Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2aSC: Linking Perception and Production (Poster Session)

2aSC23. Developmental aspects of American English diphthong trajectories in the formant space

Sungbok Lee*, Alexandros Potamianos and Shrikanth Narayanan

***Corresponding author's address: Electrical Engineering, University of Southern California, Los Angeles, California 90089, sungbokl@usc.edu**

Formant trajectories of five American English diphthongs embedded in the target words BAIT (/ei/), BITE (/ai/), POUT (/au/), BOAT (/Ou/), BOYS (/oi/) are investigated in the F1-F2 space as a function of age and gender. Age range considered is from 5 to 18 years. In this report, the focus is given to the differences in position between the start/end points of diphthongs and nine monophthons. Averaged formant data across subjects in each age group are examined for the purpose. Two findings are worth mentioning. Firstly, across all age groups, the start and end positions of diphthongs hardly match with the monophthongs that are typically used to transcribe the diphthongs across all age groups (c.f., Holbrook and Fairbanks, 1962). For instance, the start position of /ei/ is very close to /I/ than to /e/, and the end points of /ei, ai, oi/ are significantly different with respect to each other. Secondly, in addition to the areas of vowel space , an overshoot trend toward the nominal end points of diphthongs is the most prominent developmental trend. That is, formant values of diphthongs produced by younger age children are closer to the nominal monophthongs used to transcribe the diphthongs.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4798783]
Received 18 Jan 2013; published 2 Jun 2013

Introduction

This study investigates the developmental acoustic properties of five American English diphthongs as a function of age and gender. Diphthongs are commonly characterized by *formant movements* from one vowel sound (i.e., onset) to another (i.e., offset) (Lehiste and Peterson, 1961; Holbrook and Fairbanks, 1962). The movement of the second formant (F2) seems most prominent, and the *rate* of F2 transition is shown to be different from diphthong to diphthong (Gay, 1968), being a useful parameter to discriminate diphthongs (Gottfried et al., 1993). Regarding the nature of the onset and offset portions of diphthongs, the study of Holbrook and Fairbanks (1962) suggests that they hardly matches with the monophthongs that are typically used to transcribe the diphthongs. For instance, as mentioned in Gay (1968), the onset of /ai/ ("bite") can vary from /aa/ to /ae/ and the offset from /eh/ to /iy/. There seem more disagreements on the phonetic identities of the offset portions and it has shown that the offsets of diphthongs can also vary with speech rate (Lehiste, 1964; Wise, 1965; Gay, 1968). Therefore, it seems difficult to define diphthongs in one way (e.g., dual targets and transition) or in another (e.g., onset plus transition).

In Gottfried et al. (1993), as an attempt to elucidate the phonetic definition of diphthongs, three different hypotheses were tested in terms of the *classification accuracy* of diphthong segments in the F1-F2 plane produced by adult speakers as a function of stress and speaking rate. They showed that the [onset + offset] hypothesis and the [onset + slope (i.e., transition)] hypothesis show similar performances of 97.0% vs. 97.8%, respectively, while the [onset + direction (toward to the offset)] hypothesis shows a less performance of 94%. Their results suggest that if one includes the onset as a default element in the definition of diphthong, the transition rate and the offset position may be equally effective for the phonetic description of diphthongs. Therefore, in the current study, we investigate the developmental trends of the onset and offset formant positions and also the formant transition rate, as a function of age. Such investigations might help to understand the phonetic nature of diphthongs and how it vary as a function of age.

Method

A. Database

As described in a previous study (Lee et al., 1999), the speech database analyzed in this study was collected from 436 children, ages 5 through 18 with a resolution of 1 year of age, and from 56 adult speakers (ages 25-50). The speech material in the database consisted of ten monophthongal and five diphthongal vowels in American English as well as five phonetically rich meaningful sentences. The distribution of subjects by age and gender is shown in Table I. Among the 492 subjects, 316 were born and raised in the two Midwestern states of Missouri and Illinois.

The five diphthongs analyzed in the current study were produced in target words of bait (/ei/), bite (/ai/), pout (/au/), boat (/Ou/) and boys (/oi/). The target words were produced in the carrier sentence "I say uh ___ again," except for children of ages 5 and 6, who produced target words in isolation. The target utterances were produced twice in random order. No specific instructions were given to the subjects regarding the manner of production. Prior to the recording session, any target utterances that the speakers (mostly 5- and 6-year-olds) had difficulty reading were identified and elicited through imitation of a sample prerecorded by a female speech pathologist. Recordings were made in a sound-treated booth located inside a glass-panel enclosure, using a high-fidelity microphone (Bruel & Kjaer model #4179) connected to a real-time waveform digitizer with 20-kHz sampling rate and 16-bit resolution.

B. Measurements of Formant Frequencies

First, in order to isolate the diphthong segments for formant frequency estimations, an automatic segmentation procedure (i.e., a forced-alignment procedure using hidden Markov models) was utilized as described in Lee et al. (1999). In order to examine the accuracy of the automatic segmentation procedure, durations of 160 diphthongs from 16 randomly selected subjects of ages 5, 7, 9, 11, 13, 15, 17 and adults (age 39) were manually measured. One male and one female subject were selected in each age group. It was found that among 160 tokens, only 8 tokens exhibit segmentation errors larger than 50 msec. When these 8 tokens were excluded, the mean duration difference between automatic and manual segmentations averaged across all tokens was -6.97 msec (std. dev. = 13.3 msec),

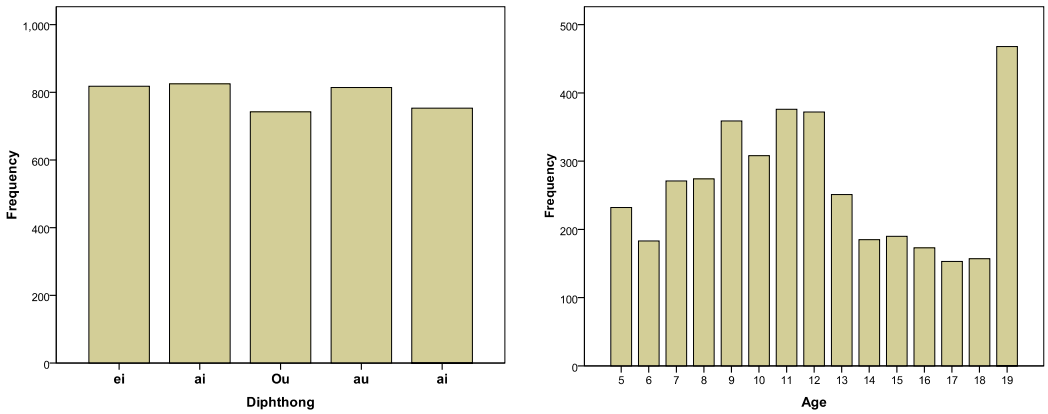

FIGURE 1. Distributions of 3,952 (male: 2046, female: 1906) data points by diphthong type(left) and by age(right).

indicating that automatic segmentation slightly underestimates the diphthong durations. Therefore, it was viewed that there exist no significant errors in the automatic estimation of diphthong durations including onset and offset positions, except for some erroneous cases of excessively short (less than 100 msec) or long durations (more than 600 msec) which were excluded from the current investigation.

The first three formant frequencies (F1-F3) of each diphthong segment were estimated using the PRAAT software [REF] with a pre-emphasis factor of 0.94 and a 12th-order linear-prediction analysis with a 10-msec window. A pilot experiment was performed and it was found that for ages 5 through 9, total 5 formant peaks in the frequency range up to 7000 Hz was appropriate in that it yields the reasonable 3rd and 4th formant values. For subjects older than the younger age groups (i.e., from age 10 to adult), the frequency range was reduced to 6000 Hz for five-formant trajectory searches. This adjustment was made to empirically accommodate widely varying harmonic frequency spacing.

The automatic pitch and formant-tracking programs yielded reasonable estimates of the F0 and F1 trajectories in most cases. The second (F2) and the third (F3) formant tracks, however, were often inaccurate for vowels produced by young children due to poor spectral resolution at high frequencies (partially caused by wider harmonic spacing and breathy voicing), spurious spectral peaks, and formant-track merging. In such cases, manual estimation of formants from the speech spectrogram was also difficult. Therefore, after the 5-point median filtering followed by a spline smoothing, the raw formant trajectory data were refined based on the outlier detection of onset and offset frequencies of each token based on SPSS statistical software package and then a heuristic and iterative formant trajectory outlier detection algorithm. Finally, the refined formant trajectories of *3,952 tokens* were processed in order to determine onset and offset positions and a maximum formant transition rate.

C. Analysis

Resulting formant trajectory onset, offset and transition rate data are organized by diphthong type, age and gender and analyzed by the SPSS software package to examine the effects of age and gender on the acoustic properties of diphthongs. Fisher's discriminant analysis is also applied to the dataset in order to investigate the effectiveness of a number of combinations of acoustic parameters in diphthong classification.

Results

In the current study we report the diphthong classification results and their implications to the phonetic description of diphthongs. We also report the aspects of the developmental relationships between onset, offset and trajectory orientation of diphthongs and the formant positions of the nine monophthongs in the F1-F2 plane by visually examining between-age group plots.

Classification of diphthongs: Results of diphthong classification based on Fisher's discriminant analysis are summarized as follow for each combination of diphthong trajectory parameters: When consider [F1,F2 onsets] only, the leave-one-out cross validation accuracy is 74.4% whereas, for [F1,F2 offsets] only, it is 65.4%. For [F1, F2 onsets and rates] it is 81.9%, whereas for [F1,F2 offsets and rates], it is 78.6%. Those results imply more offset target variability in diphthong productions, and the usefulness of the formant transition rates in diphthong classification. Furthermore, for [F1, F2 onsets and offsets], the accuracy is 92.3% and for [F1, F2 onsets, offsets and

transition rates], the accuracy was 92.5%. As references, when all F1-F3 formant frequencies considered, it may be noted that the accuracy is 95.3% for [F1,F2,F3 onsets and offsets], 91.3% for [F1,F2,F3 onsets with F1-F3 transition rates], and 95.3% again for [F1,F2,F3 onsets, offsets and F1-F3 transition rates]. This implies that offset position and transition rate may have an equivalent discriminative power, although the current results suggest a slightly better discrimination performance in the [onset and offset] description of dipthongs.

In summary, the results of classification experiments suggest that "onset and offset" positions are primarily important elements in the phonetic description of the diphthongs in the F1-F2 space. The F2 transition rate seems next relevant diphthong parameter. Since the offset position, which could be either implicit or explicit in a speaker's phonological vowel space (see Fig. 2), guides the direction of formant movements, it is plausible that the [onset + offset] is a primary factor that determines diphthong identity. It is also speculated that the F2 transition rate is a secondary factor that may not be directly related to explicit kinematic control factors manipulated by speakers. That is, the formant transition rate could be an epiphenomenon that is related to the formant distance between onset and offset formant values to be spanned under the probable constraint of the isochrony principle of similar durations (i.e., long distance, larger transition rate). In fact, although not reported in the current study, it is found that differences in duration of the five diphthongs are not statistically significant, except /oi/, in each age group.

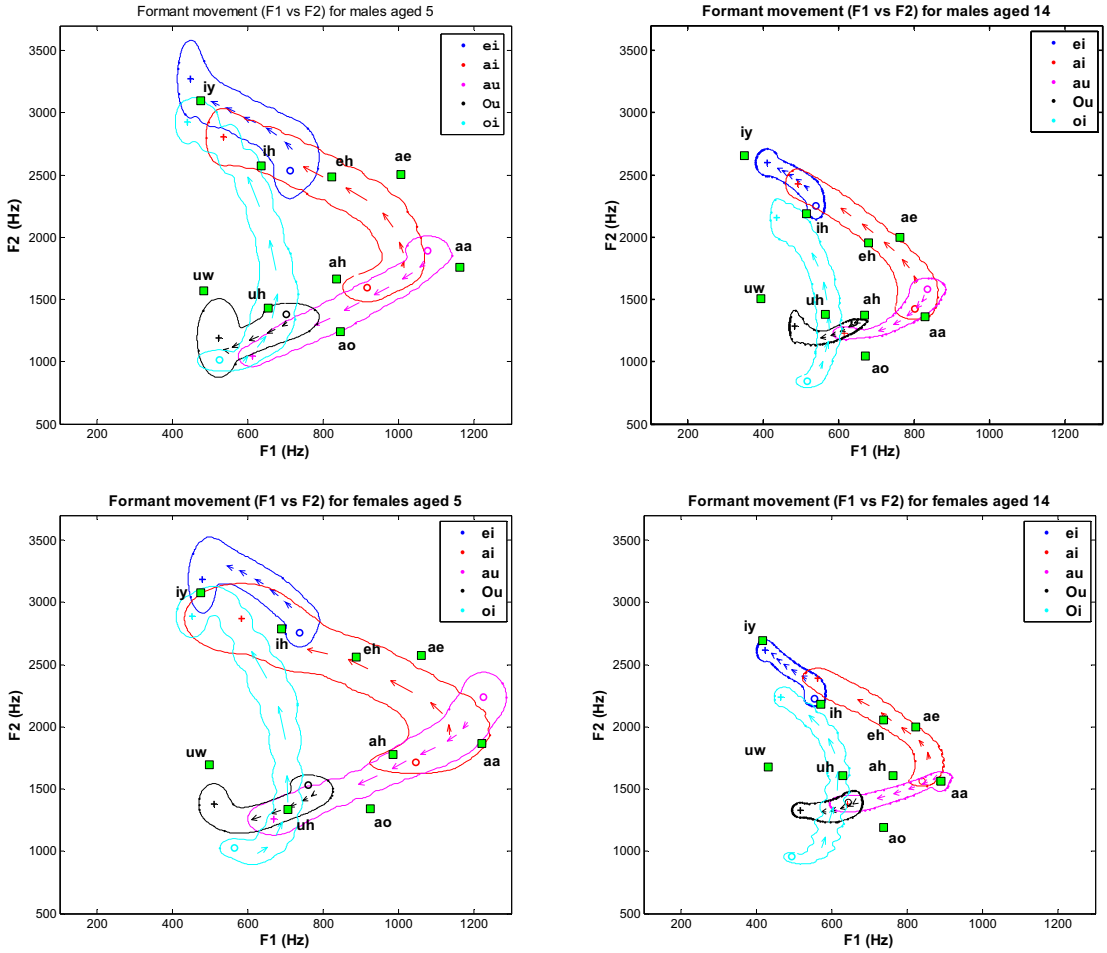

FIGURE 2. In each plot, five diphthong trajectories are represented by five colored "*strips*" (*blue*: /ei/, *red*: /ai/, *magenta*: /au/, *black*: /Ou/, *cyan*: /oi/) in which arrows represent trajectory midline segments between onset ('o') and offset ('+') positions, and *strip width* corresponds to formant variability at selected locations along the midline. Top row is for male age 5(left) and age 14(right) and bottom row is for female age 5(left) and age 14(right)). Formant positions of 9 monophthongs (green squares with two-character ARPABET vowel symbols) of each age group are also shown in background for comparison to onsets and offsets of diphthongs.

***Developmental changes in onset and offset formant positions of diphthongs w.r.t. monophthongs*:** In Figure 2, five diphthong trajectories as "strips" are plotted for two exemplary age groups (age 5 and age 14). In each strip,

arrows represent trajectory midlines between onset ('o') and offset ('+') and "strip width" represents variability at selected locations along the midline. F1 and F2 formant values of the 9 monophthongal vowels (green squares) of each age group are from a previous monophthong study (Lee et al., 1999).

Several interesting observations can be made from Fig. 2. First of all, irrespective of age, the onset and offset positions of diphthongs seems different from those of monophthongs used for the transcription of diphthongs, except the cases of the offset position of /ei/ and onset position of /ai/. Another prominent tendency is that for the younger age group (e.g., age 5), the onset and offset positions of diphthongs are much closer to those of monophthongs. For offset, this tendency can be observed clearly for the diphthongs whose formant movements are directed toward /iy/ (e.g., especially /ai/ and /oi/) and toward /uw/ (e.g., especially /Ou/). For onset, /oi/ is special in that the onset position is well separated from any adjacent monophthongs especially in the older age group (i.e., age 14).

Those aforementioned observations from the formant representation of diphthong acoustics may imply that diphthongs has their own phonological representations that are developed separately, independent of monophthongs, as speakers grow older.

Summary

Results of the diphthong classification experiments and the positional comparison of diphthong onset and offset positions against monophthongs suggest that the [onset+offset] specification, or definition, of diphthongs be reasonable and phonetically relevant. It is also speculated that diphthongs possess their own onset and offset positions and it is not phonetically necessary to relate them to monophthongs. A prominent developmental tendency in diphthong acoustics is that the onset or offset positions of diphthongs are closer to monophthongs used for the transcription of diphthongs for younger-age children (e.g., age 5). This is especially so for diphthongs /ei/, /ai/ and /oi/ whose offset positions are very close or equivalent to /i/. Probably except /ei/, the offset does not reach /i/ but show undershoots for older speakers (i.e., age 14).

Although the size of vowel space is reduced and the positional relations of monophthongs vary as a function of age as can be observed from Fig. 2, the orientation of five diphthong trajectories with respect to each other (see Fig. 2) seems to be preserved. This may suggest that the onset and offset of diphthongs may be acquired without referring to, or independently of, monophthongal vowel positions. That is, it is plausible to hypothesize that that diphthongs has their own phonological representations that are being developed separately, independent of monophthongs, as speakers grow older and the [onset+offset] definition of diphthongs may have its phonetic relevance.

Based on the results of the study, it is speculated that initially younger age speakers acquire diphthong production skill based on monophthongs as reference and then as speakers grow older, the onset and offset position is being settled down in the phonological vowel space, independently of the monophthongs position. It is thought that the onset position is more stable in the phonological vowel space and the offset itself may vary depending on the speaking environment such as speaking rate.

REFERENCES

Gay, T. **(1968).** "Effects of speaking rate on diphthong formant movements," J. Acoust. Soc. Am. **44,** 1570-1573.

Gottfried, M., Miller, J. D., and Meyer, D. J. **(1993).** "Three approaches to the classification of American English diphthongs," J. Phon. **21,** 205-229.

Holbrook, A. and Fairbanks, G. **(1962).** "Diphthong formants and their movements," J. Speech Hear. Res. **5,** 38-58.

Lehiste, I. and Peterson, G. E. **(1961).** "Transition, glides and diphthongs," J. Acoust. Soc. Am. **33,** 268-277.

S. Lee, A. Potamianos and S. Narayanan (1999). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. 105, 1455-1468.

Boersma, P. and Weenink, D. **(2009).** "Praat: Doing phonetics by computer (version 5.1.1) [computer program]," htpp://www.praat.org.