Vowel Pronunciation Accuracy Checking System Based on Phoneme Segmentation and Formants Extraction

Chanwoo Kim and Wonyong Sung
School of Electrical Engineering
Seoul National University
Shinlim-Dong, Kwanak-Gu, Seoul 151-742 KOREA
E-mail: {chan, wysung}@dsp.snu.ac.kr

Abstract – In this paper, we developed a vowel sound accuracy checking system for educational purpose in learning foreign language. We employed an HMM (Hidden Markov Model) based phoneme segmentation algorithm, and used the 1st and 2nd formants as a measure of the vowel sound quality. We tested this system for several speakers and concluded that it produces reliable results for educational purpose.

I. Introduction

In learning foreign language, it is often difficult to pronounce a vowel accurately if it is not in one's mother tongue. Specifically, there are many vowels in American English that cannot be found in Korean. There are 12 principal vowels in American English [1]. Among them, the sounds like ER, AO don't have similar counterparts in Korean.

The most important feature that characterizes a specific vowel is the formants, which are the resonant frequencies of the vocal tract. During the vowel articulation, the shape of the vocal tract remains relatively in constant shape so the formants do not change abruptly during a single vowel. We used this feature as the measure of vowel pronunciation accuracy.

There have been some researches concerning development of automatic pronunciation checking system but none of them give special attention to the vowel sound quality [2] [3].

This system consists of two main procedures. The first procedure conducts the phoneme segment. This one is based on the HMM similar to the one used in the speech recognition systems for isolated word recognition. We adopted the segmental K-means method in order to separate the input speech into phonemes [4]. Among the segmented vowel phonemes, it selects the accented one for formants checking.

Two types of formants extraction methods are commonly used [5]. They are the spectral peak picking type and the prediction polynomial root finding type. Generally, the pole extraction type methods produce much more accurate result, while the spectral peak-peaking methods sometimes miss one formant when it is close to another strong one. In spite of the advantages of the pole extraction method, the relative complexity of this technique frequently precludes them [6].

In our system, the accuracy requirement for the formants is somewhat different from the case of typical automatic formant tracking applications. First, formants extraction is done on the speech segment that the phoneme segment procedure decides as a vowel, and we only want to find the representative formants value for this vowel segment. Thus, the median smoothing technique can eliminate most of the spurious formants.

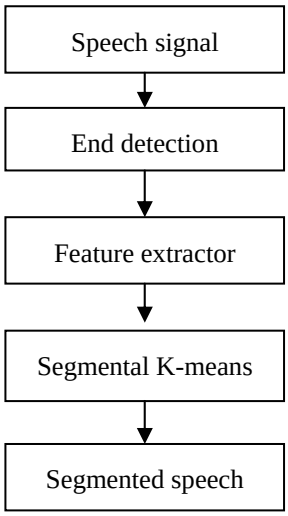


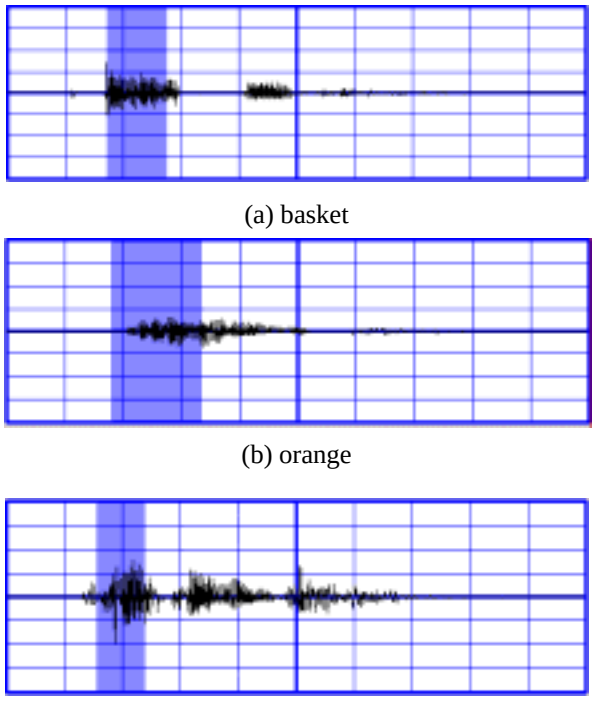
Fig. 1. Block diagram of the phoneme segmentation procedure

Considering the above conditions, we chose the spectral peak picking formants extractor similar to that of McCandless [7]. Details of our procedure will be explained in Section III.

II. Phoneme Segment

To locate the accented vowel, we used a phoneme segment procedure based on the HMM (Hidden Markov Model). In this system, we adopted the phoneme based states and each phoneme consists of 3 states. Because the test speech signals to our system are words, we used the Viterbi algorithm as in the case of the isolated word recognition. This algorithm is included as a part of the segmental K-means procedure [4]. The input feature to this procedure is a combination of cepstrum and delta cepstrum. We used the cepstrum of the order of 12. Figure 1 shows the block diagram of this phoneme segment procedure.

We tested this procedure for several words from several speakers. Our system produced reliable result in most cases. Figure 2 shows some of the resultant segmented portion corresponding to the accented vowel.



(a) basket

(b) orange

(c) coconut

Fig. 2. Phoneme segmentation result. The highlighted regions correspond to the accented vowels.

III. Formants Extraction

Vowels can be distinguished with sufficient accuracy by the first three formants [1]. But the first two of them play the more important role than the third. It is well known that there are tight relation between a vowel and its F_1 and F_2 and that these are also closely tied to the shape of the vocal-tract articulators [1]. TABLE I shows the typical formants for several American English vowels. We used these typical values as references for testing the pronunciation accuracy. Figure 3 shows the well-known vowel triangle where the x-axis is the F_1 and the y-axis is the F_2 [5]. We used this vowel triangle as the model of this system.

As briefly mentioned in Section I, the procedure we adopted for formant extraction is based on the spectral peak-peaking algorithm. Figure 4 shows the block

TABEL I

Typical formant frequencies for vowels [4]

vowel phoneme	F_1	F_2
IY	270	2290
IH	390	1990
EH	530	1840
AE	660	1720
AA	730	1090
AO	570	840
UH	440	1020
UW	300	870
ER	490	1350
AX	500	1500
AH	520	1190

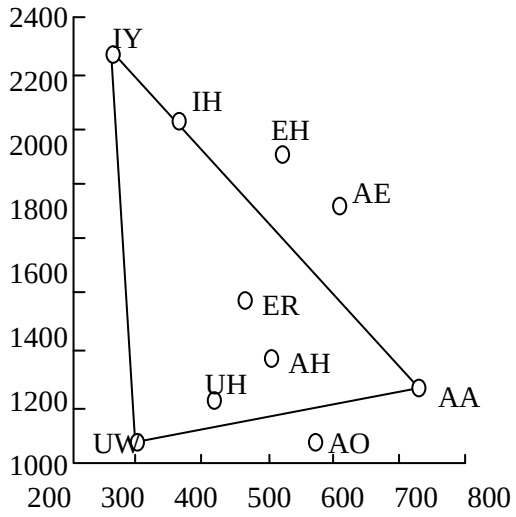


Fig. 3. The vowel triangle

diagram of this formant extraction procedure.

We used 512 point FFT to compute the LP (Linear Prediction) spectrum. To find the spectral peaks more accurately, we tested the spectral inside the unit circle in order to increase the resolution to two adjacent formants like in [7]. The LP vector for this is given by $[1 \rho a_1 \rho a_2 ... \rho a_{14}]$ where $\rho = 0.98$.

If the candidate index for the peak is denoted as k_0

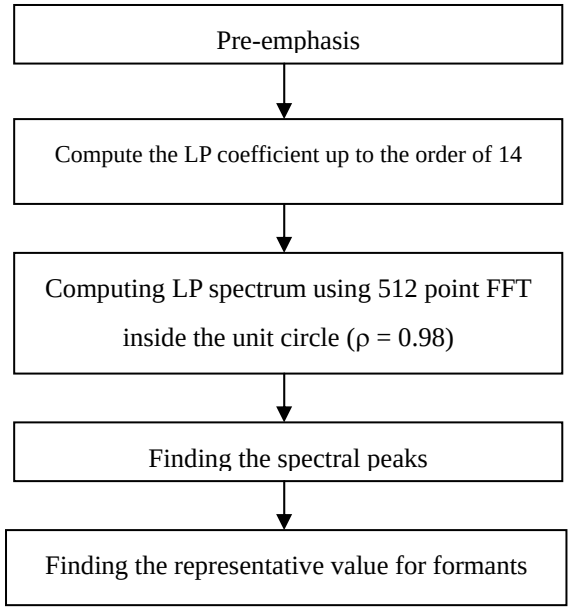
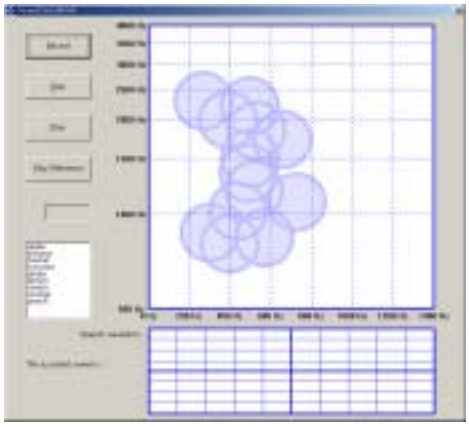


Fig. 4. Block diagram of formants extraction procedure

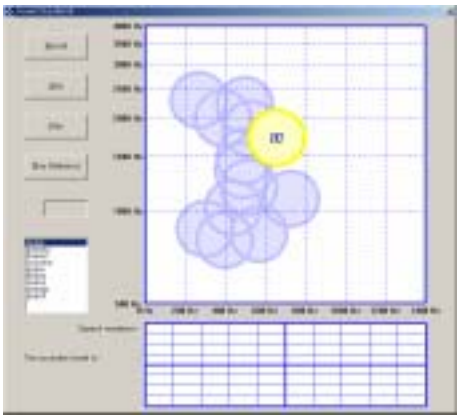
and the spectrum as V[k], the peaks selected in this procedure should satisfy the following constraints. Also, it should be noted again that we used 512 point FFT and 8 kHz sampling rate, so the conditions 1 and 2 should be altered if a different length of FFT or a different sampling rate is used.

1. $V[k_0 - 2] \le V[k_0 - 1] \le V[k_0] \ge V[k_0 + 1] \ge V[k_0 + 2]$
2. $V[k_0]/V[k_0 - 3] > 1.05$ and $V[k_0]/V[k_0 + 3] > 1.05$
3. The Second formant frequencies should be at least 150 Hz over the first formant frequency.
4. The first formant frequency should be at least 200 Hz.

We didn't adopt delicate smoothing methods, since we want to obtain the representative value for the formants of the accented vowel and don't need to know the formats tracking result. Some abrupt errors can exist in some frames even if we apply the conditions 1 ~ 4. We can eliminate some abrupt errors by finding the median values in the interval. In most cases, this method produced reasonable result, but one disadvantage of this method is that it is not always perform well when two

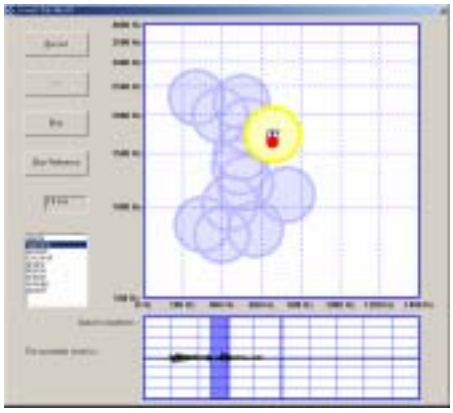


(a)

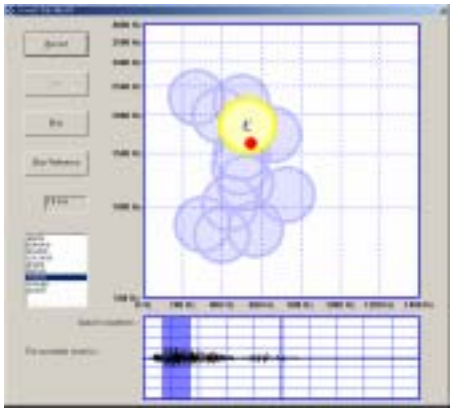


(b)

Fig. 5. The interface of the program. In (a), no word is selected. In (b) apple is selected. The formants region corresponding to the ‘AE’ sound is highlighted.



(a)



(b)

Fig. 6. The correct pronunciation case. (a) is for banana and (b) is for melon

poles corresponding to formants are very close such as the case of ‘AA’ and ‘AO’ phonemes as shown in Fig. 3.

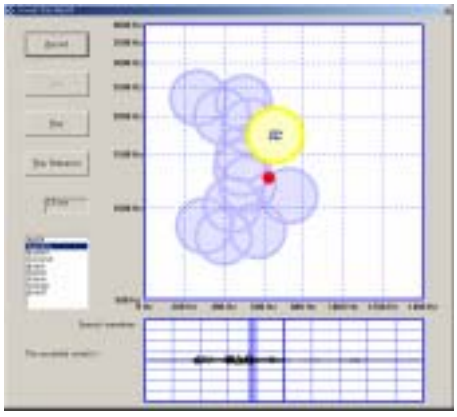
IV. Implementation

We combined the above two procedures, namely the phoneme segment and the formants extraction procedures, into a single Windows program. This system operates on the Microsoft Windows environment. We developed this system using Microsoft Visual C++ 6.0. Figure 5 shows the user interface of this program. As shown in this figure, the y-axis of the formants region is shown in log scale. This is due to the fact that the F_2 varies relatively large compared to $F1$. We can find this fact in [8]. Each circle in the F_1-F_2 plane represents the typical formants region for the corresponding vowel. This program automatically highlights the circle that corresponds to the accented vowel when we select a word in the list box that is in the bottom left corner of the dialog. We can select the words that we want to test by just clicking on it in the list box.

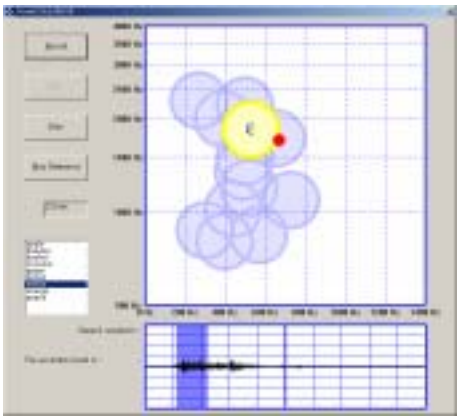
The appearance of the program when we select a specific word is shown in Fig. 5. When we pronounce this, the point determined by the formants is displayed in red.

V. Simulation Result and Analysis

Figure 6 shows the result when a speaker pronounces the



(a)



(b)

Fig. 7. The incorrect pronunciation case. (a) is for "banana" pronounced as [B-AA-N-AA-N-AA] and (b) is for "melon" pronounced as [M-AE-L-AH-N]

word "banana" and "melon". In these cases, the speaker accurately pronounced those words and the program confirms this by showing that the (F_1, F_2) points are inside the highlighted region.

Figure 7 shows another case when the speaker pronounced inaccurately. In the case of Fig. 7 (a), the speaker pronounced this word as 'B-AH-N-AA-N-AA'. And in the case of Fig. 7 (b), the speaker pronounced the word "melon" as "M-AE-L-AH-N'. In these cases, the (F_1, F_2) points are outside the highlighted circle. The circles for the "AE" and "EH" sounds are adjacent, so in the case of (b), the deviation is relatively small compared to the case of (a).

As shown in Table II, the false alarm probability of

TABLE II

Error rate for correctly spoken vowels

Test words	Total error rate(false alarm prob.)	Phoneme segment error rate	Formants extraction error rate when phoneme segment is correct
Apple	16.429 %	7.143 %	10.000 %
Banana	8.571 %	2.143 %	6.569 %
Basket	11.429%	6.429 %	5.344 %
Coconut	27.857%	12.857 %	17.213 %
Grape	12.857%	5. 000 %	8.270 %
Lemon	13.571%	2.857 %	11.194 %
Melon	13.571%	2.143 %	11.940 %
Orange	15.714%	5.714 %	10.606 %
Peach	17.143%	9.296 %	8.661 %
Average	15.238%	5.952 %	9.915 %

the word "coconut" is rather large. This result is largely due to the fact that formants extraction accuracy for this word is not so good, since the 1st and 2nd formants of the tested vowel in this word are close to each other as shown in Fig. 3.

Table II summarizes the error rate of this system due to the incorrect phoneme segmentation or formants extraction error and Table III shows the possibility that this system evaluates the input speech as being good when it is actually not so good. Both tests were conducted in noiseless environment.

VI. Conclusion

In this article, we proposed a system for checking the pronunciation accuracy of vowels. The proposed system showed reliable results in most cases and we found that this system can be efficiently used for educational purpose in learning foreign language.

TABLE III

Error rate for incorrectly spoken vowels

Inaccurately pronounced test words	misdetection prob.
apple[EH-P-L]	45.0 %
banana[B-AA-N-AA-N-AA]	17.5 %
basket[B-AA-S-K-EH-T]	5.0 %
coconut[K-AA-K-AX-N-AX-T]	52.5 %
grape[G-R-AH-P]	2.5 %
lemon[L-AE-M-AX-N]	40.0 %
melon[M-AE-L-AX-N]	32.5 %
orange[OW-R-EH-N-JH]	7.5 %
peach[P-EH-CH]	37.5 %
Average	26.67 %

We primarily worked on the non-diphongized vowels. But this method can be applied to diphthongs or diphongized vowels by tracking formants values. We are modifying our system to efficiently handle them. One of the things requiring improvement is that when the speaker pronounces much differently from the trained data, the HMM based segmentation procedure does not work properly. We will include more training data containing incorrectly pronounced case to resolve this problem. And we are designing a more robust system to increase resolution to two closely adjacent poles in LP spectrum.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Namsoo Kim and Dongguk Kim of Seoul National University for their constructive suggestions and comments, which helped to develop this system. This study was supported by the Brain Korea 21 Project and the National Research Laboratory program (2000-X-7155) funded by the Ministry of Science & Technology of Korea.

REFERENCES

[1] J. R. Deller, Jr., J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*, New York : Macmillan Publishing Company, 1993.

[2] 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.

[3] 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.

[4] 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.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* Engle wood cliffs, NJ : Prentice Hall, 1978.

[6] R. C. Snell and F. Milinazzo, “Formant location from LPC analysis data”, *IEEE Trans. Speech Audio Processing*, vol.1, no. 2, Apr. 1993.

[7] S. S. McCandless, “An algorithm for automatic formant extraction using linear prediction spectra,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22. pp.135-141, 1974.

[8] G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *J. Acoust. Soc. Am.*, vol. 24, no .2 pp. 175-184, Mar, 1952.