*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Investigating Efficient Feature Representation Methods and Training Objective for BLSTM-Based Phone Duration Prediction

*Yibin Zheng[1,3], Jianhua Tao[1,2,3], Zhengqi Wen[1], Ya Li[1], Bin Liu[1]*

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, China
[3]School of Computer and Control Engineering, University of Chinese Academy of Science, China

{yibin.zheng, jhtao, zqwen, yli, liubin}@nlpr.ia.ac.cn

**Abstract**

Accurate modeling and prediction of speech-sound durations are important in generating natural synthetic speech. This paper focuses on both feature and training objective aspects to improve the performance of the phone duration model for speech synthesis system. In feature aspect, we combine the feature representation from gradient boosting decision tree (GBDT) and phoneme identity embedding model (which is realized by the jointly training of phoneme embedded vector (PEV) and word embedded vector (WEV)) for BLSTM to predict the phone duration. The PEV is used to replace the one-hot phoneme identity, and GBDT is utilized to transform the traditional contextual features. In the training objective aspect, a new training objective function which taking into account of the correlation and consistency between the predicted utterance and the natural utterance is proposed. Perceptual tests indicate the proposed methods could improve the naturalness of the synthetic speech, which benefits from the proposed feature representation methods could capture more precise contextual features, and the proposed training objective function could tackle the over-averaged problem for the generated phone durations.

**Index Terms**: Phone duration modeling, BLSTM, feature representation methods, training objective, speech synthesis

## 1. Introduction

In text-to-speech (TTS) synthesis, phone duration determines the rhythm and tempo of synthetic speech, and thus influences quality and naturalness [1]. Generally, phone duration prediction can be viewed as a problem of estimating nonlinear prediction functions using contextual features obtained from input text as explanatory variables. From this point of view, many methods have been developed for phone duration modeling, such as decision tree (DT), gradient boosting decision tree (GBDT), Gaussian process regression (GPR), and bidirectional long short-term memory (BLSTM) recurrent neural network and etc. [2-10]. Among these methods, the state-of-the-art performance was achieved with the BLSTM [9], as BLSTM is able to capture the long-short term dependencies across the input phone sequence.

However, even the state-of-the-art method (BLSTM) is not always satisfied with reproducing phone duration, which makes synthetic speech come across as unnatural, inappropriate, and unappealing [8]. There are two reasons accounting for this. One is in the feature engineering aspect, as these methods used discrete contextual representation of phones (like part-of-speech (POS) or phoneme identity). Such a representation requires a hard classification of phones into a set of discrete classes and doesn't take into account the distributional behaviors of phones. Take the phoneme identity features used in [8-9] for example, the relatedness between two phonemes wouldn't be conveyed effectively since the phoneme identity features are encoded as a binary vector by one-hot representation. Another one is in the training objective aspect, as these methods utilized the root mean squared error (*RMSE*) as the training objective. Such a training objective doesn't always correlate well with human perception, since the correlation and consistency between the predicted utterance and the natural utterance would not be considered during training. In this way, the fluctuation of phone durations in utterances would be small, leading to the over-averaged of predicted phone duration. To address these issues, this paper focuses on both feature and training objective aspects to improve the performance of the phone duration model for speech synthesis system based on BLSTM.

The aim of this paper is to improve the naturalness of synthetic speech by improving phone duration modeling. We combine the feature representation from gradient boosting decision tree (GBDT) and phoneme identity embedding model for BLSTM to predict the phone duration. Additionally, the pseudo R-squared ($R^2$), (which indicates the correlation and consistency between the predicted utterance and the natural utterance.) is added as a part of the training objective function in this paper. To summarize, the contributions of this paper are:

- A phoneme embedded vector (PEV) is utilized to replace the one-hot representation of phoneme identity features for BLSTM based phone duration model. Thus the relatedness between surrounding phonemes would be considered. The PEV is learned under a joint training structure in the word embedding model.
- GBDT is employed to transform the traditional contextual features into more efficient discriminative features for BLSTM based phone duration model, since features represented by GBDT could be seen as an effective coding process.
- A new training objective function which gives a comprehension consideration between the *RMSE* and $R^2$ is proposed. Thus the proposed training objective not only minimizes the *RMSE* but also makes the distribution of the predicted phone durations close to that of natural utterances.

## 2. Efficient Feature Representation Methods

In this section, we will discuss how to make the feature representation using phone identity embedding and GBDT.

### 2.1. Phoneme identity embedding

Word embedding represents words as continuous vectors in a low-dimensional space based on the assumption that the



http://dx.doi.org/10.21437/Interspeech.2017-1086

semantic meaning of a word can be predicted from the external contexts with large-scale corpus. This assumption also works for the phonemes whose pronunciation is affected by the neighboring words and phonemes [11]. So we can believe that encoding the pronunciation of the phoneme as a continuous real-valued vector is accessible and impactful. Recently, various embedding models have been developed, including continuous bag-of-words model (CBOW), Skip-Gram model [12] and Global C&W [13]. Related ideas of embedding features have been employed in TTS, including the acoustic [14] and prosody model [15]. The following will show how to jointly train the phoneme embedded vector (PEV) and the word embedded vector (WEV) based on CBOW.

### *2.1.1. CBOW*

CBOW aims at predicting the target word, given context words in a sliding window. Formally, given a word sequence $D = \{x_1, \dots, x_M\}$, the objective of CBOW is to maximize the average log probability,

$$L(D) = \frac{1}{M}\sum_{i=K}^{M-K} log\ Pr(x_i | x_{i-K}, \dots, x_{i+K}) \quad (1)$$

where $K$ is the context window size of target word. CBOW formulates the probability $Pr(x_i|x_{i-K}, \dots, x_{i+K})$ using a softmax function as follows:

$$Pr(x_i|x_{i-K}, \dots, x_{i+K}) = \frac{exp(X_0^T \cdot X_i)}{\sum_{x_i' \epsilon W} exp(X_0^T \cdot X_i')} \quad (2)$$

where $W$ is the word vocabulary, $X_i$ is the vector representation of the target word $x_i$, and $X_o$ is the average of all context word vectors:

$$X_0 = \frac{1}{2K}\sum_{j=i-K,\dots,i+K, j\neq i} X_j \quad (3)$$

### *2.1.2. Joint training for PEV and WEV*

It is difficult to train the PEV directly from the large corpus because the PEV takes a non-sematic meaning of a word. But it could be learned simultaneously with the WEV in a joint training structure described in [16]. The structure is proposed to jointly train the syllable embedded vector with WEV for Mandarin where the target word is predicted by the context words and the composition syllables together. We have used this model for prosody structure prediction in Mandarin and obtained competitive results [15]. While in [14], we replaced the composition syllables with the pronunciation initials and finals, (The trained unit for the pronunciation of the word is phoneme for English, and initial or final for Mandarin.) and predicted the target word by the context words and the pronunciation initials and finals together. In this way, the PEV could be learned together with WEV. Such idea is also adopted in this paper.

We denote the Mandarin phoneme set (including initials and finals) as $P$, and the Mandarin word vocabulary as $W$. Each phoneme $p_i \in P$ is represented by vector $P_i$ and each word $x_i \in W$ *is* represented by vector $X_i$. As we learn to maximize the average log probability in Equation (1) with a word sequence $D = \{x_1, \dots, x_M\}$, we represent context words with both PEV and WEV to predict target word. So the embedded vector for the context word $x_i$ is changed from $X_i$ to $X_i^{new}$ in Equation (4).

$$X_i^{new} = X_i + \frac{1}{N_i}\sum_{k=1}^{N_i} P_k \quad (4)$$

where $X_i^{new}$ is the composed embedded vector of $x_i$, $X_i$ is the word embedded vector (WEV) of $x_i$, $N_i$ is the number of initials and finals in $x_i$, $P_k$ is the phoneme embedded vector (PEV) of *k-th* phoneme $p_k$ in $x_i$.

In order to make the model more efficient for learning, hierarchical softmax and negative sampling are used [12].

## 2.2. Feature representation using GBDT

In this section, we will review the GBDT algorithm first and then demonstrate how to use the GBDT to get the transformed features.

### *2.2.1. GBDT*

GBDT [17] is a meta algorithm for constructing multiple regression trees and takes advantages of them. It has been employed as a regression model for phone duration prediction in [5] and achieved better performance than DT based method. Given that the training process of GBDT could be seen as an effective coding process for the clustering, we believe that every leaf in the GBDT could be employed as an efficient discriminative feature to represent the traditional contextual features. Hence, we use it to make the feature representation in this paper.

Define explanatory variables and a target value as $x = (x_1, \dots, x_K)$ and y, respectively. Let $\{x_i, y_i\}_1^N$ be a set of training data including $N$ pairs. The GBDT algorithm iteratively constructs $M$ different regression tree $h(x, a_1), \dots, h(x, a_M)$ from the set of training data and construct the following additive function $F(x)$:

$$F(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h(x, a_m) \quad (5)$$

where $\beta_m$ and $a_m$ are a weight and vector of parameters for the *m-th* regression tree $h(x, a_m)$, and $\beta_0$ is an initial value. Both $\beta_m$ and $a_m$ are iteratively determined from *m=1* to *m=M* so that the loss function $L(y, F(x))$ is minimized. We define an additive function that combining from the first regression tree to the *(m-1)-th* regression tree as $F_{m-1}(x)$. The weight $\beta_m$ and $a_m$ for the *m-th* regression tree is decided by:

$$(\beta_m, a_m) = argmin_{\beta,a} \sum_{i=1}^{N} L(y_i, F_{m-1}(x) + \beta h(x_i, a)) \quad (6)$$

where $F_0(x)$ is an initial value and given by $F_0(x) = \beta_0 = argmin_\beta \sum_{i=1}^{N} L(y_i, \beta)$. And we utilize the least-square loss $L(y, F(x)) = (y - F(x))^2$ here.

### *2.2.2. GBDT based feature representation*

There are two simple ways to transform the input features in order to boost the performance. For continuous features, a simple trick for learning such transform is to bin the feature and treat the bin index as a categorical feature; for categorical features, the brute force approach consists in taking the Cartesian product [18].

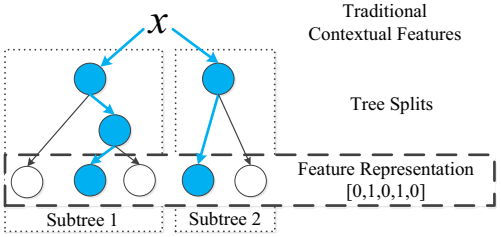


Figure 1: *Feature representation using GBDT.*

GBDT has been proved as a powerful and very convenient way to implement such non-linear and tuple transformations we just described. We treat each individual tree as a categorical feature that takes as value the index of the leaf an instance ends up falling in. For example, consider the GBDT model in Figure 1 with 2 subtrees (regression tree), where the first subtree has 3 leafs and the second has 2 leafs. If an instance ends up in leaf 2 in the first subtree and leaf 1 in the second subtree, the feature representation by the GBDT would be a binary vector [0,1,0,1,0], where the first 3 entries correspond to the leaves of the first subtree and last 2 to those of the second subtree.

## 3. Integration in the BLSTM Based Phone Duration Model

Feature representation learning (such as word/syllable embeddings and etc.) has been proved useful in describing contextual features for the acoustic and prosody model in TTS [14, 15]. This paper also adopts similar idea to represent traditional contextual features for phone duration modeling. We adopt BLSTM [19] as the modeling method in this paper. The BLSTM is very effective at modeling sequences, generating state-of-the-art performance across various dynamic tasks that involve complex contextual dependencies, including phone duration prediction [9], prosody and acoustic model in TTS [20-22]. Therefore, we combine the feature representation from GBDT and phoneme identity embedding model for BLSTM to predict the phone duration.

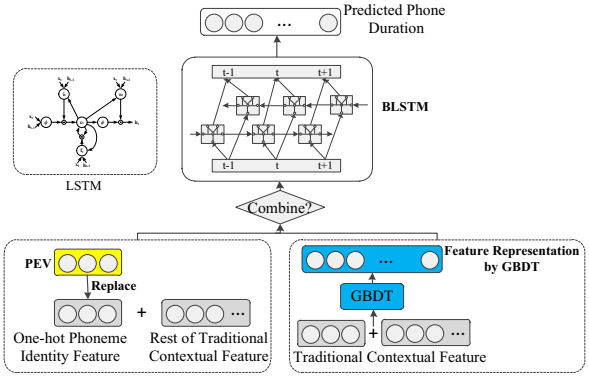


Figure 2: *Integrate feature representations for BLSTM based phone duration model*

The proposed feature representation for BLSTM based phone duration model is shown in Figure 2. For feature engineering aspect, on one hand, the one-hot representation of phoneme identity features is replaced with the PEV (which considers the relatedness between the surrounding phonemes and is jointly trained with WEV), while other remaining traditional contextual features (such as lexical tone and position-related features etc. [14]) are kept the same. On the other hand, all the traditional contextual features are used to train a GBDT based phone duration model, then the trained GBDT model is used to transform the traditional contextual features using method mentioned in Section 2.2.2. These two feature representations can be combined (just simply appended in this paper) together or independently used as the input for BLSTM based duration model. As for the modeling aspect, we employ BLSTM to model the phone duration due to its powerful sequence modeling capabilities.

## 4. Training Objective

Previous researches on phone duration modeling almost utilized the *RMSE* (or its variety squared error), between the predicted duration and the natural utterance, as the training objective. Such a training objective doesn't always correlate well with human perception, since the correlation and consistency between the predicted utterance and the natural utterance would not be considered during training. In this way, the distribution of the predicted phone durations would be far from that of natural utterances. More specifically, the fluctuation of phone durations in utterances would be small, and thus leading to the over-averaged of predicted phone duration. Therefore, in this paper, we give a comprehension consideration between the *RMSE* and the consistency between the predicted phone duration and the natural utterance by modifying the training objective function. The training objective function is defined as:

$$Loss = (1-a) * RMSE + a * (1-R^2) \quad (7)$$

where $R^2$ is the pseudo $R$-squared, which indicates the correlation and consistency between the predicted duration and the natural utterance [5], and $a \in [0,1]$. Therefore, the training objective not only minimizes the generation error but also makes the distribution of the predicted phone durations close to that of natural utterances. When $a = 0$, the training objective is equivalent to the *RMSE* (which is often used in previous researches). Since BLSTM is adopted as our phone duration model, the $R^2$ and *RMSE* could be defined as follow:

$$R^2 = 1 - \sum_{n=1}^{N} \frac{\sum_{i=1}^{T}(F(x_{ni}) - y_{ni})^2}{N \sum_{i=1}^{T}(y_{ni} - \bar{y})^2} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\sum_{i=1}^{T}(F(x_{ni}) - y_{ni})^2}{NT}} \quad (9)$$

where $N$ is the number of utterances in training data, $T$ is the number of samples in each utterance, $F(x_{ni})$, $y_{ni}$ are the predicted duration and ground-truth duration for *i-th* phone in *n-th* utterance, respectively; and $\bar{y} = \sum_{n=1}^{N}\sum_{i=1}^{T} y_{ni} / NT$ is the mean of the training samples. Lower *RMSE* and higher $R^2$ are expected for better system in objective evaluation.

## 5. Experiments and Discussions

For evaluating the effectiveness of the proposed methods, we rely on a speech synthesis corpus recorded by a professional female speaker. The corpus contains 20000 sentences and more than 400000 syllables. The standard phone duration is obtained by force alignment using HTS tool HSMMAlign [23] after conventional HMM training, since the automatically aligned phone duration is proved more effective than manually labeled one in our previous experiments [1]. The whole corpus is partitioned into training, validation and testing set for all experiments according to 8:1:1. Here the number of Mandarin units is 60, including initial, which is an initial consonant of a syllable, and final, which is the part after the initial of the syllable.

For the BLSTM-related systems, the traditional contextual features are represented as a vector for each phone with a dimension of 377 where 300 dimensions are used for the one-hot representation of the phoneme identity features. Therefore, during PEV training, both PEV and WEV dimension are also set as 300 to keep the same with the one-hot representation. The training tool provided in [16] is a joint training version of word2vec and is further adopted to train the PEV in this paper.

For objective evaluation, we utilize the following two measures: $R^2$ and *RMSE*. For subjective evaluation, we present mean opinion score (MOS) and ABX preference test in terms of the naturalness of synthetic speech. A total 50 sentences random selected from test sets are used. 10 speech experts take part in the subjective evaluation test.

### 5.1. System built

To evaluate the proposed techniques, the following systems are built. For all BLSTM-related systems, a 3-layer neural network consisting a single non-recurrent layer, followed by 2 stacks of bidirectional layers (each with 128*2 LSTM hidden units), and an output layers is used. All BLSTM-related system are trained with a momentum of 0.9, an initial learning of 0.001 for the first 10 epoch, and then decreases half after each epoch. For GBDT, the max depth and number of subtrees are set as 4 and 60 respectively. The dimension of the transformed features by GBDT is 750. A brief description of systems built is given below. The proposed new training objective function is applied only in system S5.

- BLSTM (**S1**): BLSTM based system trained with traditional contextual features (where the phoneme identity features use a one-hot representation).
- BLSTM+PEV (**S2**): BLSTM based system trained with PEV and the rest of traditional contextual features.
- BLSTM+GBDT (**S3**): BLSTM based system trained with feature representation by GBDT.
- BLSTM+PEV+GBDT (**S4**): BLSTM based system trained with feature representation by GBDT, PEV and rest traditional contextual features.
- S4+modifying training objective function (**S5**): Based on S4, and the proposed new training objective function (Equation (7)) is applied.

All the systems are trained by Theano [24] toolkit.

Table 1. *Objective evaluation results for system S1-S4.*

| Systems | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| ***RMSE(ms)*** | 31.57 | 31.06 | 30.65 | **30.34** |
| $R^2$ | 0.680 | 0.689 | 0.698 | **0.709** |

### 5.2. Evaluation of feature representation methods

To testify the availability of the feature representation methods, system S1, S2, S3 and S4 are built for comparison. Table 1 shows the objective evaluation results (in test set) of these four systems. Compare the performance of S2, S3 with S1, we could find that both S2 and S3 could achieve better performance than S1, in terms of both the *RMSE* and $R^2$. This indicates that PEV is a better substitution compared to the one-hot phoneme identity feature as PEV takes into account the relatedness of its surrounding contexts; and features represented by GBDT can serve as more effective features for the input of BLSTM since each subtree in GBDT can be viewed as a clustering process. In this way, features represented by GBDT could be seen as an effective coding process for the clustering results, and thereby are more discriminative than the traditional contextual features. Additionally, we combine the features represented by these two methods (donated as S4) to validate its effectiveness. The results in Table 1 show that S4 achieves the best performance among these four systems (including S2, S3). This means combining these two feature representation methods can further boost the performance. Table 2 shows the results of ABX preference test for S1-S4. It's seen that the listeners preferred both the two proposed feature representation methods than the traditional contextual features (S1). When combining these two feature representation methods (S4), we see that S4 received more preference. This is consistent with the objective evaluation results.

### 5.3. Evaluation of training objective function

To evaluate the effect of the proposed training objective function, system S5 is built and the weight coefficient $a$ in Equation (7) is varied from 0.000 to 0.300 in our experiments. (S5 with weight coefficients $a = 0.000$ is equal to S4, because the training objective is equivalent to the *RMSE* at this time.) The objective evaluation results (in test set) for system 5 with different weight coefficient $a$ are presented in Table 3. From which, we can see that by adding the $R^2$ into the training objective function, both *RMSE* and $R^2$ increase as the weight coefficient $a$ grows. In order to evaluate their effects to the synthetic speech, we further conduct a MOS test for S5 with different weight coefficients $a$. The MOS test results in Table 4 show that, the systems with weight coefficient $a = 0.150$ achieve the best performance in terms of the naturalness of synthetic speech. This indicates that: (1) By adding the $R^2$ into the training objective function, the *RMSE* is increased. But more importantly, the higher $R^2$ it brings about indicates the correlation and consistency between the predicted duration and the natural utterance is greater. This means the distribution of the predicted phone durations would be closer to that of natural utterances, which can tackle the over–averaged of the generated phone durations in a way. (2) When the weight coefficient $a$ reaches a significant degree (like $a > 0.150$), (though the $R^2$ increases) the *RMSE* it brings about would become too larger. This means the accuracy of phone duration prediction would become worse, thus demonstrating the superiority of the S5 with $a = 0.150$ than S5 with $a > 0.150$ in MOS test. Therefore, it is suggestive that the weight coefficient $a$ should be chosen not only by the two objective measures (*RMSE* and $R^2$), but also by the subjective evaluation.

Table 2. *Preference scores (%) of different compared pair systems. The confidence of t-test is 95%.*

| Compared System | The Former | The Latter | Neutral | *p-value* |
|---|---|---|---|---|
| ***S1 vs S2*** | 21.4 | 38.8 | 39.8 | <0.0001 |
| ***S1 vs S3*** | 18.9 | 40.5 | 40.6 | <0.0001 |
| ***S1 vs S4*** | 13.5 | 54.2 | 32.3 | <0.0001 |
| ***S2 vs S4*** | 20.9 | 42.6 | 38.5 | <0.001 |
| ***S3 vs S4*** | 24.2 | 41.8 | 34.0 | <0.001 |

Table 3. *Objective evaluation results for S5 with different coefficients a.*

| Coefficient *a* | 0.000 | 0.075 | 0.150 | 0.225 | 0.300 |
|---|---|---|---|---|---|
| ***RMSE(ms)*** | 30.34 | 30.97 | 31.76 | 32.91 | 34.03 |
| $R^2$ | 0.709 | 0.723 | 0.744 | 0.761 | 0.785 |

Table 4. *MOS for S5 with different coefficients a. The confidence of t-test is 95%, p-value <0.0001.*

| Coefficient *a* | 0.000 | 0.075 | 0.150 | 0.225 | 0.300 |
|---|---|---|---|---|---|
| ***MOS*** | 3.62 | 3.68 | **3.79** | 3.70 | 3.54 |

## 6. Conclusions and Future Works

This paper focuses on both feature and training objective aspects to improve the performance of the phone duration model. It combines the feature representation from gradient boosting decision tree (GBDT) and phoneme identity embedding model (which is realized by the jointly training of PEV and WEV) for BLSTM to predict the phone duration. The PEV is used to replace the one-hot phoneme identity, and GBDT is utilized to transform the traditional contextual features. In addition, a new training objective function which taking into account of the correlation and consistency between the predicted utterance and the natural utterance is proposed. Perceptual tests indicate the proposed methods could improve the naturalness of the synthetic speech, which benefits from the proposed feature representation methods could describe more precise contextual features, and the proposed training objective function could tackle the over-averaged problem for the generated phone durations. In future, we will extend the proposed approach in other language like English and etc.

## 7. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017， No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006) and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

## 8. References


[1] Y. Wang, M.H. Yang, Z.Q. Wen, and J.H Tao, "Combining extreme learning machine and decision Tree for Duration Prediction in HMM based speech synthesis," in *Annual Conference of the International Speech Communication Association, Interspeech,* pp. 2197-2201, 2015.

[2] O. Goubanova, and S. King, "Bayesian networks for phone duration prediction," [J]. *Speech Communication*, vol. 50, pp. 301-311, 2008.

[3] S. Sovilj-Nikic, V. Delic, I. Sovilj-Nikic, et al. "Tree-based phone duration modeling of the Serbian language," [J]. *Electronics & Electrical Engineering*, 20(3): pp.77-82, 2014.

[4] M. Riedi, M. Riedi, "A neural-network-based model of segmental duration for speech synthesis," in *European Conference on Speech Communication and Technology, Eurospeech,* 1995.

[5] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," [J]. *Speech Communication*, 50(5), pp. 405-415, 2008.

[6] A. Lazaridis, I. Mporas, T. Ganchev, et al. "Improving phone duration modelling using support vector regression fusion," [J]. *Speech Communication*, 53(1): pp.85-97, 2011.

[7] D. Moungsri, T. Koriyama, T. Kobayashi, "Duration prediction using multi-level model for GPR-Based speech synthesis," in *Annual Conference of the International Speech Communication Association, Interspeech,* pp. 1591-1595, 2015.

[8] G.E. Henter, S. Ronanki, O. Watts, et al. "Robust TTS duration modelling using DNN," in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, pp. 5130-5134, 2016.

[9] J. Tao, Y. Zheng, Z. Wen, Y. L, et al. "A BLSTM Guided Unit Selection Synthesis System for Blizzard Challenge 2016", in *Proceeding of Blizzard Challenge 2016*.

[10] D. Moungsri, T. Koriyama, T. Kobayashi, "Duration Prediction Using Multiple Gaussian Process for GPR-Based Speech Synthesis," in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, pp.5495-5499, 2017.

[11] Z.J. Wu, "The Chinese Phonetics in 'Man-Machine Dialogue'", Chinese Teaching In The World, vol 4, pp. 3-20, 1997. (In Chinese).

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS,* pp. 3111–3119, 2013.

[13] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–154, 2014.

[14] Z.Q. Wen, Y. Li, and J.H. Tao, "The parameterized phoneme identity feature as a continuous a real-valued vector for Neural network based speech synthesis," in *Annual Conference of the International Speech Communication Association, Interspeech,* 2016.

[15] Y. Zheng, Y. Li, Z. Wen, X. Ding and J. Tao, "Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach," in *Annual Conference of the International Speech Communication Association, Interspeech,* 2016.

[16] X. Chen, L. Xu, Z. Liu, M. Sun, H. Luan, "Joint Learning of Character and Word Embeddings", *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[17] T. Hastie, R. Tibshirani, J. H, Friedman, "10. Boosting and Additive Trees," *The Elements of Statistical Learning (2nd ed.). New York: Springer*. pp. 337–384, 2009.

[18] X. He, J. Pan, O. Jin, et al. "Practical lessons from predicting clicks on Ads at Facebook," *Eighth International Workshop on Data Mining for Online Advertising. ACM*, pp. 1-9, 2014.

[19] M. Schuster, K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[20] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Annual Conference of the International Speech Communication Association*, *Interspeech*, pp. 2268–2272, 2014.

[21] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Annual Conference of the International Speech Communication Association, Interspeech*, pp. 1964–1968, 2014.

[22] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Using deep bidirectional recurrent neural networks for prosodic target prediction in a unit-selection text-to-speech system," in *Annual Conference of the International Speech Communication Association, Interspeech,* pp. 1606–1610, 2015.

[23] HTS [Oneline]. Available: *http://hts.sp.nitech.ac.jp/*

[24] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions". [Online]. Available: *http://deeplearning.net/software/theano/*