IMPLEMENTATION OF AN INTONATIONAL QUALITY ASSESSMENT SYSTEM

Chanwoo Kim and Wonyong Sung

School of Electrical Engineering Seoul National University
Shinlim-Dong, Kwanak-Gu, Seoul 151-742 Korea

e-mail : chan@mpeg4.snu.ac.kr, wysung@dsp.snu.ac.kr

ABSTRACT

In this paper, we describe an intonational quality scoring system for foreign language learning. We employed the segmental *K*-means algorithm to segment the speech into syllables and a pitch detection algorithm to extract the intonational features. We classified the segmented syllables into 5 types according to the shapes of the pitch contours in them. To present visual aids to students, we displayed the classified tonal pitch type of each syllable and the overall pitch movement tendency of the test and reference sentences. We devised an algorithm to obtain a score from the spoken sentence and used this value as a measure for assessing the intonational quality.

1. INTRODUCTION

Some languages like Korean do not have any conspicuous intonational features that prevail in western languages like English. Thus it is rather arduous for natives using languages without intonation to learn the prosodic features in English [1][2]. Some researches tried to develop an efficient system for non-natives to learn the foreign language pronunciation [3][4], but most of them did not try to assess the intonational quality.

Apart from this, there have been many researches related to the study of intonational features of spoken sentences, and they can be classified into a few approaches such as the morphological, the superpositional and the structural ones [5]. Among them, a model suggested by Fujisaki is widely used in applications like TTS (Text-to-Speech) [5][6][7]. But it is not suitable to adopt this model for assessing the intonational quality is not suitable, since intonation can be varied among different speakers and the correct tendency of the pitch movement is more important than the difference with the reference pattern.

Previous research like [2] employed the DTW (Dynamic Time Warping) method to align the spoken speech to the reference one. But in this case, inter-speaker reliability cannot be very high.

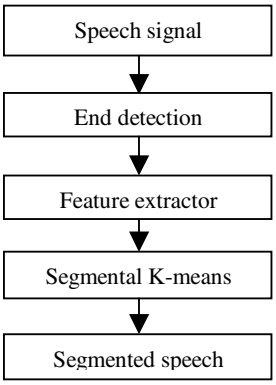


Figure 1: The block diagram of the phoneme segmentation procedure.

In the proposed system, we employed the phoneme segmentation procedure based on the segmental K-means algorithm. It is employed in the training aspect of a speech recognition system, and showed superior inter-speaker reliability compared to the DTW method. We assessed the intonational quality based on two criterions. The first one is the tendency of the pitch movement in a sentence and the second is the type of the pitch movement in a single syllable. We calculated the partial scores in each level and incorporated them into an overall score.

To extract the pitch, we employed the pitch detector in G. 729 annex A [8], and adopted the tonal pitch movement classification model found in [9] to classify the pitch movement in a single syllable.

2. BACKGROUNDS

2.1 Phoneme segmentation

We adopted the phoneme segmentation procedure to locate the phonemes in the spoken language. This procedure is based on the segmental K-means algorithm

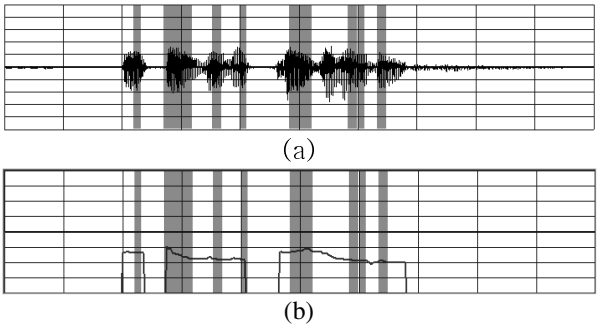


Figure 2: Phoneme segmentation result applied to a spoken sentence: "This is a refrigerator."
(a) Waveform. (b) Corresponding pitch contour.

[10] that is used in obtaining the HMM (Hidden Markov Model) model parameters for a speech recognition system. The states in the proposed system are the phoneme-based ones and three states constitute a single phoneme. This one is similar to the procedure employed in the vowel accuracy checking system in [11]. But in the latter case, this procedure only applies to a single word case and states are based on words.

In this procedure, we used a combination of the 12-th order cepstrum and the 12-th order delta cepstrum as the input feature vectors.

2.2 Intonation checking

There exist various definitions on intonation [5]. But in this system, we focus on the pitch contour movement. In fact, this element is considered to be more important than the other factors, such as phrasing that is used in construction of the prosodic structure [9]. The most significant part of the overall contour is assumed to be certain pitch movements associated with prominent syllables [9]. In order to extract the pitch contour, we employed the open-loop pitch extractor adopted in the G.729 annex A CS-ACELP system [8] incorporated with a 3-point median smoothing filter. The reason for adopting the median smoother is to reduce some possible abrupt errors like the pitch period doubling or halving.

While previous researches like [2] could only evaluate the prosodic quality within a single syllable, we consider the intonational features in the sentence level as well as in the syllable level. Thus, we conducted the scoring procedure twofold. One is performed in the sentence level and the other is done in the syllable level.

In the sentence level, we tried to check the pitch contour movement along the entire sentence. In evaluating the intonational accuracy, the tendency of the pitch movement is more important than the actual pitch values and others derived from them. This is because the actual pitch values and changes in them with respect to time are varied among different speakers. In designing this system, we computed the representative values for pitches from each syllable that is located by the phoneme segmentation procedure. And then we assign ranks to them, which are the orders when these representative values of a sentence are arranged in an increasing order. These ranks were used for assessing the intonational accuracy.

$$S_{sent} = \frac{1}{N_s^2}\sum_{i=1}^{N_s}(t_i - r_i)^2 \quad (1)$$

The score in this sentence level was computed using (1) where N_s is the number of syllables, t_i is the pitch rank of a syllable in the test sentence speech, and r_i is the pitch rank of the syllable corresponding to this one in the reference speech. Division by N_s^2 was adopted here for normalization purpose to alleviate the effect of the sentence length to the score.

In the syllable level, scoring is performed based on the tonal pitch contour model shown in Table 1. After spotting the prominent syllables using the phoneme segmentation procedure that are described in Section 2.1, we partitioned this syllabic interval into 3 sub-syllables. And we computed the median value of the pitches in these sub-syllables. We classified the tendency of the pitch movement in a single syllable according to the model shown in Table 1. But in some cases, the located sub-syllable length might be smaller than 30 ms and it is arduous to classify the contour in that case. Such cases were regarded as type 'E'. The score in the i-th syllable was given by (2) according to the tonal pitch contour types of the reference and the test speech.

$$d_i = \begin{cases} 0 : \text{Their types are the same.} \\ 2 : \text{One of the types is 'C' and the other is 'D'.} \\ 1 : \text{Other cases.} \end{cases} \quad (2)$$

The overall score for all the syllables in the sentence is given by (3).

$$S_{syl} = \frac{1}{N_s}\sum_{i=1}^{N_S} d_i \quad (3)$$

After obtaining scores for these two levels by (2) and (3), they are incorporated into an overall score according to their weights by (4).

$$S_{total} = \alpha_1 S_{sent} + \alpha_2 S_{syl} \quad (4)$$

After repeated tests, we concluded that S_{sent} plays more important role in the perceived intonational quality. Thus we let the α_1 and α_2 values be 0.8 and 0.2, respectively.

We deduce the reason for this result in two ways. First, human speakers are naturally less sensitive to the pitch movement within a syllable [5]. The other reason is the inherent inaccuracy in locating the syllable boundary.

For instance, when the actual pitch type of a syllable is 'C' and the syllable beginning position detected by the phoneme segmentation procedure is somewhat behind the real position, this system might misclassify this tonal

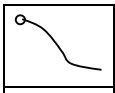
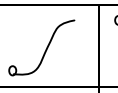
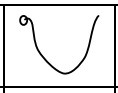
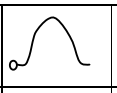
	A	B	C	D	E
(a)					
(b)	\	/	∨	∧	—
(c)	H*L	L*H	H*LH	L*HL	no single equivalent

Table 1: Classification of tonal pitch contours [9]. Tadpole transcription (a), iconic symbols (b), and autosegmental representation of tones (c).

type as 'B'. Many other similar cases are also possible, when the detected syllable beginning and ending position is not very precise.

3. SYSTEM DESIGN

Figure 3 shows the entire system structure and Fig. 4 shows the user interface of this program. We implemented this system on the Windows PC, using visual C++ 6.0. The upper lines and the lower lines in the program appearance in Fig. 4 denote the reference and test intonations respectively. These plots are intended to present graphical feedback to users by enabling them to compare their intonation with the reference. Graphical feedback is proved to be more useful, in some cases, than just presenting scores of the spoken sentences.

Figure 5 shows the devised intonation plot, which is adopted in this program. It shows both the pitch contour type in a syllable and the overall pitch movement in a sentence. This stylized plot was developed for educational effect, since the pitch plot itself is not suitable for this purpose.

This program also shows the evaluation result in the bottom part of the interface as shown in Fig. 4. These comments, such as 'good' or 'poor,' were generated based on the computed score described in Section 2.2. When the condition $S_{total} \leq 0.3$ holds, the tested speech is regarded as good in the intonational accuracy. And when S_{total} is between 0.3 and 0.5, it is evaluated as being a moderate one. In other cases, the intonational qualities were considered to be poor.

4. SIMULATION RESULTS

We tested this system on 34 sentences, spoken several times by 5 speakers. Table 2 shows the reliability of the phoneme segmentation procedure. The "good" case in this Table indicates situation where all phoneme locations are spotted without any severe errors. The "moderate" one means that one severe or several

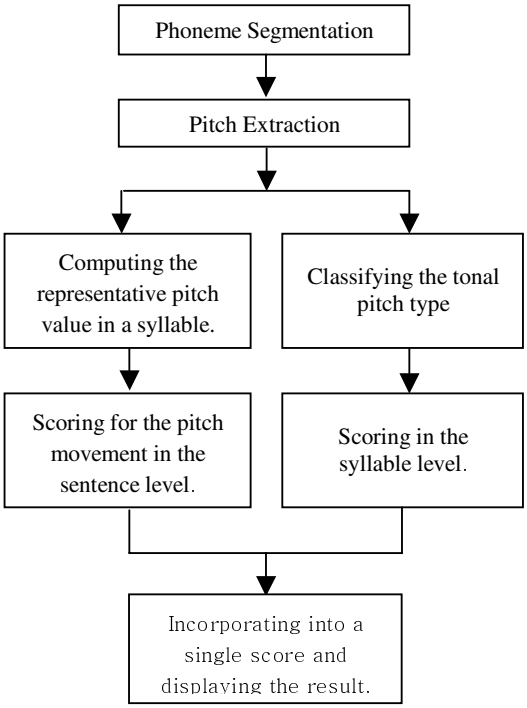
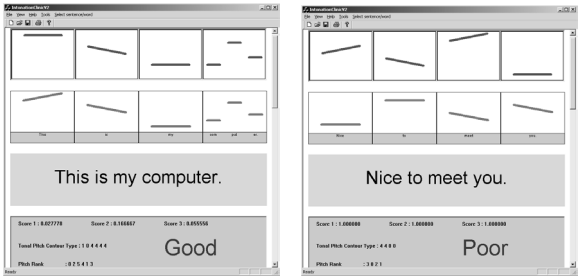


Figure 3: The block diagram of the proposed intonational quality assessment system.



(a) (b)

Figure 4: Test results for two spoken sentences.
(a) A case of a good intonation.
(b) A case of a poor intonation.

slight errors in spotting the phoneme location occurred, but they that do not affect the assessment process severely. In the "poor" case, the phoneme segmentation result has noticeable severe faults, thus the intonational quality assessment result is also unreliable due to this effect. As you can see in this Table, slight errors in the phoneme segmentation do not collapse the entire assessment result. This is because the total score is based on sum of many terms as shown in (1), (3), and (4).

In this system, there are virtually no effects from the pitch halving or doubling problems, since we do not use the actual pitch values in each frame directly. In the

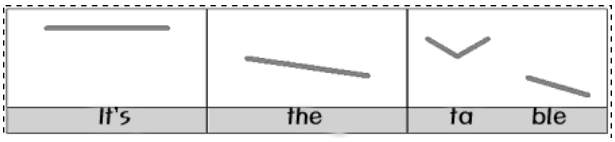


Figure 5: The devised intonation plot that is stylized from the pitch plot

	Good	Moderate	Poor
Percentages(%)	72	21	7

Table 2: The reliability of the phoneme segmentation procedure adopted in this system.

Human / Machine	Good	Moderate	Poor
Good	89	19	16
Moderate	6	16	8
Poor	7	67	77

Table 3: The intonational quality assessment result by this system and a human. (Unit : number of spoken speech sentences)

course of obtaining the representative values for pitches in a syllable or sub-syllable, some possible errors are eliminated.

Table 3 summarizes the comparison of the assessment result performed by this system and human evaluators. From this table, we can find that this system yields reliable results to some extent.

For sentences that human observers evaluated as moderate ones, the correlation between them was low, compared to the other cases.

5. CONCLUDING REMARKS

In this paper, we implemented a system that can assess the intonational quality of spoken sentences. We employed the phoneme segmentation procedure and the pitch extraction method for this purpose. We obtained the score of a spoken sentence in two levels i.e. the sentence level and the syllable level. In many cases, this system yielded desirable results, but we found that the inherent inaccuracy in locating the beginning and ending positions of a syllable degrades the performance of this system. After repeated tests, we are convinced that this system can be usefully utilized in learning foreign language intonation. We are now enhancing the algorithm to obtain more reliable result on the phoneme segmentation procedure. And we are also planning to port this system on the portable devices through fixed-point arithmetic conversion and memory optimization.

6. REFERENCES

[1] C. Kim, "Implementation of an intonation and pronunciation checking system for embedded systems", M. S. thesis, Seoul, Natl. Univ., Seoul, Korea, Feb. 2001.

[2] S. Beack and M. Hahn, "A study of accent and intonation correction in English learning", in *Proc. Int.Conf. Speech Processing,* Daejeon, Korea, pp. 443-446, Aug. 2001.

[3] C Cucchiarini, Helmer Strik, Lou Boves, "Differenct aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition alogithms," *Speech Communication,* vol. 30, no. 2-3, pp. 109-119, Feb. 2000.

[4] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality", *Speech Communication,* vol. 30, no. 2-3, pp. 83-93, Feb. 2000.

[5] A. Botinis, *Intonation: Analysis, Modelling and Technology,* Dordrecht, the Netherlands: Kluwer Academic Publishers, 2000.

[6] A. Tams and M. Tatham, "Intonation for synthesis of speaking styles", in *Proc.* IEE Seminar on state of the art in speech synthesis, 2000.

[7] S. Furui and M. M Sondhi, *Advances in Speech Signal Processing,*, New York: Marcel Dekker, 1992.

[8] Rec. ITU-T G.729, "Coding of speech at 8 k bit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," Feb. 1996.

[9] A. Wichmann, *Intonation in Text and Discourse,* Essex, UK: Pearson Education Ltd, 2000.

[10] B. H. Juang and L. R. Rabienr, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing* vol. ASSP-38, no. 9, pp.1639-1641, Sep. 1990.

[11] C. Kim and W. Sung, "Vowel pronunciation checking system based on phoneme segmentation and formants extraction," in *Proc. Int. Conf. Speech Processing 2001,* Daejeon, Korea, pp. 447-452, Aug. 2001.