
<b>4</b>	<b>Problem definition</b>	<b>9</b>
<b>5</b>	<b>Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis</b>	<b>9</b>
5.1	Overview . . . . .	9

# 1 Introduction

Deep learning models have become predominant in many fields of applied machine learning. text-to-speech (TTS), the process of synthesizing artificial speech from a text prompt, is no exception. Deep models that would produce more natural-sounding speech than the traditional concatenative approaches begun appearing in 2016. Much of the research focus has been since gathered around making these deep models more efficient, more natural, or training them in an end-to-end fashion. Inference has come from being hundreds of times slower than real-time on GPU (van den Oord et al., 2016) to possible in real-time on a mobile CPU (Kalchbrenner et al., 2018). As for the quality of the generated speech, Shen et al. (2017) demonstrate near human naturalness. Interestingly, speech naturalness is best rated with subjective metrics; and comparison with actual human speech leads to the conclusion that there might be such a thing as "speech more natural than human speech". In fact, some argue that the human naturalness threshold has already been crossed (Shirali-Shahreza and Penn, 2018).

Datasets of professionally recorded speech are a scarce resource. Synthesizing a natural voice with a correct pronunciation, lively intonation and a minimum of background noise requires training data with the same qualities. Furthermore, data efficiency often remains one of the shortcomings of deep learning. Training a common text-to-speech model such as Tacotron (Wang et al., 2017) typically requires tens of hours of speech. Yet the ability of generating speech with any voice is attractive for a range of applications be they useful or merely a matter of customization. Research has led to frameworks for voice conversion and voice cloning. They differ in that voice conversion is a form of style transfer on a speech segment from a voice to another, whereas voice cloning consists in capturing the voice of a speaker to perform text-to-speech on arbitrary inputs.

While the complete training of a single-speaker TTS model is technically a form of voice cloning, the interest rather lies in creating a fixed model that is able to incorporate newer voices with little data. The common approach is to condition a TTS model trained to generalize to new speakers on an embedding of the voice to clone (Arik et al., 2017, 2018; Jia et al., 2018). The embedding is low-dimensional and derived by a speaker encoder model that takes reference speech as input. This approach is typically more data efficient than training a separate TTS model for each speaker, in addition to being orders of magnitude faster and less computationally expensive. Interestingly, there is a large discrepancy between the duration of reference speech needed to clone a voice among the different methods, ranging from half an hour per speaker to only a few seconds.

Our objective is to achieve a powerful form of voice cloning. The resulting framework must be able to operate in a zero-shot setting, that is, for speakers unseen during training. It should incorporate a speaker's voice with only a few seconds of reference speech. These desired results are shown to be fulfilled by (Jia et al., 2018).

Their results are impressive<sup>1</sup>, but not backed by any public implementation. We reproduce their framework and make our implementation open-source<sup>2</sup>. In addition, we integrate the model of (Kalchbrenner et al., 2018) in the framework to make it run in real-time, i.e. to generate speech in a time shorter or equal to the duration of the produced speech. **add a word about our results**

The structure of this document goes as follows. We begin with a short introduction on TTS methods that involve machine learning. Follows a review of the evolution of the state of the art for TTS and for voice cloning. We then present the work of (Jia et al., 2018) as is, without mention of our own implementation. The remainder of this document focuses on the three individual parts of the framework, and on our implementation of the whole.

## 2 A review of text-to-speech methods in machine learning

### 2.1 Statistical parametric speech synthesis

Statistical parametric speech synthesis (SPSS) refers to a group of data-driven TTS methods that emerged in the late 90s. In SPSS, the relationship between the features computed on the input text and the output acoustic features is learned by a statistical generative model (called the acoustic model). A complete SPSS framework thus also includes a pipeline to extract features from the text to synthesize, as well as a system able to reconstruct an audio waveform from the acoustic features produced by the acoustic model (such a system is called a vocoder). Unlike the acoustic model, these two parts of the framework may be entirely engineered and make use of no statistical methods. While modern deep TTS models are usually not referred to as SPSS, the SPSS pipeline as depicted in figure 1 applies just as well to those newer methods.

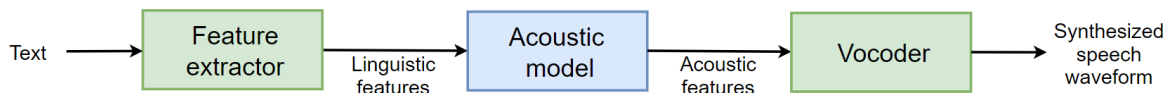


Figure 1: The general SPSS pipeline. The blue box is purely a statistical model while the green boxes can be engineered processes or/and statistical models.

The role of the feature extractor is to provide data that is more indicative of what the speech produced by the model is expected to sound like. Speech is a complex process, and directly feeding characters to a weak acoustic model will prove not to be effective. Providing additional features from natural language processing (NLP) techniques may greatly reduce the extent of the task to be learned by the acoustic

<sup>1</sup>[https://google.github.io/tacotron/publications/speaker\\_adaptation/index.html](https://google.github.io/tacotron/publications/speaker_adaptation/index.html)

<sup>2</sup>**repo link**

model. It may however result in trade-offs when it comes to naturalness, especially for rare or unknown words. Indeed, manually engineered heuristics do not quite fully characterize all intricacies of spoken language. For this reason, feature extraction can also be done with trained models. The line between the feature extractor and the acoustic model can then become blurry, especially for deep models. In fact, a tendency that is common across all areas where deep models have overtaken traditional machine learning techniques is for feature extraction to consist of less heuristics, as models become able to operate at higher levels of abstraction.

A common feature extraction technique is to build frames that will integrate surrounding context in a hierarchical fashion. For example, a frame at the syllable level could include the word that comprises it, its position in the word, the neighbouring syllables, the phonemes that make up the syllable, ... The lexical stress and accent of individual syllables can be predicted by a statistical model such as a decision tree. To encode prosody, a set of rules such as ToBI (Beckman and Elam, 1997) can be used. Ultimately, there remains a work of feature engineering to present a frame as a numerical object to the model, e.g. categorical features are typically encoded using a one-hot representation.

One could wonder why the acoustic model does not directly predict an audio waveform. Audio happens to be difficult to model: it is a particularly dense domain and audio signals are typically highly nonlinear. A representation that brings out features in a more tractable manner is the time-frequency domain. Spectrograms are much less dense than their waveform counterpart and also have the benefit of being two-dimensional, thus allowing models to better leverage spatial connectivity. Unfortunately, a spectrogram is a lossy representation of the waveform that discards the phase. There is no unique inverse transformation function, and deriving one that produces natural-sounding results is not trivial. When referring to speech, this generative function is called a vocoder. The choice of the vocoder is an important factor in determining the quality of the generated audio.

Talk about evaluation metrics (MOS and A/B testing)?

## 2.2 Evolution of the state of the art in text-to-speech

The state of the art in SPSS has for long remained a hidden Markov model (HMM) based framework (Tokuda, 2013). This approach, laid out in figure 2, consists in clustering the linguistic features extracted from the input text with a decision tree, and to train a HMM per cluster (Yoshimura et al., 1999). The HMMs are tasked to produce a distribution over spectrogram coefficients, their derivative, second derivative and a binary flag that indicates which parts of the generated audio should contain voice. With the maximum likelihood parameter generation algorithm (MLPG) (Tokuda et al., 2000), spectrogram coefficients are sampled from this distribution and eventually fed to the MLSA vocoder (Imai, 1983). It is possible to modify the voice generated by conditioning the HMMs on a speaker or tuning the generated speech parameters with

adaptation or interpolation techniques (Yoshimura et al., 1997). Note that, while this framework used to be state of the art for SPSS, it was still inferior in terms of the naturalness of the generated speech compared to the well-established concatenative approaches.

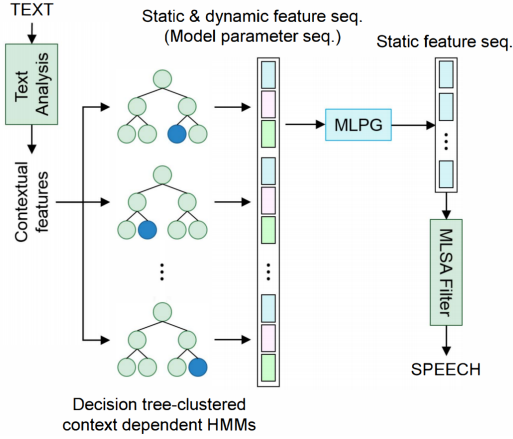


Figure 2: The general HMM-based TTS pipeline.

Method	MOS
HMM+MLPG	3.08 ( $\pm 0.12$ )
HMM+DNN	2.86 ( $\pm 0.12$ )
<b>DNN+MLPG</b>	<b>3.53 (<math>\pm 0.12</math>)</b>
DNN+DNN	3.17 ( $\pm 0.12$ )

Table 1: MOS of the different methods explored in (Hashimoto et al., 2015). The first line is the HMM-based framework. For the second and fourth line, the MLPG algorithm is replaced by a fully-connected neural network.

Improvements to this framework were later brought by feed-forward deep neural networks (DNN), as a result of progress in both hardware and software. Zen et al. (2013) proposes to replace entirely the decision tree-clustered HMMs in favor of a DNN. They argue for better data efficiency as the training set is no longer fragmented in different clusters of contexts. They demonstrate improvements over the speech quality with a number of parameters similar to that of the HMM-based approach. Later researches corroborate these findings (Qian et al., 2014; Hashimoto et al., 2015). The MOS of different model combinations tried by (Hashimoto et al., 2015) are reported in Table 1

(Fan et al., 2014) support that RNNs make natural acoustic models as they are able to learn a compact representation of complex and long-span functions. As RNNs are fit to generate temporally consistent series, the static features can directly be determined by the acoustic model, alleviating the need for dynamic features and MLPG. They compare networks of bidirectional LSTMs against the HMM and DNN based approaches described previously. Their A/B testing results are conclusive, we report them in figure 3.

59% Hybrid_B	19% Neutral	22% HMM
55% Hybrid_B	25% Neutral	20% DNN_B

Figure 3: A/B testing of the models. Hybrid\_B is a network with fully-connected and bidirectional LSTM layers.

The coming of WaveNet (van den Oord et al., 2016) made a substantial break-

through in TTS. WaveNet is a deep convolutional neural network that, for a raw audio waveform, models the distribution of a single sample given all previous ones. It is thus possible to directly generate audio by predicting samples one at a time in an autoregressive fashion. WaveNet leverages stacks of one-dimensional dilated convolutions with a dilation factor increasing exponentially with the layer depth, allowing for the very large receptive field and the strong nonlinearity needed to model raw audio. Conditioning the model on linguistic features is required to perform TTS. WaveNet acts thus both as an acoustic model and as a vocoder. Note that without the local conditioning, a trained WaveNet generates sound alike the training data but without structure or semantics (essentially babbling). The authors compare WaveNet to an older parametric approach and to a concatenative approach, the results are reported in Figure 4. The parametric approach is an LSTM-based system while the other is an HMM-driven unit selection concatenative system (not detailed in this document). Notice how the results vary between US English and Mandarin Chinese, showing that TTS performance is not language agnostic.

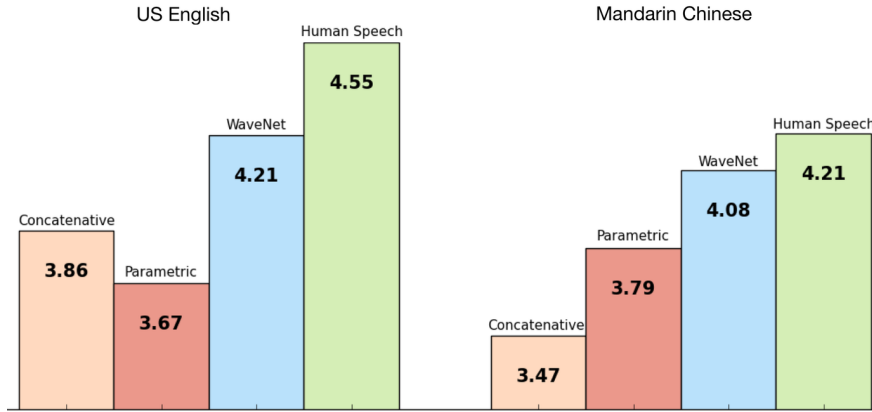


Figure 4: MOS of WaveNet’s performance compared with a parametric and concatenative approach, as well as with natural speech.

Follows Tacotron (Wang et al., 2017), a sequence-to-sequence model that produces a spectrogram from a sequence of characters alone, further reducing the need for domain expertise. In this framework, the vocoder is the Griffin-Lim algorithm. Tacotron uses an encoder-decoder architecture where, at each step, the decoder operates on a weighted sum of the encoder outputs. This attention mechanism, described in (Bahdanau et al., 2014), lets the network decide which steps of the input sequence are important with respect to each step of the output sequence. Tacotron achieves a MOS of 3.85 on a US English dataset, which is more than the 3.69 score obtained in the parametric approach of (Zen et al., 2016) but less than the 4.09 score obtained by the concatenative approach of (Gonzalvo et al., 2016). The authors mention that Tacotron is merely a step towards a better framework. Subsequently, Tacotron 2 is published (Shen et al., 2017). The architecture of Tacotron 2 remains that of an encoder-decoder with attention although several changes to the type of layers are made. The main difference with Tacotron is the use of a modified WaveNet as vocoder. On the same dataset, Tacotron 2 achieves a MOS of 4.53, which compares to the 4.58 for human speech (the difference is not statistically significant), achieving the all-time



highest MOS for TTS. With A/B testing, Tacotron 2 was found to be only slightly less preferred on average than ground truth samples. These ratings are shown in figure 5.

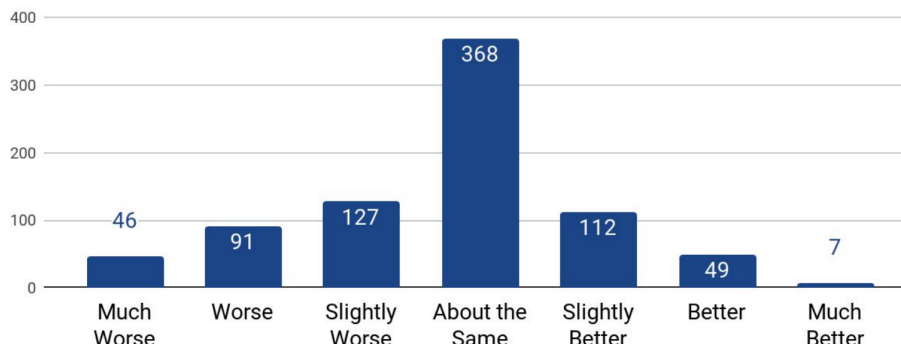


Figure 5: Preference ratings between Tacotron 2 and ground truth samples. There are 800 ratings from 100 items. The labels are expressed with respect to Tacotron 2.

Efficient neural audio synthesis?

### 3 Related voice cloning methods

### 4 Problem definition

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

### 5.1 Overview

Our solution to real-time voice cloning is largely based on (Jia et al., 2018) (referred to as SV2TTS throughout this document). It describes an approach to zero-shot voice cloning that only requires 5 seconds of reference speech. This paper is only one of the many publications from the Tacotron series<sup>3</sup> authored at Google. Interestingly, the SV2TTS paper does not bring much innovation of its own, rather it is based on three major earlier works from Google: the GE2E loss (Wan et al., 2017), Tacotron (Wang et al., 2017) and WaveNet (van den Oord et al., 2016). The complete framework is a three-stage pipeline, where the steps correspond to the models listed in order previously. Many of the current TTS tools and functionalities provided by Google, such as the Google assistant<sup>4</sup> or the Google cloud services<sup>5</sup>, make use of these same

<sup>3</sup><https://google.github.io/tacotron/>

<sup>4</sup><https://assistant.google.com/>

<sup>5</sup><https://cloud.google.com/text-to-speech/>

models. While there are many open-source reimplementations of these models online, there is none of the SV2TTS framework to our knowledge (as of May 2019).

The three stages of the framework are as follows:

- A speaker encoder that derives an embedding from the short utterance of a single speaker. The embedding is a meaningful representation of the voice of the speaker, such that similar voices are close in latent space. This model is described in (Wan et al., 2017) (referred as GE2E throughout this document) and (Heigold et al., 2015).
- A synthesizer that, conditioned on the embedding of a speaker, generates a spectrogram from a text. This model is the popular Tacotron 2 (Shen et al., 2017) without WaveNet (which is often referred to as just Tacotron because it is largely similar to the first iteration).
- A vocoder that infers an audio waveform from the spectrograms generated by the synthesizer. The authors used WaveNet (van den Oord et al., 2016) as a vocoder, effectively reapplying the entire Tacotron 2 framework.

At inference time, the speaker encoder is fed a short reference utterance of the speaker to clone. It generates an embedding that is used to condition the synthesizer, and a text processed as a phoneme sequence is given as input to the synthesizer. The vocoder takes the output of the synthesizer to generate the speech waveform. This is illustrated in figure 6.

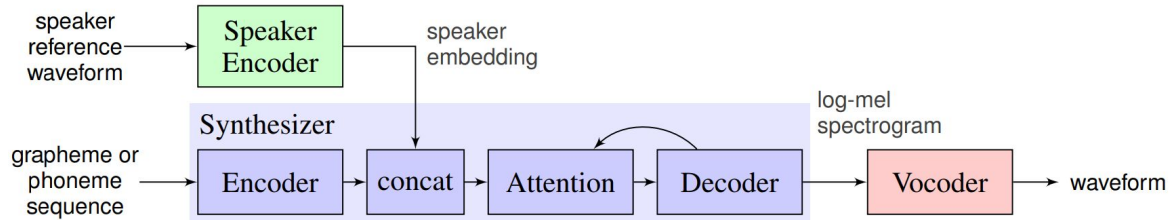


Figure 6: The SV2TTS framework. The blue blocks represent a high-level view of the Tacotron architecture modified to allow conditioning.

A particularity of the SV2TTS framework is that all models can be trained separately and on distinct datasets. For the encoder, one seeks to have a model that is robust to noise and able to capture the many characteristics of the human voice. Additionally, the encoder is trained with the GE2E loss which requires no labels other than the speaker identity! Therefore, a large corpus of many different speakers would be preferable to train the encoder, without any transcripts needed nor any strong requirements on the noise level of the audios. For the synthesizer and the vocoder however, transcripts are required and the quality of the generated audio can only be as good as that of the data. Higher quality datasets are thus required, which often means they are smaller in size.

encoder (or any):- a word about digital audio  
encoder: - meaning of the paper title  
synth: - the need of intermediate audio features  
synth: - advantages of conditioning on embedding rather than

## References

- Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech, 2017.
- Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples, 2018.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL <http://arxiv.org/abs/1409.0473>.
- Mary E. Beckman and Gayle Ayers Elam. Guidelines for tobi labelling, 03 1997.
- Y Fan, Yuqian Qian, Feng-Long Xie, and Frank Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. pages 1964–1968, 01 2014.
- Xavi Gonzalvo, Siamak Tazari, Chun-an Chan, Markus Becker, Alexander Gutkin, and Hanna Silen. Recent advances in google real-time hmm-driven unit selection synthesizer. In *Interspeech*, pages 2238–2242, 2016.
- K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. The effect of neural networks in statistical parametric speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4455–4459, April 2015. doi: 10.1109/ICASSP.2015.7178813.
- Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. *CoRR*, abs/1509.08062, 2015. URL <http://arxiv.org/abs/1509.08062>.
- S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 93–96, April 1983. doi: 10.1109/ICASSP.1983.1172250.
- Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *CoRR*, abs/1806.04558, 2018. URL <http://arxiv.org/abs/1806.04558>.
- Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis, 2018.
- Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Fast wavenet generation algorithm, 2016.

- Y. Qian, Y. Fan, W. Hu, and F. K. Soong. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833, May 2014. doi: 10.1109/ICASSP.2014.6854318.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017. URL <http://arxiv.org/abs/1712.05884>.
- S. Shirali-Shahreza and G. Penn. Mos naturalness and the quest for human-like speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 346–352, Dec 2018. doi: 10.1109/SLT.2018.8639599.
- K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1315–1318 vol.3, June 2000. doi: 10.1109/ICASSP.2000.861820.
- Yoshihiko; Toda Tomoki; Zen Heiga; Yamagishi Junichi; Oura Keiichiro Tokuda, Keiichi; Nankaku. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101, 05 2013. doi: 10.1109/JPROC.2013.2251852.
- Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL <http://arxiv.org/abs/1609.03499>.
- Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. *CoRR*, abs/1711.10433, 2017. URL <http://arxiv.org/abs/1711.10433>.
- Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2017.
- Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: A fully end-to-end text-to-speech synthesis model. *CoRR*, abs/1703.10135, 2017. URL <http://arxiv.org/abs/1703.10135>.
- Takayoshi Yoshimura, Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Tadashi Kitamura. Speaker interpolation in hmm-based speech synthesis system. In *EUROSPEECH*, 1997.

- Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *EUROSPEECH*, 1999.
- H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966, May 2013. doi: 10.1109/ICASSP.2013.6639215.
- Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemyslaw Szczepaniak. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. *CoRR*, abs/1606.06061, 2016. URL <http://arxiv.org/abs/1606.06061>.