한국음향학회지 제32권 제2호 pp. 167~173 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.2 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.2.167

pISSN : 1225-4428
eISSN : 2287-3775

Classification of Diphthongs using Acoustic Phonetic Parameters

음향음성학 파라메터를 이용한 이중모음의 분류

Suk-Myung Lee† and Jeung-Yoon Choi

(이석명†, 최정윤)

School of Electrical and Electronic Engineering, Yonsei University

(Received September 13, 2012; revised November 22, 2012; accepted December 5, 2012)

ABSTRACT: This work examines classification of diphthongs, as part of a distinctive feature-based speech recognition system. Acoustic measurements related to the vocal tract and the voice source are examined, and analysis of variance (ANOVA) results show that vowel duration, energy trajectory, and formant variation are significant. A balanced error rate of 17.8% is obtained for 2-way diphthong classification on the TIMIT database, and error rates of 32.9%, 29.9%, and 20.2% are obtained for /aw/, /ay/, and /oy/, for 4-way classification, respectively. Adding the acoustic features to widely used Mel-frequency cepstral coefficients also improves classification.
Key words: Diphthong, Diphthong classification, Acoustic phonetic parameter, Speech recognition
PACS numbers: 43.72.Ar, 43.72.Kb, 43.72.Ne

초 록: 본 논문은 이중모음을 분류하기 위한 음향음성학적 파라메터를 연구하였다. 음향음성학적 파라메터는 성도를 통해 음성이 발성될 때 나타나는 특징을 기반으로 하여 분산분석(ANOVA) 방법을 통해 선별한 모음의 길이, 에너지 궤적, 그리고 포먼트의 차이를 이용하였다. TIMIT 데이터 베이스를 사용하였을 때, 단모음과 이중모음만을 구분하는 실험에서는 17.8% 의 밸런스 에러율(BER)을 얻을 수 있었고, /aw/, /ay/, 그리고 /oy/를 단모음과 분류하는 실험에서는 각각 32.9%, 29.9%, 그리고 20.2%의 에러율을 얻을 수 있었다. 추가적으로 진행한 실험에서, 음향음성학적 파라메터와 음성인식에 널리 쓰이고 있는 MFCC를 함께 사용하였을 경우 역시 성능향상이 나타나는 것을 확인하였다.
핵심용어: 이중모음, 이중모음 분류, 음향음성학 파라메터, 음성인식

I. Introduction

A knowledge-based speech recognition system described by Stevens[1] outlines procedures to find linguistic units termed distinctive features from the speech signal. Distinctive features include three broad classes, the articulator-free features, articulator features, and articulator-bound features. Articulator-free features [or manner features] describe the type of sound being produced, and include the features [vowel], [glide], and [consonant], along with the features [continuant], [sonorant] and [strident], which further specify the consonant types. Articulator features indicate which articulator is used, and articulator-bound features describe the different ways the articulator can be used.

Of the class of sounds that are specified by the articulator-free feature [vowel], two subtypes are possible. Monophthongs are produced with an open vocal tract in a steady configuration, and specification of the articulator-bound features [high], [low], [back] and [tense] is sufficient for distinguishing among the vowels. In contrast, diphthongs are characterized by a changing vocal tract shape, which includes narrowing of the vocal tract starting from an initial open configuration. A diphthong can therefore be defined as a smooth transition between two

†**Corresponding author:** Suk-Myung Lee (pooh390@dsp.yonsei.ac.kr)
School of Electrical and Electronic Engineering, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea
(Tel: 82-2-2123-4534; Fax: 82-2-364-4870)

target vowel configurations.[2,3] It can also be defined as a sequence of a vowel onset and an offglide, which can be represented by two articulator-free features consisting of a [vowel] and a [glide], where each part can be further described by its associated articulator-bound features. In English, the three vowels /aw/, /ay/ and /oy/ are considered to be diphthongs. For a complete description of a vowel segment in a distinctive-feature based speech recognition system, it is necessary to distinguish diphthongs from the monophthongs.

Much research has been conducted on the acoustic characteristics of diphthongs, including a well-known study by Lehiste and Peterson.[2] More recently, Yang[4] reports on an extended study of diphthong acoustics. In these and other studies, diphthongs are shown to correlate with longer durations and varying formant trajectories, in contrast to monophthongs. Carlson et al.[5] includes diphthong classification in a study on classification of vowels using acoustic characteristics, and a 71% correct classification rate can be derived from the reported confusion matrix data. However, studies that specifically describe classification experiments for diphthongs are rare.

Therefore, this study aims to investigate diphthong characteristics, and to use the associated acoustic phonetic parameters for diphthong classification for a distinctive feature-based speech recognition system. It is assumed that vowel detection has been completed, so that diphthong classification is carried out on vowel segments only. Acoustic measurements that describe characteristics of diphthongs are investigated, along with Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used in statistical speech recognition systems. Analysis of variance (ANOVA)[11] tests are used to assess the significance of measurements for diphthong classification, and results for 2-way discrimination of monophthongs versus diphthongs, and 4-way discrimination of monophthongs and /aw/, /ay/, and /oy/, are presented.

II. Description of Acoustic Measurements

A number of acoustic measurements have been investigated for describing diphthongs. Since diphthongs consist of a vowel onset and a following offglide, measurements that reflect this characteristic are chosen.

Diphthongs are usually longer in duration compared to monophthongs, so that vowel duration is expected to be a significant acoustic cue. Vowel duration may be found from voice activity detection or probability of voicing measures, but in this paper, it is assumed that the presence of a vowel, and its start and end points, are found in advance, so that vowel durations are directly found from phone labels.

Espy-Wilson[6] points out that glides usually have less energy in the low- to mid-frequency range compared to vowels. Energy trajectories of monophthongs and diphthongs are expected to show different patterns. To access the difference in the energy trajectory between monophthongs and diphthongs, we used to band-limited energies in the frequency rages 300-900 Hz, 640-2800 Hz and 2000-3000 Hz. The frequency range 640-2800 Hz and 2000-3000 Hz are examined because Espy-Wilson[6] reported that the lower F1 for glides is expected to cause a decrease in the amplitudes of the formants in these region. Also, first formant region, nominally about 300 to 900 Hz, is measured.

In addition, features related to the voice source are investigated, such as fundamental frequency (F0), open quotient and spectral tilt. Open quotient is calculated as the amplitude of the first harmonic relative to that of the second harmonic (H1-H2), and spectral tilt is calculated as the amplitude of the first harmonic relative to that of the third formant spectral peak (H1-A3). Although articulatory movements for producing diphthongs are mainly in vocal tract shape, it is hypothesized that these movements may affect the voice source as well.

In order to capture the time variation characteristics of these acoustic measurements, range, slope, and convexity

of the contours are found. Range is the difference between maximum and minimum values, and slope is calculated as the ratio of the difference of start and end values to duration. Convexity is calculated as the sum of the difference between each signal point and the linear interpolation between the start and end values of a segment. That is,

$$convexity = \frac{\sum_{t=t_1}^{t_2}(s(t)-h(t))}{t_2-t_1},$$

where t_1 and t_2 are respectively the start and end times of the vowels, $s(t)$ is the value of the measurements at time t, and $h(t)$ is the linear interpolated function,

$$h(t) = \frac{s(t_2)-s(t_1)}{t_2-t_1}(t-t_1),$$

for $t_1 \le t \le t_2$, and $t_1 \le t_2$, respectively.

These time variation measures are found for all acoustic measurements, except duration. In addition, dip and peak locations of overall RMS energy are found, in order to capture energy change locations in the signal.

In this paper, RMS energy, formant frequencies and amplitudes, and F0 are found using the Snack program package.[7] First and second harmonic amplitudes used in calculating open quotient are found by measuring amplitudes at the fundamental, and twice the fundamental frequency, respectively.

Also, in order to compare with widely used spectral measures, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. 13th-order MFCCs are extracted at start and end positions of vowels, and delta MFCCs are found as the difference between the MFCCs at start and end positions. In total, 39th-order MFCCs are used in the experiments.

III. Experimental Results

3.1 Database

The TIMIT[8] corpus contains 6300 utterances spoken by 630 speakers from 8 different dialect regions in the United States, and includes word and phone labels. For diphthong classification, vowel stimuli are extracted, with no restrictions in phonetic environment. All three diphthongs (/aw/, /ay/ and /oy/), along with 17 monophthongs, are included. The excised vowel database consists of 66944 tokens, with 48395 tokens included in the training set (44595 monophthongs, 3800 diphthongs), and 18549 tokens included in the test set (17213 monophthongs, 1336 diphthongs). The numbers of /aw/, /ay/ and /oy/ tokens in

Table 1. ANOVA results (F-values) for 11 acoustic measurements for the training data set. Entries with probabilities greater than P > 0.05 are not significant and marked with a dash (-).
Sec

	Measurements	mono/diph	mono/ay	mono/aw	mono/oy
Duration	Duration	2191	1831	1323	1263
Energy property	RMS slope	335	419	142	-
	RMS convexity	296	617	214	291
	2000-3000 Hz energy slope	911	761	-	710
	2000-3000 Hz energy convexity	1142	1220	221	1302
Formant property	F1 range	2093	2574	470	-
	F1 slope	561	810	-	-
	F1 convexity	1491	1733	488	-
	F2 range	2987	2624	651	2202
	F2 slope	1377	2123	613	527
	F2 convexity	350	191	-	641

the training and test sets are, 729, 2387, and 684, and 216, 852, and 268, respectively. In this paper, TIMIT phone labels are used to find locations for extracting features, at the vowel onset and the offglide. In order to reduce endpoint effects, start and end locations where features are extracted are at 10% and 90% of total duration, measured from the beginning of the vowel.

3.2 ANOVA Analysis

The measurements obtained for diphthong classification in the TIMIT training set are first examined using ANOVA. One-way analysis is performed for each of the acoustic measurements, and significant features with P<0.05 are found. Results show that measurements for band energy in the 300-900 Hz and 640-2800 Hz ranges are not significant. Likewise, voice source measurements, including F0, open quotient, and spectral tilt measurements are all found to be not significant. This implies that vocal tract movements do not significantly affect voice source characteristics in the case of diphthongs. In all, 11 significant features are found, and F-values for monophthong versus diphthong discrimination, and for discriminating each diphthong from monophthongs are shown in Table 1. The F-value is computed as the ratio of the between-group variance in the data over within-group variance, and indicates relative discriminative power between features. Entries that are not significant are marked with a dash(-). From the results, it can be seen that duration and F2 range parameters are significant indicators for all cases. F1 slope is discriminative only for /ay/, and F2 convexity is discriminative for /oy/. Among the band energy measurements, 2000- 3000 Hz energy slope and convexity seem to be significant indicators for /ay/ and /oy/.

3.3 Experimental Results

Using acoustic phonetic parameters and/or cepstral features, Gaussian Mixture Models (GMMs) with 8 mixtures which showed optimal performance are trained for each task from TIMIT training data. For performance evaluation, Balanced Error Rate (BER)[9] is found, in addition to overall classification rates.

The Balanced Error Rate(BER) is the mean of the error-rates for each class, and is defined as

$$BER = \frac{1}{C}\sum_{i}\frac{(\sum_{j} M_{ij}) - M_{ii}}{\sum_{j} M_{ij}}.$$

where C is the number of classes and M is the $C \times C$ confusion matrix, i.e. $M(i,j)$ is the number of times that the vowel of class C_i is mis-classified as class C_j.

The 11 acoustic phonetic parameters, which are listed in Table 1, are then used to classify diphthongs in the TIMIT test set. Three configurations are considered. First, 2-class classification between monophthongs and diphthongs is carried out. Next, concurrent 4-class classification between monophthongs, and /aw/,/ay/, and /oy/, is conducted, and finally, 4-class classification following a tree procedure is carried out, where diphthongs are separated from monophthongs in the first step, and then classified into /aw/, /ay/, and /oy/ in the second step.

First, results of classification of monophthongs versus diphthongs are presented. Using the 11 acoustic phonetic parameters results in a BER of 17.8% and 82.0% classification rate, which is better than that using 39th-order MFCCs (with 18.1% and 81.6%, respectively). However, using acoustic phonetic parameters in addition to MFCCs improves performance, to 14.8% BER and 84.7% classification rate. This implies that acoustic phonetic parameters and MFCCs provide complementary

Table 2. Balanced Error Rates (BERs) for acoustic property of 11 acoustic phonetic parameter. The results of duration, energy property, formant property, and 11 acoustic phonetic parameter (all) are represented. Entries are in percent (%).

	BER
Duration	26.5
Energy property	29.0
Formant property	21.2
All	17.8

information in detecting diphthongs. Also, experiments were performed to examine the effect of acoustic property. 11 acoustic phonetic parameters are divided by three properties (duration, energy property and formant property) depend on its acoustical characteristic. Energy property includes RMS slope, RMS convexity, 2000-3000 Hz energy slope and 2000-3000 Hz energy convexity. And formant property include F1 range, F1 slope, F1 convexity, F2 range, F2 slope and F2 convexity. BERs are calculated for each property and results are represented in Table 2. Formant property showed best performance with BER of 21.2%, as predicted by ANOVA results. BERs for duration and formant property are 26.5 % and 29.0 %, respectively. These results indicate that duration and energy property are useful for diphthong distinction.

To explore adjacent phoneme effects on diphthong discrimination, classification error rates are analyzed depending on context. All phones in the TIMIT database are divided into four manner classes, i.e. vowels, glides, nasals and obstruents, and classification results are

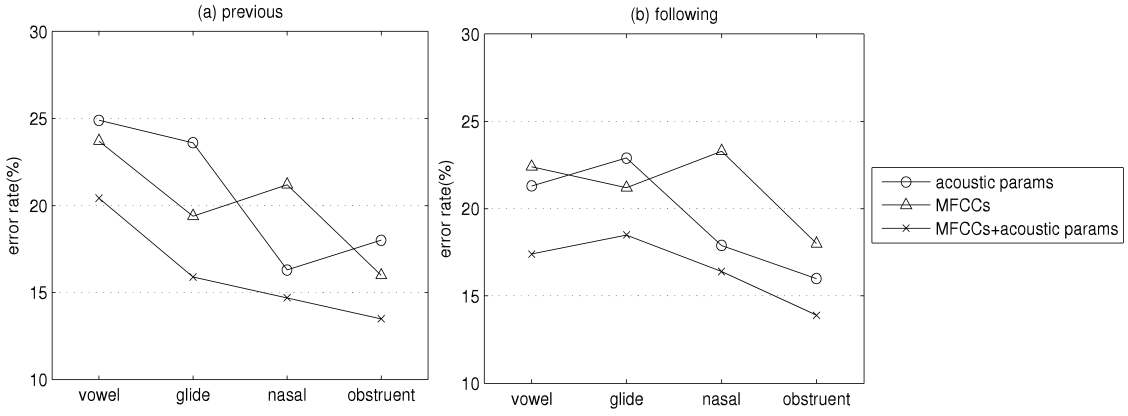

Fig. 1. Context effects on diphthong error rate depending on adjacent (previous and following) phoneme class: vowel, glide, nasal and obstruent consonant.

Table 3. Confusion matrices and Balanced Error Rates (BERs) for 4-way (monophthong, /aw/, /ay/, and /oy/) concurrent and tree classification methods using acoustic phonetic parameters (a,d), MFCCs (b,e), and acoustic phonetic parameter with MFCCs (c,f), respectively. Monophthongs are denoted mono. Entries are in percent (%).

(a) acoustic phonetic parameters						(b) MFCCs						(c) MFCCs + acoustic phonetic parameters					
	mono	aw	ay	oy	rate		mono	aw	ay	oy	rate		mono	aw	ay	oy	rate
mono	81.4	8.0	6.1	4.5	81.4	mono	81.2	8.0	6.2	4.6	81.2	mono	85.2	6.3	5.0	3.5	85.2
aw	27.3	67.1	3.7	1.9	67.1	aw	25.9	68.9	3.3	1.9	68.9	aw	26.9	69.0	3.7	0.5	69.0
ay	19.7	4.0	70.1	6.3	70.1	ay	16.1	3.2	75.3	5.4	75.3	ay	14.3	4.9	73.4	7.4	73.4
oy	9.1	1.1	9.9	79.8	79.8	oy	15.2	1.9	16.7	66.2	66.2	oy	7.2	0.8	6.1	85.9	85.9
			Total BER		25.4				Total BER		27.1				Total BER		21.6
(d) acoustic phonetic parameters						(e) MFCCs						(f) MFCCs + acoustic phonetic parameters					
	mono	aw	ay	oy	rate		mono	aw	ay	oy	rate		mono	aw	ay	oy	rate
mono	81.2	8.3	5.3	5.2	81.2	mono	82.1	9.8	3.6	4.5	82.1	mono	85.4	8.6	4.1	1.9	85.4
aw	35.1	59.7	4.1	1.1	59.7	aw	37.6	58.9	1.3	2.2	58.9	aw	28.8	65.1	3.8	2.3	65.1
ay	15.6	2.8	66.8	14.8	66.8	ay	12.8	2.8	79.4	5.0	79.4	ay	12.6	3.7	79.0	4.7	79.0
oy	21.2	9.1	2.6	67.1	67.1	oy	22.3	6.2	8.8	62.7	62.7	oy	10.3	3.4	6.7	79.6	79.6
			Total BER		31.3				Total BER		29.2				Total BER		22.7

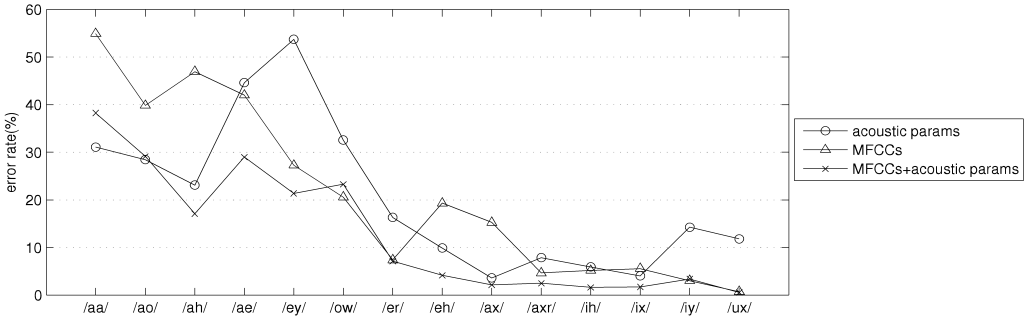

Fig. 2. Error rates for monophthongs from 4-way concurrent diphthong classification using acoustic phonetic parameters, MFCCs, and acoustic phonetic parameters with MFCCs.

analyzed depending on phoneme class of preceding or following segment. Results of context effects are shown in Fig. 1. Overall, the highest error rates occur with adjacent vowels, and the lowest for adjacent obstruents. Classification rates with MFCCs+acoustic phonetic parameters are less affected by adjacent phonemes; using only acoustic phonetic parameters shows about 9% difference depending on adjacent phoneme.

In the next experiment, 4-way classification is carried out to distinguish between monophthongs and the 3 diphthongs /aw/, /ay/, and /oy/. Tables 3 (a) through (c) show confusion matrix results using acoustic phonetic parameters, MFCCs, and MFCCs in addition to acoustic phonetic parameters, respectively. Classification rates using acoustic phonetic parameters for /aw/, /ay/, and /oy/ are 32.9%, 29.9%, and 20.2%, respectively, while classification rates using acoustic phonetic parameters with MFCCs shows 3 to 6% performance improvement for all diphthongs. Overall, diphthongs with a /y/ offglide show better performance compared to diphthongs with a /w/ offglide. Also, more errors occur between monophthongs and diphthongs, and less among the diphthongs.

Next, each diphthong is classified using a tree procedure. First, diphthongs are separated from monophthongs, and are then classified into one of the three diphthongs. Tables 3 (d) through (f) show the resulting confusion matrices. Overall, classification rates are slightly lower than concurrent 4-way classification.

Finally, error analysis is performed for concurrent 4-way classification, and error rates for each monophthong vowel are shown in Fig. 2. The analysis is limited to monophthongs with more than 200 tokens in the TIMIT database, so that /ax-h/, /uw/ and /uh/ are excluded. Results show vowels with longer durations[10] such as /aa/, /ey/, /ah/ and /ao/, have greater error rates. Also, high vowels such as /ih/,/iy/ and /ux/ show lower error rates compared to low vowels.

IV. Conclusions

This work examines acoustic phonetic parameters for classification of diphthongs in English, as part of a distinctive feature-based speech recognition system. Time variation characteristics of acoustic measurements related to the vocal tract and the voice source are examined, along with widely used cepstral coefficient features. From ANOVA tests, duration and formant range are found to be significant measurements, along with RMS and 2000-3000 Hz band energy trajectories. Measurements related to the voice source are found to be not significant.

In the two-class experiments (monophthongs versus diphthongs), an overall 17.8% balanced error rate is obtained using the proposed acoustic phonetic parameters, and 32.9%, 29.9%, and 20.2% error rates are obtained for /aw/, /ay/, and /oy/, in the four class experiments (discriminating between monophthongs, /aw/, /ay/ and /oy/). Concurrent 4-way classification is found to be more effective than a tree procedure, where diphthongs are first

separated from monophthongs, and are then classified into one of the three diphthongs. In addition, adding the acoustic phonetic parameters to MFCCs shows performance improvement in all cases.

In this paper, the experiments did not take into account contextual information. However, results show that the manner class of the previous or following phoneme is significant, especially if vowels or glides are adjacent. Therefore, normalization methods or compensation for adjacent phoneme effects may be necessary. The results of this study are expected to be included in an overall vowel detection module, as part of a distinctive feature-based speech recognition system.

References

1. K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111,** 1872-1891 (2002).
2. I. Lehiste and G. E. Peterson, "Transitions, glides, and diphthongs," J. Acoust. Soc. Am. **33**, 268-277 (1961).
3. A. Holbrook and G. Fairbanks, "Diphthong formants and their movements," J. Speech and Hearing Res. **5**, 38-58 (1962).
4. B. Yang, "An acoustic study of English diphthongs produced by American males and females," Phonetics and Speech Sciences, **2**, 43-50 (2010).
5. R. Carlson and J. Glass, "Vowel classification based on analysis-by-synthesis," in Proc. Int. Conf. Spoken Language Processing, 575-578 (1992).
6. C. Y. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels in American English," J. Acoust. Soc. Am. **92**, 736-757 (1992).
7. J. Gustafson and K. Sj¨olander, "Educational tools for speech technology," in Proc. Fonetik, 176-179 (1998).
8. J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," Linguistic Data Consortium (1993).
9. I. Read and S. Cox, "Automatic pitch accent prediction for Text-To-Speech synthesis," in Proc. Interspeech, 482-485 (2007).
10. J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am .**97**, 3099-3111 (1995).
11. R. G. Miller, *Beyond ANOVA: Basics of Applied Statistics*. (Chapman & Hall, New York, 1997).

저자 약력

▸ Suk-Myung Lee

He received the B.S. degree in electrical & electronic engineering from Yonsei University, Seoul, Korea, in 2007. He is currently pursuing the Ph.D. degree in electrical & electronic engineering at Yonsei University, Seoul, Korea. His current research interests include speech recognition and speech signal processing.

▸ Jeung-Yoon Choi

She received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1992 and 1994, respectively. She received her Ph.D in electrical engineering and computer science from Massachusetts Institute of Technology in 1999. She was a postdoctoral researcher at MIT from 1999 to 2001. She continued her studies as a postdoctoral researcher and as a visiting scholar at the University of Illinois, Urbana-Champaign from 2001 to 2005. After a year as a research professor at the Biometrics Engineering Research Center at Yonsei University, from 2005 to 2006, she joined the faculty of the Department of Electrical and Electronic Engineering at Yonsei University, as an assistant professor from 2006 to 2012. She is now a research scientist at MIT. Her current research interests include knowledge-based speech signal processing, prosody detection and speech communication.