INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improving Prosodic Boundaries Prediction for Mandarin Speech Synthesis by Using Enhanced Embedding Feature and Model Fusion Approach

Yibin Zheng[1], *Ya Li*[1], *Zhengqi Wen*[1], *Xingguang Ding*[1], *Jianhua Tao*[1,2]

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
{yibin.zheng, yli, zqwen, xingguang.ding, jhtao}@nlpr.ia.ac.cn

Abstract

Hierarchical prosody structure generation is an important but challenging component for speech synthesis systems. In this paper, we investigate the use of enhanced embedding (joint learning of character and word embedding (CWE)) features and different model fusion approaches at both character and word level for Mandarin prosodic boundaries prediction. For CWE module, the internal structures of words and non-compositional words are considered in the word embedding, while the character ambiguity is addressed by multiple-prototype character embedding. For model fusion module, linear function (LF) and gradient boosting decision tree (GBDT), are investigated at the decision level respectively, with the important features selected by feature ranking module used as its input. Experiment results show the effectiveness of the proposed enhanced embedding features and the two model fusion approaches at both character and word level.

Index Terms: prosodic boundaries prediction, model fusion, BLSTM, enhanced embedding features, speech synthesis

1. Introduction

Prosody structure plays an important role in both naturalness and intelligibility of speech [1]. It splits an utterance into prosodic units which can be easily understood by people. Therefore, identifying the phrase boundaries of different prosodic units from text is crucial in speech synthesis.

In mandarin speech synthesis systems, a typical hierarchical prosodic structure is widely employed to distinguish different levels of pauses between words in speech. Normally, the prosodic boundaries are often classified into prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH) [2]. Traditional methods including classification and regression tree, memory based learning, conditional random field (CRF) and deep recurrent network are adopted to predict prosodic boundaries with linguistic class features (such as part-of-speech (POS), word-terminal syllables etc.) [3-10]. However, the linguistic class features are discrete linguistic representations of words, which don't take into account the distributional behavior of words [11]. And this issue has been addressed by word embedding (also known as distributed word representation) [12][13], which encodes a word as a real-valued low-dimensional vector. Related ideas of word embedding have taken effect for statistics parameter based and unit selection based speech synthesis system [14][15]. Recently, the embedding features are employed for prosodic prediction [16]. In [17], character embedding are applied to substitute for linguistic class features in bi-directional long-short term memory (BLSTM) [18] recurrent network for Mandarin prosodic boundaries prediction. Similar work can be found in [19], which utilizes word embedding to argument, rather than replace, linguistic class features for English prosodic phrasing and prominence prediction.

Note that [17] represents one character with only one vector, which is insufficient for Mandarin characters that are much more ambiguous. While in [19], the internal structures of words are ignored when learning word embedding. However, in Mandarin, a word, usually composed of several characters, contains rich internal information. Hence an intuitive idea is to take into account internal characters for learning word embedding. Besides, not all Mandarin words are semantically compositional, such as transliterated words. Thus, a character-enhanced word embedding model (CWE) and a multiple-prototype character embedding model [20] are employed to address these issues in this work. Meanwhile, it would be meaningful to investigate the effects of word embedding for Mandarin prosodic boundaries prediction since the word is often used as the ideographic unit in Mandarin, while only character embedding is considered in [17].

Instead of simply combining embedding with linguistic class features [19], we conduct a complementary research inquiry to focus on model fusion at the decision level. In our work, we fuse the results from CRF [9] and BLSTM [10], which shows the best reported results with linguistic class features and embedding features respectively. Indeed, this manifests the idea of ensemble learning (i.e. boosting, results fusion), which is designed to improve prediction accuracy of single predictors [21] [22].

In this paper, we explore the novel use of ensemble learning (for model fusion module) and enhanced embedding features into Mandarin prosodic boundaries prediction. There are three main contributions. (1) LF-based and GBDT-based model fusion methods are proposed to predict prosodic boundaries. By model fusion, the dependency of the final result on each single classifier can be figured out. (2) The enhanced embedding features which take into account the internal of words, character ambiguity and non-compositional words are investigated. (3) Different operated levels for Mandarin prosodic boundaries prediction are compared.

2. Enhanced embedding features

Word embedding represents words as continuous vectors in a low-dimensional space based on the distributional hypothesis that words in similar contexts have similar meaning. Based on

http://dx.doi.org/10.21437/Interspeech.2016-1060

this hypothesis, various embedding models have been developed, including continuous bag-of-words model (CBOW), Skip-Gram model [12] and Global C&W [13]. We will take CBOW for example and demonstrate the framework of CWE on CBOW.

2.1. Basic embedding features

We use CBOW, which takes a word or a character as basic unit as basic model to generate basic embedding features. The training objective of CBOW is to combine the embedding features of context words to predict the target words. Formally, given a word sequence $D = \{x_1, \dots, x_M\}$, the objective of CBOW is to maximize the average log probability,

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

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:

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

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:

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

An example of CBOW is shown in Figure 1(A), where yellow boxes are word embedding of context words, which are combined together to get the embedding (the orange box) for prediction of the target word.

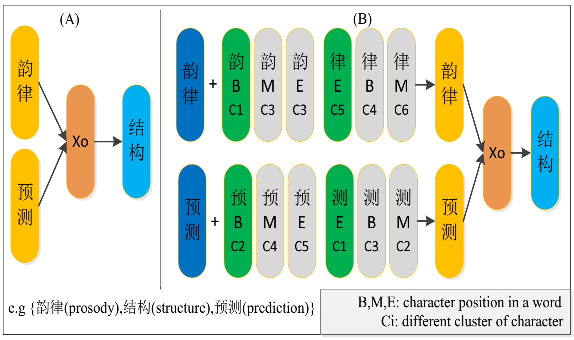

Figure 1: (A) CBOW; (B) position-cluster-based multiple-prototype character embedding for CWE.

2.2. Enhanced embedding features

2.2.1. *Character-enhanced word embedding (CWE)*

Character-enhanced word embedding (CWE) considers character embedding in an effort to improve word embedding. We denote the Mandarin character set as C and the Mandarin word vocabulary as W. Each character $c_i \in C$ is represented by vector C_i *and* each word $w_i \in W$ *is* represented by vector W_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 character embedding and word embedding to predict target words. Formally, a context word x_j is represented as:

$$X_j = \frac{1}{2}(W_j + \frac{1}{N_j}\sum_{k=1}^{N_j} C_k) \qquad (4)$$

where W_j is the word embedding of x_j, N_j is the number of characters in x_j, C_k is the embedding of k-th character c_k in x_j.

Note that multiplying $\frac{1}{2}$ is crucial because it maintains similar length between embedding of compositional and non-compositional words. A non-compositional word list is built manually and their characters are not considered when learning these words.

2.2.2. *Multiple-prototype character embedding*

Mandarin characters are highly ambiguous. Here we employ multiple-prototype character embedding to address this issue. The idea is that, we keep multiple vectors for one character, each corresponding to one of the meanings.

As demonstrated in Figure 2(B), we keep three embedding for each character c, (C^B, C^M, C^E), corresponding to its three types of position in a word (Begin, Middle, and End). Hence, Equation (4) can be rewritten as:

$$X_j = \frac{1}{2}(W_j + \frac{1}{N_j}(C_1^B + \sum_{k=2}^{N_j-1} C_k^M + C_{N_j}^E) \qquad (5)$$

Motivated by the position-based character embedding, for each character c, we can also cluster all its occurrences into N_c cluster and build one embedding for each cluster, shown in Figure 1(B). Take context word $x_j = \{c_1, \dots, c_N\}$ for example, $C_k^{r_k^{max}}$ will be used to get x_j. Define $S()$ as consine similarity, then

$$r_k^{max} = \arg\max_{r_k} S(C_k^{r_k}, V_{context}) \qquad (6)$$

$$V_{context} = \sum_{t=j-K}^{j+K} X_t$$
$$= \sum_{t=j-K}^{j+K} \frac{1}{2}(W_t + \frac{1}{N_t}\sum_{C_u \epsilon x_t} C_u^{most}) \qquad (7)$$

C_u^{most} is the character embedding most frequently chosen by x_t in the previous training. After obtaining the optimal cluster assignment collection $R = \left\{r_1^{max}, \dots, r_{N_j}^{max}\right\}$,we can get the embedding X_j of x_j as

$$X_j = \frac{1}{2}(W_j + \frac{1}{N_j}\sum_{k=1}^{N_j} C_k^{r_k^{max}}) \qquad (8)$$

which can be further used to obtain X_o using Equation (3) for optimization.

In this work, for each position of a character (B, M, E), we learn multiple embedding to solve the possible ambiguity issue confronted in this position. We designate it as position-cluster-based multiple-prototype character embedding. An example of joint learning of enhanced embedding in CEW is showed in Figure 1(B), where the word embedding (blue boxes in figure) and character embedding (green boxes) are composed together to get new embedding (yellow boxes).

3. Model fusion approach

As a modelling approach we adopt BLSTM that involves complex contextual dependencies and has been recently shown to provide state-of-the-performance across various dynamic modeling tasks [24][25][26] as one of single classifiers. Another single classifier is the best shallow model CRF.

The flowchart of proposed framework of model fusion is shown in Figure 2. Firstly, two single classifiers CRF and BLSTM, which use linguistic class features and embedding features respectively are trained, then the probability of Breaks (PW, PPH and IPH) can be obtained by these two single classifiers; Next, by a feature ranking module, we can acquire the importance of each feature, which is realized by ranking the F-Measure [27] value that promotes by this feature. Finally, the output probabilities, together with the important features are consisting of the inputs for model fusion module. During model fusion, two different methods, LF and GBDT are employed to make the final prediction.

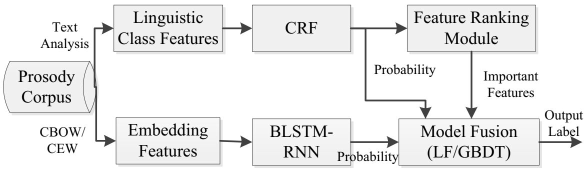

Figure 2: *Flowchart of model fusion.*

3.1. LF-based model fusion

LF-based model fusion has been applied to fuse multiple parameterizations for high quality speech synthesis [28]. For LF-based model fusion, the predictions from all the single classifies are combined through a weighted majority vote to produce the final prediction:

$$f(x) = sign(\sum_{m=1}^{M} \alpha_m f_m(x)) \qquad (9)$$

where α_m is the weight contribution of each respective classifier $f_m(x)$. Their effect is to give higher influence to the more accurate classifiers in the sequence. In order to clearly analyze the dependency of results on each single classifies, we set the sum of the weight coefficient to 1.

3.2. GBDT-based model fusion

GBDT is an additive classification (or regression) model consisting of an ensemble of trees, fitted to current residuals (gradient of the loss function), in a forward step-wise manner. A decision tree partitions the space of all joint predictor variant into disjoint regions R_j,*j=1,2,…,J* as represented by the terminal nodes of the tree. A constant γ_j is assigned to each such region. Thus a tree can be formally expressed as：

$$T(x,\theta) = \sum_{j=1}^{J} \gamma_j \, Index(x \in R_j) \qquad (10)$$

$$Index = \begin{cases} 1, (x \in R_j) \\ 0, others \end{cases} \qquad (11)$$

Here, x is the input feature set. Then GBDT is a sum of such trees, where M is the numbers of trees in GBDT,

$$f_M(x) = \sum_{m=1}^{M} T(x, \theta_m) \qquad (12)$$

Training process consists of fitting an ensemble of decision trees, each of them are trained separately in a sequential manner. Predicting is accomplished by adding the predictions of each decision tree, as Equation (12) suggests. For more details, refer to [29]. Shrinkage techniques [29] are employed in our work to prevent over fitting.

4. Experiments and results analysis

For evaluating the effectiveness of the proposed approaches, we rely on a speech synthesis corpus recorded by a professional female speaker. The corpus contains 20000 sentences and more than 400000 syllables. Prosodic boundaries (PW, PPH and IPH) are annotated by two expert annotators who have access to the audio and their text transcriptions and the labelling consistence is ensured. Word segmentation and POS tagging are carried out by a front-end preprocessing tool. The accuracy of word segmentation is 96.6% and the accuracy of POS tagging is 96.4%. The whole corpus is partitioned into training, validation and test set for all experiments according to 8:1:1.

We collect 15 G Mandarin text corpus relevant to our prosody corpus for embedding training [30]. The word and character vocabulary size is 514,703 and 83,608 respectively. Both character and word embedding dimension is set as 100 and context window size is set as 5 during training. For optimization, we use both hierarchical softmax and 10-word negative sampling.

In all the experiments, PW, PPH and IPH are predicted hierarchically. The predicted boundary from the lower level is used as an input feature for the current boundary for labeling decision.

4.1. Systems built

All prosodic boundaries prediction systems are built at both word and character level in our experiments. For all BLSTM-based systems, a 3-layer neural network consisting a single non-recurrent layer, followed by 2 stacks of bidirectional layers (each with 256*2 LSTM hidden units), and a binary output softmax layers is used. All networks are trained with a momentum of 0.9, an initial learning of 0.001 for the first 10 epoch, and then decreases by 20% after each epoch. While in model fusion, the best performance of each single classifier is employed. Based on these, the following systems are built.

1. **CRF**: Linguistic class features used for prosodic boundaries prediction based on CRF. These features include POS tags, the length of words and word position in sentence, etc. For character-based CRF, the character position in the word is also included in features set. A greedy algorithm is employed to optimize the feature templates that used for CRF. Specifically speaking, if F-Measure [27] is improved by cross validation on test sets, then this template is supposed to be a part of final feature template. The CRF++ toolkit [31] is used for the CRF-based prosodic prediction system.

2. **BLSTM_CBOW**: Basic embedding used for prosodic boundaries prediction based on BLSTM. The basic embedding is generated by CBOW model using wer2vec toolkit [23].

3. **BLSTM_CEW**: Enhanced embedding used for prosodic boundaries prediction based on BLSTM. The enhanced character embedding is generated by multiple-prototype character embedding based on CEW, while enhanced word embedding is generated by CEW model.

4. **LF**: LF-based model fusion used for prosodic boundaries prediction based on the output probability of the two single classifiers (CRF and BLSTM), while other features that generated from feature ranking module are ignored.

5. **GBDT1**: GBDT-based model fusion used for prosodic boundaries prediction based on the output probability of the two single classifiers (CRF and BLSTM), while other features that generated from feature ranking module are ignored. The depth and number of trees in GBDT1 is set as 2 and 36 respectively.

6. **GBDT2**: GBDT-based model fusion used for prosodic boundaries prediction based on the output probability of the two single classifiers (CRF and BLSTM), together with the features that generated from feature ranking module as input. The depth and number of trees in GBDT2 is set as 4 and 36 respectively.

Table 1 and 2 shows the performance of all six systems described in 4.1. We report our results in terms of F-Measure [27], which is defined as the harmonic mean of precision and recall. We analyze the results below.

Table 1. *Performance of F-Measure at character level.*

Systems	CRF	BLSTM CBOW	BLSTM _CEW	LF	GBDT1	GBDT2
PW	95.39	95.49	95.65	95.96	96.73	**96.65**
PPH	82.01	81.38	81.79	82.84	83.28	**83.53**
IPH	72.55	73.70	74.31	75.22	76.13	**76.85**

Table 2. *Performance of F-Measure at word level.*

Systems	CRF	BLSTM CBOW	BLSTM CEW	LF	GBDT1	GBDT2
PW	95.52	95.79	95.90	96.19	96.43	**96.79**
PPH	82.25	82.95	83.36	84.17	84.82	**85.25**
IPH	79.51	81.08	81.74	82.88	83.67	**84.73**

4.2. Evaluation of operated level

Table 1 and 2 clearly shows that word-based systems achieve superior performance than character-based systems on all three prosodic boundaries, especially in higher boundaries (IPH). The contribution factor for this improvement is that the word is often used as the ideographic unit in Mandarin, which carries more semantic information than isolated character. Therefore, it is more interpretable word embedding (rather than character embedding in [17]) can achieve superior performance than linguistic class features.

4.3. Evaluation of enhanced embedding features

To evaluate the effects of enhanced embedding features, we compare the results of BLSTM_CBOW with BLSTM_CEW. Table 1 and 2 shows that by using the enhanced embedding features (at both character and word level), the performance on all three prosodic boundaries is improved, which proves the effectiveness of proposed enhanced embedding for Mandarin prosodic boundaries prediction. Such results indicate the necessity of considering non-compositional words and words' internal information for word representation, as well as considering different character representation according to its position and clusters.

4.4. Evaluation of LF-based model fusion

More detailed results of system LF are presented in table 3, where α_1, α_2 are the weight coefficients of CRF and BLSTM respectively, *Fc* and *Fw* are the F-Measure achieved at character and word level respectively.

Table 3. *The results of system* LF.

Boundary	α_1	α_2	*Fc*	α_1	α_2	*Fw*
PW	0.38	0.62	95.96	0.42	0.58	**96.19**
PPH	0.36	0.64	82.84	0.35	0.65	**84.17**
IPH	0.41	0.59	75.22	0.43	0.57	**82.88**

Compare with the performance of single classifier systems (CRF and BLSTM) from Table 1 and 2, system LF boosts the performance on all three prosodic boundaries, which proves the effectiveness of proposed model fusion method. Moreover, from the weight coefficients, we could see the dependency of results (at both character and word level) on BLSTM is greater than that of CRF. This can be explained by system BLSTM achieves superior performance than system CRF.

4.5. Evaluation of GBDT-based model fusion

Table 1 and 2 shows that system GBDT1 achieves superior performance than system LF. It can be explained by LF is just a linear combination of two single classifiers, while GBDT exploits the strength of ensemble learning. And system GBDT2 shows an absolute increase of around 5% than system CRF for IPH prediction (at both character and word level). Such improvement at word level (5.22%) for IPH is more obvious than the best absolute improvement in [19] (2.19%) where the author just combines the word embedding features with linguistic class features. (We also conduct the approach in [19] on our corpus, where absolute improvement is only 3.09% for IPH prediction). Such result indicates model level fusion is more effective than feature level fusion. This may be caused by such two features may not suitable to fuse with each other as they are two different representation of words and characters, while model fusion can avoid this problem well.

Also system GBDT2 achieves the best performance over all six systems-including GBDT1, which shows the effectiveness of feature ranking module. To evaluate the effects of two single classifiers for system GBDT2, we further calculate the contribution of each feature by Gini importance [32], which is used as a general indicator of feature importance. Take IPH prediction for example, the degree of the top five features contributions for IPH prediction on word-based GBDT2 are listed in Table 4, where the top 2 features are the output of BLSTM and CRF. This means the output from two single classifiers are playing the dominant role rather than other features selected from feature ranking module.

Table 4. *The degree of features Contribution for IPH prediction on word-based GBDT2.*

Rank	Features description	Contribution (%)
1	**Output of BLSTM**	**42.8**
2	**Output of CRF**	**32.5**
3	POS	6.28
4	The predicted boundary from the lower level	5.67
5	The syllable distance to the end of the sentence	3.01

5. Conclusions

In this paper, we investigate the effects of enhanced embedding features and two model fusion approaches at both character and word level, as well as show their effectiveness for Mandarin prosodic boundaries prediction. Our results also show that word-based approach is more suitable than character-based approach and model fusion level is more effective than feature level fusion for Mandarin prosodic boundaries prediction. Meanwhile, the model fusion results indicate that the dependency of results on BLSTM is greater than CRF, and the features generated from feature ranking module can further boost the performance of prosodic boundaries prediction.

In the future, we wish to explore the use of our proposed enhanced embedding features and models fusion approaches for other aspects of prosody prediction such as pitch contour at the word or phrase level, as well as for predicting the spectral parameters for Mandarin speech synthesis.

6. 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.61305003, 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).

7. References

[1] Z. Chen, G. Hu, W. Jiang, “Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2010, pp. 1421-1424.

[2] M. Chu, Y. Qian, “Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts,” *Computational Linguistics and Chinese Language Processing,* vol. 6, no.1, pp.61-82, 2001.

[3] C. W. Wightman, M. Ostendorf, “Automatic labeling of prosodic patterns,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no.4, pp.469–481, 1994.

[4] A. Rosenberg, R. Fernandez, and B. Ramabhadran, “Phrase boundary assignment from text in multiple domains,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2012, pp. 2558-2561.

[5] S. Ananthakrishnan, S. S. Narayanan, “An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model,” in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2005, pp. 269–272.

[6] M. Hasegawa-Johnson, K. Chen, J. Cole, S. S. Cohen, A., et al, “Simultaneous recognition of words and prosody in the boston university radio speech corpus,” *Speech Communication*, vol. 46, no.3, pp. 418–439, 2005

[7] V. R. Sridhar, S. Bangalore, S. S. Narayanan, “Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework,” *IEEE Transactions on Audio, Speech and Language Processing,* vol. 16, no. 4, pp. 797–811, 2008.

[8] G. J. Busser, W. Daelemans, A. Van den Bosch, “predicting phrase breaks with memory-based learning”, *Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire Scotland, August 29th-September 1s*t, 2001.

[9] R. Fernandez and B. Ramabhadran, “Driscriminative training and unsupervised adaptation for labeling prosodic events with limited training data,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2010, pp. 1429-1432.

[10] A. Rosenberg, R. Fernandez, and B. Ramabhadran, “Modeling phrasing and prominence using deep recurrent learning,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2015, pp. 3066－3070.

[11] O. Watts, J. Yamagishi, and S. King, “Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2011, pp. 2157–2160.

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

[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,* 2014, pp. 1532–1543.

[14] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, “Word embedding for recurrent neural network based TTS synthesis,” in in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP,* 2015, pp. 4879–4883.

[15] Merritt T, Yamagishi J, Wu Z, et al, “Deep neural network context embeddings for model selection in rich-context HMM synthesis,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2015, pp.2207-2211.

[16] A. Vadapalli and K. Prahallad, “Learning continuous-valued word representations for phrase break prediction,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2014, pp. 41–45.

[17] C. Ding, L. Xie, J. Yan, et al, “Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features,” *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE*, 2015.

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

[19] A. Rendel, R. Fernandez, R. Hoory and B. Ramabhadran, “Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end,” in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2016.

[20] 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.

[21] C.-L. Liu, “Classifier combination based on confidence transformation,” *Pattern Recognition*, vol. 38, no. 1, 11-28, 2005.

[22] L. I. Kuncheva, “Combining Pattern Classifiers: Methods and Algorithms,” *Wiley Interscience*, 2004.

[23] “word2vec: the original word2vec using CBOW architectures,” *https://code.google.com/p/word2vec/*, 2013.

[24] 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,* 2014, pp. 2268–2272.

[25] 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,* 2015, pp. 1606–1610.

[26] 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,* 2014, pp. 1964–1968.

[27] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.

[28] Q. Hu, Z. Wu, K. Richmon, etc, “Fusion of multiple parameterizations for DNN-based sinusoidal speech synthesis with multi-task learning,” in *Annual Conference of the International Speech Communication Association, Interspeech,* 2015, pp. 854-858.

[29] 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.

[30] S. Lai, K. Liu, L. Xu, et al, “How to Generate a Good Word Embedding,” *Credit Union Times*, 2015.

[31] “CRF++: The original toolkit of version 0.58 CRF++,” *http://taku910.github.io/crfpp/,* 2013.

[32] Menze, H. Bjoern, et al, “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” *Bmc Bioinformatics,* 2008.