INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

UnitNet-Based Hybrid Speech Synthesis

Xiao Zhou, Zhen-Hua Ling, Li-Rong Dai

NELSLIP, University of Science and Technology of China, Hefei, P.R.China
xiaozh@mail.ustc.edu.cn, {zhling,lrdai}@ustc.edu.cn

Abstract

This paper presents a hybrid speech synthesis method based on UnitNet, a unified sequence-to-sequence (Seq2Seq) acoustic model for both statistical parametric speech synthesis (SPSS) and concatenative speech synthesis (CSS). This method combines CSS and SPSS approaches to synthesize different segments in an utterance. Comparing with the Tacotron2 model for Seq2Seq speech synthesis, UnitNet utilizes the phone boundaries of training data and its decoder contains autoregressive structures at both phone and frame levels. This hierarchical architecture can not only extract embedding vectors for representing phone-sized units in the corpus but also measure the dependency among consecutive units, which makes UnitNet capable of guiding the selection of phone-sized units for CSS. Furthermore, hybrid synthesis can be achieved by integrating the units generated by SPSS into the framework of CSS for the target phones without appropriate candidates in the corpus. Experimental results show that UnitNet can achieve comparable naturalness with Tacotron2 for SPSS and outperform our previous Tacotron-based method for CSS. Besides, the naturalness and inference efficiency of SPSS can be further improved through hybrid synthesis.

Index Terms: speech synthesis, text-to-speech, unit selection, sequence-to-sequence

1. Introduction

Nowadays, there are two main approaches to text-to-speech (TTS), which are statistical parametric speech synthesis (SPSS) [1] and concatenative speech synthesis (CSS) [2]. SPSS utilizes an acoustic model to represent the relationship between linguistic and acoustic features [3], together with a vocoder to render speech waveforms given predicted acoustic features. Recently, neural sequence-to-sequence (Seq2Seq) acoustic models, such as Tacotron [4, 5], and neural vocoders, such as WaveNet [6] have been proposed. However, it is still a challenge for SPSS to develop low computational complexity and high quality neural vocoders [7, 8].

The basic idea of CSS is to select a series of candidate units from a prerecorded speech corpus and then concatenate the waveforms of them to produce synthesized speech. After 2016, deep neural networks (DNNs) and recurrent neural networks (RNNs) have been utilized to guide unit selection by either predicting target acoustic features or deriving unit embeddings [9–14]. Our previous work proposed a Tacotron-based CSS method [14], which utilized the outputs of Tacotron2 [5] encoder as unit embeddings and achieved better performance than learning unit embeddings from acoustic features using a DNN [13] and a hidden Markov model (HMM) based baseline [15]. One disadvantage of this method is that the Tacotron model is not specifically designed for CSS. Thus, HMMs are still necessary for calculating some cost components and a separate neural network is required for deriving the concatenation cost function. Besides, there still exist glitches in synthetic speech due to the lack of appropriate candidates for some target phones.

In this paper, we present a unified Seq2Seq acoustic model named *UnitNet* which can support SPSS, CSS and hybrid speech synthesis. Comparing with the frame-level autoregressive structure in the decoder of Tacotron2, the decoder of UnitNet contains autoregressive structures at both phone and frame levels. The frame-level auto-regression predicts frame-sized mel-spectrograms for SPSS. The phone-level auto-regression utilizes the phone boundaries of training data and tracks the long-term dependencies among consecutive phone units, which achieves the calculation of target and concatenation costs for CSS. No HMM models are used for calculating costs, which significantly simplifies its composition. More details of UnitNet can be found here [16]. For hybrid synthesis, the units generated by SPSS are integrated into the framework of CSS to compensate the lack of appropriate candidates in the corpus for some target phones. Experimental results show that this hybrid approach can obtain better naturalness than pure SPSS or CSS, and improve the inference speed of SPSS.

2. Proposed Methods

2.1. UnitNet

UnitNet is composed of an encoder and a decoder as shown in Fig. 1. Its encoder is similar to the one of Tacotron2. Its decoder is composed of three components, in which the phone boundaries of training data are utilized and the attention-based alignment in the decoder of Tacotron2 is discarded.

1) ***Encoder*** The linguistic representations of phone sequences pass through a stack of 3 convolutional layers and a single BiLSTM layer. The two directional hidden state vectors $\overrightarrow{\boldsymbol{h}_n^c}$ and $\overleftarrow{\boldsymbol{h}_n^c}$ of the BiLSTM are concatenated to obtain a fixed-length *context unit embedding* $\boldsymbol{h}_n^c$ for the n-th phone. For assigning a constant $\boldsymbol{h}_n^c$ to each phone unit in the corpus, the dropout [17] operation in the encoder of Tacotron2 is not used.

2) ***Phone-level representation*** This component summarizes the frame-sized acoustic features within a phone to obtain the acoustic unit embedding for each phone-sized unit. The mel-spectrogram of previous frame is first processed by a DNN composed of two fully connected layers with ReLU activation and dropout, named PreNet. Then, a frame-level LSTM is adopted to model temporal dependencies. In order to ignore the influence of neighboring phones, we reset LSTM states at phone boundaries. Finally, generalized pooling [18] is adopted to transform the LSTM outputs to a fixed-length *acoustic unit embedding* $\boldsymbol{h}_n^a$ for the n-th phone in the speech. In order to unify $\boldsymbol{h}_n^a$ and $\boldsymbol{h}_n^c$, an additive attention module [19] is adopted. Here, the query is $\boldsymbol{h}_n^a$ and the keys are $\{\boldsymbol{h}_1^c, ..., \boldsymbol{h}_N^c\}$. We minimize the cross-entropy between the attention weights and one-hot true labels at the training stage.

3) ***Phone-level prediction*** Since the states of the frame-level LSTM in the phone-level representation component are

http://dx.doi.org/10.21437/Interspeech.2021-1092

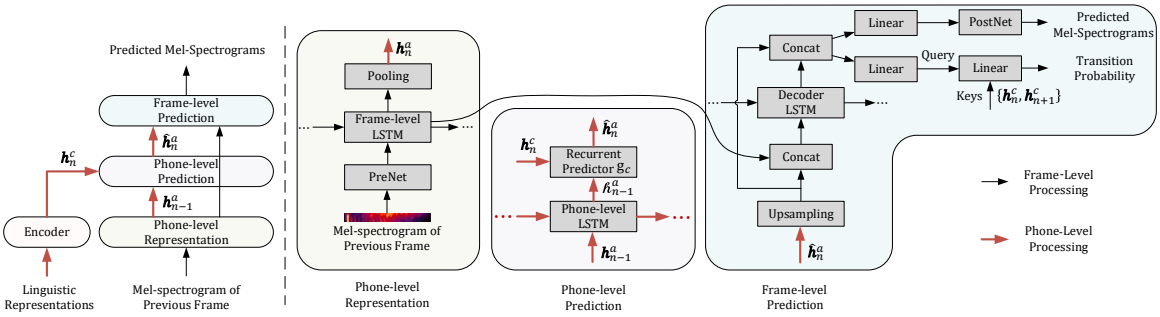

Figure 1: *The architecture of our proposed UnitNet model. The model consists of an encoder and a decoder. The decoder contains three components, phone-level representation, phone-level prediction and frame-level prediction.*

truncated at phone boundaries, the information of previous phones cannot be utilized when decoding the current phone. Besides, the phone-level representation component cannot output the acoustic unit embedding of a phone before all its frames have been decoded. In order to address these issues, the phone-level prediction component adopts a phone-level autoregressive structure to predict the acoustic unit embedding and to describe the dependencies among consecutive phones. This component consists of a phone-level LSTM and a recurrent predictor g_c. The phone-level LSTM converts the acoustic unit embedding vectors of preceding phones $\{\boldsymbol{h}_1^a, \boldsymbol{h}_2^a, \cdots, \boldsymbol{h}_{n-1}^a\}$ into an acoustic history vector $\boldsymbol{\hbar}_{n-1}^a$, i.e., $\boldsymbol{\hbar}_{n-1}^a = lstm\left(\boldsymbol{h}_1^a, \cdots, \boldsymbol{h}_{n-1}^a\right)$, where $\boldsymbol{\hbar}_0^a$ is set as a zero vector. The recurrent predictor g_c is a DNN, which concatenates $\boldsymbol{\hbar}_{n-1}^a$ and $\boldsymbol{h}_n^c$ as input to predict the acoustic unit embeddings $\hat{\boldsymbol{h}}_n^a$ of current phone, i.e., $\hat{\boldsymbol{h}}_n^a = \mathrm{g}_c\left(\boldsymbol{\hbar}_{n-1}^a, \boldsymbol{h}_n^c\right)$. At the training stage, we minimize the mean squared error (MSE) between $\hat{\boldsymbol{h}}_n^a$ and the reference one $\boldsymbol{h}_n^a$ to realize phone-level auto-regression. Finally, $\hat{\boldsymbol{h}}_n^a$ is upsampled repetitively to frame-level ones and then sent into the next frame-level prediction component.

4) ***Frame-level prediction*** This component predicts frame-level acoustic features via a decoder LSTM. Its input includes the phone-level $\hat{\boldsymbol{h}}_n^a$ and the frame-level hidden states of the frame-level LSTM. Similar to Tacotron2, a linear projection is used to predict current mel-spectrogram, then a PostNet with 5 convolutional layers is applied for refining it. We minimize the summed MSEs of the mel-spectrograms before and after the PostNet to aid convergence at the training stage. The decoder also predicts a phone-level transition probability p_{tra} at each frame, which describes how likely that current frame is the last frame of a phone. As shown in Fig. 1, an attention module [20] is used to calculate p_{tra}. Here, the query is the linear transformation of the concatenation of upsampled $\hat{\boldsymbol{h}}_n^a$ and current hidden state of the decoder LSTM. The keys for attention are $\boldsymbol{h}_n^c$ and $\boldsymbol{h}_{n+1}^c$. An end-of-sentence embedding vector is learnt to replace $\boldsymbol{h}_{n+1}^c$ when n is index of the last phone. The purpose of this attention is to utilize the weight of the key $\boldsymbol{h}_{n+1}^c$ as transition probability. We minimize the cross-entropy between the predicted transition probabilities and the true ones determined by phone boundaries at the training stage.

2.2. UnitNet-based SPSS

The inference process of applying UnitNet to SPSS is just like other Seq2Seq acoustic models, e.g., Tacotron2. The linguistic representations of phone sequences are sent into the model to predict mel-spectrograms directly without relying on other models, such as HMMs and duration predictors. The difference is that there is no attention-based alignment in the decoding process of UnitNet. As introduced in Section 2.1, the prediction of transition probabilities is an implicit phone duration model. Once transition probability exceeds a threshold of 0.5, the decoder resets the frame-level LSTM states in the phone-level representation component and starts to decode the next phone. Thus, a "hard alignment" between input phone sequences and output acoustic feature sequences can be obtained, which meets the strict criteria of the TTS attention mechanism [21].

2.3. UnitNet-based CSS

Given a trained UnitNet model, the context unit embedding and the acoustic unit embedding of each phone unit in the corpus can be determined. The context unit embeddings of corpus units are obtained by sending the linguistic representations of corpus texts into the encoder. The acoustic unit embeddings are derived in a teacher-forcing way, i.e., the true mel-spectrogram history and true phone boundaries are used in the decoder.

In the synthesis stage, the context and acoustic unit embeddings of each target phone to be synthesized are obtained by sending input text into the UnitNet model. Assume the test sentence to be synthesized is composed of N phone units, and $\boldsymbol{C}^c = \{\boldsymbol{t}_1^c, \boldsymbol{t}_2^c, \cdots, \boldsymbol{t}_N^c\}$ and $\boldsymbol{C}^a = \{\boldsymbol{t}_1^a, \boldsymbol{t}_2^a, \cdots, \boldsymbol{t}_N^a\}$ represent their context unit embeddings and acoustic unit embeddings respectively. Let $\boldsymbol{U} = \{u_1, u_2, \cdots, u_N\}$ denote a sequence of phone-sized candidate units to synthesize this sentence, with $\boldsymbol{d}_n^c$ and $\boldsymbol{d}_n^a$ denoting the context and acoustic unit embeddings of u_n. The target cost function for unit selection is defined as $D_{\text{targ}}\left(u_n, \boldsymbol{c}_n\right) = \frac{1}{2}\left(\sqrt{\|\boldsymbol{d}_n^c - \boldsymbol{t}_n^c\|^2} + \sqrt{\|\boldsymbol{d}_n^a - \boldsymbol{t}_n^a\|^2}\right)$, which utilizes both the context representations and the acoustic representations of phone units.

In the phone-level prediction component, the phone-level LSTM keeps track of the acoustic unit embeddings of preceding candidate units $\{\boldsymbol{d}_1^a, \boldsymbol{d}_2^a, \cdots, \boldsymbol{d}_{n-1}^a\}$ and converts them into an acoustic history vector $\boldsymbol{\hbar}_{n-1}^a$, i.e., $\boldsymbol{\hbar}_{n-1}^a = lstm\left(\boldsymbol{d}_1^a, \cdots, \boldsymbol{d}_{n-1}^a\right)$. Then, the predicted acoustic unit embedding $\hat{\boldsymbol{d}}_n^a$ of the n-th phone is $\hat{\boldsymbol{d}}_n^a = \mathrm{g}_c\left(\boldsymbol{\hbar}_{n-1}^a, \boldsymbol{t}_n^c\right)$, Thus, we can defined $D_{\text{con}}\left(u_1, \cdots, u_n, \boldsymbol{c}_n\right) = \sqrt{\|\boldsymbol{d}_n^a - \hat{\boldsymbol{d}}_n^a\|^2}$ as a concatenation cost function. Since the phone-level LSTM and the recurrent predictor g_c are estimated simultaneously with other model parameters, UnitNet combines learning and modeling unit embeddings in a single model, which helps to find more appropriate unit representations.

In order to boost the calculation efficiency of unit selection, the top K candidate units with the lowest D_{targ} are pre-selected for each target phone. A pruning search strategy originally

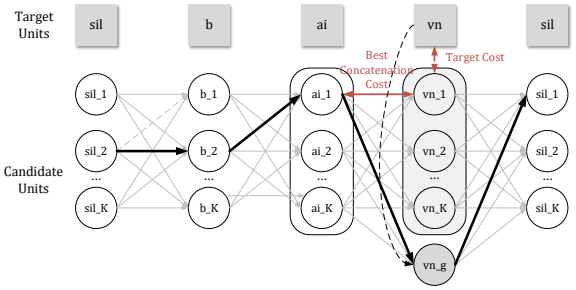

Figure 2: *An example of the DP search in hybrid speech synthesis. The text is "白云" (white cloud) with a target phone sequence "sil-b-ai-vn-sil". The thick line represent the optimal unit sequences obtained by DP search, where "vn_1" to "vn_K" are corpus units and "vn_g" is the unit generated by SPSS.*

designed for frame-sized unit selection [22] is also applied to reduce the complexity of dynamic programming (DP) search to $O\left(NK^2\right)$. Finally, the optimal unit sequence is determined by DP search and their waveforms are concatenated to produce the synthetic speech.

2.4. UnitNet-based hybrid speech synthesis

In hybrid speech synthesis, the phone-sized units generated by SPSS are integrated into the framework of CSS to compensate the lack of appropriate corpus candidates for some target phones. During the unit selection procedure of CSS, a local cost for each candidate is calculated, which is the sum of its target cost and its best concatenation cost. If the lowest local cost among all candidates for a target phone is above a threshold, a unit generated by SPSS is added to the candidate set of this target phone for DP search.

Specifically, the acoustic features of the generated unit are just the acoustic features of this target phone for SPSS, which have been predicted when deriving the acoustic unit embeddings of target phones for CSS. The waveforms of this generated unit are further reconstructed using the SPSS vocoder. Then, this unit is treated as an additional candidate of the target phone and participates the DP search with other candidates. Its context and acoustic unit embeddings are copied from those of the target phone.

An example of the DP search in hybrid speech synthesis is shown in Fig. 2. In our implementation, only voiced target phones are considered for adding generated candidates. In this example, the lowest local cost of the target phone "vn" is above the threshold and a generated unit "vn_g" is added to its candidate set for DP search.

3. Experiments

3.1. Experimental setup

A Chinese corpus pronounced by a female speaker was used in our experiments. The scripts were selected from newspapers, and the recordings were sampled at 16kHz with 16 bits resolution. The total 12,319 utterances ($\approx$ 17.51 hours) were split into a training set of 11,608 utterances, a validation set of 611 utterances, and a test set of 100 sentences. The training set was used to train acoustic models, the validation set was used to tune hyperparameters. The training set and the validation set were merged to provide candidates for unit selection, and the total number of phone instances was 459,750.

When training Seq2Seq acoustic models, 80-band mel-spectrograms were used as acoustic features. The frame length was 64 ms and the frame shift was 15 ms. Phone sequences composed of the initials and finals in Chinese were adopted as model input. A phoneme embedding vector, a tone embedding vector and a prosodic position embedding vector were concatenated to represent each phone.

Finally, Tacotron-based SPSS and CSS systems, UnitNet-based SPSS, CSS and hybrid systems were built for comparison in our experiments.[1]

Tacotron2_CSS The *Prop_All* system in our previous work on Tacotron-based CSS [14] was adopted. A publicly available implementation of Tacotron2[2] was used. The dimension of unit embeddings was 256 and K was set as 25 for unit pre-selection, same as our previous work [14].

Tacotron2_SPSS The Tacotron2 model used here was the same as the one used by Tacotron2_CSS. A publicly available implementation of WaveNet[3] was adopted to build its vocoder.

UnitNet_SPSS The UnitNet model was built following the introduction in Section 2.1. The phone boundaries used at the training stage were obtained by HMM-based forced alignment. The dimensions of context and acoustic unit embeddings were both set as 256. Either the WaveNet vocoder in Tacotron2_SPSS or a Parallel WaveGAN vocoder [8] with a publicly available implementation[4] was adopted to reconstruct waveforms.

UnitNet_CSS The UnitNet model in UnitNet_SPSS was employed to build UnitNet_CSS following Section 2.3. K was set as 25 for unit pre-selection. The final concatenation cost function used for DP search was the weighted sum of D_{con} and the negative log-likelihoods of the differential mel-spectrograms at phone boundaries [14]. The weights of different costs were tuned manually by informal listening tests on the validations set.

UnitNet_Hybrid It was build based on UnitNet_CSS. The threshold of the lowest local cost in Section 2.4 was empirically set as 4. In the synthetic speech of test set, about 3% target phones were synthesized by generated units, excluding silences and short pauses. The waveforms of generated units were rendered by the Parallel WaveGAN vocoder used by UnitNet_SPSS.

3.2. Experimental results

To compare UnitNet with Tacotron2, an AB preference test was first conducted between Tacotron2_CSS and UnitNet_CSS. 20 sentences in the test set and were synthesized by these two systems to form pairs in random order. 12 Chinese native listeners participated the test and were asked to judge which sentence in each pair sounded more natural. The results are shown in the first row of Table 1, which shows the superiority of UnitNet over Tacotron2 for CSS.

Besides, a mean opinion score (MOS) test was conducted to measure the naturalness of UnitNet_CSS, UnitNet_SPSS, Tacotron2_SPSS, together with natural recordings. The two

[1] Audio samples are available at `https://xiaozhah.github.io/UnitNet-Hybrid-TTS_demos`.

[2] `https://github.com/NVIDIA/tacotron2`

[3] `https://github.com/r9y9/wavenet_vocoder`

[4] `https://github.com/kan-bayashi/ParallelWaveGAN`

Table 1: *Subjective preference scores (%) among the four systems, where N/P denotes "No Preference" and p means the p-value of t-test between two systems.*

Tacotron2_CSS	UnitNet_CSS	UnitNet_SPSS*	UnitNet_Hybrid	N/P	p
23.75	**55.00**	-	-	21.25	< 0.001
-	25.83	-	**61.25**	12.92	< 0.001
-	-	30.00	**45.00**	25.00	0.007

Table 2: *Real-time factors (RTFs) of different systems with 95% confidence intervals. Here, RTF indicates the time required to generate waveforms of 1 second.*

	Acoustic Model	Vocoder (Parallel WaveGAN)	Unit Pre-selection	Unit Selection	Waveform Concatenation	Total
UnitNet_Hybrid	0.134 ± 0.003	0.021 ± 0.004	0.024 ± 0.001	0.089 ± 0.002	0.216 ± 0.003	0.484 ± 0.006
UnitNet_CSS	0.134 ± 0.003	-	0.024 ± 0.001	0.087 ± 0.002	0.214 ± 0.002	0.459 ± 0.004
UnitNet_SPSS*	0.134 ± 0.003	1.223 ± 0.069	-	-	-	1.360 ± 0.070

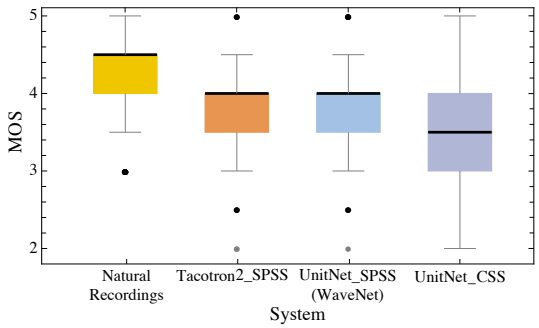

Figure 3: *The boxplot of MOS evaluation results on the naturalness of several CSS and SPSS systems together with natural recordings.*

SPSS systems adopted the same WaveNet vocoder. 30 test sentences synthesized by the three systems were evaluated by 14 native Chinese listeners. The listeners were asked to rate each synthetic utterance using a score from 1 (very unnatural) to 5 (very natural) with an interval of 0.5. Fig. 3 shows the boxplot of the MOS evaluation results. It should be noticed that the median score of natural recordings was 4.5 due to the score interval of 0.5. We adopted the Bonferoni-corrected pairwise Wilcoxon signed rank test [23] to determine whether the MOS difference between two systems was significant. In Fig. 3, UnitNet_SPSS (MOS=3.80) achieved similar naturalness score to Tacotron2_SPSS (MOS=3.83), and the difference between these two systems was insignificant (p=0.435). However, the naturalness of UnitNet_CSS (MOS=3.56) was not as good as UnitNet_SPSS and Tacotron2_SPSS ($p<0.001$).

Furthermore, in order to evaluate the performance of UnitNet-based hybrid synthesis, two AB preference tests were conducted among UnitNet_CSS, UnitNet_SPSS and UnitNet_Hybrid. Here, the UnitNet_SPSS system adopted the Parallel WaveGAN vocoder to achieve comparable inference efficiency with other two systems. The test configurations were the same as the one between Tacotron2_CSS and UnitNet_CSS. The results are shown in the last two rows of Table 1. We can also see that UnitNet_Hybrid achieved significantly better naturalness than both UnitNet-based CSS and SPSS systems. The reasons can be attributed to that hybrid synthesis reduced the glitches of CSS and improved the audio quality of SPSS by using natural waveforms for most target phones.

The inference efficiencies of the three UnitNet-based systems in Table 1 were also evaluated. The UnitNet_SPSS adopted the Parallel WaveGAN vocoder. Real-time factor (RTF), which indicates the time required to generate waveforms of 1 second, was adopted as the metric. The evaluation was conducted on a Linux server with two Intel Xeon CPUs of 20 cores using test sentences. The results of total RTFs and the RTFs of system components are shown in Table 2. Because no neural vocoders were used, the RTF of UnitNet_CSS was only one third of that of UnitNet_SPSS. The RTF of UnitNet_Hybrid was slightly higher than that of UnitNet_CSS but still much lower than UnitNet_SPSS because only about 3% target phones were rendered by the neural vocoder in the output of UnitNet_Hybrid. It should be noted that the inference efficiencies of all three systems can be further improved, such as employing more efficient neural vocoders and optimizing the algorithm of waveform concatenation. This is worth further investigation in the future.

4. Conclusions

This paper has presented a unified sequence-to-sequence acoustic model, named UnitNet, for parametric, concatenative and hybrid speech synthesis. Its decoder contains auto-regressive structures at both phone-level and frame-level, and utilizes a transition probability prediction instead of the attention module in Tacotron2 to learn the mapping relationship between phone sequences and acoustic feature sequences. The UnitNet model is capable of jointly learning and modeling unit embeddings for deriving the cost functions in unit selection. Evaluation results demonstrated the effectiveness of our proposed model, which outperformed Tacotron2 for CSS and achieved comparable naturalness with Tacotron2 for SPSS. Furthermore, the UnitNet-based hybrid synthesis system obtained better naturalness than pure SPSS or CSS with high inference efficiency. To achieve a further unification of SPSS and CSS methods under the framework of UnitNet will be a task of our future work.

5. Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303001, and in part by the National Nature Science Foundation of China under Grant 61871358.

6. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 373–376.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[4] 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 *Proc. Interspeech 2017*, 2017, pp. 4006–4010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[6] A. V. Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop (SSW9)*, 2016, pp. 125–125.

[7] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[8] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[9] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5145–5149.

[10] V. Wan, Y. Agiomyrgiannakis, H. Silen, and J. Vít, "Google's Next-Generation Real-Time Unit-Selection Synthesizer Using Sequence-to-Sequence LSTM-Based Autoencoders," in *Proc. Interspeech 2017*, 2017, pp. 1143–1147. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1107

[11] T. Capes, P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston, M. Hunt, J. Li, M. Neeracher, K. Prahallad, T. Raitio, R. Rasipuram, G. Townsend, B. Williamson, D. Winarsky, Z. Wu, and H. Zhang, "Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System," in *Proc. Interspeech 2017*, 2017, pp. 4011–4015. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1798

[12] V. Pollet, E. Zovato, S. Irhimeh, and P. Batzu, "Unit Selection with Hierarchical Cascaded Long Short Term Memory Bidirectional Recurrent Neural Nets," in *Proc. Interspeech 2017*, 2017, pp. 3966–3970. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-428

[13] X. Zhou, Z.-H. Ling, Z.-P. Zhou, and L.-R. Dai, "Learning and modeling unit embeddings for improving hmm-based unit selection speech synthesis," in *Proc. Interspeech 2018*, 2018, pp. 2509–2513. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1198

[14] ——, "Extracting unit embeddings using sequence-to-sequence acoustic models for unit selection speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*. IEEE, 2020, pp. 1–5.

[15] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1245.

[16] X. Zhou, Z. Ling, and L.-R. Dai, "UnitNet: A Sequence-to-Sequence Acoustic Model for Concatenative Speech Synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3093823

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] Q. Chen, Z.-H. Ling, and X. Zhu, "Enhancing Sentence Embedding with Generalized Pooling," in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe, USA: ACL, August 2018.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[21] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proc. Interspeech 2019*, 2019, pp. 1293–1297. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1972

[22] Z.-H. Ling and Z.-P. Zhou, "Unit selection speech synthesis using frame-sized speech segments and neural network based acoustic models," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 1053–1062, 2018.

[23] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," *Proc. BLZ3-2007 (in Proc. SSW6)*, 2007.