LETTER

Study of Prominence Detection Based on Various Phone-Specific Features

Sung Soo KIM[†a)], **Chang Woo HAN**[†b)], *Nonmembers*, *and* **Nam Soo KIM**[†c)], *Member*

SUMMARY In this letter, we present useful features accounting for pronunciation prominence and propose a classification technique for prominence detection. A set of phone-specific features are extracted based on a forced alignment of the test pronunciation provided by a speech recognition system. These features are then applied to the traditional classifiers such as the support vector machine (SVM), artificial neural network (ANN) and adaptive boosting (Adaboost) for detecting the place of prominence.
key words: *prominence, phone-specific, SVM, ANN, Adaboost*

1. Introduction

An important application of speech recognition technologies is the computer-assisted language learning (CALL) system. A key component of a CALL system is the pronunciation evaluation module which assesses the test speaker's pronunciation in terms of intonation, prominence and articulation. This pronunciation evaluation is generally achieved by comparing the variation in pitch, energy and duration observed in the utterance with those of the given reference pronunciation.

It is difficult to give a simple but rigorous definition of prominence. However, in the related area, it is generally accepted that the prominence corresponds to words or syllables that are perceived as standing out from their environment [1]. A number of attempts have been made to detect the prominences in the given pronunciation. Tamburuni [2] proposed an algorithm in which an unsupervised learning approach is applied to a set of acoustic features extracted from the test pronunciation. In [3], a system utilizing acoustic measures was proposed without applying any speech recognition algorithms. To our knowledge, there have been very few, if any, studies in prominence detection that are based on the speech recognition technologies.

In this letter, we propose a prominence detection algorithm which is based on an extended class of features. We apply the hidden Markov model (HMM) to segment the given pronunciation into phone-sized units. This forced alignment enables us to extract phone-specific features and improves the performance of prominence detection. Once the phone-specific features are extracted, we apply the conventional classifiers to detect the prominences. As the classifiers, we apply the support vector machine (SVM), artificial neural network (ANN) and adaptive boosting (Adaboost), which are trained in a supervised manner. From a number of prominence detection experiments, it has been observed that the proposed features are effective to improve the performance.

Manuscript received February 18, 2010.
Manuscript revised April 12, 2010.
[†]The authors are with the School of Electrical Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 151–742, Korea.
a) E-mail: sskim@hi.snu.ac.kr
b) E-mail: cwhan@hi.snu.ac.kr
c) E-mail: nkim@snu.ac.kr
DOI: 10.1587/transinf.E93.D.2327

2. Features for Prominence Detection

The first step in our prominence detection algorithm is to apply the HMM for the purpose of segmenting the given utterance into phone-sized units. If the transcript of the utterance is available, forced alignment is carried out with reference to the model concatenating the corresponding phone HMMs. If, on the other hand, the transcript is not given, traditional phone recognition is performed in conjunction with the phone boundary information. It is noted that our purpose is to segment the pronunciation into phone-sized units not to find the exact phone identity of each segment. Selecting good features is very important to build a successful prominence detection system. It has been reported that the duration, energy and pitch are useful to spot the prominence. In this section, we present a number of features extracted from each phone-sized segment.

2.1 Duration

In general, the duration of a prominent phone tends to be longer than that of a non-prominent phone. Since, however, the duration of a phone is affected by the learner's mother tongue and speaking rate, it needs to be normalized. In order to normalize the phone duration, we apply a measure of rate of speech (ROS), which indicates the average number of phones uttered by the test speaker per a unit time [4]. Let $\hat{d}(i)$ denote the normalized duration of the i-th phone segment. Then,

$$\hat{d}(i) = d(i) \times ros_a \tag{1}$$

where $d(i)$ represents the phone duration and ros_a is the ROS for the speaker a.

2.2 Energy

People usually tend to speak a phone louder when they want to emphasize it. For that reason, the energy of the pronunciation serves a good acoustic parameter together with the

duration. In our study, we extract three energy-related features.

First, the root mean square (RMS) energy is considered as a parameter. The RMS energy has been employed to detect the prominence in many previous researches, including the studies conducted by Tamburini [2]. It is defined as follows:

$$E_{RMS}(i) = \sqrt{\frac{1}{d(i)}\sum_{k=1}^{d(i)} s_i^2(k)} \tag{2}$$

where $E_{RMS}(i)$ denotes the RMS energy of the i-th phone segment, $s_i(k)$ represents the k-th sample of the speech waveform in the i-th phone segment and $d(i)$ is the duration of the i-th phone segment.

Secondly, a log-scaled energy is also used as a parameter. Since the sensitivity of human auditory perception is not linear but log-scaled with respect to the signal strength, it is worth taking the logarithm. The log-scaled energy is given by

$$E_{log}(i) = \log\left(\frac{1}{d(i)}\sum_{k=1}^{d(i)} s_i^2(k)\right) \tag{3}$$

where $E_{log}(i)$ stands for the log-scaled energy of the i-th phone segment.

Finally, Teager energy is considered as another feature for signal. The Teager energy has been applied to various speech applications, including endpoint detection and prosody recognition. It was proposed by Kaiser [5] and a remarkable property of this feature is that it is affected not only by the current signal, but also by the past and future samples. Let $E_{Teager}(i)$ be the Teager energy of the i-th phone segment. Then,

$$E_{Teager}(i) = E(i)^2 - E(i-1) \times E(i+1) \tag{4}$$

where $E(i)$ is the energy computed from the i-th phone segment.

2.3 Spectral-Temporal Correlation

A spectral-temporal correlation is thought as a promising feature for prominence, which is affected by the spectral density proposed by Wang [3]. Advantages of this feature are to boost up the points which have more energy and to select the sonorant frequency bands in the segment. A procedure for obtaining the spectral-temporal correlation is given in Fig. 1. First, the input signal is passed through the 19 channel filter bank where the pass band of each bandpass filter is specified as in [7]. Secondly, the temporal correlation is computed by

$$C_t^m(n) = \frac{1}{K(K-1)}\sum_{j=0}^{K-2}\sum_{p=j+1}^{K-1} e^m(n+j)e^m(n+p) \tag{5}$$

where $C_t^m(n)$ represents the temporal correlation of the signal at time n for the m-th subband and $e^m(n)$ is the m-th subband energy at time n, and K is the number of filter banks. Then subband selection is performed by following the procedures proposed in [6]. The purpose of this subband selection module is to focus on the sonorant frequency bands. After subband selection, the spectral correlation is computed by

$$C_s(n) = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} C_t^i(n)C_t^j(n) \tag{6}$$

where $C_s(n)$ is the value of spectral correlation, N represents the number of selected subbands, which is set to 12 in our work, and M means the number of pairs of non-overlapping subbands given by $M = \frac{N(N-1)}{2}$. Finally, the spectral-temporal correlation of the i-th phone segment is obtained by summing the correlation among the selected subbands over the interval of the segment. Interested readers are referred to [6] for a more detailed information of the procedures.

2.4 Spectral Emphasis Features

It has been reported that all the frequency bands do not have significant effects on the prominence. Tamburini [2] divided the whole frequency range into three bands: 0-500 Hz, 500-2000 Hz and 2000-4000 Hz. The 500-2000 Hz band was found very useful for prominence detection while the other two bands 0-500 Hz and 2000-4000 Hz were less informative [2]. Based on this observation, we extract all the aforementioned energy-related features including RMS energy, log-scaled energy, Teager energy and spectral-temporal correlation after passing the signal through a bandpass filter of which pass band is 500-2000 Hz as depicted in Fig. 2.

2.5 Pitch Information

The pitch information, which is related to the fundamental frequency, has been a popular research issue for a long time. Unlike the parameters mentioned above, it is difficult to find a fixed-dimensional feature that summarizes the pitch

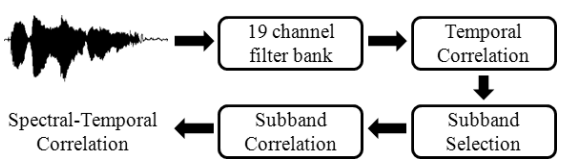

Fig. 1 Flowchart for spectral-temporal correlation.

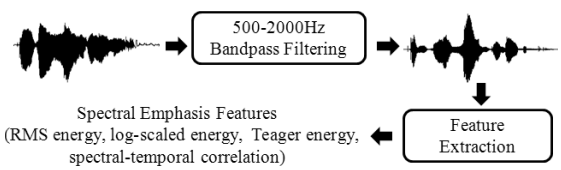

Fig. 2 Procedure to obtain spectral emphasis features.

Table 1 Measures for pitch information of the i-th phone segment.

MAX(i)	the maximum of pitch values
MIN(i)	the minimum of pitch values
MEDIAN(i)	the median of pitch values
MEAN(i)	the mean of pitch values
STD(i)	the standard deviation of pitch values
FLUC(i)	the fluctuation of pitch values

characteristic of a phone-segment. For this reason, we use several simple statistics of the pitch obtained from the given phone segment as the features. The pitch statistics we apply are listed in Table 1 where **FLUC**(i) is defined as

$$\mathbf{FLUC}(i) = \frac{|\mathbf{MAX}(i)| - |\mathbf{MIN}(i)|}{|\mathbf{MAX}(i)| + |\mathbf{MIN}(i)|}, \tag{7}$$

and it gives a rough measure of pitch fluctuation over the given segment.

2.6 Delta and Acceleration Parameters

In speech recognition technologies, a dramatic performance improvement has been achieved with the incorporation of the dynamic features which are basically the time derivatives of the original static features. Inspired by this success, we also extract the delta and acceleration parameters from the static features which are introduced in this section. Let $c(i)$ be a feature parameter extracted from the i-th phone segment. Then, its delta parameter is calculated according to

$$\Delta(c(i)) = \frac{\sum_{k=1}^{K} k(c(i+k) - c(i-k))}{2\sum_{k=1}^{K} k^2} \tag{8}$$

where $\Delta(c(i))$ is the delta parameter of $c(i)$, and $K = 1$ in our work. Once the delta parameter $\Delta(c(i))$ is obtained, the acceleration parameter $\Delta\Delta(c(i))$ can be derived from $\Delta(c(i))$ in a similar way by applying (8).

3. Experimental Results

The speech material used in our experiments was selected from Intonational Variation in English (IViE) corpus. This corpus contained recordings of nine urban dialects of English spoken in the British Isles. It has been used in two research projects: IViE and Oxigen projects [8].

A total of 924 utterance files were used in our experiments and the information on the locations of prominence was also available from the database. For the experiments, we divided the utterance files into training and test groups consisting of 660 and 264 utterances, respectively. Speech data were sampled at 16000 Hz, and the HMM-based speech recognition system was applied to segment each pronunciation waveform into phone-sized units. The units for recognition were composed of 40 context-independent phones including silence. The parameters of this HMM were trained based on the TIMIT database.

An overall block diagram of proposed prominence detection algorithm is shown in Fig. 3. First, the phone boundaries are identified by the forced alignment performed in the HMM-based speech recognition. Based on this segmentation, we extract the phone-specific features as summarized in Table 2. By appending the delta and acceleration parameters, a 45 dimensional feature vector is extracted for each phone segment.

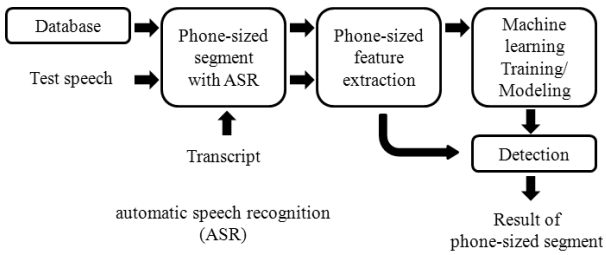

Fig. 3 Flowchart of the prominence detection algorithm.

Table 2 Features used in test.

duration	ROS
energy	RMS energy, log-scaled energy, Teager energy
spectral-temporal correlation (STC)	STC
spectral emphasis (SE)	RMS of SE, log-scaled of SE, Teager of SE, STC of SE
pitch information	MAX, MIN, MEDIAN, MEAN, STD, FLUC

A test was performed to measure the detection accuracy with various different combinations of the features. For this test, we applied the SVM algorithm as a basic classifier and the detection performance was described by a receiver operating characteristic (ROC) curve. The obtained ROC curves are shown in Fig. 4 where SVM15 applied all the 15 static features listed in Table 2, SVM45 used all the 45 features including static, delta and acceleration parameters, SVM11 applied the features of SVM15 except for the 4 features related to SE. We also implemented the techniques proposed by Wang and Narayanan [3] and Streefkerk et al. [1] for performance comparison. In [1], acoustic features accounting for duration, energy and pitch were applied to the ANN, and in [3], similar features were directly extracted from the input speech without a need for a speech recognition technique. Since the proposed algorithm in this letter is based on the extended phone-specific features generated based on a speech recognition technique, the comparison with these techniques is considered meaningful. From the results, it is seen that the incorporation of proposed features improves the performance as compared with those used in [1] and [3]. It is also found that the features related to spectral emphasis have no significant effect on detecting prominence. Comparing between SVM45 and SVM15, it can be found that using both delta and acceleration parameters is superior to applying only the static features.

Next, an experiment was carried out to see the performance when the exact transcript of the utterance was unavailable. In this case, the phone recognition result provided by the HMM system should be treated as the reference transcript for alignment. The test result is given in Table 3

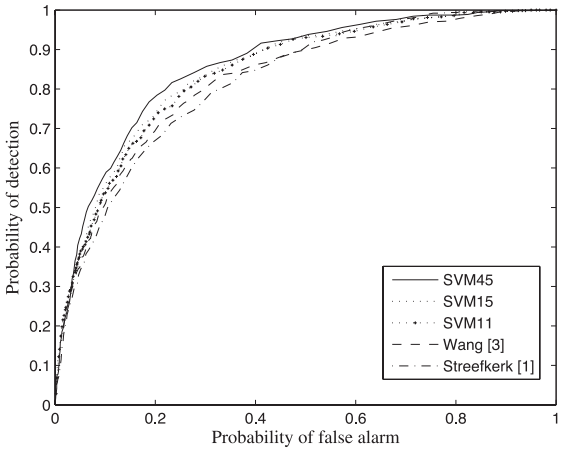

Fig. 4 Comparison according to feature incorporations.

Table 3 Performances with and without transcript.

	Error rate	Precision rate	Recall rate
With transcript	21.79	78.16	78.45
Without transcript	21.75	77.94	80.10

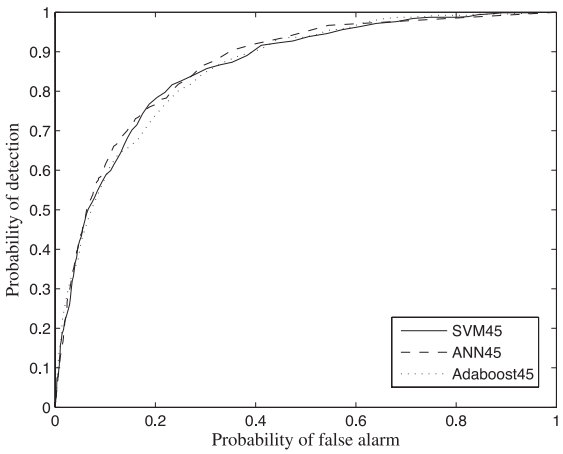

Fig. 5 Comparison among three algorithms. (SVM, ANN, Adaboost)

where the performance is compared with that obtained when the exact transcript was provided. From the results, it can be seen that the proposed prominence detection algorithm works well even without the exact transcript and is robust to the recognition errors of the HMM system.

Finally, we applied three different classifiers: SVM, ANN and Adaboost. Adaboost is a technique combining multiple weak classifiers whose performance can be significantly better than that of any of the weak classifiers [9]. For this experiment, we used all the 45 phone-specific features. The ANN was composed of single output node and 3 hidden units, and the Adaboost consisted of 400 weak classifiers designed based on decision tree structure, which showed good performance in our experiments. The ROC curves obtained from the three classifiers are compared in Fig. 5 where we can see that SVM, ANN and Adaboost show a very similar performance. In [2], it is reported that the accuracy of the manual tagging of prominence is around 80%. The result with the proposed features is comparable with those obtained by human taggers.

4. Conclusions

In this letter, we proposed a prominence detection algorithm based on phone-specific features. These phone-specific features showed high performance in a number of experiments on prominence detection. In addition, it has also been demonstrated that the proposed algorithm can be efficiently applied even when the transcript is unavailable.

Acknowledgements

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2009-0083044) and by the Advanced Industrial Technology Development Program funded by the Ministry of Knowledge Economy (No.10031489).

References

[1] B.M. Streefkerk, L.C.W. Pols, and L.F.M. ten Bosch, "Acoustical features as predictors for prominence in read aloud dutch sentences used in ANNs," Proc. Eurospeech, pp.551–554, 1999.
[2] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: An unsupervised system," Proc. Eurospeech 2003, pp.129–132, 2003.
[3] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," IEEE Trans. Acoust. Speech Signal Process., vol.15, no.2, pp.690–701, Feb. 2007.
[4] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," Proc. ICSLP96, pp.1457–1460, 1996.
[5] J.F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp.381–384, 1990.
[6] S. Narayanan and D. Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," Proc. ICASSP 2005, pp.413–416, March 2005.
[7] Speech Filing System website. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/
[8] The IViE Corpus, Phonetics Laboratory, University of Oxford, website. [Online]. Available: http://www.phon.ox.ac.uk/IViE/
[9] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.