



# Learning and Modeling Unit Embeddings for Improving HMM-based Unit Selection Speech Synthesis

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

National Engineering Laboratory for Speech and Language Information Processing,  
University of Science and Technology of China, Hefei, P.R. China

xiaozh@mail.ustc.edu.cn, zhling@ustc.edu.cn

## Abstract

This paper presents a method of learning and modeling unit embeddings using deep neural networks (DNNs) to improve the performance of HMM-based unit selection speech synthesis. First, a DNN with an embedding layer is built to learn a fixed-length embedding vector for each phone-sized candidate unit in the corpus from scratch. Then, another two DNNs are constructed to map linguistic features toward the extracted unit vector of each phone. One of them employs the unit vectors of preceding phones as model input. At synthesis time, the  $L_2$  distances between the unit vectors predicted by these two DNNs and the ones derived from candidate units are integrated into the target cost and the concatenation cost of HMM-based unit selection speech synthesis respectively. Experimental results demonstrate that the unit vectors estimated using only acoustic features display phone-dependent clustering properties. Furthermore, integrating unit vector distances into cost functions, especially the concatenation cost, improves the naturalness of HMM-based unit selection speech synthesis in our experiments.

**Index Terms:** speech synthesis, hidden Markov model, deep neural networks, unit selection, unit embedding

## 1. Introduction

Speech synthesis aims to make machines speak like human beings full of emotions and it benefits a lot of voice interactive applications, such as intelligent personal assistants and robots. Nowadays, there are two mainstream speech synthesis approaches, which are unit selection and waveform concatenation [1] and statistical parametric speech synthesis (SPSS) [2]. In SPSS, statistical acoustic models are built to predict acoustic parameters from input texts [3]. Then, the predicted acoustic features are sent into a vocoder to reconstruct speech waveforms. Although this approach is able to produce smooth sound flexibly, the quality of synthetic speech is always degraded due to the inaccuracy of acoustic parameter prediction and the usage of vocoders [2].

Unit selection and waveform concatenation is another approach of speech synthesis which became popular before SPSS [1]. The basic idea of this approach is to select a sequence of candidate units from a prerecorded speech corpus and then concatenate the waveforms of the selected units to produce the synthetic speech. The basic units for selection could be syllables, phones or frames. Two cost functions, i.e., target cost and concatenation cost [4] are usually adopted to measure the appropriateness of the selected unit sequence. The target

cost describes the difference between a candidate unit in the corpus and a target unit to be synthesized. The concatenation cost measures the continuity between two consecutive candidate units. Based on these two cost functions, dynamic programming (DP) search [5] is employed to determine the optimal sequence of candidate units. Due to the using of natural speech segments for synthesis, the unit selection and waveform concatenation approach can usually achieve better naturalness of synthetic speech than SPSS if a corpus with sufficient and high quality recordings is available [2].

In the last decade, a hidden Markov model (HMM) based unit selection method has been proposed [6, 7, 8, 9]. This method applies the statistical acoustic modeling techniques developed in HMM-based SPSS [10] to unit selection. It derives the target and concatenation cost functions from statistical acoustic models. Thus, it is also named a “hybrid” approach for speech synthesis [2]. The speech synthesis systems developed based on this method achieved good performance in Blizzard Challenge evaluations of recent years [11, 12]. On the other hand, deep learning models, such as deep neural networks (DNNs) and recurrent neural networks (RNNs) have been successfully introduced into SPSS to replace HMMs for improving the accuracy of acoustic modeling [13, 14, 15]. Accordingly, DNNs and RNNs have also been utilized to guide the unit selection in “hybrid” speech synthesis by either predicting target acoustic features or deriving context embeddings [11, 16]. The DNNs and RNNs used in these methods are similar to the ones used in SPSS, which predict the acoustic parameters of each frame from the input context descriptions.

This paper proposes to utilize DNNs for unit selection in a way different from these methods. It first derives unit embeddings using a DNN to represent the acoustic characteristics of phone-sized candidate units with fixed-length vectors. Then, phone-level DNN acoustic models instead of traditional frame-level ones are built to predict the embedding vectors and are integrated into unit selection criterion. Compared with frame-level DNNs designed for SPSS, phone-level models can better capture the dependencies among consecutive candidate units and are expected to be more appropriate for the unit selection task. The idea of deriving unit embeddings for unit selection speech synthesis has been investigated very recently [17, 18]. In these methods, the unit embeddings are learnt either by combining a context-to-acoustic regression predictor and an acoustic-to-acoustic autoencoder [17] or by deriving bottleneck features from context inputs [18].

This paper estimates unit embeddings only using acoustic features which makes it easy for implementation and also studies the effects of using unit embeddings for target cost and concatenation cost calculation specifically by experiments.

The paper is organized as follows. Section 2 briefly review the HMM-based unit selection method. Section 3 introduces our proposed methods. Sections 4 and 5 are the experiments

Zhi-Ping Zhou is now with Baidu Speech Department, Baidu Technology Park, Beijing, 100193, China. This work was partially supported by the National Natural Science Foundation of China under Grants U1613211 and the Key Science and Technology Project of Anhui Province under Grant No. 17030901005.

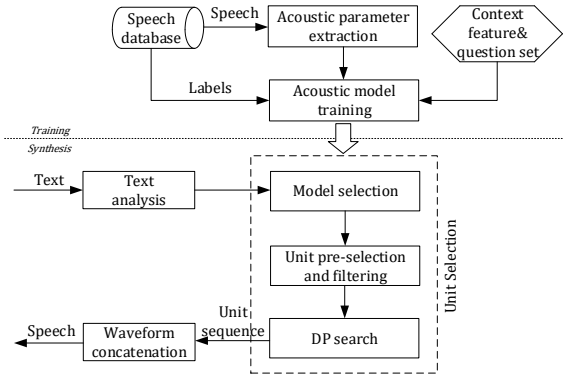


Figure 1: Flowchart of conventional HMM-based unit selection and conclusions respectively.

## 2. HMM-Based Unit Selection

### 2.1. Selecting phone-sized speech segments using HMMs

The HMM-based unit selection method using phones as basic units [7] is adopted to build the baseline system in this paper. The flowchart of this method is shown in Fig. 1. It includes two stages, training and synthesis. At the training stage,  $M$  kinds of acoustic features, which are used to measure the performance of unit selection synthesis systems, are first designed based on some prior knowledge. Given a speech corpus with context annotations, these features are extracted to train context-dependent models  $\{\Lambda_1, \Lambda_2, \dots, \Lambda_M\}$ . These models could be HMMs or just Gaussian distributions according to the characteristics of different acoustic features. At the synthesis stage, the context descriptions  $C$  of an input sentence are first derived by text analysis. Let  $\mathbf{U} = \{u_1, u_2, \dots, u_N\}$  denotes a sequence of phone-sized candidate units to synthesis the input sentences. The optimal unit sequence  $\mathbf{U}^*$  is determined as

$$\mathbf{U}^* = \arg \max_{\mathbf{U}} \sum_{m=1}^M \omega_m [\ln (P_{\Lambda_m}(X(\mathbf{U}, m)|C)) - \omega_{\text{KLD}} D_{\Lambda_m}(C(\mathbf{U}), C)], \quad (1)$$

where  $X(\mathbf{U}, m)$  represents the  $m$ -th kind of acoustic features extracted from the unit sequence  $\mathbf{U}$ ,  $\ln (P_{\Lambda_m}(\mathbf{X}|C))$  denotes the log likelihood of observing  $\mathbf{X}$  given model  $\Lambda_m$  and context information  $C$ ,  $C(\mathbf{U})$  represent the context descriptions of the candidate unit sequence  $\mathbf{U}$ ,  $D_{\Lambda_m}(C(\mathbf{U}), C)$  calculates the Kullback-Leibler divergence (KLD) [7] between two distributions in the model set  $\Lambda_m$  with context information  $C(\mathbf{U})$  and  $C$ ,  $\omega_m$  and  $\omega_{\text{KLD}}$  are the weights of each model and the KLD component which are set manually. Eq.(1) can be rewritten into the conventional form of a sum of target costs and concatenation costs [9] and then DP search can be applied to find the optimal sequence of candidate units.

When building the baseline system in our experiments, acoustic features are selected (i.e.,  $M = 5$ ), which are

1. the frame-level mel-cepstral coefficients (MCCs) with dynamic components,
2. the frame-level F0s with dynamic components,
3. the phone durations,
4. the differentials of MCCs between the first frame of current phone and the last frame of previous phone [19],
5. the differentials of F0s between the first frame of current phone and the last frame of previous phone [19].

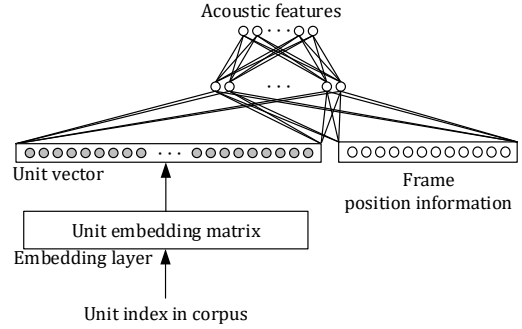


Figure 2: Flowchart of the Unit2Vec model for learning unit embeddings.

The first two kinds of acoustic features are modeled using the context-dependent HMMs for SPSS. The last three kinds of features are modeled by context-dependent Gaussian distributions with decision-tree based model clustering. Only the KLDs of the models corresponding to the first three kinds of features are used to calculate Eq.(1) in our implementation.

### 2.2. Unit pre-selection and unit filtering

In order to increase the computation efficiency of unit selection, unit pre-selection and unit filtering procedures are conducted before DP search. The KLD components in Eq.(1) describe the similarity between the context information of two units, and do not rely on their acoustic features. Therefore, they can be calculated efficiently at synthesis time by off-line computation [19]. In unit pre-selection, the  $K'$ -best candidate units with the minimum sum of KLDs are determined for each target phone.

The unit filtering procedure utilizes the distances between the acoustic features of the candidates and the acoustic features predicted by HMM-based SPSS. The top- $K$  candidate phone units in the unit pre-selection results with minimum acoustic distances toward the prediction results are finally determined and used in the DP search.

## 3. Proposed Methods

### 3.1. Learning unit embedding using a DNN

Inspired by the word embedding techniques developed for natural language processing, such as Word2Vec [20], a DNN named *Unit2Vec* is designed to learn a fixed-length vector for each phone unit in the corpus for unit selection from scratch. The flowchart of the Unit2Vec model is shown in Fig. 2. The dimension of the unit embedding matrix is  $R \times D$ , where  $R$  is the total number of candidates in the corpus and  $D$  is the length of the embedding vector for each candidate. All unit vectors are stored in the weight of the embedding layer<sup>1</sup> as an embedding matrix. Given a row index, we can extract corresponding unit vector from the matrix. The phone boundaries in the corpus are given by HMM-based force-alignment. For each frame in the corpus, a unit vector is determined by selecting the row index of the embedding matrix corresponding to the phone unit that the frame belongs to. Then the unit vector is concatenated with frame position information to predict the acoustic features (i.e., MCCs and F0s) of this frame. The unit embedding matrix is learnt by minimizing the mean square error (MSE) between the predicted and the natural acoustic features. Shuffling the

<sup>1</sup>The implementation of the embedding layer can be found at <https://mxnet.incubator.apache.org/api/python/gluon/nn.html#mxnet.gluon.nn.Embedding>

Table 1: The 12-dimension frame position information used in our implementation.

Dim.	Description
1-5	5-dim. one-hot vector indicating current state index
6	duration of current state
7	duration of current phone
8	forward position of the frame in current state
9	backward position of the frame in current state
10	forward position of the frame in current phone
11	backward position of the frame in current phone
12	the duration proportion of current state in phone

frame-level training data is necessary, which makes the data corresponding to a same unit randomly distribute in the training set and helps to improve the estimation of unit vectors. A 12-dimension vector is adopted to represent the frame position information in our implementation as shown in Table 1. The learnt unit vector of each phone unit is expected to describe the overall acoustic characteristics of the unit, which will be further modeled to derive the cost functions for unit selection.

### 3.2. Modeling unit vectors for HMM-based unit selection

Assume that the sentence to be synthesized is composed of  $N$  phones units and  $C = \{c_1, c_2, \dots, c_N\}$  represents the context descriptions given by text analysis. For a sequence of candidate units  $U = \{u_1, u_2, \dots, u_N\}$ , the unit vector and the context features corresponding to  $u_n$  are written as  $v_n$  and  $w_n$  respectively. Then, two functions for calculating target costs and concatenation costs are designed as

$$C_{\text{targ}}(u_n, c_n) = \|f_t(w_n) - v_n\|^2, \quad (2)$$

$$C_{\text{con}}(u_{n-T}, \dots, u_{n-1}, u_n, c_n) = \|f_c(w_n) - v_n\|^2, \quad (3)$$

where  $f_t$  and  $f_c$  are two phone-level DNN models which predict unit vectors from input features.

The  $f_t$  model maps the context features of each phone unit toward its unit vector directly, its structure is similar to the left part of Fig. 3, but has no bottleneck (BN) layer to squeeze the input. The parameters of  $f_t$  are trained using the context features and the learnt unit vectors of all units in the corpus under minimum MSE criterion. Thus, the  $C_{\text{targ}}$  function measures the overall acoustic difference between a candidate unit and a target unit, which can be used as a part of the target cost for unit selection.

The  $f_c$  model predicts the unit vector of current phone given the unit vectors of preceding phones and the context features of current target as shown in the right part of Fig. 3. Considering that the dimension of context features is usually much higher than that of unit vectors, another network is constructed to extract BN features from original context features as shown in the left part of Fig. 3. The parameters of  $f_c$  are trained using the corpus with context features and learnt unit vectors. At training stage, the unit vectors learned in Section 3.1 are adopted as the references for calculating the MSE losses of the two predicted unit vectors in Fig. 3. At synthesis stage, only the unit vectors predicted by right part of Fig. 3 are used as the output of the  $f_c$  model. The  $C_{\text{con}}$  function measures the appropriateness of concatenating  $u_n$  with  $\{u_{n-T}, \dots, u_{n-1}\}$ , which can be used as a part of the concatenation cost for unit selection. Comparing with the concatenation cost derived from Eq.(1),  $C_{\text{con}}$  is better at measuring the long-term dependencies among consecutive candidate units when increasing the history length  $T$ .

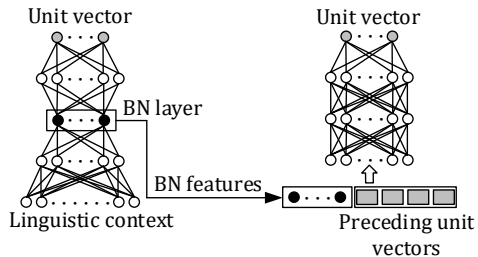


Figure 3: The DNN structure of modeling unit vectors for concatenation cost calculation.

When using  $C_{\text{con}}$  for concatenation cost calculation, the computation complexity of DP search becomes  $O(NK^{T+1})$ . Therefore, a pruning search strategy designed for frame-sized unit selection [21] is applied here to reduce the search complexity to  $O(NK^2)$  when  $T > 2$ .

## 4. Experiments

### 4.1. 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 12,219 utterances ( $\approx 17.25$  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  $f_t$  and  $f_c$  for cost function calculation. The validation set was used to evaluate the performance of predicting unit vectors using  $f_t$  and  $f_c$  in our experiments. The test set was adopted for the subjective evaluation on the naturalness of synthetic speech. Speech signals were analysed at 5ms frame shift by STRAIGHT [22] and 12-dimensional MCCs and a  $\log F_0$  were extracted at each frame. When building the baseline HMM-based unit selection system, 5-state left-to-right HMMs were estimated for each context-dependent phone<sup>2</sup>. The frame-level acoustic features for HMM modeling were composed of the extracted MCCs and  $\log F_0$  together with their delta and delta-delta coefficients.

The training set and the validation set were merged together to provide candidate units for unit selection and to train the Unit2Vec model. The total number of phone instances was  $R = 459,753$ . The acoustic features used in this model were the same as the ones used in HMMs except that continuous  $F_0$  trajectories were adopted and an extra dimension of binary voiced/unvoiced (U/V) flag was added. The dimension of unit vectors was set as  $D = 32$ . The Unit2Vec model had one hidden layer with 64 sigmoid units.

When training the  $f_t$  and  $f_c$  models, 523-dimension context features were used. The  $f_t$  model had three hidden layers and 128 ReLU units per layer. The BN feature extractor in Fig. 3 had three hidden layers and the BN layer was the second one. The BN layer had 64 hidden units and the other two hidden layers had 128 units. In the  $f_c$  mode, the history length was set as  $T = 4$ . The  $f_c$  model had three hidden layers with 256 ReLU units per layer. The dropout regularization with a probability of 0.1 was applied when training  $f_t$  and  $f_c$ . The MXNet C++ API<sup>3</sup> was adopted to calculate  $f_t$  and  $f_c$  at synthesis time.

Finally, two systems using our purposed methods were built to compare with the baseline HMM-based unit selection system

<sup>2</sup> The initials and finals of Chinese were treated as phones in our experiments for simplification.

<sup>3</sup>[https://mxnet.incubator.apache.org/doxygen/c\\_\\_predict\\_\\_api\\_8h.html](https://mxnet.incubator.apache.org/doxygen/c__predict__api_8h.html)

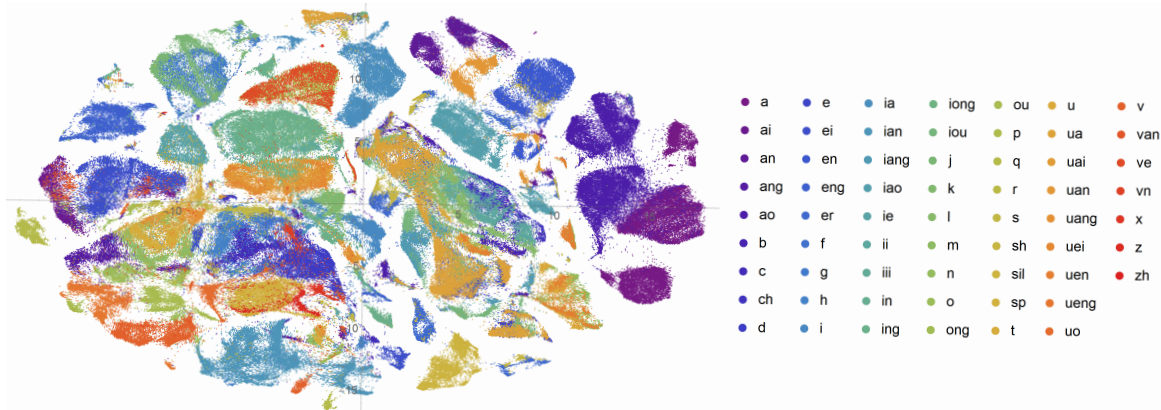


Figure 4: Visualization of the phone-dependent distributions of learnt unit vectors using t-SNE.

Table 2: Average reconstruction errors of learnt unit vectors, where **MCD**, **RMSE**, **CORR** and **UV** denote mel-cepstral distortion, F0 RMSE, F0 correlation and U/V error percentage.

MCD(dB)	RMSE(Hz)	CORR	UV(%)
2.1338	18.4240	0.9673	0.7049

shown in Fig. 1.

**Prop\_TC** When calculating target costs,  $C_{\text{targ}}$  was adopted to replace the  $\ln P_{\Lambda_m}(X|C)$  components in Eq.(1) which corresponded to frame-level MCCs and F0s;

**Prop\_All** Based on Prop\_TC,  $C_{\text{con}}$  was further added to the concatenation costs derived from Eq.(1).

These two systems shared the same unit pre-selection and filtering results as the baseline system where  $K' = 200$  and  $K = 100$ . The weights in Eq.(1) were tuned by informal listening for each system.

#### 4.2. Performance of the learning and modeling unit vectors

The visualization of the phone-dependent distributions of the learnt unit vectors is shown in Fig. 4 by t-SNE [23]. From this figure we can see that the learnt unit vectors displayed phone-dependent clustering properties although they were estimated using only acoustic features.

We further calculated the errors of reconstructing acoustic features from the learnt unit vectors on the training and validation sets. Natural frame position information was adopted here. The average reconstruction errors of different acoustic features are shown in Table 2. We can see that there still existed inaccuracy when representing the overall acoustic features of candidates using 32-dimension unit vectors.

We also evaluated the prediction accuracy of  $f_t$  and  $f_c$  by reconstructing acoustic features from the predicted unit vectors. This experiment was conducted on the validation set which was not used for training and natural frame position information was also adopted. The average prediction errors of different acoustic features are shown in Table 3. We can see that the prediction errors were acceptable considering the general performance of conventional SPSS systems. Besides, the prediction accuracy of  $f_c$  was better than  $f_t$  because the unit vectors of natural preceding units were adopted as history when evaluating  $f_c$ .

#### 4.3. Subjective evaluation

Thirty sentences were randomly selected from the test set and synthesized by the three systems. Three groups of ABX preference tests were conducted by 10 Chinese native listeners

Table 3: Average prediction errors of  $f_t$  and  $f_c$  on the validation set.

model	MCD(dB)	RMSE(Hz)	CORR	UV(%)
$f_t$	3.4267	38.2256	0.8524	5.2502
$f_c$	3.3449	35.1925	0.8746	4.9525

Table 4: Subjective preference scores (%) among the three systems, where N/P denotes "No Preference" and  $p$  means the  $p$ -value of t-Test between two systems.

Baseline	Prop_TC	Prop_All	N/P	$p$
40.00	31.67	-	28.33	0.0882
15.33	-	<b>55.67</b>	29.00	<0.001
-	15.00	<b>52.67</b>	32.33	<0.001

and each one was to make a comparison between two systems. The listeners were asked to judge which sentence in each pair sounded more natural. The results are summarized in Table 4<sup>4</sup>. We can see that there was no significant preference between the baseline system and the Prop\_TC system ( $p > 0.05$ ). According to the comments from the listeners, the Prop\_TC system made the synthetic speech sound more expressive than the baseline system while it introduced more glitches. Furthermore, the Prop\_All system performed significantly better than the baseline and the Prop\_TC systems. This could be attributed to the  $f_c$  model built using unit vectors which captured the long-term dependencies among consecutive candidate units.

## 5. Conclusions

In this paper, a method of learning and modeling unit vectors has been proposed to improve the performance of HMM-based unit selection speech synthesis. The DNN-based *Unit2Vec* model learns fixed-length unit vectors for candidate phone units. The unit vectors are modeled by another two DNNs to derive the functions for target cost and concatenation cost calculation. Subjective evaluation results demonstrate the effectiveness of our proposed methods and show that the DNN-based concatenation cost helped to handle the long-term dependencies among candidate units. To improve the reconstruction accuracy, find better phone-dependent distributions, add more features into unit vectors will be the tasks of our future work.

<sup>4</sup>The speech examples can be found at <http://home.ustc.edu.cn/~xiaozh/Interspeech2018/>.

## 6. References

- [1] 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.
- [2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [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] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *EUROSPEECH. International Speech Communication Association*, 1995, pp. 581–584.
- [5] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [6] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *Ninth International Conference on Spoken Language Processing*, 2006.
- [7] —, "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.
- [8] Y. Qian, F. Soong, and Z.-J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 280–290, 2013.
- [9] X.-J. Xia, Z.-H. Ling, Y. Jiang, and L.-R. Dai, "HMM-based unit selection speech synthesis using log likelihood ratios derived from perceptual data," *Speech Communication*, vol. 63, pp. 27–37, 2014.
- [10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999, pp. 2347–2350.
- [11] L.-H. Chen, Y. Jiang, M. Zhou, Z.-H. Ling, and L.-R. Dai, "The USTC System for Blizzard Challenge 2016," in *Blizzard Challenge Workshop*, 2016.
- [12] L.-J. Liu, C. Ding, Y. Jiang, M. Zhou, and S. Wei, "The IFLYTEK System for Blizzard Challenge 2017," in *Blizzard Challenge Workshop*, 2017.
- [13] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.
- [14] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4460–4464.
- [15] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.
- [16] 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.
- [17] V. Wan, Y. Agiomyriannakis, H. Silen, and J. Vit, "Googles next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders," in *Proc. Interspeech*, 2017, pp. 1143–1147.
- [18] V. Pollet, E. Zovato, S. Irhimeh, and P. Batzu, "Unit selection with hierarchical cascaded long short term memory bidirectional recurrent neural nets," *Proc. Interspeech 2017*, pp. 3966–3970, 2017.
- [19] Z.-H. Ling, H. Lu, G.-P. Hu, L.-R. Dai, and R.-H. Wang, "The USTC system for Blizzard Challenge 2008," *Blizzard Challenge Workshop*, 2008.
- [20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [21] 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*, pp. 1–10, 2018.
- [22] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.
- [23] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.