Problem-Agnostic Speech Embeddings for Multi-Speaker Text-to-Speech with SampleRNN

David Álvarez, Santiago Pascual, Antonio Bonafonte[*]

Universitat Politècnica de Catalunya

davidalvarezdlt@gmail.com, santi.pascual@upc.edu

Abstract

Text-to-speech (TTS) acoustic models map linguistic features into an acoustic representation out of which an audible waveform is generated. The latest and most natural TTS systems build a direct mapping between linguistic and waveform domains, like SampleRNN. This way, possible signal naturalness losses are avoided as intermediate acoustic representations are discarded. Another important dimension of study apart from naturalness is their adaptability to generate voice from new speakers that were unseen during training. In this paper we first propose the use of problem-agnostic speech embeddings in a multi-speaker acoustic model for TTS based on SampleRNN. This way, we feed the acoustic model with speaker acoustically-dependent representations that enrich the waveform generation more than embeddings unrelated to these factors. Our first results suggest that the proposed embeddings lead to better quality voices than those obtained with one-hot embeddings. Furthermore, as we can use any speech segment as an encoded representation during inference, the model is capable to generalize to new speaker identities without retraining the network. We finally show that, with a small increase of speech duration in the embedding extractor, we dramatically reduce the spectral distortion to close the gap towards the target identities.

Index Terms: speech synthesis, text-to-speech, problem-agnostic speech embeddings, speaker adaptation.

1. Introduction

The fast development of new and deeper neural architectures has allowed the improvement of all tasks related to complex non-structured data. Speech synthesis is one of those tasks whose performance has improved dramatically with the latest trends in deep generative models. TTS is the technique by which a machine maps a text into audible speech. The task is especially difficult when we want to be able to generate speech from different speakers, forcing the system to learn how to differentiate among them but sharing some core information on the linguistic to acoustic signals correspondence.

Classic TTS systems are based on statistical models that extract linguistic and prosodic features from text and map them into an acoustic representation to obtain the speech waveform [1, 2]. These models work in a two-stage methodology: (1) linguistic and prosodic features are used to predict the duration of phonetic units with a duration model (predicting number of acoustic frames to be predicted); and (2) these same linguistic, prosodic and duration features are then fed into an acoustic model to generate the actual acoustic units that will finally conform a waveform. These acoustic units usually take the form of spectral or cepstral features, and to make these statistical

mappings in both stages we either rely on hidden Markov models (HMMs) or deep neural networks (DNNs) [2, 3]. The latter outperform previous statistical systems, and as such most successful statistical models apply some type of neural component in both stages of the TTS [4, 5]. Remarkably, different neural architectures and optimization schemes can be applied to increase the generation efficiency and voice naturalness [6, 7, 8, 9, 10, 11], especially to palliate over-smoothing effects that produce "buzziness" and muffled sounds [2].

The introduction of the WaveNet and SampleRNN systems showed that working on acoustic models at waveform level is plausible and currently outperforms acoustic feature-based systems in terms of naturalness [12]. WaveNet was first used as a plain acoustic mapping conditioned on linguistic and prosodic features, whereas SampleRNN was first used as a vocoder in the Char2wav [13] TTS system. Nonetheless, both systems are flexible enough to be conditioned on many different features that either express linguistic contents or spectral features out of which waveforms are generated, as WaveNet showed in the Tacotron [14] TTS. Also SampleRNN has been used as an acoustic model [15] for voice conversion with linguistic conditionings.

An important feature of statistical TTS systems, especially related to our work, is their flexibility in terms of modeling different speaker characteristics to mimic them in the generation. Acoustic models and neural vocoders have the capacity to model different speaker identities within the same neural structure. If we model all of them in the same training session, the way in which these multiple identities can be controlled is with a specific multi-output structure, where each output generates the acoustics of its respective speaker [16, 17], or with some input codes or embeddings expressing the modeled identity in a single model output [12, 18]. In case of having new identities to model after the model is already trained, there exist a number of speaker adaptation techniques to tune models to new identities with the least amount of data possible. These techniques usually rely on some form of model fine-tuning [19, 16, 17], but adapting identities with tiny amounts of data should potentially rely on more descriptive speaker embeddings than one-hot ones. This way, we could extract the embeddings from an example utterance of a target identity, and we could avoid fine-tuning the models by injecting the extracted acoustic descriptors of the target speaker [20, 21]. Even though this is a desirable case, the best results are still obtained after a fine-tuning stage of the model parameters even in data scarcity situations.

In this paper, we first build a multi-speaker acoustic model with a SampleRNN conditioned on linguistic-prosodic features and speaker identity embeddings. We then propose the use of problem-agnostic acoustic embeddings extracted from chunks of speech that serve as seed examples of speaker identity traits. This brings us the possibility of generating new identities by obtaining the acoustic vectors for new speakers without fine-

---

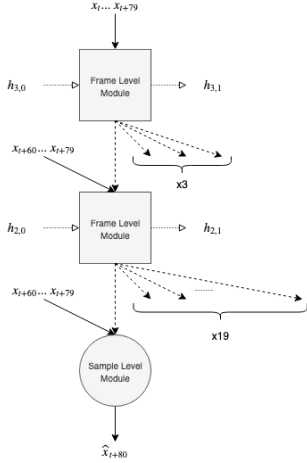[*]A. Bonafonte is currently at Amazon Research, Cambridge, UK.

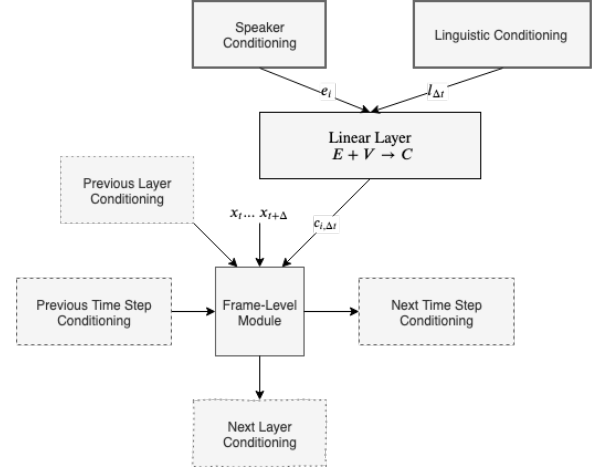Figure 1: *Original SampleRNN architecture with 2 frame-level layers and upsampling ratios {4, 20}*

Figure 2: *Combination of speaker and linguistic features along with other SampleRNN model-specific conditioning inputs*

tuning the network. Finally, we present an initial study of speaker adaptation without fine-tuning by varying the amount of available seed speech for different test speakers that remain unseen during training.

This work is structured as follows: we first review SampleRNN in section 2 as our waveform generator with which we build the acoustic model. In section 3 we review the problem-agnostic speech encoder (PASE), with which we obtain the speaker identity embeddings. The experimental setup is described in section 4, where datasets, modeling features and hyper-parameters are described. Results and conclusions are discussed in sections 5 and 6, respectively. The code of this project is publicly available online [1].

2. SampleRNN for Text-to-Speech

The original design of SampleRNN (see Fig. 1) is an auto-regressive model with data that propagates both from upper to bottom layers and from previous to next time steps. In each layer, it uses different scales of previous samples to condition the lower layer until it predicts the sample itself in the last layer. In the original work, the lowest layer is called sample-level layer and the others are called frame-level layers. As we want to control both the content of the speech and the speaker identity, we must somehow inject conditioning information to the system in addition to previous samples. We first experimented injecting the conditionings at the top-level layer (as in existing vocoding applications of SampleRNN [22]), however this solution had difficulties to generate intelligible speech in our acoustic model for multiple identities. To solve that, we use a similar strategy as in [15], combining both the original inputs with a global feature vector composed of speaker and linguistic features. These new feature vectors are used as inputs in the frame and sample level layers and are then concatenated with the other inputs before getting into the recurrent units. Hence, given a speaker i with its feature vector $\boldsymbol{e}_i \in \mathbb{R}^E$ and the linguistic feature vector of a certain time interval $\boldsymbol{l}_{\Delta t} \in \mathbb{R}^V$, we use a linear layer to obtain a global conditioning vector $\boldsymbol{c}_{i,\Delta t} \in \mathbb{R}^C$. Fig. 2 gives an overall scheme of this conditioning methodology along with other inputs of frame-level layers. The initial states of the

frame-level layers $h_{z,0}$ are learned via back-propagation. For the case of the sample-level module, it follows a similar structure but without time-step conditioning. Speaker identity can be introduced either with embedded one-hot codes as in previous approaches [12], or with some feature extractor out of speech signals that serve as seeds [20, 21]. In this work we propose to use a problem-agnostic speech encoder to follow the latter scheme, as described next.

3. Problem-Agnostic Acoustic Embeddings

As mentioned earlier, the identity conditionings injected to SampleRNN can be one-hot codes, as typical; or acoustically relevant features extracted from some spoken examples from the modeled identities that serve as seeds.

To extract these acoustic features, we use the problem-agnostic speech encoder (PASE) [23]. PASE is an encoder that transforms raw waveforms $\boldsymbol{x}$ into an intermediate representation $\boldsymbol{z}$ with a deep fully convolutional structure. This converts a waveform into a collection of 100-dimensional frames with frame rate 100Hz. The $\boldsymbol{z}$ representation is learned out of a collection of self-supervised tasks that predict different factors of the input waveform: the waveform itself, its log-power spectrum, the MFCC coefficients, etc. Each prediction is perform by a small multi-layer perceptron (MLP), called worker.

We decide to use PASE for its proven ability to capture identity and prosodic cues, as shown in the original paper with the results on speaker identification and emotion recognition [23]. In our work, we use the publicly available pre-trained encoder to extract our acoustic seeds, which are replaced by the usual one-hot embeddings.

Let the function $\Phi(\boldsymbol{x})$ represent the output latent representation $\boldsymbol{z}$ of the PASE encoder. We average this output vectors in time to obtain a summary of the speaker characteristics for a certain seed signal length T, thus obtaining the embeddings $\boldsymbol{e}_i$ as:

$$\boldsymbol{e}_i = \frac{1}{L}\sum_{n=0}^{L-1}\Phi(\boldsymbol{s}_i)[n] \qquad (1)$$

Where $\boldsymbol{s}_i$ corresponds to the i-th speaker seed signal, and L is the number of PASE frames coming out of the encoding (a

---

16 kHz signal the waveform is decimated to a 100 Hz signal after $\Phi(\boldsymbol{s}_i)$).

4. Experimental Setup

4.1. Datasets

For our experiments, we use utterances from two different datasets: VCTK [24] and CMU Arctic [25]. We decide to use speakers from both datasets as, while CMU Arctic contains more recorded speech per speaker, it does not contain enough speakers for our experiments. To avoid excessively modelling of silences, we trim them to a maximum length of 100 ms with the help of a voice activity detector.

To train the base models, we select the 20 speakers of each gender with most speech duration, allocating a total of 40 speakers for this purpose. Then, for each of those speakers, we select a total of 45 s of speech for the validation split and another 45 s to perform objective tests. The remaining data of each speaker is then allocated to the training split, obtaining an unbalanced training dataset where the less representative speaker (VCTK speaker) has about 10 minutes of speech and the most representative one (CMU Arctic speaker) reaches the 64 min. In total, the sum of the individual contributions of the 40 speakers reaches the 12 h of training data and 30 min. of validation and test data.

For the adaptation phase, we select 5 different random speakers of each gender (a total of 10 speakers) and allocate enough data in the training split to experiment with seed signal lengths T up to 120 s. Note that we do not adjust the model weights, but we still require a training set of utterances to have enough T samples to build the acoustic seed. We then assess the adapted speaker similarity towards the target identity using a different test split with 180 s of data per speaker and the metrics described in section 4.5.

4.2. Linguistic-prosodic features

As mentioned in section 2, one of the conditioning features that we input in every frame and sample level module is related to the linguistic and prosodic contents, as we are building an acoustic model. In our experiments, we will compare the performance difference when using two different sets of features:

- Linguistic features with a forced duration model.
- The previous linguistic features with $\log F0$ and unvoiced/voiced (UV) binary flags (similar to [12, 15]).

Given the set of utterances of a certain speaker, we use Merlin [26] to extract a total of 53 linguistic-prosodic features per phonetic unit jointly with its forced duration. Then, we repeat these features in time to fit the waveform time resolution Δt using the time-stamps as reference to include the duration information to the feature vector:

- Absolute duration: the total duration of the phoneme.
- Relative duration: the position of the generated Δt inside the phoneme, ranging from 0 to 1.

We use the Ahocoder vocoder to extract $\log F0$ and V/UV features [27]. All real-valued features (either linguistic or prosodic) are z-normalized with speaker-dependent statistics except for the relative duration, which is already bounded.

4.3. Experiments

First, we want to assess the performance difference between using the PASE embeddings against plain one-hot embeddings.

Secondly, we want to make a first evaluation of the adaptation capability of the PASE embeddings by picking different seed signal lengths T to infer the identity embedding $\boldsymbol{e}_i$.

For the first experiment, four SampleRNN variations are trained for 50 epochs, as the result of using $i)$ one-hot vs. PASE embeddings and $ii)$ including or not $\log F0 + $ UV

The training PASE embeddings are extracted from $T = 60\,s$ of seed speech per speaker. Then, we analyze the performance of each model in terms of the negative log-likelihood loss and finally extract the objective distortion metrics of section 4.5.

On the other hand, to assess the speaker adaptation, we pick the 10 randomly sampled speakers described in section 4.1 and the PASE embedding experiment without $\log F0$ information from the previous list. For each new speaker, we randomly sample chunks of lengths $T = \{1, 10, 60, 120\}$ seconds to build their time-dependent acoustic PASE embedding. Then, we compute the objective distortion of these speakers' synthesized test utterances for each one of the 4 embeddings against their own ground-truth utterances.

4.4. Hyperparameter tuning

The hyper-parameter configuration used throughout all the experiments is summarized in Table 1. Note that, depending on the experiment, only the PASE seed or the speaker embedding is used to represent the speaker identity. The learning rate is reduced on validation plateaus. The optimizer is Adam with all default parameters except for the learning rate. To have a reference on the computational cost per experiment, each model is ran on a Titan X GPU for approximately 8 days to reach the 50 epochs.

Table 1: *Hyperparameters used in the experiments*

Speech sampling frequency	16 kHz
Speech quantization	8 bits (μLaw)
Speaker embedding size (E)	100
Global features size (C)	50
Categorical linguistic features embedding size	15
Top frame-level seq. length	13
Top frame-level inp. size (Δt)	80
Upsampling Ratios	4, 20
GRU hidden size	1024
Batch size	128
Initial learning rate	10^{-4}
Learning rate patience	3
Learning scaling factor	0.5

4.5. Objective Metrics

We objectively evaluate the performance of our experiments by comparing the original utterances with the ones generated by our systems in two terms. First we have direct likelihood metrics as SampleRNN is a classifier computing the probability of the next waveform sample given the previous ones, which means that we get a score correlated to the generated quality in the likelihood validation and test curves. Secondly, we have spectral distortion measures that depict how good is the waveform generation embedding speaker, prosodic and content characteristics in the speech (without an explicit modelling of spectral features themselves). Hence, we measure the mel cepstral distortion (MCD) (in dB) and the RMSE of F0 (in Hz), as usually done in acoustic models assessment for TTS [28, 5].
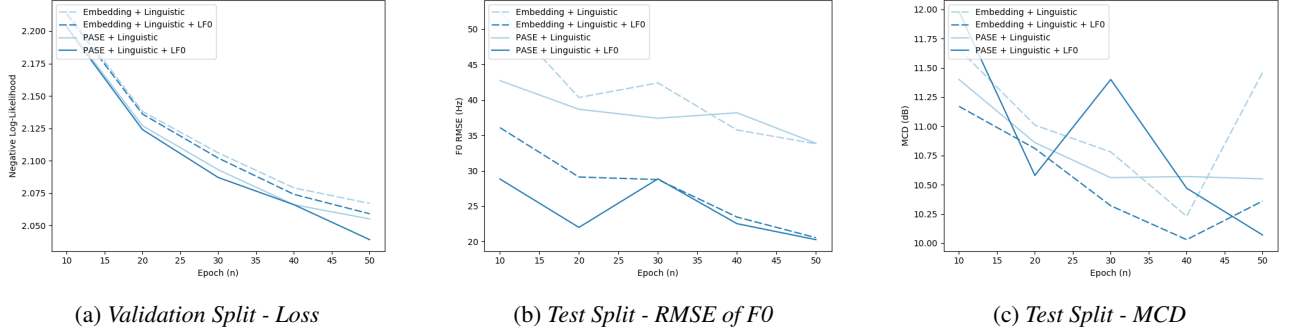
(a) *Validation Split - Loss*   (b) *Test Split - RMSE of F0*   (c) *Test Split - MCD*

Figure 3: *Objective measures of the acoustic multi-speaker TTS model*

5. Results

Fig. 3 illustrates the likelihood convergence in validation and the acoustic distortion metrics in test for the four different TTS variations explained in section 4.3. We chose to train each model for 50 epochs (as stated in section 4.3) for comparability purposes, but the different models can converge further as shown in these results. Nonetheless, some evidences appear at this point worth commenting. First, we can clearly see that the two experiments using our acoustic embeddings converge quicker than with the standard approach overall both in terms of validation likelihood and acoustic distortions. The peaks we observe at some points in the spectral distortions are probably a reflect of under-optimized models, but we can still observe a clear convergence acceleration trend in the validation loss for models involving PASE embeddings that make these TTS variations comparable.

Moreover, the distortions of the models with $\log F0$ contours as inputs are lower than those that lack them. This is expected because these contours supply long-term information that is useful for SampleRNN to retain far-past information better than it might in its internal states, as observed first in [12]. Nevertheless, the experiment with PASE embeddings that lacks $\log F0$ generally obtains better results than those of the one-hot embeddings with the $\log F0$ contours. We hypothesize that this happens due to PASE embeddings capturing some bias in each speaker prosodic traits, but we still have an averaged representation in this case so it does perform worse than when we use PASE embeddings with the prosodic contours where dynamic long-term information is still fully present.

For the sake of simplicity, in our first adaptability experiment we only take the PASE embeddings (without $\log F0$ contours) model to proceed to the next experiment. After inferring four embeddings per speaker identity (one per seed length as introduced in section 4.3) we use them to synthesize the test utterances. Then, the 10 different speaker distortion curves for both RMSE and MCD are averaged and depicted in Fig. 4. There we observe that the more seed signal we have, the less distortion we obtain across speakers, especially for MCD. The decreasing distortion trend depending on the amount of seed speech shows so far a promising direction towards generalizing speaker identities without fine-tuning TTS models. However, better optimized models are a potential need as shown in these results as models could converge further, hence biasing the adaptation curves to lower distortion rates with the same seed lengths. Qualitative results are available online in an audio samples webpage [2]. Infor-
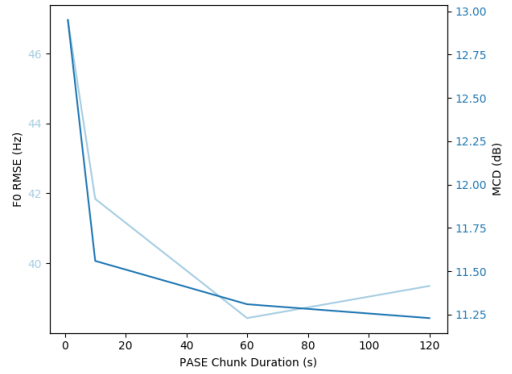
_____

Figure 4: *MCD and RMSE of $F0$ for new speakers with respect to the length of speech used for the embedding generation*

mal listenings indicate that PASE embedding improve slightly the quality and that systems are significantly better if F0 is included as input. The system proposed allows to generate speech with the identity of unseen speakers providing just the PASE embedding. As a general rule, 10 s are required to capture the identity. On the other hand, for the quality, it benefits from embeddings computed using longer speech segments (10-60 s).

6. Conclusions

In this work we have proposed the use of problem-agnostic speech embeddings as an acoustically meaningful representation of speaker identities for a multi-speaker SampleRNN based acoustic model for TTS. We have concluded that the use of these new embeddings help the network to converge quicker and obtain better objective results in terms of likelihood and spectral distortions. Additionally, we also perform a first step towards speaker adaptation by inferring new identity embeddings for speakers unseen during training from some seed speech without any fine-tuning involved. Future lines of research further developing the contents presented here include training the models for longer periods and to fine-tune the PASE encoder simultaneously to the TTS training such that VCTK and CMU recording conditions can be captured along with the TTS learning.

7. Acknowledgements

8. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *Proceedings of ISCA Workshop on Speech Synthesis*, 2007.

[2] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[3] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[4] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[5] S. Pascual de la Puente, "Deep learning applied to speech synthesis," Master's thesis, Universitat Politècnica de Catalunya, 2016.

[6] S. Pascual, A. Bonafonte, and J. Serrà, "Self-attention linguistic-acoustic decoder," in *Proceedings of IberSPEECH*, 2018.

[7] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[8] C. Valentini-Botinhao, Z. Wu, and S. King, "Towards minimum perceptual error training for dnn-based speech synthesis," in *Proceedings of International Speech Communication Association Conference (Interspeech)*, 2015.

[9] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[10] B. Uria, I. Murray, S. Renals, C. Valentini-Botinhao, and J. Bridle, "Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[11] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016.

[13] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of International Speech Communication Association Conference (Interspeech)*, 2017.

[15] C. Zhou, M. Horgan, V. Kumar, C. Vasco, and D. Darcy, "Voice conversion with conditional SampleRNN," in *Proceedings of International Speech Communication Association Conference (Interspeech)*, 2018.

[16] S. Pascual and A. Bonafonte, "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2016.

[17] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[18] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[19] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proceedings of International Speech Communication Association Conference (Interspeech)*, 2015.

[20] Y. Chen, Y. M. Assael, B. Shillingford, D. Budden, S. E. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, Ç. Gülçehre, A. van den Oord, O. Vinyals, and N. de Freitas, "Sample efficient adaptive text-to-speech," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[21] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[22] O. Barbany, A. Bonafonte, and S. Pascual, "Multi-Speaker Neural Vocoder," in *Proceedings of IberSPEECH*, 2018.

[23] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *Proceedings of International Speech Communication Association Conference (Interspeech)*, 2019.

[24] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2012.

[25] J. Kominek, A. W. Black, and V. Ver, "CMU ARCTIC databases for speech synthesis," 2003.

[26] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2016.

[27] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2013.

[28] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PacRim)*, 1993.