Using Tilt for Automatic Emphasis Detection with Bayesian Networks

Yishuang Ning[1,2], *Zhiyong Wu*[1,2,3], *Xiaoyan Lou*[4], *Helen Meng*[1,3], *Jia Jia*[1,2,*], *Lianhong Cai*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[4] Beijing Samsung Telecom R&D Center, Beijing 100081, China

ningys13@mails.tsinghua.edu.cn, {zywu,hmmeng}@se.cuhk.edu.hk, xiaoyan.lou@samsung.com, {jjia,clh-dcs}@tsinghua.edu.cn

Abstract

This paper proposes a new framework for emphasis detection from natural speech, where emphasis refers to a word or part of a word perceived as standing out from its surrounding words. Labeling emphatic words from speech recordings plays a significant role not only in human-computer interactions, but also in building speech corpus for expressive speech synthesis. Many previous researches use the global features to train their models, neglecting the efficiency of the local ones. In this paper, we introduce the tilt parameters which correspond to the phonetic prominence of an intonation event to our task. Besides, traditional approaches such as emphasis detection with support vector machines (SVMs) neglect the correlations between features, thus degrading the accuracy of emphasis detection. In this paper, we use Bayesian networks (BNs) which consider the dependency between features as detector. Experimental results demonstrate that BNs outperform the baseline and SVMs for the task. Specifically, by combining the tilt feature with the traditional segmental features and semitone, the proposed method yields an 11.6% improvement in emphasis detection accuracy as compared with the baseline and 2.2%-3.1% improvement with other feature combinations.

Index Terms: emphasis, emphasis detection, tilt, Bayesian networks (BNs), support vector machines (SVMs)

1. Introduction

Emphasis is an important feature of prosody and plays a very important role in human communications. Emphasis detection is to perceive or recognize the emphasized speech segments (that may correspond to a word or part of a word) from natural speech. Currently, the study of automatic emphasis detection has become a hot topic and represents one of the main streams in the related areas of human-computer interactions. Moreover, the automatic construction of large scale emphasis-annotated language resources has attracted great interest from both research purposes and industry perspectives. Automatic detection of emphasis has broad prospects, such as emphatic speech synthesis, automatic summarization of spoken discourse, content spotting, identification of focal elements, generation of improved prosody, language learning, and improved facial animation generation for interactive tutors.

Although manual emphasis annotation can obtain high accuracy, the annotating process not only is labor intensive and time-consuming, but also involves labelers' subjective interpretation of the sentences. To address the problems, a variety of automatic emphasis detection approaches have been proposed in recent years. Many previous researches on emphasis detection have focused on the automatic recognition of pitch accents (lexical stress), most of which have used traditional acoustic features such as logarithm of fundamental frequency (F0), duration and energy, as well as lexical features such as part-of-speech [1][2], contextual features [3]-[6] etc. [7] calculated the F0 difference between original speech and synthetic speech and then used pre-determined threshold to label emphasis. However, the selection of threshold affects the emphasis detection accuracy very much. [8] investigated the correlation between the pitch range of the second accent and its perceived prominence. There are also some literatures aiming at predicting word prominence in spontaneous speech with features like spectral emphasis or RASTA-PLP [9]. Although using segmental features can obtain good performance for some applications, they still cannot satisfy the need of other ones requiring high accuracy. One of the main problems is that these features are global ones which are statistically averaged at word level or syllable level. However, local features have rarely been considered. Previous work have shown that emphatic words usually have local prominence in speech. The work in [10] shows high accuracy for lexical stress detection using tilt parameters (a kind of local feature). Besides, [11] also demonstrates that pitch plateau (the extension of Taylor's rise/fall model) is found to outperform the traditional pitch statistics. In this paper, we attempt to use tilt parameters for emphasis detection.

In [12], emphasis (or prominence) detection is a strict two-class problem. Hence, various classification models such as naïve Bayes [6], SVMs [13] or ensemble machines [14] have been used. However, most of these researches have rarely considered the dependency and complementarity between different features. In this work, Bayesian networks [15], which can consider the correlation between features, are used as classifier for emphasis detection. Experimental results demonstrate the effectiveness of the proposed approach.

2. Corpus

To consider different situations of emphasis occurrences, a set of text prompts are carefully designed and their corresponding speech utterances are recorded [16][17].

The corpus includes 350 text prompts. Each of the text prompt contains one or more emphatic words. These emphatic words are located at different positions in the sentences. For the emphatic words, they might be monosyllabic or polysyllabic, with the primary stressed syllables at different

places in the words. Besides, the design of these text prompts considers all kinds of pronunciation mechanisms of phones. The context characteristics of the phones are also covered by the text prompts as many as possible.

For each text prompt, its corresponding speech utterance is recorded with expressive intonation to place proper emphasis on the emphatic words in the sentence. A female speaker with a high level of English proficiency was invited to record the contrastive speech utterances in a sound proof studio and the recorded utterances are saved in the wav format as sound files (16 bit mono, sampled at 16 kHz).

To extract F0 (pitch) contour, we used the ESPS get_f0 program which is derived from the algorithm presented in [18] and integrated into the HMM-based Speech Synthesis System (HTS) [19]. Smoothing is then performed in the F0 trajectory with the Edinburgh Speech Tools Library (EST) [20]. Finally, phone boundaries are located automatically by means of forced alignment with the HMM framework; and the prosodic events including pitch accents and boundary tones are labeled by Wavesurfer [21].

3. Features

The methods for emphasis detection described in the next section are based on the computation of the following traditional segmental features including log F0, energy and duration of phonemes, new segmental features such as semitone and intonational features such as tilt.

3.1. Traditional segmental features

Previous work shows that emphatic words usually have high F0, large energy and long duration. The work conducted in [22] also presents that F0 maximum, duration, energy and F0 range have high correlations with emphasis categories. Hence, we decide to include F0 related features (mean, min, max, range of log F0), energy related features (mean, min, max of energy) and duration.

Pitch tracking is done by the ESPS get_f0 program and smoothed by the EST tools. Phone boundaries are located automatically by means of forced alignment with the HMM framework. Duration for each phoneme is then calculated from the phone boundary information. We also computed the mean, min, max and range of log F0 for each phoneme. For energy, we first extracted the 13 dimensional Mel-frequency cepstrum coefficients (MFCCs) from which the 1st dimension (energy) for each frame is retrieved. We then calculated the mean, min and max of energy for each phoneme.

3.2. New segmental features

Research shows that the change of semitone is consistent with the distance of auditory perception. This indicates semitone may be more suitable for human auditory perception than original F0. Thus, we also choose semitone as one feature. Semitone can be calculated from F0 as follows:

$$S = 69 + 12\log_2\left(\frac{f}{440}\right) \tag{1}$$

where S is the semitone and f is the F0 value.

3.3. Intonational features

Though the global segmental features can generally detect emphasis with good performance, it is not enough for some applications that require high accuracy. One reason is that the acoustic characteristics of emphasis speech is not only influenced by the global segmental features, but also correlated with the local intonational ones.

In the long tradition of studies dealing with intonation profiles, people have proposed numerous models. The work in [23] introduced a two-level categorization of pitch profiles enriched by a wide combination of symbols and diacritics to represent all possible intonation contours and pitch accents. However, such a categorization, as well as the famous ToBI [24] labeling scheme, appears to be difficult to implement in an automatic system. Although this problem can be solved by the Fujisaki model [25], the model is still not perfect. For example, only minor gradient variation is allowed in the underlying phrase component, which is harmful for detection accuracy. To address the problem, [26] proposed a different view of intonation events with the rise/fall/connection (RFC) model. Our work follows this model and uses the tilt parameters to describe intonational characteristics.

The RFC model is a phonetic model of intonation that represents intonation as a sequence of continuously parameterized events. In the RFC model, each event is modeled by a rise part followed by a fall part. Each part has an amplitude and duration, and two parameters are used to give the time position of the event in the utterance and the F0 height of the event. However, although the RFC model can describe F0 contours accurately, the mechanism is not ideal in that the RFC parameters for each contour are not easy to interpret and manipulate since there are two rises and falls for each event. Therefore, starting from the RFC model, [27] proposed the tilt model, which defines a set of parameters capable of uniquely describing events in the pitch contour. The set consists of five parameters defined as follows:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{2}$$

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{3}$$

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2\left(|A_{rise}| + |A_{fall}|\right)} + \frac{D_{rise} - D_{fall}}{2\left(D_{rise} + D_{fall}\right)} \tag{4}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \tag{5}$$

$$D_{event} = D_{rise} + D_{fall} \tag{6}$$

where A_{rise}, A_{fall}, D_{rise}, D_{fall} are the amplitude and duration of the rise and fall segments of the intonation event.

The other two parameters are F0 position and time position, which are extracted with the EST tools.

4. Emphasis detection with Bayesian networks

4.1. Bayesian networks

4.1.1. Definition

A Bayesian network (BN) B is a network structure B_S, which is a directed acyclic graph (DAG) over a set of variables $U = \{x_1, x_2, ..., x_n\}$ and a set of probability tables $B_P = \{p(u \mid pa(u)), u \in U\}$ where $pa(u)$ is the set of parents of u in B. The purpose is to calculate the joint probability distribution (JPD) of a number of variables. However, doing

such computations directly would involve a potential very large joint probability table (JPT). One solution is to encode independence given its parents: each variable x_i is independent of its non-descendants given its parents. Thus, the JPD of B can be calculated as:

$$P(U) = \prod_{u \in U} p(u \mid pa(u)) \quad (7)$$

The learning algorithm of B involves two steps. Firstly, the network structure B_S is learned from the score metrics and search algorithms. Secondly, the probability tables are learned from maximum likelihood estimates.

4.1.2. *Score metrics*

There are various approaches, including local score metrics, global score metrics and fixed structure, to learn B_S, and in this work we used the local score metrics. The quality measure can be based on a Bayesian approach, minimum description length (MDL), information entropy and other criteria.

Let the entropy metric $H(B_S, D)$ of a network structure and database be defined as:

$$H(B_S, D) = -N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}} \quad (8)$$

and the number of parameters K as:

$$K = \sum_{i=1}^{n} (r_i - 1) q_i \quad (9)$$

where r_i $(1 \le i \le n)$ is the cardinality of x_i and $q_i = \prod_{x_j \in pa(x_i)} r_j$ is the cardinality of the parents set of x_i in B_S. Let N_{ij} $(1 \le i \le n, 1 \le j \le q_i)$ be the number of records in D for which $pa(x_i)$ takes its jth value and N_{ijk} $(1 \le i \le n, 1 \le j \le q_i, 1 \le k \le r_i)$ be the number of records in D for which $pa(x_i)$ takes its jth value and for which x_i takes its kth value.

Then the MDL metric $Q_{MDL}(B_S, D)$ of a Bayesian network structure B_S can be defined as:

$$Q_{MDL}(B_S, D) = H(B_S, D) + \frac{K}{2} \log N \quad (10)$$

and the Bayesian metric:

$$Q_{Bayes}(B_S, D) = P(B_S) \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \quad (11)$$

where $P(B_S)$ is the prior on the network structure and $\Gamma(.)$ the gamma function. N'_{ijk} and N'_{ijk} represent choices of priors on counts restricted by $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

4.1.3. *Search algorithms*

There are a variety of search algorithms for local score metrics.

- K2: hill climbing adding arcs with a fixed ordering of variables.
- LAGD hill climbing (LAGD): hill climbing with look ahead on a limited set of best scoring steps.
- Tabu search (TBS): using adding and deleting arrows.
- TAN: tree augmented naïve Bayes where the tree is formed by calculating the maximum weight spanning tree using Chow and Liu algorithm [28].

These algorithms attempt to maximize the score metrics to learn the network structure and we used the Weka [29] implementation of these algorithms. In the previous work, experiments showed that the TAN model works well in that it yields good classifiers compared to other search algorithms. Therefore, we choose the TAN search algorithm to learn our classifier.

4.2. Emphasis detection with Bayesian networks

As described in the introduction section, emphasis detection is actually a strict two-class problem. The classification task consists of classifying the class variable y given a set of feature variables $\mathbf{x}=x_1,x_2,\ldots,x_n$. Previous research indicates that a more accurate modeling of the dependencies between features leads to improved classification [15]. This means the performance of a classifier may be improved if the learning procedure takes into account the correlations between features. BNs offer a useful framework for learning such kind of structure.

Fortunately, the features used in our work also reveal this kind of correlation dependency. Firstly, as can be seen from Equation (1), semitone is tightly correlated with log F0. Secondly, the definition of tilt parameters indicates that there exists close relationships between tilt parameters and log F0 as well. The learned network also demonstrates this perspective. Figure 1 shows part of the network that is learned from all the features with the TAN search algorithm. In this network, each feature is a node and edges between pairs of nodes represent direct correlations between the features. Suppose x_i and x_j represent max of log F0 (maxlf0) and F0 position (f0position) respectively. An edge from x_i to x_j implies the influence of x_j on the assessment of the class variable also depends on the value of x_i where x_i is the parent of x_j. Besides, from Figure 1, we can also see the dependency between log F0 (meanlf0) and energy (meanenergy). Finally, this structure can be learned from data using BN search algorithms.

To use BN as classifier, we simply calculate $argmax_y P(y|\mathbf{x})$ using $P(U)$:

$$\begin{aligned} P(y \mid \mathrm{x}) &= P(U) / \mathrm{P}(\mathrm{x}) \\ &\propto P(U) \\ &= \prod_{u \in U} p(u \mid pa(u)) \end{aligned} \quad (12)$$

Figure 2 illustrates the framework of emphasis detection with BNs. Firstly, the emphatic text is processed into emphatic labels, including their contexts and whether current word is emphatic. Secondly, we do forced alignment with the emphatic labels and speeches through the HTS framework. The traditional segmental features such as F0 and energy are also extracted from the emphatic speeches. Then the semitone and tilt parameters are calculated from the F0 contour. Since there are only few emphatic words for each utterance, the emphatic phonemes will be very sparse. To deal with the imbalanced distribution between the positive and negative samples, the synthetic minority over-sampling technique (SMOTE) [30] is used to over-sample the positive instance class. Finally, the emphatic phonemes are recognized with the Bayesian networks classification procedure.

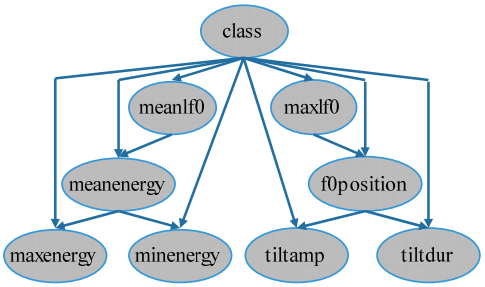

Figure 1: *Part of the Bayesian network (BN) learned from all the features with the TAN search algorithm*

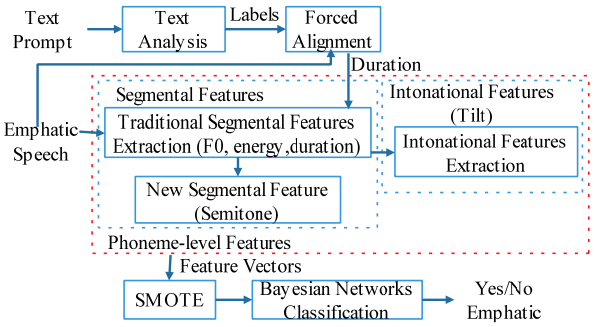

Figure 2: *The framework of emphasis detection with BNs*

5. Experiments and results

To evaluate proposed approach, we conducted two objective experiments for phoneme classification (i.e. emphasized or not) on the emphatic corpus and adopted the 10-fold cross-validation method (randomly use one fold for testing and the remaining 9 folds for training, repeat until all folds are used for testing) to avoid over-fitting. The first experiment is the comparison between the SVM and BN using the features introduced in Section 3. The second one is the validation of the newly proposed features by comparing the performance of different features using BN. We used SMOTE to handle the imbalanced distribution between the positive and negative samples and adopted the Weka implementation for SVM and BN in our experiments. In SMOTE, the number of the nearest neighbors is set to be 5. For each classification experiment, we compute emphasis detection accuracy, precision and recall of phoneme classification as the performance measurement and set up a simple baseline which always predicted the instances as negative. The data in each cell of the tables is the weighted average value of the two classes.

In these experiments, two feature sets with or without new features are used to compare the performance between SVM and BN. One feature set, called **traditional feature set**, includes all traditional segmental features, while the other, called **new feature set**, contains the new segmental feature (i.e. semitone) and the intonational ones (i.e. tilt) in addition to the traditional feature set. Table 1 indicates the results of the first experiment. It can be seen that all the three measurements for BN are much higher than SVM with and without SMOTE for the two feature sets. By using the BN classifier without and with SMOTE, the accuracy is improved by 4.1% and 17.4% respectively for traditional feature set. This trend is almost the same on the new feature set. The results demonstrate that BN performs significantly better than SVM. The reason for this is the correlation between features has been considered in BN, thus promoting the detection performance. Furthermore, with the help of SMOTE, the performance is improved significantly.

For the purpose to justify whether the local intonational features such as tilt can help detect emphasis from utterances, we conducted another experiment to compare the performance of different feature combinations using BN with SMOTE. Table 2 summarizes the results of this experiment. As can be seen that all the features are effective, especially the log F0 and semitone. The performance of the lf0 and semitone is almost the same and higher than tilt. However, when the tilt feature is combined with all the other features, it achieves the best performance and yields an 11.6% improvement as compared with the baseline and 2.2%-3.1% improvement as compared with the other feature combinations. Therefore, the experimental results verify that the tilt feature can compensate the deficiency of the global segmental features and help detect emphasis in our problem.

Table 1. *Performance of SVM and BN based detector with/without SMOTE on traditional and new feature sets.*

Model	Feature Set	Accuracy	Precision	Recall
Baseline	---	75.5%	---	---
$SVM_{Original}$	traditional	75.5%	57.1%	75.5%
$BN_{Original}$	traditional	79.6%	78.2%	79.6%
SVM_{SMOTE}	traditional	66.8%	67.4%	66.8%
BN_{SMOTE}	traditional	84.2%	84.4%	84.2%
$SVM_{Original}$	new	79.4%	79.7%	79.4%
$BN_{Original}$	new	82.9%	82.6%	82.9%
SVM_{SMOTE}	new	71.3%	71.4%	71.3%
BN_{SMOTE}	new	**87.1%**	**87.1%**	**87.1%**

Table 2. *Performance of emphasis detection based on BN with SMOTE using different feature combinations. Traditional (nolf0) represents traditional feature set without log F0.*

Feature Combinations	Accuracy	Precision	Recall
Baseline	75.5%	---	---
traditional	84.2%	84.4%	84.2%
traditional(nolf0)+semitone	84.4%	84.8%	84.4%
tilt	80.2%	80.2%	80.2%
traditional(nolf0)+tilt	84.0%	84.0%	84.0%
traditional+tilt	84.9%	84.9%	84.9%
traditional(nolf0)+semitone+tilt	84.5%	84.5%	84.5%
traditional+semitone	84.6%	84.7%	84.6%
traditional+semitone+tilt	**87.1%**	**87.1%**	**87.1%**

6. Conclusions and future work

This paper focuses on the problem of emphasis detection from natural speech. To investigate the effecificy of the local features, we introduce the tilt parameters which represent the intonational characteristics within the utterance into our task. The tilt parameters are originated from the RFC model and are an abstract description of the F0 shape of an event which can be divided into pitch accents and boundary tones. Using the combination of this feature with the traditional segmental features and semitone, the emphasis detection performance is improved significantly. Besides, the experiments demonstrate that BNs outperform SVMs experimentally. The reason is that BNs can take advantage of the correlations between features. To further improve detection accuracy, SMOTE is introduced. Experimental results validate the effectiveness of our approach.

Our future work will be committed to the automatic recognition of intonation events and incorporate the deep learning methodology into our framework to learn these features for further improving the detection accuracy. We will also conduct additional experiments on other materials/corpora.

7. Acknowledgements

This work is supported by National Basic Research Program of China (2012CB316401, 2013CB329304) and National High Technology Research and Development Program of China (2015AA016305). The work is also partially supported by Hong Kong SAR Government's Research Grants Council (N-CUHK414/09), National Natural Science Foundation of China (61375027, 61433018 and 61370023) and Major Program for National Social Science Foundation of China (13&ZD189).

8. References

[1] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proc. 42nd Meeting of the Association for Computational Linguistics (ACL)*, pp. 677-683, 2004.

[2] A. Rosenberg, "Automatic detection and classification of prosodic events," *Ph.D thesis*, Columbia University, 2009.

[3] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, 2005.

[4] J.M. Brenier, D.M. Cer and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 3297-3300, 2005.

[5] L.Y. Chen and J. S. Jang, "Stress detection of English words for a CAPT system using word-length dependent GMM-based Bayesian classifiers," *Interdisciplinary Information Sciences*, vol. 18, no. 2, pp. 65-70, 2012.

[6] J.Y. Chen and L. Wang, "Automatic lexical stress detection for Chinese learners' of English," in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 407-411, 2010.

[7] Y. Maeno, T. Nose and T. Kobayashi, "HMM-based emphatic speech synthesis using unsupervised context labeling," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 1849-1852, 2011.

[8] D.R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 725, pp. 313-342, 1997.

[9] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 2640-2644, 2014.

[10] J.H. Zhao, H. Yuan, J. Liu and S.H. Xia, "Automatic lexical stress detection using acoustic features for computer-assisted language learning," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2011.

[11] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 690-701, 2007.

[12] F. Tamburini, "Prosodic prominence detection in speech," in *Proc. International Symposium on Signal Processing and its Applications (ISSPA)*, Paris, pp. 385-388, 2003.

[13] J.F. Wang, G.M. Chang, J.C. Wang and S.C. Lin, "Stress detection based on multi-class probabilistic support vector machines for accented English speech," *Computer Science and Information Engineering*, vol. 7, pp. 346-350, 2009.

[14] X.J. Sun, "Pitch accent prediction using ensemble machine learning," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, pp. 16-20, 2002.

[15] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131-163, 1997.

[16] Z.Y. Wu, Y.S. Ning, X. Zang, J. Jia, F.B. Meng, H. Meng and L.H. Cai, "Generating emphatic speech with hidden Markov model for expressive speech synthesis," *Multimedia Tools and Applications*, Springer, 2014, DOI: 10.1007/s11042-014-2164-2.

[17] F.B. Meng, Z.Y. Wu, J. Jia, H. Meng and L.H. Cai, "Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 463-489, DOI: 10.1007/s11042-013-1601-y, Springer, 2014.

[18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, New York, Elsevier, pp. 495–518, 1995.

[19] HMM-based Speech Synthesis System (HTS) [OL].[2014-03-28]. http://hts.sp.nitech.ac.jp/.

[20] Speech Tools [OL]. [2014-03-29]. http://www.cstr.ed.ac.uk/projects/speech_tools/index.html.

[21] WaveSurfer User Manual [OL]. [2014-11-25]. http://www.speech.kth.se/wavesurfer/man.html.

[22] F.B. Meng, H. Meng, Z.Y. Wu and L.H. Cai, "Synthesizing expressive speech to convey focus using a perturbation model for computer aided pronunciation training," in *Proc. Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, pp. 22-27, 2010.

[23] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," *Language Sound Structure*, 1984.

[24] J. Pitrelli, M. Beckman and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Japan, pp. 123-126, 1994.

[25] H. Fujisaki, K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," *Journal of the Acoustical Society of Japan*, vol. 5, no. 4, pp. 233-241, 1984.

[26] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, pp. 169-186, 1994.

[27] P. Taylor, "The tilt intonation model," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Syndey, 1998.

[28] C.K. Chow and C.N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. on Infomation Theory*, vol. 14, pp. 426-467, 1968.

[29] Weka 3-Data Mining with Open Source Machine Learning Software in Java [OL]. [2014-10-19]. http://www.cs.waikato.ac.nz/ml/weka/.

[30] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.