# Computer Aided Evaluation of Intonation for Language Learning Based on Prosodic Unit Segmentation

Sixuan ZHAO[1], Kang Kwong LUKE[2], Soo Ngee KOH[1], Yang ZHANG[1]
[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2] School of Humanities and Social Sciences, Nanyang Technological University, Singapore

**Abstract—This paper considers computer aided intonation evaluation based on prosodic unit segmentation for the learning of English. To evaluate the intonation of utterances of a foreign English learner, the learner's intonation pattern is normally compared with that of the teacher's utterance. The evaluation score can be obtained by measuring the "distance" between the two intonation patterns. One salient feature of most current computer-aided language learning (CALL) systems is segmentation of an utterance according to word or syllable boundaries for distance measurement in the evaluation process. This method may lead to inaccuracy of the evaluation, since the prosody of natural speech always corresponds to the boundaries of prosodic units, such as foot and intonation unit, rather than the word boundaries in an utterance. In this paper, the pronounced sentences are segmented according to prosodic unit boundaries. Dynamic Time Warping (DTW) and Mahalanobis Distance are then used to measure the difference between the learner's and the teacher's intonation, leading to a machine score.**

***Index Terms*—prosody, intonation, prosodic unit, foot, segmentation.**

## I. INTORODUCTION

Computer-aided language learning systems have been in existence for some time. However, their limitations hinder their widespread adoption. More research and development effort is being made by various groups in the world to improve their reliability. With proliferation of computer hardware and software plus the increasing demand for foreign language learning, CALL is experiencing some revival in interest.

Compared to traditional classroom English learning, a CALL system has three major advantages: 1. Self-adapting learning: It enables language learners to pace their study according to their own schedule and ability. 2. Anxiety-free practice: It allows learners practice their second language without direct interaction with other people and thus reducing the level of anxiety. 3. Personalized study: Learners can choose the most suitable learning method from a host of different methods offered by a CALL system. Considering these merits, it is potentially efficient and effective to use a CALL system for English training and learning, especially for the acquisition of speaking skills. Currently, most CALL systems for speech training mainly target pronunciation training from phoneme to word level. However, prosody is also very important because it not only expresses the emotion and intention of a speaker, but also reflects his phrasing skills. Furthermore, for CALL systems that have prosody evaluation features, the evaluation accuracy is still not very high. Therefore, issues pertinent to prosody evaluation for language learning deserve to be studied.

Prosody can be roughly divided into intonation and rhythm, which reflect phrasing and prominence information, respectively [1]. In this paper, we mainly focus on the intonation of an utterance. For intonation evaluation, one important issue is the segmentation of uttered sentences. Before the evaluation, all the pronounced sentences should be segmented into small units appropriately. This step is critical because much details of prosody will be lost and the result will be inaccurate if the intonation is evaluated at the sentence level. Appropriate segmentation is therefore an important aspect in designing a CALL system.

In current systems, segmentation is mainly based on word or syllable boundaries [2], [3], [4]. This means that sentences are segmented according to the words and syllables rather than prosody. Though such a segmentation method is suitable for pronunciation evaluation, it may not be suitable for intonation evaluation. Unlike lexical units which include phoneme, syllable and word, prosody is a supra-segmental unit in a sentence and it may not be related to the lexical boundaries strictly. Hence, if we segment a sentence according to lexical or syllabic units, the evaluation results may fail to reflect the learner's mastery of phrasing skills.

One logical solution is to shift the segmentation unit from the lexical domain to the prosodic domain, i.e., the basic unit for segmentation should be a prosodic unit. Unlike lexical units which mainly affect the lexical and syntactic meaning of an utterance, prosodic units can reflect the emotion, intention and rhythm of an utterance. Therefore, segmentation based on prosodic units seems more reasonable for prosody evaluation. In this paper, segmentation based on two different prosodic units, namely foot and intonation unit (IU) are discussed and implemented for evaluation of the intonation of English utterances.

In section II, the definitions and hierarchy of prosodic units will be discussed. Foot and IU will be considered and reasons to select them will be presented. In section III, the basic framework of the system and the distance measurement method will be given. In section IV, some experimental



*Proceedings of the Second APSIPA Annual Summit and Conference*, pages 788–793,
Biopolis, Singapore, 14-17 December 2010.

conditions and results will be reported. Finally, the conclusion of this paper and proposals for further research will be given.

## II. PROPOSED PROSODIC UNITS

To introduce prosodic units to the segmentation process, it is necessary to discuss the hierarchy of prosodic units. Prosody is a supra-segmental layer of speech which consists of pitch, duration and intensity. It is generally used by a speaker to organize phonetic segments (vowels and consonants) into systematic units of various sizes. Since prosodic features are supra-segmental, many different levels exist, e.g., syllable, foot, phonological word, clitic, intonation unit, declination unit and utterance (from the lowest level to the highest level) [5]. Hence, when adopting prosodic units in the segmentation of utterances, the first problem is to select the most appropriate units. In this paper, two different prosodic units are considered, namely foot and intonation unit.

### A. Foot

In English, the foot is defined as a phonological unit consisting of an accented syllable followed by any number of unaccented syllables. Specifically, a foot starts from the beginning of a stressed syllable to the beginning of the next stressed syllable. A phenomenon which should be noticed is that feet are mainly delimited by the stresses in a sentence. As a result, foot boundaries may not correspond to word boundaries. Thus, one word may contain multiple feet or one foot may consist of multiple words. For instance, the sentence "*I felt that I might never stop the machine from running*" and its pitch and intensity contours (extracted by *Praat*) are shown in Fig.1 (with the blue line indicating the pitch contour and the green line indicating the intensity contour). The foot segmentation is "*I/ felt that I /might/ never s/top the ma/chine from/ running*" (delimited by vertical lines). Clearly, in this sentence, the boundary of a foot may locate inside a lexical word.

One exception to the foot level definition is anacrusis. Since each foot is combined by a stress and the following unstressed parts, it is necessary to define the component before the first stress. Anacrusis is defined as one or more unstressed syllables preceding the first stressed syllable in an utterance. For example, in the sentence "*We/only/spoke for a/short/time*", the word "*we*" is defined as the anacrusis here. Because anacrusis is unstressed and always short, it should be assigned a smaller weight when doing distance measurement between two intonation patterns.

### B. Intonation Unit

The intonation unit is defined as a part of a speech utterance which possesses a continuous pitch contour. It is located at a higher prosodic level than that of the foot. However, such a definition is unclear in many situations. One problem is that due to current pitch detection technique, it is difficult to use the continuity of the pitch contour to make a decision. Even in the most accurate pitch detection system, erroneous pitch values may appear sometimes and thus interrupt the continuity. As a result, a more accurate definition from a linguistic perspective should be considered. In linguistics, an intonation unit usually corresponds to a sense group (or word group), which may contain several syllables, with some of them stressed and some unstressed. The nucleus, the most prominent syllable of the intonation unit and always the last one, is usually a marker of the highest importance and has focal stress. Hence, the segmentation on intonation can be done according to the location of a nucleus. Two components, intensity value and pitch variation, are always considered in the detection of a nucleus.

However, there is also some ambiguity for such a definition of the IU due to different variations of the pitch contour. For instance, in Fig.2, the sentence "*for the doctor was meeting some friends for dinner soon after*" and its pitch contour are shown. We can find that there are three prominences (all indicated by vertical lines) which can be considered as candidates of nucleus. As a result, there can be two definitions of IU in such a situation. The first one is the "narrow" definition. Based on this definition, all the three prominences are taken as nucleus. For each nucleus, there can be an intonation unit in the "narrow" sense of the definition (delimited by slashes in the transcription). The second one is the "broad" definition. Contrary to the "narrow" definition, only the last prominence in the whole utterance is considered as nucleus. Therefore, the whole sentence can be seen as one "broad" intonation unit with the nucleus located at the last pitch variation. Hence, there is an inconsistency on whether

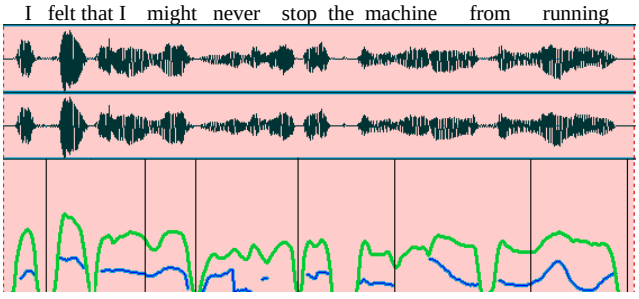


Fig.1. Waveform, Pitch and Intensity of *I/ felt that I/ might/ never s/top the ma/chine from/ running*

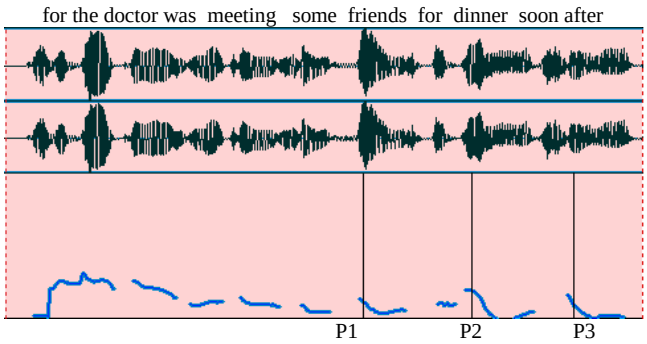


Fig.2. Waveform and Pitch of "*for the doctor was meeting some friends/ for dinner/ soon after/*"

the "narrow" or "broad" definition of IU should be adopted. In our system, we propose to use the "narrow" definition, because it can provide a more detailed segmentation of the sentence. Therefore, the IU segmentation of the sentence in Fig.2 is "for the doctor was meeting some friends/ for dinner/ soon after/", which is different from its foot level segmentation, "for the/ doctor was/ meeting some/ friends for/ dinner soon/ after".

### C. Discussion on Foot and Intonation Unit

Foot is selected for its status in the prosodic domain and correlation with stresses. First, to segment a pronounced sentence appropriately, it is obvious that the segmentation units should not be too long; otherwise, it may give rise to problems for short sentences and reduce the accuracy of evaluation results. The level of foot in the prosodic domain generally coincides with that of word in the lexical domain. It means that foot is a suitable unit to measure intonation. Second, the definition of foot pertains to stresses, which contain a lot of rhythm information. An important phenomenon is that, in the same sentence spoken by the same speaker, there is a tendency to keep the length of each foot not far away from the norm (relative to the tempo at the moment of the utterance). Accordingly, feet in a sentence can express significant rhythm information, which may contribute to the accuracy of evaluation results. Finally, when automatic segmentation is done at the foot level, many developed techniques can be used since much of previous researches have been done in the stress detection field [6].

Unlike the foot which reflects the intonation in a limited region, an intonation unit contains more information about global pitch variation and intensity change. It thus reflects higher level prosodic information of pronounced sentences. Therefore, it may function as a supplement for foot level segmentation in some situations.

According to the discussion above, foot segmentation is adopted in this paper for its appropriate length and facilitation for the design of an automatic segmentation system. Intonation unit segmentation may be further explored and adopted in future for evaluation of very long sentences which are difficult for human evaluation.

## III. PRINCIPLE AND METHOD

### A. Framework of Evaluation System

The basic flowchart of intonation evaluation is shown in Fig.3. In this paper, intonation evaluation can be done as follows.

First, for each learner's utterance, the corresponding teacher's utterance is selected. Foot or intonation unit segmentation is then performed on both of the teacher's and the learner's utterances. Currently, to guarantee the accuracy of such new segmentation, faculties from School of Humanities and Social Sciences in Nanyang Technological University (NTU) are invited to do the segmentation manually. In future, stress detection techniques [6] may be employed to develop automatic segmentation system. Since the prosodic unit boundaries are determined by the teacher's intonation, segmentation of the learner's utterances should follow the way

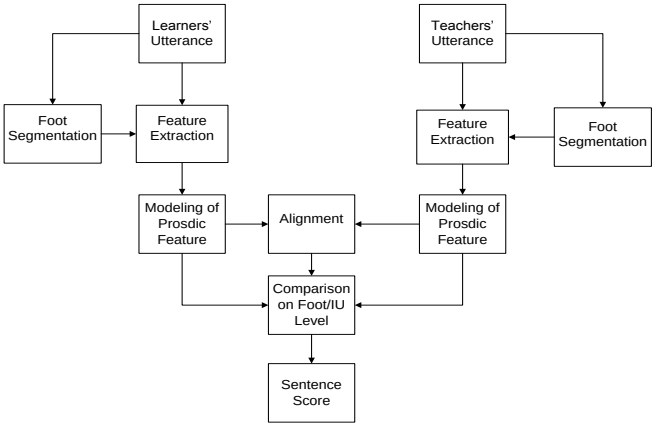


Fig.3. Outline of Evaluation System

of segmentation done on the corresponding teacher's utterances.

Second, feature extraction should be implemented on each unit. In this paper, a 4-dimensional feature vector is adopted. It includes normalized log pitch, normalized log intensity and their first derivatives. The function of log is to simulate the human auditory system. As the human auditory system perceives tones in a logarithmic way (instead of in a linear way), the logarithmic scale is a more appropriate measure of the perceptual consequences of differences in intonation. In addition, average and maximum values are subtracted from log pitch and log intensity for normalization.

After obtaining the intonation sequences of the teacher and the learner, DTW is implemented to align the two sequences. Then, the accumulative distance between the teacher's and the learner's intonation sequences is calculated unit by unit to obtain the unit score. Finally, the overall sentence score can be obtained by averaging all the unit scores.

### B. Distance Measurement

The intonation score is calculated based on the distance between the learner's and the teacher's intonation sequences. Rather than using the well-known Euclidean distance, the Mahalanobis distance is used in the calculation.

In previous papers [2], [3], Euclidean distance is used as a distance measure. However, it is noticed that the variances of the features in the feature vectors are different, e.g., the variance of log intensity is larger than those of the other three features. To resolve this problem, Mahalanobis distance is used as proposed in [7]. The Mahalanobis distance between two feature vectors $s$ and $t$ with covariance matrix C is calculated as follows:

$$d_M(s,t,C) = \sqrt{(s-t)^T C^{-1}(s-t)} \qquad (1)$$

Let $s_j(n)$ and $t_j(n)$ be the *n*-th feature vector of the *j*-th prosodic unit (e.g. foot) of the learner's and the teacher's utterance, respectively. Assuming $M_j(n)$ is the Mahalanobis distance of the *n*-th frame of the *j*-th prosodic unit, the machine score can be calculated as follows:

$$M_j(n) = d_M(s_j(n), t_j(n), C) \quad (2)$$

$$U_j = \sum_{n=1}^{N_j} M_j(n) \quad (3)$$

$$S = \sum_{j=1}^{J} U_j \quad (4)$$

where $N_j$ and $J$ are the number of frames of the *j*-th prosodic unit and the number of units in a sentence, respectively. $U_j$ and $S$ represent the accumulative distance of the *j*-th prosodic unit and that of the whole sentence. The covariance matrix C can be calculated as:

$$C = \frac{1}{N-1} \sum_{j=1}^{J} \sum_{n=1}^{N_j} (t_j(n) - \bar{t})(t_j(n) - \bar{t})^T \quad (5)$$

where $\bar{t}$ is the average of the teacher's feature vector.

In this way, the distance between each pair of the teacher's and the learner's intonations can be calculated. The distance can be used to indicate the similarity of two intonations, thus giving the evaluation score of the learner's prosody. To test the accuracy of the evaluation results, subjective scores of each utterance done by human evaluators should be obtained. The correlation coefficient between human evaluation scores and machine measured distance can then be calculated to assess the performance of such a system.

## IV. Evaluation Experiment and Analysis

### A. Experimental Conditions and Results

To test the validity of our proposal, a preliminary experiment based on foot level segmentation with a limited corpus has been performed. In the experiment, sentences from an audio book pronounced by a native speaker are used as the teacher's utterances. The experimental conditions are described in Table I.

Six unique sentences are used. They include two long sentences and four short sentences; five of which are statements and one is a question. Pronounced sentences from NTU students whose native language is not English are recorded for evaluation. Eight students were invited to record their utterances which are taken as the learner samples.

Table I
Conditions of the Evaluation Experiment

| | |
|---|---|
| Database | Audio Book from Native Speakers |
| Learners | 8 students from School of Humanities and Social Science and School of Electrical and Electronic Engineering in NTU |
| No. of Sentences | 6 different sentences with 2 long sentences and 4 short sentences |
| Intonation Feature Vector | normalized log pitch, normalized log intensity, and their first derivatives |
| Subjective Evaluation Score | 10 point evaluation score (from 1 to 10) |

All the recorded utterances are segmented manually according to the foot boundaries of the teacher's utterances. After segmentation, feature vectors of each prosodic unit are extracted for distance measurement. One of the features is the pitch contour of voiced speech which could be obtained using a sample based technique [8] or a frame based technique [9].

Subjective scoring was done by a faculty from School of Humanities and Social Sciences and an engineering faculty in NTU. The subjective evaluators were asked to evaluate each of the utterances of the students and to give a score ranging from 1 being the worst to 10 being the best.

With the obtained accumulative distance and the subjective scores, the human-machine correlation is calculated and given in Fig.4. The correlation between human evaluators is also given in the same figure. Since smaller distance value means higher evaluation score, the correlation coefficient should always be negative. In the figure the correlation coefficients are shown as positive for convenience.

The correlation coefficients based on foot level segmentation are around 0.39. In contrast, experiments with the same conditions performed at word level give the correlation coefficients of around 0.36. Also, compared to other intonation evaluation systems based on word level segmentation [3], [7], and in the situation where word important factor is not used, the correlation coefficient obtained by foot level segmentation is slightly better. The obtained results also show that foot level segmentation outperforms word level segmentation in our experiment. However, the correlation is obtained with a limited corpus. More unique sentences and utterances will be used to perform a thorough test to assess the effectiveness of the intonation evaluation system based on prosodic unit segmentation in future.

### B. Analysis

The main difference between the proposed evaluation method in this paper and the previously reported ones lies in the way the utterances are segmented before the distance measures are obtained. Reasons and advantages for using foot level segmentation are analyzed and discussed below.

Segmentation based on prosodic unit imitates the way human experts segment an utterance in their minds. To human listeners, the prosody information is naturally obtained from the pitch, intensity and duration of each part of an utterance, and it is only partially aligned with the lexical information.

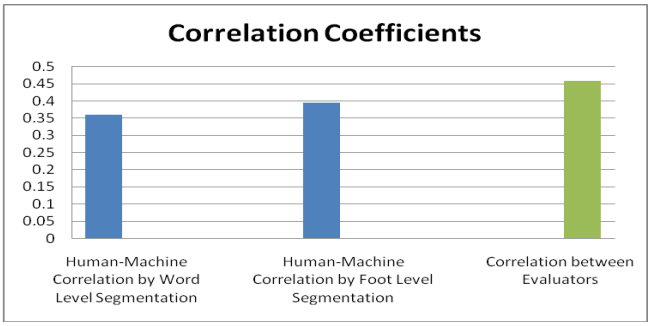


Fig.4. Evaluation Result

For example, when listening to a foreign language, the listener can still extract the prosody information via stress and rhythm to gauge the intentions and emotions of a speaker though he may not understand the meaning of each uttered word. Therefore, segmentation based on foot could be a good simulation of the segmentation in human minds. As a result, the human-machine correlation might be improved.

Second, from section II, it is clear that the definition of foot correlates to the locations of stresses tightly. Also, the location of stress is always determined by the intensity and pitch values. In the paper [10] on stress and pitch accent detection, a stress or pitch accent can be estimated from the combination of pitch, intensity and duration. In the stressed parts of a speech utterance, the pitch and intensity variations always exist. Therefore, variations of intonation features are always found in each prosodic unit. With such variations, the alignment between two intonations by DTW can be improved, since more shape information is provided.

For instance, we consider the sentence "*I/ felt that I/ might/ never s/top the ma/chine from/ running*" (slashes delimit foot boundaries), which is one of the sentences used in the evaluation experiment. The segmented pitch and intensity contours in both foot and word level are shown in Fig.5.

From this figure, it is clear that the word boundaries and foot boundaries may not coincide with each other. Besides, there is always a pitch variation (falling, rising or both) in each foot. In contrast, in some words, such as "that" and the second "I", the pitch contours are relatively flat, which reduce the accuracy of alignment. On the other hand, foot level segments contain more information on the variation of features in each unit. As discussed in section III, DTW is used in the alignment between two intonation patterns. Besides, the feature vector of intonation consists of pitch, intensity and their first derivatives. Therefore, alignment in this case also takes into account the shape contours of both pitch and intensity of each segment. If the feature contours of both intonations are flat, the derivative of features may fail to contribute to the alignment process, leading to inaccuracy of alignment.

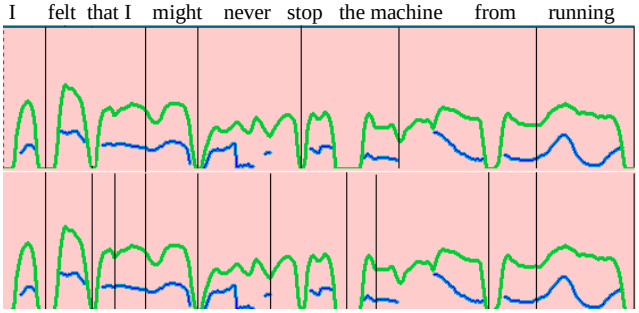


Fig.5. Segmented Pitch and Intensity contour of "*I/ felt that I/ might/ never s/top the ma/chine from/ running"* (the upper one is foot level segmentation and the lower one is word level segmentation, all delimited by vertical lines)

## V. Conclusion

In this paper, segmentation based on prosodic units for intonation evaluation is proposed. Since each prosodic unit corresponds to the phrasing information appropriately, such segmentation is similar to the way human experts evaluate the intonation. Also, as the variation of intonation features in each prosodic unit contributes to the alignment process, evaluation results are thus improved. In our preliminary experiment with a limited corpus, the human-machine correlation obtained by segmentation based on foot outperforms the correlation by word level segmentation.

Even though we have discussed the reasons for choosing foot and IU in part II of this paper, other prosodic units should also be considered to reflect the phrasing information at different levels. In particular, the evaluation scores obtained by segmentations based on different prosodic units can be combined to give an overall evaluation of the intonation.

In addition, an important factor which could be included in our future system is the pitch accent. From [10], the accentuation of words can be estimated from pitch, duration and intensity of a speech utterance. Since the features to be used are the same and the accented words correlate to stresses in an utterance, it is helpful to consider accented words in prosody evaluation. One straightforward and effective way of accented word detection is logistic regression based on pitch, duration and intensity. The next step is to combine word accent detection results with the proposed scheme presented in this paper to improve the human-machine correlation.

Other existing schemes for prosody evaluation could also be used to enhance the performance of our scheme. This includes the use of DTW alignment incorporating MFCC [1] and word important factor [3].

## Ackownledgement

The authors would like to thank the Institute of Media Innovation of Nanyang Technological University for providing the research grant and support for this project.

## References

[1] Juan Pablo Arias a, Nestor Becerra Yoma a, Hiram Vivanco, "Automatic Intonation Assessment for Computer Aided Language Learning", *Speech Communication*, vol. 52, no. 3, pp. 254-267, Mar. 2010.

[2] Akinori Ito, Tomoaki Konno, Masashi Ito and Shozo Makino," Evaluation of English Intonation based on Combination of Multiple Evaluation Scores", *INTERSPEECH-2009*, pp. 596-599.

[3] Motoyuki Suzuki, Tatsuki Konno, Akinori Ito and Shozo Makino, "Automatic Evaluation System of English Prosody for Japanese Learner's Speech", *Proc. 5th Int. Conf. Education and Information Systems*, Technologies and Applications (EISTA), 2007.

[4] Kim and W. Sung. "Implementation of intonational quality assessment system", *INTERSPEECH-2002*, pp. 1225–1228, Sept. 2002.

[5] Anthony Fox, "*Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*", Oxford: Oxford University Press, 2000.

[6] Kazunori Imoto, Yasushi Tsubota, Antoine Raux, Tatsuya Kawahara, Masatake Dantsuji, "Modeling and automatic detection of English sentence stress for computer-assisted English prosody

learning system", *Acoustical science and technology* 24(3), 159-160, May 2003.

[7] Konno, T. Ito, A. Ito, M. Makino, S. Suzuki, M. Grad. Sch. of Eng., Tohoku Univ., Sendai, "Intonation evaluation of English utterances using synthesized speech for Computer-Assisted Language Learning", *Int. Conf. Natural Language Processing and Knowledge Engineering*, pp. 1-7, Oct. 2008.

[8] Qiu L J, Yang H Y and Koh S N, "Fundamental-frequency determination based on instantaneous frequency estimation", *Signal Processing*, vol. 44, no. 2, pp. 233-241, June 1995.

[9] Sun, X., "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proc. of ICASSP2002*, Orlando, Florida, vol. 1, pp. 333-336.

[10] Andrew Rosenberg and Julia Hirschberg, "Detecting pitch accents at the word, syllable and vowel level", *Proc. of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, pp. 81-84.