Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Vowel place detection for a knowledge-based speech recognition system

Sukmyung Lee and Jeung-Yoon Choi

Yonsei University, 134 Sinchon-dong, Seodaemun-gu, 120-749 Seoul, Republic of Korea
pooh390@dsp.yonsei.ac.kr

This work aims to detect vowel place as part of a knowledge-based speech recognition system. Vowel place was classified into 6 groups based on tongue advancement [Front/Back] and height [High/Mid/Low]. Experiments were performed using 660 /hVd/ utterance data from Hillenbrand [J. Acoust. Soc. Am. 97, 3099-3111] and 6600 TIMIT vowels. Features used include fundamental frequency (F0) and formant value (F1~F3), where formant measurements were classified into separate groups using F0 measurements. The nearest class was found using a simple Mahalanobis distance measure, and yielded a 92.0% classification rate for the /hVd/ data. The results for the TIMIT data were 65.7%, and error analysis with regard to adjacent segment manner and place was carried out to observe the effects of coarticulation, which was not observed in the /hVd/ data.

1 Introduction

A knowledge-based speech recognition procedure can be considered as a type of distinctive feature based speech recognition, which has been considered by Stevens [10] and by Espy-Wilson [4] as an event-based speech recognition. In a knowledge-based approach, the primary purpose is modelling the human perception process. Current statistically based recognition methods face performance degradation under mismatched conditions, and a knowledge-based approach offers an alternative attempt. Because knowledge sources are made in a directed, meaningful manner, if they can be made to work well, they should be more robust against variability. From this point of view, the goal of this work is to detect vowel place as part of a knowledge-based speech recognition system.

Numerous efforts have been made on analyse of vowels. Peterson and Barney (PB) [8] studied the acoustic characteristics of vowels using formant frequencies (F1~F3) and fundamental frequency (F0); also, Hillenbrand *et al* [2] extended the vowel acoustics in a similar way. In addition, Stevens [9] examined the acoustic correlation between formant frequency and vocal tract shape using resonator models. These results show vocal tract shape due to tongue movements are strongly related to vowel production and perception. Also in the past, Meng *et al* [5] attempted to classify vowels using distinctive features. Although Meng reported good performance, the study used many spectral and cepstral coefficients in their of knowledge-based approach.

The purpose of this study, then, is to detect vowel place using primary acoustic features such as formant and fundamental frequency. Vowel place is represented using the following distinctive features: [high, low, back] , and minimum distance measure was used to detect vowel place in 6600 vowels extracted from the TIMIT corpus and Hillenbrand's 660 /hVd/ data [2].

This paper will report preliminary work on vowel place detection using formant and fundamental frequencies. Firstly, we will describe the experimental methodology in detail. We will then present the results of vowel place detection and discuss the results. Finally, we will summarize the paper and consider future work.

2 Experimental methodology

2.1 Test signals

Two different types of databases were used for these experiments. The first test signals consisted of 660 /hVd/ utterances recorded by Hillenbrand *et al* [2]. And 6600 vowels from the TIMIT corpus were also used in this experiment. The vowels chosen for these experiments are 11 monophthongs in American English such as /ɪ/,/i/,/ɛ/,/æ/,/ɑ/,/ɔ/,/ʌ/,/o/,/ʊ/,/u/ and /ɜ^{r}/. The talker of /hVd/ data consist of 30 men and 30 women, so each vowel has 30 signals ($30\times2\times11=660$). 6600 vowels are randomly selected from the TIMIT corpus equally from each gender, and each vowel has 150 male data and 150 female data ($300\times2\times11=6600$). The diphthongs and schwa are excluded here.

2.2 Acoustic measurements

The formant tracking methods for F1 ~ F3 were similar to the Entropic ESPS formant program, spaced every 10 ms, with linear predictive coding (LPC) resonances to find formant frequencies, and include dynamic programming. The formant frequencies are found at 50% of vowel duration which was found from the labels.

Fundamental-frequency (F0) was measured using conventional autocorrelation method every 10 ms using 25 ms Hamming window. It was also sampled at 50% of vowel duration.

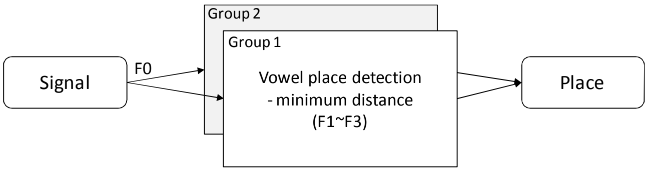

Fig. 1 Vowel place detection process. The input signal is divided into two groups by F0.

	ɪ	i	ɛ	æ	ɑ	ɔ	ʌ	o	ʊ	u	ɝ
HIGH	+	+	-	-	-	-	-	-	+	+	-
LOW	-	-	-	+	+	-	-	-	-	-	-
BACK	+	+	+	+	-	-	-	-	-	-	-

Table 1 The distinctive feature set of 11 vowels.

2.3 Feature analysis

The features chosen to detect vowel place are $[\pm \text{high}]$, $[\pm \text{low}]$ and $[\pm \text{back}]$, which are tongue body features [10]. Then, for each tongue body feature, the tokens are divided into two classes – $[+ \text{feature}]$ and $[- \text{feature}]$. In this paper, every vowel is classified into one of 6 groups depending on tongue body features.

The vocal tract shape can be approximated roughly as a concatenation of tubes. The articulator movement, which can be modeled as concatenated tubes, lead to formant-frequency changes resulting from perturbations (local constrictions) of a tube resonator. The frequency of F1 is inversely related to tongue height (e.g., high vowels have a low F1 frequency), and the frequency of F2 is related to tongue advancement (e.g., front vowels have high F2 frequency).

The $[\pm \text{high}, \pm \text{low}]$ features are related to tongue height. $[+ \text{high}, - \text{low}]$, $[- \text{high}, - \text{low}]$ and $[- \text{high}, + \text{low}]$ represent vowel height that are pronounced with high, mid and back tongue position, respectively. Fig. 2 shows the Gaussian distribution of F1 of high/mid/low vowels for TIMIT data and /hVd/ data. As we expected, Fig. 2 representing tongue height (high/mid/low) is inversely related to F1.

Tongue advancement is connected to $[\pm \text{back}]$ features. The front/back vowel, which are pronounced with front/back tongue position, is represented as $[- \text{back}]$ and $[+ \text{back}]$. Fig. 3 shows the Gaussian distribution of F2 of front/back vowels for TIMIT data and /hVd/ data. As we expected, tongue advancement (front/back) is related to F2. The feature values for these vowels are summarized in Table 1.

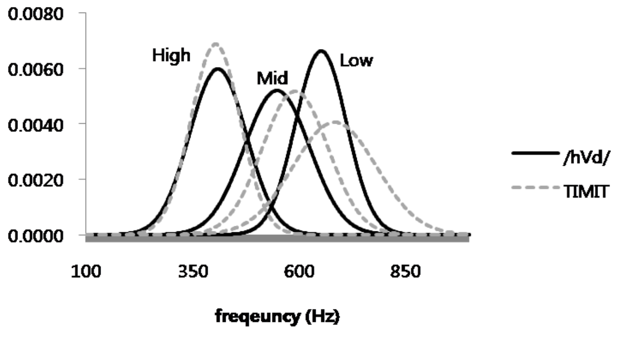

Fig. 2 Gaussian distribution of F1 of high/mid/low vowels for TIMIT data and /hVd/ data

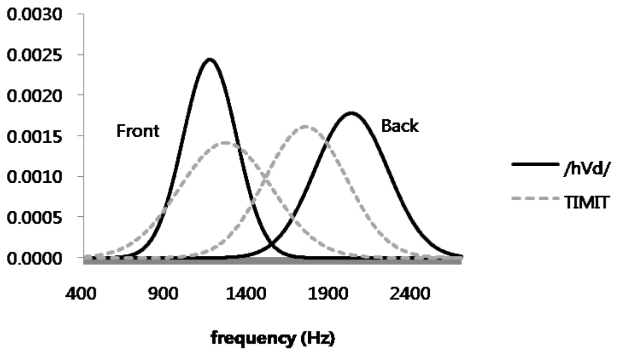

Fig. 3 Gaussian distribution of F2 of front/back vowels for TIMIT data and /hVd/ data.

2.4 Classification strategy

The classification strategy for the vowel place is divided into two steps: grouping and vowel place classification as shown in Fig.1. The grouping process separates input signals into two sets depending on F0. This process can compensate the differences of formant frequency due to the vocal tract length between male and female. The nearest class was found using a simple Mahalanobis distance measure in the vowel place classification process with F1 and F2. The Mahalanobis distance is defined as:

$$D_M(f) = \sqrt{(f-\mu)^T \Sigma^{-1} (f-\mu)} \tag{1}$$

where f is a formant vector, $f = (F1, F2)$, μ is formant mean, $\mu = (\mu_{F1}, \mu_{F2})$, and covariance matrix Σ for a vector f.

In addition, retroflexed vowel processing is performed using F3. Each of reference means and covariance matrices was calculated by training set of database.

3 Result

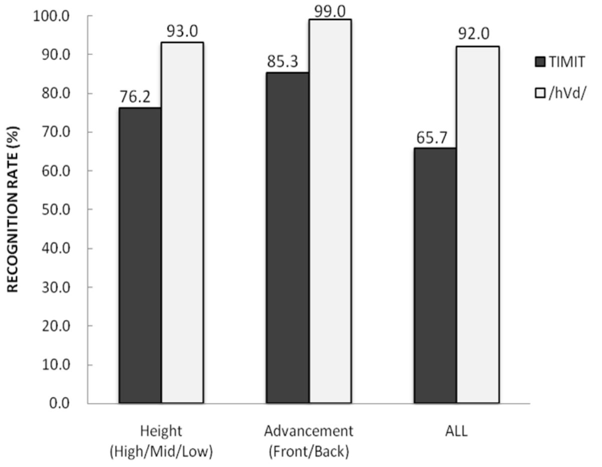

Fig. 4 The overall detection results for TIMIT data and /hVd/ data.

	High	Mid	Low
High	684	109	9
Mid	108	700	97
Low	8	191	294

(a)

	High	Mid	Low
High	75	0	1
Mid	5	78	6
Low	0	2	33

(b)

Table 2 The confusion matrix of tongue height features of (a) TIMIT data and (b) /hVd/ data.

	Front	Back
Front	673	196
Back	127	1204

(a)

	Front	Back
Front	79	1
Back	1	119

(b)

Table 3 The confusion matrix of tongue advancement features of (a) TIMIT data and (b) /hVd/ data.

The vowel place detector is evaluated both for /hVd/ data and TIMIT data. The detection rate was determined by comparing the output of the detector with the labeled data. The overall detection results for the databases are summarized in Fig. 4.

The detection results for tongue height are 76.2% and 93.0% in TIMIT data and /hVd/ data, respectively. The tongue height, which can be represented as [± high, ± low] , are determined by F1, and it is classified into three classes: high/mid/low. Table 2 shows confusion matrix of tongue height. Most of the errors are found between high/low and mid with a few exceptions.

The detection results for tongue advancement are 85.3% and 99.0% in TIMIT data and /hVd/ data, respectively. The tongue advancement, which can be represented as [± back] , are determined by F2, and it is classified into two classes: front/back. Table 3 shows confusion matrix of tongue advancement. By comparing the results for tongue height and advancement, [± back] features show better performance compared to [± high, ± low] features.

The overall detection results for vowel place are 65.7% and 92.0% in TIMIT and /hVd/ database, respectively. Every vowel is classified into six different classes based upon vowel place – High/Front, High/Back, Mid/Front, Mid/Back, Low/Front, and Low/Back. Table 4 and table 5 show confusion matrix of vowel place for TIMIT data and /hVd/ data, respectively. With a few exceptions, most of errors are made between adjacent classes.

	Front/High	Front/Mid	Front/Low	Back/High	Back/Mid	Back/Low
Front/High	280	20	8	0	15	73
Front/Mid	44	94	22	3	59	11
Front/Low	7	44	154	5	29	1
Back/High	0	7	7	128	111	0
Back/Mid	10	29	9	63	518	43
Back/Low	59	6	0	1	68	272

Table 4 The confusion matrix of vowel place of TIMIT data

	Front/High	Front/Mid	Front/Low	Back/High	Back/Mid	Back/Low
Front/High	39	0	1	0	0	1
Front/Mid	0	19	5	0	0	0
Front/Low	0	1	14	0	0	0
Back/High	0	0	0	19	1	0
Back/Mid	0	0	0	1	59	5
Back/Low	1	0	0	0	0	34

Table 5 The confusion matrix of vowel place of /hVd/ data

4 Discussion

To summarize briefly, the main purpose of this study was to detect vowel place using formant frequency and fundamental frequency as a part of a knowledge-based speech recognition system. The nearest class was found using a simple Mahalanobis distance measure, and yielded a 92.0% classification rate for the /hVd/ data from Hillenbrand. The results for the TIMIT data were 65.7%.

Our research was partly motivated by the use of distinctive features for knowledge-based approach. From this point of view, acoustic characteristic that we have chosen are formant frequency and fundamental frequency. These features are no guarantee of detection performance, but they are intuitively reasonable and directly measurable.

Fig 2 shows that the overall detection rate of TIMIT data is worse than /hVd/ data; it is mainly due to coarticulation effect in formant pattern. Hillenbrand have already pointed out that vowel formant patterns are strongly related to phonetic environment [3]. Since /hVd/ data was recorded h-V-d syllables, we can not observe coarticulation effects in formant pattern. TIMIT vowels, however, was extracted from various phonetic environment. Therefore, formant frequency was affected by adjacent phonetic environment. This result suggests that consonant environment on vowel is also significant cues to detect vowel place.

Important areas for further study will comprise three issues. First, phonetic environment on vowel will be considered to overcome the formant pattern changes due to coarticulation effect. Second, the temporal movements of formant frequency will give more information, however one point sampled at 50% of vowel was used in this paper. Furthermore, the speaker normalization that can compensate the inter-speaker variability can be applied, even if fundamental frequency was used in this paper. This work has shown that use of acoustic attributes for vowel place detection is feasible. Although previous study [6] have shown that vowel detection with spectral coefficients, this paper limits to detecting vowel place with acoustic parameters. With a modification of the detection strategy using contextual information it would be expected that performance can be improved.

References

[1] Chomsky, N. and Halle, M., *The Sound Pattern of English*, Harper and Row, (1968)

[2] Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. "Acoustic characteristics of American English vowels", *J. Acoust. Soc. Am.* 97, 3099–3111 (1995)

[3] Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. "Effects of consonant environment on vowel formant patterns", *J. Acoust. Soc. Am.* 109, 748-763 (2001)

[4] Juneja, A., and Espy-Wilson, C. Y. " An Event-based Acoustic phonetic Approach for Speech Segmentation and E-set Recognition", *Proceedings of the International Congress of Phonetic Sciences*, 1333-1336 (2003)

[5] Kent, R. D. and Read, C., *The Acoustic Analysis of Speech 2nd edition*, Thomson Learning. (2001)

[6] Meng, H. and Zue, V., "Signal representation comparison for phonetic classification", ICASSP, 285-288 (1991)

[7] Park, C., "Recognition of English vowels using top-down method", Master Thesis, M.I.T. (2004)

[8] Peterson, G. E., Barney, H. L. "Control methods used in a study of the vowels", *J. Acoust. Soc. Am*. 24, 175-184 (1952)

[9] Stevens, K. N. *Acoustic Phonetics*. The MIT Press, (1991)

[10] Stevens, K. N. "Toward a model for lexical access based on acoustic landmarks and distinctive features", *J. Acoust. Soc. Am.* 111, 1872-1891 (2002)