Developmental acoustic study of American English diphthongs[a)]

Sungbok Lee,[b)] Alexandros Potamianos,[c)] and Shrikanth Narayanan
Viterbi School of Engineering, University of Southern California, Los Angeles, California 90089

(Received 4 April 2014; revised 4 August 2014; accepted 17 August 2014)

Developmental trends of durational and spectral parameters of five American English diphthongs are investigated by age and gender. Specifically, diphthong durations, the fundamental frequency (F0), and the first three formant (F1, F2, F3) trajectories as well as formant transition rates are analyzed as a function of age, gender and diphthong type. In addition, the distance between diphthong onset and offset positions and those of nearby monophthongs in the formant space is computed and age-dependent trends are presented. Furthermore, a spectral transition mid-point is estimated for a given diphthong trajectory and normalized time durations from onsets to mid-points are analyzed as a function of age and diphthong type. Finally, diphthong classification results using formant-related parameters are reported. Results show the expected age-dependent reductions of diphthong duration, fundamental frequency, onset and offset formant values, and formant transition rate. More interestingly, it is evident that speakers adjust onset and offset positions of diphthongs with respect to monophthongs as a function of age. Normalized duration of the first demisyllable segment is found to be different among diphthongs and that younger children spend more time in the first segment. The implications for diphthong development and the onset-offset definition of diphthongs are discussed in detail. © *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4894799]

PACS number(s): 43.70.Ep [SAF] Pages: 1880–1894

I. INTRODUCTION

This study investigates the acoustic properties of five diphthongs in American English as a function of age and gender. In a previous developmental study of American English monophthongs,[1] age-dependent trends of a set of temporal and spectral parameters of the ten monophthongal vowels were investigated by analyzing speech recordings obtained from 492 participants, including 436 children aged 5 through 18 years old and 56 adults. The current study can be regarded as an extension of this earlier work on monophthongs since it is performed on the same speaker population. As was the case for the study of monophthongs, the purpose of the current study is two-fold: (1) to characterize the developmental acoustic trends of the five American English diphthongs and (2) to provide a developmental acoustic description of the diphthongs as a function of age and gender. It is hoped that the results of the current study on the diphthongs, together with those of the previous study on the monophthongs, will provide a set of comprehensive reference data on the developmental trends of the vowel acoustics of American English speakers from age 5 to 18 years old, as well as a comparison with adult speakers (ages 25–50 years old). It is noted that "Arpabet" (also known as, CMU) phonetic symbols are used throughout the study to transcribe monophthongs and diphthongs.

Diphthongs are commonly characterized by formant movements from one vowel sound to another.[2,3] The movement of the second formant (F2) is most prominent, and the rate of F2 transition is shown to be different from diphthong to diphthong,[4] being a useful parameter to discriminate diphthongs.[5] Regarding the phonetic characteristics of the initial (onset) and final (offset) portions of diphthongs, the study of Holbrook and Fairbanks[2] suggests that they hardly match those of the monophthongs that are typically used to transcribe each diphthong. For instance, as mentioned in Gay,[6] the onset of /AY/ can vary from /AA/ to /AE/ and the offset from /IH/ to /IY/. There seems to be more disagreement on the phonetic identity of the offset portions; diphthong offsets have been also shown to vary with speech rate.[4] Therefore, there seems to be no consensus on the definition of the phonetic identity of diphthongs, e.g., dual targets and transition, or onset plus transition.

Gottfried *et al*.[5] and Sánchez Miret[7] provide an excellent review and discussion on this issue. In Gottfried *et al*.,[5] three different hypotheses on the phonetic description of diphthongs were tested in terms of the classification accuracy of diphthong segments produced by adult speakers as a function of stress and speaking rate: (1) the [onset + offset] hypothesis which emphasizes the dual target nature of diphthongs, (2) the [onset + slope] hypothesis which emphasizes the relative stability of onset with respect to the offset and the F2 transition rate of diphthongs, and (3) the [onset + direction] hypothesis emphasizing the onset stability and the direction toward the offset or final target in the F1–F2 space. Note that onset is included in the all three hypotheses. The classification results indicate that the [onset + offset] hypothesis and the [onset + slope] hypothesis show similar performances of 97.0% vs 97.8%, respectively, whereas the [onset + direction] hypothesis shows a lower performance of 94%. The results suggest that if one includes the onset as a default element in

[a)]Portions of this work were presented in "A developmental acoustic characterization of English diphthongs," Acoustical Society of America meeting, New York, NY, May 2004 and in "Developmental aspects of American English diphthong trajectories in the F1-F2 plane," ICA Meeting Program 2013, Montreal, Canada, June 2013.

[b)]Author to whom correspondence should be addressed. Electronic mail: sungbokl@usc.edu

[c)]Also at School of Electrical and Computer Engineering, National Technical University of Athens, Zografou 15780, Greece

the definition of diphthong, the transition rate and the offset position may be equally effective for the phonetic realization of diphthongs.

Regarding the acquisition of vowel production skills, early investigations on vowel production by children suggest that normally developing children exhibit few vowel errors after 2.0 years of age, showing 91% average correct production.[8] For consonants, the production skills mature later, up to age 5 to 7 years old, especially for fricatives such as /s/ and /zh/ (cf. Prather *et al.*[9]). Regarding diphthongs, the study reports that 21–24 month-old children produced the diphthongs /AW/, /OY/, /AY/, /IU/ with over 97% accuracy. Paschall[10] has also found that 20 children aged 16–18 months produced the diphthongs /AW/, /OY/ and /AY/, /IU/ with accuracy ranging from 60% to 62% and 40% to 47%, respectively. The results from both studies imply that children acquire the basic production skills for most monophthongal vowels and diphthongs up to around 2 years old. Therefore, children of age 5 years and older must have already learned how to articulate diphthongs. However, this does not mean that they have achieved a mature or adult-like vowel production skill at such an early age. For instance, as shown in the previous study of monophthongs,[1] durational and spectral variability associated with vowel production reduces continuously until 11 or 12 years old and reaches adult levels thereafter. Previous studies[11,12] have also shown that the development of the vocal tract continues throughout childhood and adolescence and does not reach full maturity until the late teens or early twenties. Therefore, even after the acquisition of vowel production skill at around age 2 years old or so, speakers need to continuously adapt their speech motor control skills to the growing size of the vocal tract.

Accordingly, the main focus in this study is to characterize the developmental trends of the diphthong acoustics by analyzing magnitude and variability of temporal and spectral acoustic parameters of diphthongs as a function of age and gender. In addition to the traditional acoustic developmental study of vowel sounds, two novel measurements of diphthongs acoustics are investigated in this study: (1) Spectral distances between diphthong onset and the offset positions and nearby monophthongs in the F1–F2 plane are computed as a function of age. Our goal is to better understand the co-development and co-evolution of monophthongs and diphthongs. (2) A diphthong landmark position is defined that corresponds to a mid-point of the spectral transition between onset and offset. The landmark divides a diphthong trajectory into two demisyllabic segments, the first segment from the onset to the landmark, and the other from the landmark to the offset. It will be shown that the relative duration of the first demisyllabic segment is dependent not only upon speaker's age but also the diphthong-type.

The paper is organized as follows: In Sec. II, we describe briefly the speech database used in this study and methods used for the automatic phonetic alignment of speech waveforms so that target diphthong segments can be delimited. In Sec. III, we describe methods used for durational and spectral measurements of diphthongs. Results are presented in Sec. IV and discussed in terms of developmental aspects in Sec. V. Concluding remarks are given in Sec. VI.

II. SPEECH DATABASE AND DIPHTHONG SEGMENTATION

A. Speech database

As described in previous studies,[1,13] the speech database analyzed in this study was collected from 436 children, ages 5 through 18 years old with a resolution of 1 year of age, and from 56 adult speakers (ages 25–50 years). 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. Of the 492 subjects, 316 were born and raised in the two Midwestern states of Missouri and Illinois.

The five diphthongs in target words were /EY/(bait), /AY/(bite), /AW/(pout), /OW/(boat), and /OY/(boys) with IPA symbols eɪ, aj, aw, ow, ɔj, respectively. 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 and 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 years old) had difficulty reading were identified and elicited through imitation of production samples 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. Automatic segmentation of diphthongs

In order to isolate diphthong segments from surrounding consonants for durational and spectral measurements, an automatic segmentation procedure based on forced-alignment was utilized as described in the previous study of monophthongs.[1] In order to examine the accuracy of the automatic segmentation procedure, durations of 160 diphthongs from 16 randomly selected children of ages 5, 7, 9, 11, 13, 15, 17 year old and adults (age 39 years old) were manually measured and compared to the corresponding automatically measured durations. It is noted that one male and one female subjects were selected in each age group. It was found that out of 160 tokens examined, only eight tokens exhibited segmentation errors larger than 50 msec. When these eight tokens were excluded, the mean duration difference between automatic and manual segmentations (averaged across all tokens) was −6.97 msec (with a standard deviation of 13.3 msec), indicating that automatic segmentation slightly underestimated the diphthong durations. It was

TABLE I. Distribution of subjects by age and gender.

Age	5	6	7	8	9	10	11	12	13	14	15	16	17	18	5–18	25–50
Male	19	11	11	25	23	25	24	22	16	11	11	11	10	10	229	29
Female	13	16	24	11	25	14	19	21	13	10	11	11	9	10	207	27
Total	32	27	35	36	48	39	43	43	29	21	22	22	19	20	436	56

empirically observed that typically the preceding consonant duration was over-estimated by the forced alignment algorithm. Therefore, it was deemed that there existed no significant errors in the automatic estimation of diphthong durations including onset and offset portions. In any case, diphthongs with excessively short or long durations were considered outliers and excluded from the duration analysis, as detailed in the next section.

III. TEMPORAL AND SPECTRAL MEASUREMENTS OF DIPHTHONGS

A. Duration

The duration of each token was obtained from the corresponding label file produced by a forced-alignment segmentation procedure. The segmentation procedure sometimes erroneously yielded excessively short or long segmental length, however, an effort was made to minimize the inclusion of such outliers using the duration histogram by excluding duration values which are less that 100 msec and larger than 700 msec. Outlier detection reduced the number of diphthong tokens from the original number of 4920 down to 4633. Subsequently, there were 921 tokens for /EY/ (439 tokens for female, 482 for male), 914 tokens for /AY/ (429 for female, 485 for male), 934 tokes for /AW/ (445 for female, 489 for male), 939 tokens for /OW/ (448 for female, 491 for male), and 925 tokens for /OY/ (442 for female, 483 for male).

B. Estimation of formant onset, offset, and transition rate

The trajectories of the fundamental frequency (F0) and the first three formant frequencies (F1–F3) of a given diphthong segment were estimated using the PRAAT software.[14] A 12th-order linear-prediction analysis was used with a pre-emphasis factor of 0.94, a 20 msec analysis window duration and a 10 msec windows update. A pilot experiment was performed, and it was found that for ages 5 through 9 years old, a total of five formant peaks in the frequency range up to 7000 Hz was an appropriate parameter choice for the formant tracking algorithm to yield reasonable third and fourth formant frequency values. For subjects older than 10 years old the frequency range was reduced up to 6000 Hz for five formant trajectory searches. This empirical adjustment was made to accommodate widely varying harmonic frequency spacing and improve formant tracking estimates.

The automatic pitch and formant-tracking programs yielded reasonable estimates of the F0 and F1 trajectories in most cases. The second (F2) and 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 five-point median filtering followed by a spline smoothing, the raw formant trajectory data were refined using the following procedure: All formant tracks were resampled and smoothed to 20 linearly spaced intervals in time. Then outliers due to formant tracking were removed from the data before statistics were computed.

The outlier detection algorithm used heuristics to identify formant track points or segments that significantly deviated from the average F1–F2 formant values for each gender and age group. Formant tracks were evaluated at three time scales: Formant values (each of the 20 points), short formant track segments consisting of four points (20% of total formant track length), and long formant track segments consisting of eight points (40% of total formant track length). For short formant track segments both deviation from the mean and deviations in formant track direction in the F1–F2 space were identified. The total number of deviations was summed up to come up with an outlier score heuristically. Specifically whenever one of the following conditions was met one point was added to the outlier score: (1) For each F1–F2 point, distance from the mean value of their corresponding age and gender group was larger than 3.5 standard deviations. (2) Average point-wise distance from the mean for the short F1–F2 tracks was larger than 2.5 standard deviations. (3) Absolute angle difference between short F1–F2 tracks and the mean tracks was larger than 60° (the track direction was determined using linear regression). (4) Average point-wise distance from the mean for the long F1–F2 tracks was larger than 2.0 standard deviations. The total outlier scores for each F1–F2 track were added up and F1–F2 tracks with scores larger than 20 were automatically labeled as "outliers." For example, an outlier score of 13 could be achieved by having 6 (out of possible 20) outliers of type 1, 4 (out of possible 16) outliers of type 2 and 3 (out of possible 12) outliers of type 4. Note that the process was run iteratively, i.e., once outliers were identified they were removed from the data, the mean, and variance statistics were recomputed for each age and gender group, excluding the outliers, and the outlier detection process was repeated until no more outliers were found. Overall 197 outliers were automatically detected after four iterations. Manual inspection showed that 99% of automatically labeled outliers were actual outliers.

Furthermore, F1–F2 tracks with total outlier score between 10 and 20 were labeled as "candidate outliers" and manually inspected. Out of 371 candidate outliers, 109 were labeled as actual outliers, resulting in a total of 306 outliers (197 automatically detected outliers plus 109 manually detected ones) by this procedure. The number of outliers was higher for the 7–13 age group and for the diphthong /OW/. Other than that the outlier distribution was relatively uniform across ages, diphthong identity and age groups.

Following the automatic cleaning procedure and further manual removal of outliers a total 3952 tokens were retained for the measurements of onset and offset values of the pitch contour and formant trajectories, as well as formant transition rates. For specifying onset and offset values (for pitch and the formant frequencies), the averaged values of the first two and the last two measurements of the uniformly sampled pitch and formant trajectories were used. Note that two sample points correspond to 10% of the total diphthong duration. For formant transition rate, we first computed sample-by-sample differences

in a given smoothed and duration normalized formant trajectory, and then computed the average transition rate per track (or alternatively the maximum transition rate).

C. Positional relationship between the onsets and offsets of diphthongs and nearby monophthongs

The onset and offset phonetic quality of diphthongs has been traditionally transcribed using monophthongal symbols that are assumed to have similar phonetic qualities. Although it is acknowledged that such notations are of a matter of convenience and do not fully capture true phonetic qualities, it would be nevertheless interesting to examine how diphthong onset and offset formant values are positioned with respect to those of nearby (in terms of distance in the F1–F2 space) monophthongs and how they evolve as a function of age.

In order to effectively visualize diphthong trajectories in the F1–F2 plane, a so-called "solenoid" plot was created for each diphthong that shows mean values and associated standard deviations along the diphthong trajectory. Transition rates are shown at duration-normalized locations with arrows of proportional length. In order to form a solenoid or a strip along the mean trajectory in the F1–F2 space, the standard deviation ellipsis is computed at 20 uniformly sampled positions and the non-convex hulls formed by overlapping ellipses are plotted. Finally, in order to compare onset and offset formant positions of diphthongs to those of monophthongs for a given age and gender group, averaged F1 and F2 positions of nine American English monophthongs examined in the previous study[1] on the same corpus [i.e., /IY/ (bead), /IH/ (bit), /EH/ (bet), /AE/ (bat), /AA/ (pot), /AH/ (but), /AO/ (ball), /UH/ (put), /UW/ (boot)] were overlaid on the plot.

To examine in more detail how the distance between the onset and offset positions of diphthongs and monophthongs evolves as a function of age, we computed Euclidean distances in the F1–F2 space between onset or offset of a given diphthong and nearby monophthongs. In order to minimize distance biases due to differences in harmonic spacing and the size of the vocal tract, formant differences between onset or offset positions of a given diphthong and a nearby monophthong were first scaled by the formant scaling factor before computing Euclidean distances. Scaling factors were estimated for each age and gender group with respect to adult subjects and are also plotted independently. The selection of nearby monophthongs was based on visual inspection of formant positions in the F1–F2 space (see Fig. 7). The selected monophthongs are explicitly listed in Sec. IV E 2.

D. Estimation of a transitional landmark point in diphthong trajectory

Diphthong trajectories are also analyzed using the mel-frequency cepstral coefficient (MFCC) representation of speech signals, computed as the discrete cosine transform of the smooth spectral envelope. Our main motivation is to determine a landmark point in each diphthong trajectory to analyze age-dependent developmental trends in diphthong production for the transition segment. One such a landmark introduced in this study is the mid-point of a diphthong transitional portion between onset and offset. This landmark divides a given diphthong trajectory into two demisyllable segments. The elapsed time from onset to landmark can be interpreted as the minimal articulatory action duration required to successfully produce perceptually relevant diphthong sounds. The spectral mid-point may be regarded as a via-point of the diphthong trajectory in the F1–F2 space and a control point in diphthong production.[15] It will be shown that onset to landmark duration is different across diphthong types and it is generally longer for younger-age speakers.

The landmark was estimated using the following procedure depicted also in Fig. 1 for an example diphthong token /AY/: First, 12th-order mel-frequency cepstral coefficients (MFCCs) were estimated using a 20 msec analysis window (frame) with a 2 msec window update. A "forward" Euclidean MFCC (cepstral) distance was estimated between each analysis frame and the (average of the two) first frames, shown as the line with increasing values in Fig. 1. Similarly a "backward" cepstral distance was computed between each analysis frame and the average of the two last frames (line with decreasing values in Fig. 1). The crossing point of the two curves, where the forward and backward distances are equal, is defined as the landmark. It is observed that the landmark (i.e., the frame with equal cepstral distance to the beginning and end of the diphthong) is located close to the point of maximal spectral transition in the diphthong.

E. Classification of diphthongs based on formant trajectories

Finally, in order to assess the effectiveness of diphthong trajectory parameters of onset, offset and the maximum formant transition rate in distinguishing five diphthongs, Fisher's discriminant analysis with leave-one-out cross-validation was applied to a set of different combinations of diphthong formant trajectory parameters: (1) F1–F3 onsets only; (2) F1–F3 offsets only; (3) maximum F1–F3 rates

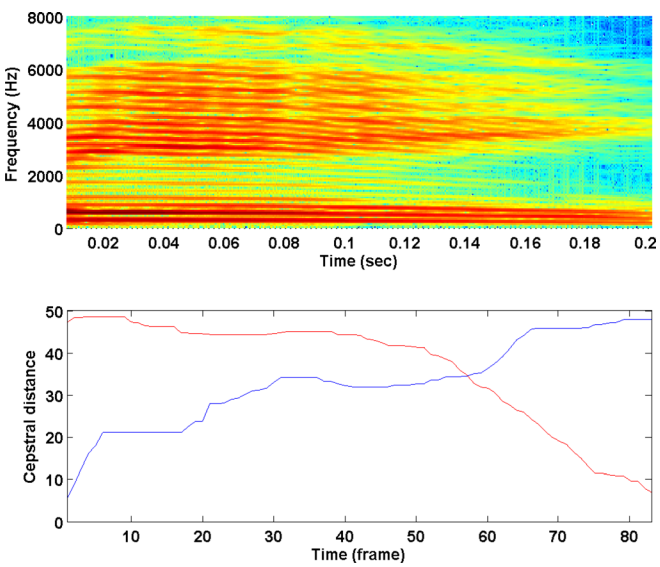

FIG. 1. (Color online) Example of front-back cepstral distances to estimate a spectral transition mid-point (bottom) together with the corresponding narrow-band spectrogram (top). Token is /AY/produced by a child of age 5 years old.

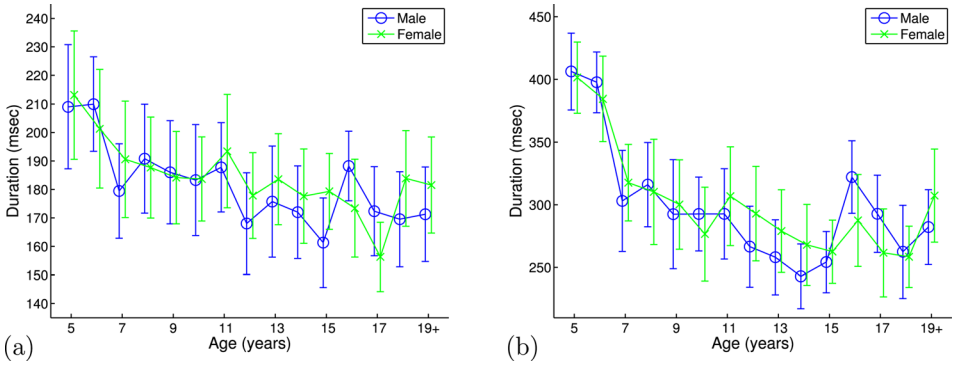

FIG. 2. (Color online) (a) Averaged durations for diphthongs /EY/, /AY/, /AW/, /OW/ for male and female subjects in each age group. (b) Averaged duration for diphthong OY/ per gender and age group. Note that the *y* axis scale is different due to the difference in dynamic range of values in the two plots.

only; (4) Onsets plus offsets; (5) Onsets plus maximum F1–F3 rates; (6) Offsets plus maximum F1–F3 rates; and (7) Onsets, offsets, and maximum F1–F3 rates combined.

IV. RESULTS

In this section, we present results obtained by the measurement procedures described in Sec. III. All statistical significance tests are performed using two-factor (analysis of variance) ANOVAs (e.g., age and gender, age and diphthong type) and Bonferroni *post hoc* tests using the SPSS software package. Because of the nature of the acoustic data, age-dependent or gender-dependent effects are usually either clearly observable or non-existent, thus "*p*-value" is reported only when necessary. It is also noted that results on intra-speaker variability associated with the two repetitions of same target diphthong are not presented here because they exhibit similar trends with monophthong production examined in the previous study.[15]

A. Diphthong duration

Diphthong durations for male and female speakers as a function of age are shown in Fig. 2. Specifically, in Fig. 2(a) duration values averaged over diphthongs /EY/, /AY/, /AW/, /OW/ are shown together and in Fig. 2(b) duration values for /OY/ only, since it displays a distinct trend. Error bar length equals one standard deviation. ANOVA indicates that there is no significant gender difference in diphthong durations for each of the five diphthongs (e.g., $p = 0.06$ for /OW/, $p = 0.83$ for /AW/). A two-factor (i.e., diphthong type and age) ANOVA on duration indicates that both diphthong and age effects are significant ($p < 0.01$). *Post hoc* analysis indicates that diphthong /OY/ exhibits a significantly large duration when compared to the other four diphthongs. Among the four remaining diphthongs, /AW/ exhibits a statistically significant longer duration than the other three diphthongs.

As can be observed in Figs. 2(a) and 2(b), diphthong duration does not decrease monotonically with age but shows a nonlinear oscillatory pattern. A statistically significant reduction of duration is observed between ages 11 and 12, followed by a significant increase between ages 15 and 16. The analysis of group variability in duration also indicates that the variability decreases in an oscillatory fashion, reaches its minimum at around age 15 and then it increases again toward adult levels.

For reference, the average durations for each age group and diphthong are shown in Tables II and III for male and female subjects, respectively (standard deviation in parentheses).

B. Fundamental frequency

Since diphthong sounds are transitional in nature from onset to offset, the fundamental frequency was analyzed

TABLE II. Mean and standard deviation (in parentheses) of duration in msec for male speakers.

Age	/EY/	/AY/	/AW/	/OW/	/OY/
5	206 (39)	203 (49)	215 (44)	213 (41)	406 (61)
6	206 (29)	200 (36)	219 (34)	214 (33)	398 (48)
7	168 (28)	193 (39)	176 (36)	183 (28)	303 (81)
8	199 (37)	193 (43)	179 (36)	193 (35)	316 (67)
9	186 (33)	184 (40)	184 (37)	191 (36)	292 (87)
10	182 (40)	177 (37)	187 (45)	188 (34)	293 (59)
11	187 (33)	182 (30)	186 (30)	196 (32)	293 (72)
12	170 (38)	161 (29)	166 (36)	175 (39)	267 (64)
13	176 (37)	167 (36)	180 (45)	179 (38)	258 (60)
14	171 (36)	168 (30)	168 (34)	180 (30)	243 (52)
15	161 (31)	159 (32)	163 (35)	162 (29)	254 (49)
16	196 (29)	188 (26)	182 (21)	187 (20)	322 (58)
17	174 (31)	167 (31)	178 (35)	170 (30)	293 (62)
18	167 (36)	156 (25)	174 (37)	181 (32)	262 (74)
19+	175 (31)	167 (28)	173 (40)	170 (34)	282 (60)

TABLE III. Mean and standard deviation (in parentheses) of duration in msec for female speakers.

Age	/EY/	/AY/	/AW/	/OW/	/OY/
5	218 (40)	207 (55)	217 (44)	210 (44)	401 (57)
6	191 (32)	202 (45)	214 (49)	198 (38)	384 (68)
7	186 (36)	188 (43)	191 (43)	198 (42)	318 (61)
8	185 (36)	189 (37)	178 (37)	197 (32)	310 (84)
9	184 (31)	182 (32)	182 (35)	189 (32)	300 (71)
10	183 (26)	187 (28)	173 (26)	192 (35)	277 (75)
11	191 (40)	191 (35)	195 (43)	197 (42)	307 (79)
12	183 (30)	175 (30)	170 (30)	184 (30)	293 (75)
13	183 (26)	181 (24)	186 (42)	184 (36)	279 (66)
14	180 (38)	170 (25)	170 (29)	191 (37)	268 (65)
15	177 (23)	174 (26)	177 (25)	189 (32)	263 (50)
16	178 (34)	169 (29)	166 (34)	182 (40)	288 (73)
17	162 (19)	153 (23)	151 (28)	159 (26)	262 (70)
18	183 (30)	181 (34)	179 (39)	193 (32)	258 (49)
19+	192 (32)	173 (31)	175 (34)	187 (35)	307 (74)

separately at the two target positions of onset and offset as a function of age for each gender. Results of the onset F0 are shown in Fig. 3(a) for male and female subjects (error bar length equals 1 standard deviation) and the offset F0 in Fig. 3(b). The expected decreasing trend of onset F0 is observed for each gender. Two-factor ANOVA indicates both age and diphthong effects are significant ($p < 0.00$) in each gender.

For male subjects, it is clear that on average the pubertal pitch change starts from age 12 and ends at age 15. Intermediate F0 values at age 13 and 14 might be originated from mixed speakers of before and after individual pubertal period. After age 15, F0 reaches adult level, although the differences in magnitude are not statistically significant.

For female subjects there exist several age boundaries where sudden increase or decrease of onset F0 occur. For instance, there exist a sudden overshoot of F0 at age 12 to age 13 and an immediate F0 undershoot from age 13 to age 14 ($p < 0.00$). There also exist a sudden undershoot at age 16 to age 17 ($p < 0.00$), and an immediate overshoot from age 17 to age 18 ($p < 0.00$). Results for the offset F0 are also shown in Fig. 3(b) for male and female subjects. The behaviors of the offset F0 are similar to those of onset F0 as a function of age and gender, except the lower F0 values for children of age 5 and 6.

As mentioned previously, the onset F0 values are significantly different among diphthongs and on average the offset F0 values are smaller than the onset F0 values. In general, /AW/ shows the largest onset F0 value as well as onset-offset F0 difference (21.9 Hz for male, 28.4 Hz for female) and /AY/ the lowest F0 value and smallest onset-offset difference (7.5 Hz for male and 8.0 Hz for female).

Regarding the group variability of F0 onset and offset values, it decreases as a function of age for both genders. The F0 variability for male subjects suddenly increases at age 13, reaches minimum value at around age 16 and then increases toward adult levels. For female subjects, F0 variability reaches minimum levels around age 14 and then increases toward adult levels. Also, the variability trends across age appear more oscillatory when compared to male speakers.

For reference, the average fundamental frequency for each age group and diphthong is shown in Tables IV and V for male and female subjects, respectively (standard deviation in parentheses).

C. Onset and offset formant frequencies

In Fig. 4, the formant (F1, F2, F3) scaling factors per age and gender group averaged over all diphthongs are shown. Scaling factors are computed relative to adult male formant values. Results are shown for diphthong onset in (a) and offset in (b), The plots serve as a summary of formant change as a function of age and gender. Overall, the scaling factors for all formants follow the expected decreasing trend toward adult values for both diphthong onsets and offsets. The only exceptions are the F2 offset scaling values for female subjects that are significantly higher than expected for ages 10 and older.

In Fig. 5(a)–5(d), diphthong onset and offset frequencies of the first two formants (F1 and F2) are shown for female subjects. Both genders show similar tendencies so only results for female subjects are shown here. Averaged onset (on the left) and offset (on the right) formant frequencies for each diphthong are shown as a function age. Multivariate ANOVA results indicate that the age and diphthong effects are statistically significant ($p < 0.00$), confirming the visual observations. It can be seen that the onset or offset formant frequencies are generally different (and in some cases significantly different) across the five diphthongs. For instance, F1 onset values are significantly different between /AY/ and /AW/, as well as between /OW/ and /OY/, although each diphthong pair is supposed to have similar phonetic qualities. For F1 offset, the formant values are significantly different between /EY/ and /AY/. This is even more clear between /AW/ and /OW/.

F2 onset and offset frequencies show similar trends as in the case of F1. For both males and females [shown in Fig. 5(c) and 5(d)], the F2 onset is different for /OW/ and /OY/ and also for /AY/ and /AW/ to a lesser degree. For F2 offset, such differences seem somewhat reduced especially for /OW/ and /AW/ but difference between /EY/ and /OY/ is statistically significant ($p < 0.01$).

In the case of F3, there are not much onset and offset differences among diphthongs and age-dependent trends are mainly observed (so these plots are omitted). In general, age (formant scaling) trends can be observed much clearer for F3 for all five diphthongs for both genders.

Onset and offset variability of the first three formant frequencies was also analyzed across age groups. As expected, variability decreases as a function of age. For F2, offset

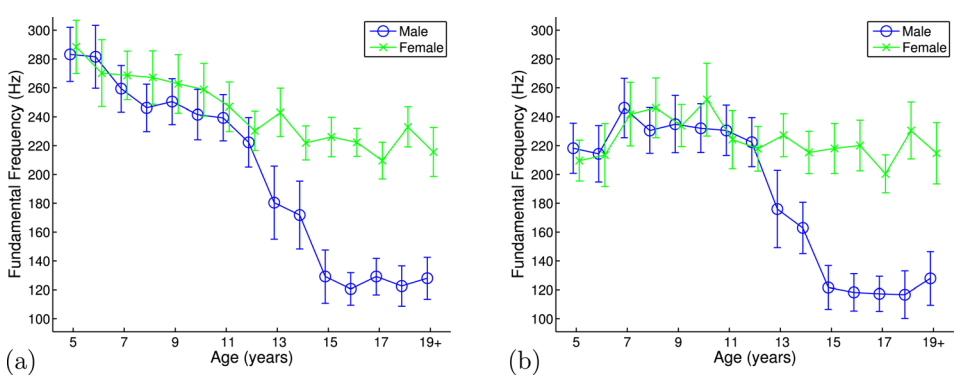

FIG. 3. (Color online) Average fundamental frequency for male and female subjects measured at (a) diphthong onset, (b) diphthong offset.

TABLE IV. Mean and standard deviation (in parentheses) of fundamental frequency and formant frequency values for male speakers (# is the number of tokens).

Age		EY onset	EY offset	AY onset	AY offset	AW onset	AW offset	OW onset	OW offset	OY onset	OY offset
5	#	29	29	26	26	24	24	27	27	25	25
	F0	282 (41)	225 (37)	279 (37)	224 (35)	288 (34)	220 (29)	289 (41)	221 (30)	277 (35)	199 (38)
	F1	702 (91)	455 (68)	932 (119)	612 (188)	1070 (146)	635 (80)	675 (99)	508 (86)	535 (93)	453 (86)
	F2	2601 (335)	3370 (326)	1579 (235)	2740 (398)	1968 (262)	1090 (176)	1364 (161)	1164 (540)	989 (192)	2833 (342)
	F3	3723 (303)	3987 (367)	3497 (358)	3674 (320)	3395 (461)	3612 (383)	3623 (255)	3688 (364)	3633 (340)	3742 (355)
6	#	16	16	15	15	19	19	18	18	17	17
	F0	279 (39)	224 (44)	278 (45)	226 (42)	279 (51)	209 (35)	292 (47)	205 (38)	278 (36)	211 (38)
	F1	666 (79)	473 (62)	853 (111)	548 (74)	1028 (140)	636 (97)	701 (92)	515 (66)	575 (80)	431 (54)
	F2	2561 (234)	3396 (324)	1541 (171)	3056 (307)	1935 (286)	1180 (152)	1439 (201)	1110 (293)	928 (130)	2959 (382)
	F3	3507 (253)	3935 (252)	3362 (265)	3630 (354)	3256 (248)	3365 (409)	3482 (279)	3496 (276)	3629 (235)	3859 (184)
7	#	17	17	17	17	20	20	17	17	17	17
	F0	261 (32)	248 (41)	245 (26)	240 (38)	277 (31)	256 (42)	254 (29)	252 (41)	256 (36)	233 (42)
	F1	565 (79)	465 (61)	850 (112)	541 (76)	988 (127)	626 (141)	682 (97)	480 (103)	575 (114)	467 (106)
	F2	2588 (170)	3133 (242)	1565 (175)	2746 (470)	1772 (194)	1282 (227)	1361 (174)	1308 (343)	967 (247)	2590 (346)
	F3	3307 (118)	3749 (296)	3302 (291)	3490 (315)	3131 (303)	3425 (296)	3465 (244)	3445 (279)	3610 (329)	3494 (209)
8	#	34	34	44	44	40	40	38	38	40	40
	F0	246 (27)	237 (29)	239 (28)	227 (31)	260 (44)	231 (36)	244 (31)	230 (35)	242 (28)	228 (27)
	F1	636 (87)	452 (49)	860 (94)	527 (66)	966 (114)	609 (84)	668 (80)	474 (73)	534 (58)	449 (55)
	F2	2456 (294)	2993 (365)	1472 (164)	2801 (264)	1689 (232)	1248 (163)	1309 (154)	1170 (225)	940 (144)	2533 (302)
	F3	3360 (307)	3537 (298)	3268 (310)	3433 (361)	3121 (354)	3354 (259)	3358 (396)	3455 (327)	3540 (279)	3437 (353)
9	#	38	38	38	38	42	42	34	34	30	30
	F0	247 (27)	238 (48)	243 (28)	236 (36)	259 (38)	230 (34)	250 (32)	234 (40)	252 (30)	237 (42)
	F1	583 (69)	430 (79)	874 (99)	547 (109)	974 (135)	612 (109)	662 (62)	487 (89)	546 (54)	452 (88)
	F2	2505 (212)	3104 (254)	1518 (172)	2767 (262)	1780 (236)	1202 (197)	1308 (118)	1201 (176)	953 (119)	2514 (467)
	F3	3343 (232)	3646 (277)	3128 (296)	3389 (289)	3090 (258)	3366 (326)	3428 (206)	3497 (281)	3580 (248)	3473 (248)
10	#	34	34	43	43	39	39	41	41	43	43
	F0	233 (43)	228 (37)	240 (33)	233 (31)	247 (35)	230 (32)	246 (33)	238 (36)	241 (30)	231 (33)
	F1	615 (81)	452 (56)	858 (81)	554 (94)	969 (96)	620 (98)	670 (72)	478 (72)	518 (50)	458 (59)
	F2	2442 (195)	2839 (388)	1543 (176)	2657 (348)	1812 (201)	1236 (197)	1382 (137)	1175 (186)	933 (107)	2522 (224)
	F3	3247 (200)	3506 (242)	3025 (292)	3323 (361)	2986 (327)	3160 (298)	3155 (225)	3099 (293)	3198 (296)	3271 (253)
11	#	39	39	44	44	40	40	38	38	40	40
	F0	236 (34)	229 (33)	238 (29)	237 (34)	244 (38)	226 (39)	239 (29)	231 (39)	240 (29)	229 (29)
	F1	577 (54)	440 (55)	842 (79)	531 (77)	908 (120)	602 (71)	621 (57)	458 (54)	508 (49)	462 (57)
	F2	2421 (194)	2795 (265)	1393 (133)	2616 (212)	1702 (231)	1240 (180)	1239 (137)	1143 (161)	861 (87)	2405 (312)
	F3	3121 (239)	3356 (262)	2932 (313)	3303 (277)	2837 (321)	2951 (326)	2973 (202)	2943 (230)	3069 (368)	3217 (308)
12	#	42	42	37	37	41	41	36	36	39	39
	F0	220 (32)	225 (34)	216 (34)	221 (32)	229 (40)	219 (34)	222 (31)	226 (34)	225 (34)	221 (38)
	F1	553 (70)	428 (52)	779 (75)	550 (106)	888 (92)	596 (102)	620 (59)	458 (51)	479 (40)	441 (43)
	F2	2308 (219)	2759 (233)	1380 (127)	2368 (302)	1562 (193)	1181 (125)	1247 (120)	1111 (158)	873 (74)	2269 (283)
	F3	3103 (201)	3309 (306)	2893 (233)	3041 (279)	2786 (303)	2905 (328)	2971 (226)	2891 (273)	3042 (251)	3134 (281)
13	#	29	29	30	30	28	28	25	25	26	26
	F0	176 (50)	173 (53)	173 (51)	175 (54)	194 (53)	185 (55)	182 (50)	173 (56)	178 (51)	173 (54)
	F1	554 (52)	395 (59)	758 (58)	511 (86)	801 (85)	589 (73)	588 (68)	440 (63)	511 (51)	438 (51)
	F2	2109 (199)	2514 (241)	1364 (165)	2221 (232)	1504 (193)	1163 (137)	1204 (115)	1179 (200)	842 (114)	2104 (265)
	F3	2813 (187)	3005 (231)	2662 (220)	2815 (201)	2672 (215)	2637 (158)	2757 (185)	2690 (222)	2827 (241)	2848 (306)
14	#	21	21	20	20	21	21	16	16	19	19
	F0	165 (50)	159 (39)	172 (43)	164 (35)	174 (49)	159 (37)	183 (51)	170 (37)	167 (45)	163 (33)
	F1	538 (46)	410 (51)	794 (101)	492 (64)	841 (102)	604 (61)	625 (60)	477 (47)	518 (48)	444 (50)
	F2	2224 (291)	2513 (417)	1442 (181)	2431 (236)	1557 (182)	1198 (86)	1256 (119)	1201 (216)	831 (110)	2136 (326)
	F3	2881 (194)	3059 (215)	2730 (168)	2918 (190)	2573 (200)	2727 (234)	2846 (151)	2778 (157)	2823 (196)	2825 (233)
15	#	20	20	21	21	18	18	19	19	18	18
	F0	128 (38)	121 (27)	125 (33)	122 (31)	136 (42)	124 (32)	132 (38)	123 (32)	127 (36)	118 (33)
	F1	517 (53)	367 (42)	706 (71)	461 (45)	747 (79)	593 (44)	581 (67)	418 (61)	518 (74)	391 (49)
	F2	1964 (144)	2363 (155)	1362 (141)	2150 (171)	1484 (197)	1195 (83)	1235 (157)	1248 (234)	772 (135)	2025 (190)
	F3	2687 (155)	2886 (189)	2545 (163)	2714 (181)	2526 (122)	2546 (192)	2664 (134)	2714 (215)	2897 (348)	2799 (189)
16	#	20	20	20	20	13	13	13	13	9	9
	F0	121 (22)	115 (27)	114 (21)	116 (24)	124 (28)	119 (33)	124 (21)	118 (24)	125 (19)	128 (20)
	F1	498 (54)	354 (54)	727 (93)	413 (76)	782 (107)	550 (40)	581 (41)	470 (131)	494 (72)	383 (28)
	F2	1955 (165)	2319 (207)	1358 (182)	2201 (206)	1498 (160)	1113 (163)	1231 (122)	1465 (512)	968 (489)	2040 (207)

TABLE IV. (*Continued*)

Age		EY onset	EY offset	AY onset	AY offset	AW onset	AW offset	OW onset	OW offset	OY onset	OY offset
	F3	2705 (120)	2846 (167)	2643 (235)	2777 (195)	2617 (194)	2646 (244)	2677 (310)	2747 (292)	2779 (234)	2848 (223)
17	#	19	19	18	18	16	16	13	13	6	6
	F0	128 (23)	117 (25)	125 (25)	117 (25)	136 (29)	120 (25)	131 (29)	116 (29)	124 (17)	115 (6)
	F1	525 (44)	327 (52)	737 (78)	485 (130)	738 (80)	567 (58)	588 (31)	436 (127)	504 (71)	360 (44)
	F2	1909 (177)	2346 (130)	1381 (236)	2073 (226)	1427 (192)	1204 (203)	1181 (222)	1534 (461)	755 (57)	2019 (181)
	F3	2681 (106)	2953 (142)	2707 (212)	2708 (121)	2633 (124)	2650 (294)	2710 (269)	2807 (352)	2908 (206)	2711 (84)
18	#	19	19	16	16	13	13	11	11	8	8
	F0	123 (26)	119 (30)	122 (25)	123 (30)	122 (32)	110 (34)	126 (34)	121 (40)	119 (27)	105 (36)
	F1	476 (40)	329 (37)	732 (82)	462 (55)	752 (50)	526 (73)	549 (84)	391 (71)	455 (103)	365 (59)
	F2	2035 (103)	2329 (93)	1368 (119)	2155 (130)	1466 (221)	1102 (319)	1239 (421)	1352 (424)	732 (74)	2085 (169)
	F3	2722 (121)	2876 (190)	2578 (147)	2765 (213)	2460 (233)	2647 (102)	2749 (193)	2749 (199)	2795 (135)	2738 (344)
19+	#	55	55	53	53	50	50	34	34	31	31
	F0	127 (28)	135 (40)	123 (29)	126 (37)	136 (32)	124 (35)	127 (26)	133 (34)	127 (30)	119 (37)
	F1	506 (50)	324 (49)	728 (66)	411 (83)	781 (92)	567 (101)	599 (84)	431 (109)	536 (83)	367 (59)
	F2	1936 (161)	2308 (148)	1271 (186)	2136 (180)	1447 (208)	1142 (404)	1439 (528)	1431 (361)	1006 (651)	2002 (212)
	F3	2686 (202)	2886 (226)	2564 (216)	2753 (227)	2551 (242)	2659 (246)	2823 (353)	2709 (346)	2768 (318)	2765 (332)

frequency values show significantly larger group variability compared to onsets across all age groups for all diphthongs except /AW/. For F1, variability is larger for onsets than offsets for the cases of /AW/, /AY/, and /EY/. For F3, the onset and offset variability shows no significant differences.

The average formant frequency values for the first three formants (F1, F2, F3) for each age group and diphthong is shown in Tables IV and V for male and female subjects, respectively (standard deviation in parentheses). The total number of diphthong instances in each group is also shown (different than those shown in Table I due to outlier removal).

D. Formant transition rate

Since the gender effect on F1, F2, and F3 transition rates is barely significant (e.g., $p = 0.06$) or not significant at all for all five diphthongs, each formant transition rate was pooled over across gender in each age group. As a representative result, error bar plots of averaged F2 transition rate for each age group are shown in Fig. 6(b) as a function of diphthong. The effects of both age and diphthong type on the F2 formant transition rates are significant ($p < 0.00$).

The average transition rates for F1 and F3 are shown in Fig. 6(a) and 6(c), respectively. It is clear from Fig. 6 that F1–F3 transition rates are different for different diphthongs. Specifically, in the case of F1, /AY/ and /AW/ exhibit relatively larger transition rate among the five diphthongs and the rest shows a similar level. In the case of F2, /AY/ and /OY/ show relatively larger transition rates when compared to the rest. In the case of F3, the transition rate seems to exhibit similar ranges of values across diphthongs, although /AY/ and /OY/ again show a statistically significant and larger transition rate than the rest.

One interesting observation is that the age dependent trends of all diphthongs show inflection points (i.e., minima) for both average values and standard deviation. For instance, such inflection points are most clearly observable in the F3 transition rates and the age boundary of inflection points occur at around age 14 and 15. For F1 and F2, the inflection points occur at about age 15 or 16 years old. Before the inflection points (i.e., ages 5 to 14 or 15 years old) the formant transition rates (and standard deviations) keep decreasing, and after that, they increase again toward adult levels.

It is also noted that very similar age- and diphthong-dependent trends have been observed for the maximum transition rate for F1–F3 (not shown here); the main difference being that the maximum transition rate is in absolute terms about twice the value of the average transition rates for all formants.

E. Positional relationship between onsets and offsets of diphthongs and monophthongs

1. Comparison of diphthong onset and offset positions with monophthong positions

In Fig. 7, five diphthong trajectories are illustrated as strips (i.e., solenoid plots) for two age groups (age 5 and age 14 years old). In each strip, arrows represent trajectory midlines between onset ("○") and offset ("+") and "strip width" represents variability at selected locations along the midline. In order to compare onset and offset positions of diphthongs with those of monophthongs, averaged F1 and F2 positions of nine monophthongs (i.e., /IY/, /IH/, /EH/, /AE/, /AA/, /AH/, /AO/, /UH/, /UW/), as squares, are also shown in Fig. 7.

Several interesting observations can be made from Fig. 7. First of all, it is confirmed that the onset and offset positions of diphthongs are different from those of monophthongs used for the transcription of diphthongs. For instance, in the case of /EY/, the onset position is very close to /IH/, not /EH/, and in the case of /AY/ the onset position is somewhat different from the monophthong /AA/, closer to /AH/ or in between. In the case of /AW/, the onset is close to /AA/, but the tongue body constriction may occur further back when compared to /AA/. For /OW/, the onset position is much closer to /UH/, than to /AO/. For /OY/, the onset

TABLE V. Mean and standard deviation (in parentheses) of fundamental frequency and formant frequency values for female speakers (# is the number of tokens).

Age		EY onset	EY offset	AY onset	AY offset	AW onset	AW offset	OW onset	OW offset	OY onset	OY offset
5	#	23	23	18	18	22	22	20	20	18	18
	F0	284 (36)	205 (29)	278 (35)	211 (36)	295 (33)	214 (24)	299 (43)	214 (25)	285 (37)	203 (28)
	F1	707 (93)	484 (69)	1017 (185)	560 (183)	1242 (152)	687 (142)	757 (107)	538 (83)	592 (79)	482 (83)
	F2	2773 (248)	3324 (450)	1737 (178)	3028 (524)	2146 (408)	1326 (231)	1527 (253)	1321 (461)	1048 (153)	2957 (303)
	F3	3759 (244)	4023 (267)	3512 (356)	3834 (217)	3650 (420)	3712 (282)	3715 (229)	3696 (329)	3707 (418)	3897 (305)
6	#	20	20	13	13	22	22	20	20	23	23
	F0	270 (51)	197 (35)	278 (49)	222 (51)	277 (46)	218 (36)	271 (54)	215 (55)	259 (34)	217 (42)
	F1	659 (97)	461 (48)	936 (112)	582 (267)	1103 (171)	682 (82)	711 (93)	555 (68)	573 (55)	471 (73)
	F2	2791 (178)	3207 (406)	1732 (206)	3023 (404)	2123 (318)	1229 (176)	1577 (243)	1220 (240)	1038 (134)	2730 (483)
	F3	3737 (313)	3906 (381)	3694 (333)	3811 (371)	3612 (322)	3659 (377)	3739 (240)	3625 (293)	3845 (248)	3778 (276)
7	#	39	39	39	39	35	35	31	31	39	39
	F0	274 (37)	242 (47)	262 (33)	245 (42)	274 (37)	242 (44)	272 (31)	240 (45)	264 (29)	240 (44)
	F1	674 (88)	513 (91)	949 (85)	599 (103)	1050 (125)	699 (129)	707 (75)	548 (62)	596 (69)	489 (91)
	F2	2660 (213)	3161 (274)	1705 (281)	2860 (240)	1879 (242)	1421 (223)	1476 (158)	1345 (286)	1048 (162)	2710 (377)
	F3	3607 (243)	3864 (309)	3449 (325)	3576 (319)	3493 (428)	3514 (393)	3576 (261)	3557 (320)	3661 (367)	3642 (270)
8	#	15	15	16	16	17	17	13	13	17	17
	F0	263 (27)	241 (36)	251 (28)	243 (37)	283 (43)	249 (45)	273 (36)	239 (24)	266 (44)	256 (56)
	F1	652 (106)	456 (45)	936 (92)	553 (160)	1090 (157)	664 (87)	727 (125)	511 (70)	563 (89)	494 (114)
	F2	2594 (258)	3186 (385)	1623 (191)	2659 (525)	1933 (222)	1370 (285)	1447 (182)	1263 (188)	1002 (115)	2515 (353)
	F3	3464 (309)	3939 (313)	3231 (415)	3537 (403)	3160 (396)	3377 (438)	3662 (353)	3730 (395)	3637 (202)	3631 (127)
9	#	37	37	35	35	36	36	34	34	35	35
	F0	256 (31)	235 (29)	245 (24)	237 (30)	286 (50)	233 (24)	263 (40)	234 (34)	264 (43)	231 (29)
	F1	634 (73)	465 (60)	938 (130)	567 (72)	1064 (121)	668 (87)	715 (51)	528 (80)	581 (75)	492 (92)
	F2	2542 (216)	3103 (278)	1672 (177)	2859 (239)	1885 (253)	1347 (180)	1393 (121)	1269 (142)	972 (133)	2617 (362)
	F3	3435 (225)	3685 (275)	3254 (291)	3478 (354)	3201 (301)	3361 (283)	3458 (242)	3439 (239)	3570 (257)	3542 (275)
10	#	18	18	22	22	21	21	24	24	23	23
	F0	270 (40)	255 (50)	252 (36)	251 (49)	266 (39)	258 (49)	258 (29)	251 (54)	251 (39)	247 (53)
	F1	645 (81)	491 (51)	945 (93)	602 (101)	1019 (96)	681 (111)	705 (88)	552 (89)	543 (84)	511 (103)
	F2	2437 (242)	2798 (406)	1605 (150)	2613 (290)	1849 (201)	1369 (220)	1462 (153)	1287 (182)	1020 (118)	2492 (273)
	F3	3234 (334)	3436 (307)	3021 (250)	3182 (304)	3028 (286)	3136 (278)	3242 (290)	3178 (305)	3312 (288)	3379 (245)
11	#	34	34	37	37	37	37	34	34	33	33
	F0	242 (31)	222 (39)	243 (31)	229 (39)	254 (41)	226 (42)	247 (32)	225 (46)	247 (34)	218 (38)
	F1	620 (71)	457 (57)	879 (81)	576 (86)	1008 (107)	672 (106)	696 (62)	511 (73)	548 (57)	476 (66)
	F2	2461 (160)	2942 (210)	1614 (163)	2637 (175)	1874 (187)	1362 (220)	1463 (131)	1265 (143)	999 (117)	2454 (206)
	F3	3245 (187)	3402 (223)	3079 (187)	3229 (187)	2986 (200)	3020 (266)	3133 (159)	3115 (177)	3221 (266)	3205 (181)
12	#	35	35	38	38	37	37	31	31	36	36
	F0	233 (24)	222 (32)	224 (21)	221 (28)	238 (31)	217 (29)	229 (34)	220 (38)	228 (25)	210 (29)
	F1	621 (68)	458 (42)	877 (111)	562 (63)	911 (129)	650 (85)	658 (58)	480 (62)	522 (62)	444 (56)
	F2	2277 (272)	2687 (412)	1580 (157)	2484 (304)	1714 (265)	1298 (189)	1381 (159)	1266 (194)	932 (100)	2362 (276)
	F3	3116 (273)	3285 (238)	2875 (353)	3132 (246)	2851 (399)	2917 (378)	3045 (261)	2975 (259)	3216 (314)	3168 (257)
13	#	24	24	22	22	24	24	22	22	21	21
	F0	241 (41)	226 (37)	241 (33)	228 (25)	253 (35)	232 (27)	237 (29)	230 (31)	242 (28)	220 (29)
	F1	629 (90)	439 (57)	854 (84)	561 (74)	960 (92)	637 (92)	675 (61)	529 (52)	511 (43)	474 (46)
	F2	2375 (187)	2811 (250)	1569 (143)	2558 (177)	1754 (181)	1310 (194)	1382 (201)	1251 (228)	895 (101)	2330 (196)
	F3	3139 (231)	3309 (230)	2881 (221)	3237 (249)	2854 (194)	2888 (268)	3062 (149)	3066 (166)	3148 (153)	3156 (198)
14	#	19	19	19	19	15	15	17	17	18	18
	F0	221 (23)	214 (28)	212 (31)	207 (34)	229 (22)	221 (24)	228 (19)	220 (26)	223 (19)	216 (33)
	F1	563 (70)	431 (50)	838 (66)	560 (84)	894 (90)	624 (81)	651 (43)	506 (66)	498 (46)	459 (34)
	F2	2250 (130)	2618 (166)	1568 (115)	2382 (192)	1613 (140)	1353 (134)	1426 (141)	1371 (191)	947 (78)	2231 (159)
	F3	2893 (177)	3060 (182)	2730 (360)	2986 (192)	2734 (365)	2753 (193)	2855 (218)	2842 (223)	2997 (139)	3017 (168)
15	#	19	19	19	19	20	20	18	18	18	18
	F0	224 (27)	215 (34)	218 (26)	223 (35)	229 (28)	214 (29)	231 (27)	220 (32)	228 (29)	218 (45)
	F1	584 (60)	406 (39)	816 (55)	519 (75)	913 (97)	623 (106)	650 (60)	491 (43)	518 (39)	471 (56)
	F2	2194 (190)	2603 (218)	1494 (103)	2425 (224)	1627 (189)	1324 (140)	1394 (183)	1328 (142)	905 (122)	2163 (178)
	F3	2897 (141)	2998 (251)	2752 (205)	2922 (248)	2715 (175)	2758 (239)	2774 (195)	2760 (181)	2852 (215)	2823 (124)
16	#	19	19	21	21	18	18	21	21	19	19
	F0	221 (22)	217 (36)	220 (19)	227 (37)	228 (21)	216 (34)	223 (18)	221 (33)	221 (20)	218 (38)
	F1	567 (61)	434 (36)	843 (75)	544 (59)	875 (91)	641 (84)	666 (53)	486 (83)	527 (63)	462 (44)
	F2	2295 (117)	2661 (285)	1541 (130)	2429 (153)	1623 (155)	1318 (119)	1395 (203)	1402 (183)	871 (117)	2168 (156)

TABLE V. (*Continued*)

Age		EY onset	EY offset	AY onset	AY offset	AW onset	AW offset	OW onset	OW offset	OY onset	OY offset
	F3	2990 (182)	3158 (221)	2785 (200)	3000 (182)	2686 (292)	2697 (289)	2861 (197)	2880 (229)	3005 (191)	3014 (230)
17	#	17	17	16	16	16	16	14	14	18	18
	F0	203 (29)	199 (36)	204 (15)	198 (18)	211 (35)	198 (34)	218 (22)	209 (20)	214 (21)	199 (19)
	F1	553 (53)	434 (55)	784 (56)	524 (79)	886 (93)	649 (124)	631 (45)	478 (64)	503 (52)	452 (48)
	F2	2316 (165)	2680 (163)	1528 (161)	2427 (181)	1585 (163)	1273 (165)	1283 (260)	1284 (146)	909 (98)	2239 (188)
	F3	2970 (204)	3029 (266)	2780 (158)	2932 (164)	2667 (163)	2732 (147)	2787 (168)	2743 (123)	2867 (230)	2951 (222)
18	#	18	18	16	16	19	19	17	17	20	20
	F0	228 (32)	236 (42)	224 (34)	228 (44)	245 (14)	234 (34)	235 (18)	230 (35)	230 (33)	225 (43)
	F1	601 (79)	440 (72)	860 (43)	566 (74)	932 (66)	629 (87)	652 (44)	498 (55)	503 (29)	467 (57)
	F2	2288 (137)	2695 (146)	1520 (111)	2425 (196)	1626 (187)	1272 (212)	1320 (134)	1223 (189)	914 (53)	2284 (230)
	F3	2930 (134)	3133 (210)	2862 (138)	2906 (309)	2768 (133)	2792 (125)	2914 (114)	2830 (152)	2986 (190)	2949 (183)
19+	#	49	49	52	52	51	51	46	46	47	47
	F0	211 (29)	216 (34)	204 (30)	218 (41)	233 (44)	217 (47)	216 (27)	217 (40)	213 (31)	204 (49)
	F1	577 (53)	394 (60)	830 (69)	546 (89)	887 (91)	613 (121)	622 (59)	448 (60)	522 (70)	421 (61)
	F2	2245 (154)	2714 (246)	1433 (144)	2332 (231)	1639 (187)	1192 (194)	1245 (160)	1149 (212)	855 (102)	2281 (221)
	F3	2909 (144)	3143 (244)	2829 (180)	2895 (202)	2769 (258)	2876 (285)	2843 (149)	2840 (173)	2917 (164)	2947 (161)

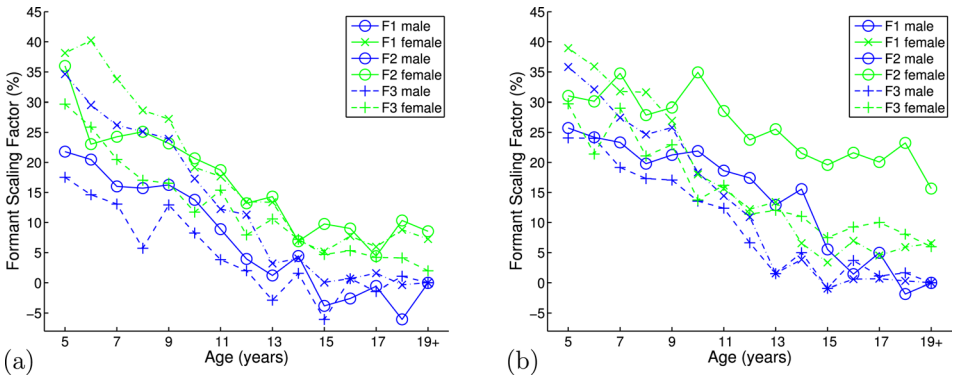

FIG. 4. (Color online) Average formant scaling factors for F1–F3 male and female subjects measured at (a) diphthong onset, (b) diphthong offset. Scaling is computed relative to adult male formant frequency values.

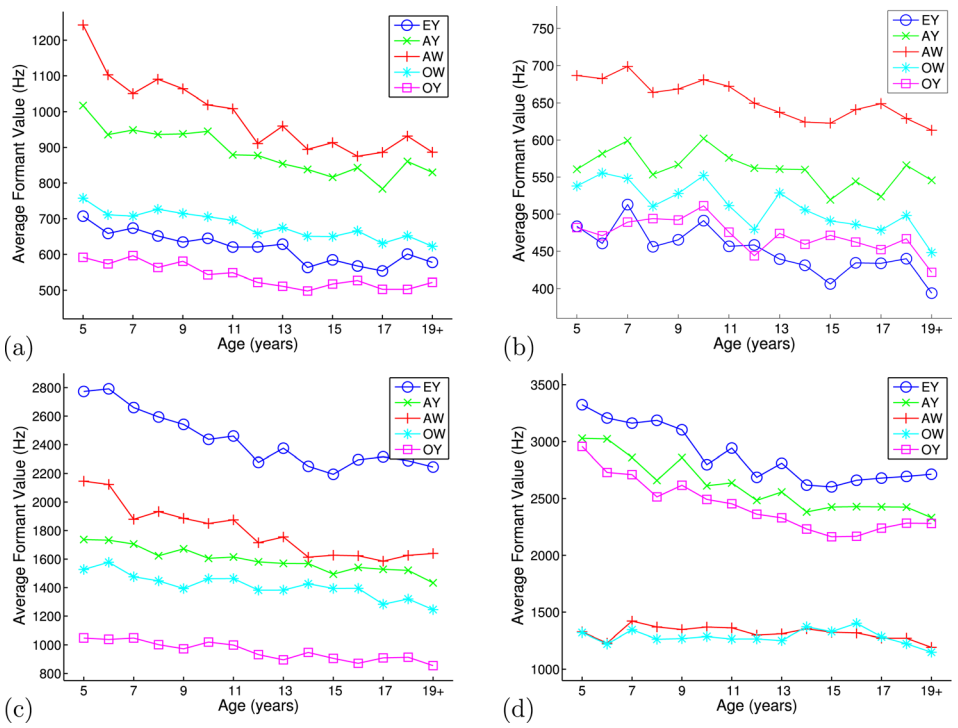

FIG. 5. (Color online) Average formant values for female subjects as a function of age group and diphthong: (a) F1 at diphthong onset, (b) F1 at diphthong offset, (c) F2 at diphthong onset, and (d) F2 at diphthong offset.

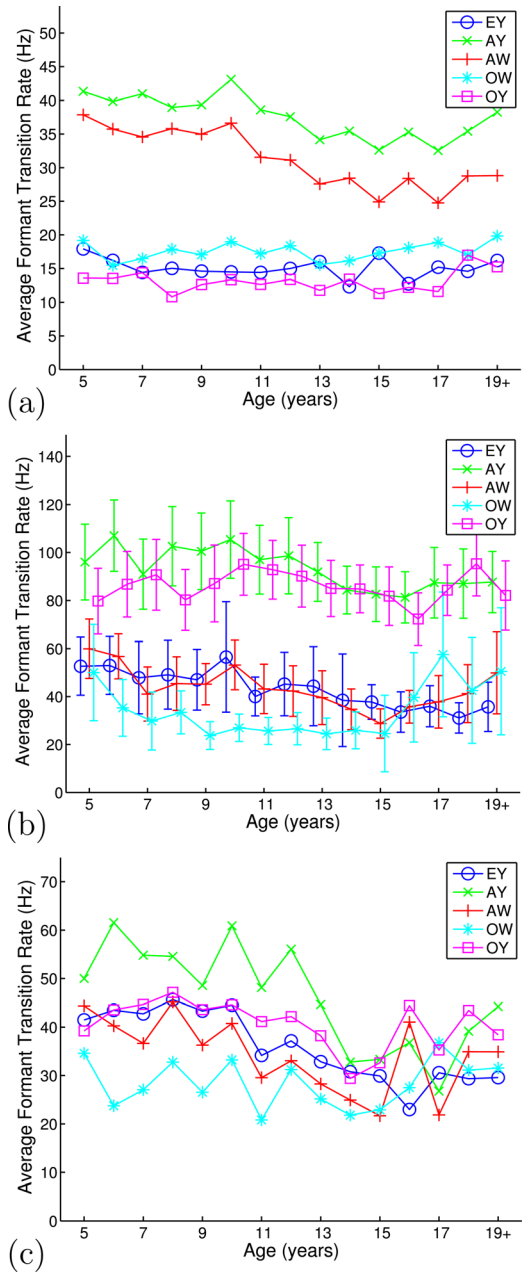

FIG. 6. (Color online) Average formant transition rates for F1–F3 for each diphthong and age group (averaged over male and female subjects): (a) F1 average rate, (b) F2 average rate and standard deviation, and (c) F3 average rate.

position seems well separated from the back and round vowels, especially for the older age groups.

As for the offset, it can be observed that only /EY/ may reach its supposed monophthongal offset target, i.e., /IY/. For /AY/, its offset position may end at around /IH/, or in between /IY/ and /IH/ at best. In the cases of /AW/ and /OW/, their offset positions are quite different with respect to each other. In addition, the offset position of /AW/ ends at near /UH/, not /UW/. In the case of /OW/, the offset may exist relatively far from both /UH/ and /UW/.

In addition to the aforementioned diphthong onset and offset positional properties, some age-dependent characteristics can be observed by comparing the plots of the two age groups. It is clear that the vowel space delineated by diphthong stripes is much smaller for the older age group. It is also observed that for /EY/, /AY/, and /OY/ the offsets end much closer to /IY/ or even exceed it for the younger age group than for the older age group. For /AY/, the onset starts very near /AH/, not /AA/, for the younger age group. The onset is much close to /AA/ for the older age group. Finally, note that the /OY/ onset position evolves from near /UH/ to a farther position. These observations suggest that diphthong onset or offset position are adjusted by speakers with respects to nearby monophthongs as speakers grow older.

2. Distance evolution between onset and offset of diphthongs and nearby monophthongs as a function of age

To further analyze the positional relationship between onset or offset of diphthongs and nearby monophthongs, we compute Euclidean distances in the F1–F2 space. Distances are first computed for each individual speaker and then averaged across subjects in each age group. In Fig. 8(a), we present distances averaged over diphthongs /EY/, /AY/, /AW/, /OW/ as a function of age group separately for diphthong onset and offset. Specifically, we show with the "○" marker distances between diphthong onsets and monophthongs averaged over the following pairs: (/EY/,/EH/), (/EY/,/IH/), (/AY/,/AA/), (/AY/,/AH/), (/AW/,/AA/), (/AW/,/AH/), and (/OW/,/AH/). For offsets ("×" marker), we averaged over the pairs: (/EY/,/IY/), (/AY/,/IY/), (/AW/,/UW/), (/AW/,/UH/), (/OW/,/UW/), and (/OW/,/UH/). The main finding is an inflection point around ages 14 or 15 years old for both the onset and offsets. That is, the distance decreases up to age 14 or 15 years old and then increases again reaching adult levels. The trend is consistently observed in all onset and offset diphthong-monophthong pairs averaged in Fig. 8(a). In general, onset positions of diphthongs exhibit clearer age-dependent trends than offsets in terms of distance to nearby monophthongs. Further, the match between diphthongs and related monophthongs is on average much better for diphthong onset than for offset points; Euclidean distances for offsets are 50%–70% higher for offsets than for onsets.

In Fig. 8(b), we show the distance between diphthong /OY/ and related monophthongs specifically for the onset distance of the pairs (/OY/,/UH/), (/OY/,/UW/) and the offset distance of the pair (/OY/,/IY/). The distance between /OY/ and nearby monophthongs stands out with respect to the rest of the diphthongs. Both the onset is progressively distanced from /UH/, /UW/ and the offset position from /IY/; there is no clear inflection point. These observations can be verified also by visual inspection of Fig. 7.

F. Behavior of the diphthong landmark

Plots of the normalized time point of the diphthong transitional mid-points across age and diphthong-type are shown in Fig. 9. The effects of both age and diphthong-type are significant ($p < 0.01$). It is observed that each diphthong has different normalized landmark position, which can be interpreted as the point in time where diphthongal spectral transition occurs are different for different diphthong. For instance, /EY/ shows earlier spectral transition mid-point

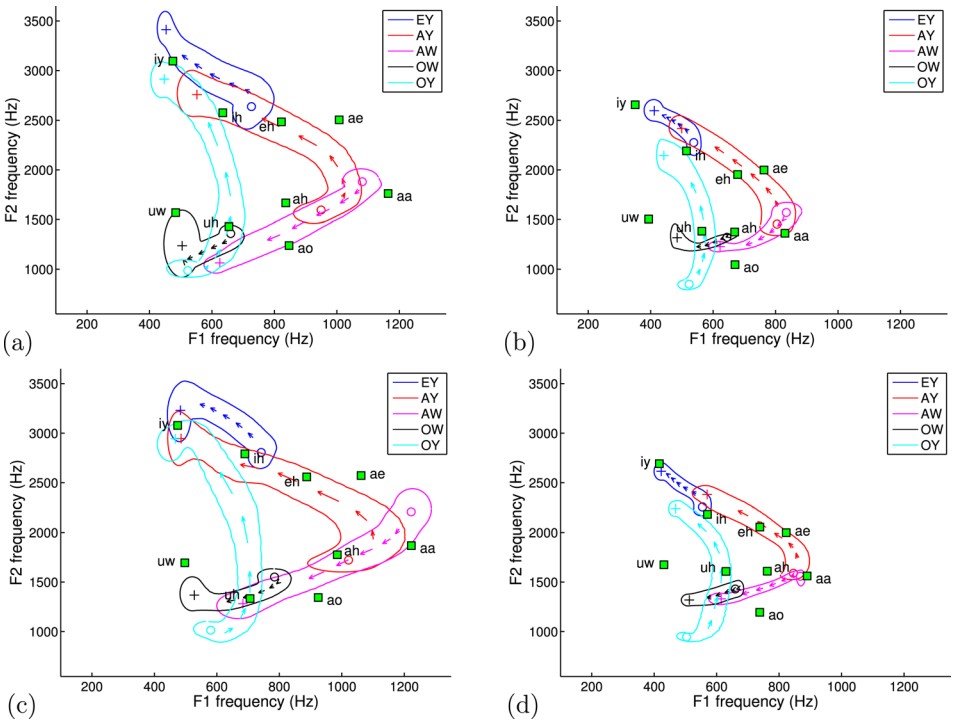

FIG. 7. (Color online) Five diphthong trajectories are represented by five colored "strips" in which arrows represent trajectory midline segments between onset ("○") and offset ("×") positions, and strip width corresponds to formant variability at selected locations along the midline. Plots are for (a) male age 5, (b) male age 14, (c) female age 5, and (d) female age 14. Formant positions of nine monophthongs (squares with two-character ARPABET vowel symbols) of each age group are also shown in background for comparison to onsets and offsets of diphthongs. It is clear that the positional relationship between diphthong onsets and offsets and nearby monophthongs are different between the two age groups.

when compared to the rest and /OW/ takes the longest to reach the transitional mid-point. That is, the time duration in which diphthongs remain at the first demisyllabic segment is shortest for /EY/ and longest for /OW/. /AY/, /AU/, and /OY/ show similar locations of the landmark position, in between /EY/ and /OW/. Regarding age-dependent behaviors, it is interesting to observe that in general younger age groups start diphthong transition relatively late when compared to the older age groups. That is, they spend more time in the first demisyllabic segment of diphthong. One exception seems the case of /OW/ where younger age groups spend less time compared to older children or adults.

G. Classification accuracies of diphthongs

Results of diphthong classification based on the Fisher's discriminant analysis are summarized in Table VI for each set of diphthong trajectory features examined. When considering features F1–F3 [onset] only, the accuracy is at 79.0%, while for [offset] only, it is at 67.9%. These results imply more noisy or diffused offset target positions in diphthong production than for onset. For [onset + rate], the accuracy is at 91.0% and for [offset + rate] 89.0%. These results demonstrate the effectiveness of formant onset values in diphthong classification. The best accuracy of 95.3% is obtained either for F1–F3 [onset + offset] or for the whole trajectory information, i.e., combination of [onset + offset + rate]. Overall, the classification results suggest that "onset" formant positions are a necessary feature in diphthong classification. It is interesting to note that "offset" formant positions are the next most relevant features when used in combination with onsets, while the F1–F3 transition rates are more effective than offsets as stand-alone features.

It may be noted that when discriminant analysis is applied to male and to female data separately, the classification accuracy is slightly higher, yielding an average classification accuracy of 95.8%. We also have applied discriminant analysis to the formant data set of ten monophthongs[1] of all age groups and obtained an accuracy of 76.8%. It is also worth noting that, for age group discrimination, the [F1–F3 onsets + offsets] feature set exhibit the best performance of 19.9%, which is significantly higher than the random performance of 6.7% (= 100/15 age-groups). Examination of the structured matrices (i.e., correlations

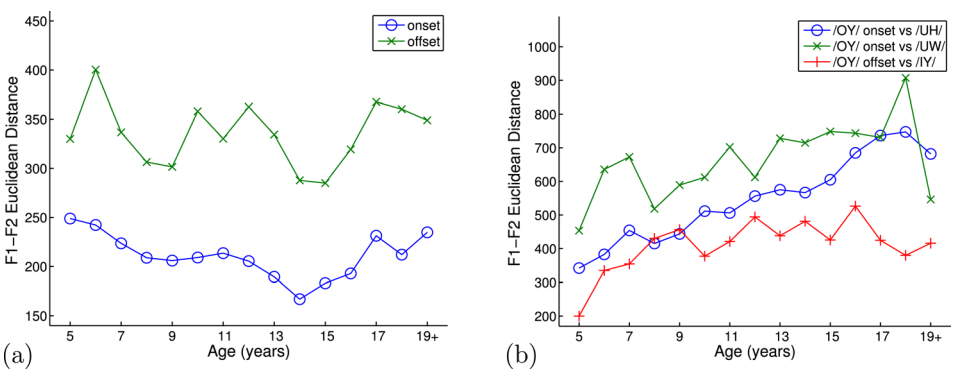

FIG. 8. (Color online) Average F1–F2 Euclidean distances between onsets and offsets of diphthongs and corresponding monophthongs as a function of age: (a) averages over diphthongs /EY/,/AY/, /AW/, /OW/ and (b) /OY/.

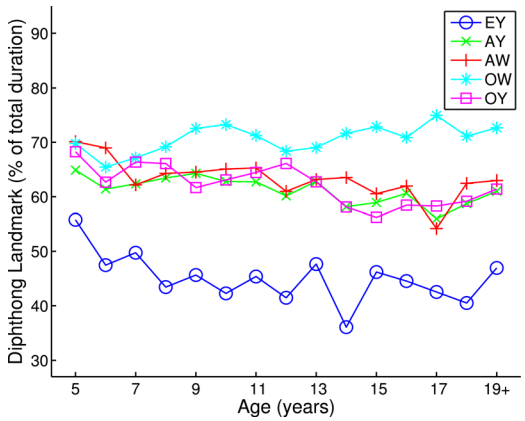

FIG. 9. (Color online) Diphthong landmark (transitional mid-points) for each age group and diphthong.

between predictor variables and standardized discriminant functions) of each parameter set reveals that the first discriminant function explains 96.2% of variance in the data and it is heavily correlated with F3 onset and offset (absolute correlation coefficients of $r = 0.78$ and 0.76, respectively). This implies that among the first three formant frequencies, F3 reflects best the age-dependent change of the vocal tract length (see also formant scaling factors as a function of age in Fig. 4).

V. DISCUSSION

Developmental trends in the magnitude and variability of temporal and spectral speech parameters in diphthong production are similar to those of monophthongs.[1] Duration, fundamental and formant frequencies, as well as associated variability (both for individuals and groups) typically decreases as a function of age. The main underlying developmental factors here are (1) the growth of the vocal tract size, (2) the growth and maturation of the vocal folds in terms of lengthening and thickness, and (3) the improvement in the speech motor control skill. For example, the main developmental factor that explains the F1–F3 age trend is the growth of the vocal tract, while for duration the age trend can be mainly attributed to improved motor skill. This age trend also holds for "derived" acoustic patterns of diphthongs. For instance, duration-normalized formant transition rate is smaller for older age groups. In a sense, older children realize a minimal perceptual contrast during diphthong production more effectively than younger children. The normalized landmark location that divides a given diphthong trajectory into two demisyllabic segments is also advanced as a function of age, implying that older age groups start diphthong transition earlier than younger age children do. That is, younger age children may emphasize the second demisyllabic portion of diphthong more (with the probable exception of /OW/).

TABLE VI. Diphthong classification accuracy for various feature sets.

Feature Set	Accuracy (%)
F1–F3 onsets only	79.0
F1–F3 offsets only	67.9
F1–F3 rates only	72.1
F1–F3 onsets and offsets	95.3
F1–F3 onsets and rates	91.0
F1–F3 offsets and rates	89.0
F1–F3 onsets, offsets, and rates	95.3

A notable observation in both this and the monophthong study[1] is that the speech parameters may vary in an nonlinear oscillatory fashion as a function of age. This oscillatory trend is noticeable both for the age-dependent change in diphthong duration (see Fig. 2) and also in F0 change (see Fig. 3), especially above age 12 years old. Group variability also shows a similar tendency. The age-dependent fluctuation (i.e., up and down movements) occur in a relatively short time period compared to the long-term averaged trends (i.e., spanning 5 to 18 years of age) and may be interpreted as the alternating undershoot and overshoot of neuromuscular adaptive behaviors to the changing vocal tract size. Socio-linguistic effects also can be a factor here. Further research using longitudinal data is required to verify and explain this trend.

As we have shown in Fig. 8, the onset and offset positions of diphthongs vary with respect to nearby monophthongs as a function of age. For all diphthongs, there is significant distance between onsets and nearby monophthongs, and significantly larger distance (almost double) between offsets and monophthongs. It is interesting however that the relative distances between diphthong onsets/offsets and associated monophthongs do not vary wildly with age, in fact diphthongs onset/offsets and monophthongs seem to co-develop and co-evolve with age. The notable exception here is the diphthong /OY/ whose onset position moves farther and below from /UH/ as a function of age, suggesting further back constriction. Similar observations can be made for /OY/ offset. This and other observations imply that the evolution of relative positions of diphthong onset or offset positions with respect to nearby monophthongs is only one facet of developmental aspects of diphthong production in the terms of vowel spacing in the formant plan (cf. Liljencrants and Lindblom[16]). Some observations relevant for diphthong production development are (1) the positions of diphthong onset and offset are not necessarily closely related to those of monophthongs, (2) the relative onset and offset positions of diphthongs with respect to the monophthongs vary as speakers improve their diphthongs production skill, and (3) monophthongs and diphthongs co-develop and co-evolve in the acoustic space but not always in a consistent manner (e.g., /OY/ onset and offset move away from nearby monophthongs). The aforementioned acoustic behavior of diphthong trajectories including onset and offset suggest that as a long-term developmental behavior, speakers adjust diphthong onset and offset positions with respect to nearby monophthongs in their vowel space (e.g., the formant space). The principle of maximal perceptual contrast[16] may have a role on the behavior of diphthongs and their co-evolution with monophthongs.

The demisyllabic division of diphthong segments shown in Fig. 9 reveals the following aspects of diphthong production: (1) each diphthong may have different time instances at

which it reaches approximately maximal or near maximal spectral transition (see also Fig. 6), (2) for each diphthong the mid-point landmark is age-dependent, and (3) younger children may spend more time at the initial demisyllabic portion of diphthong (with the possible exception of /OW/). It is thus possible that the relationship between time to reach spectral transition mid-point and spectral transition rate (e.g., F2 transition rate) is conditioned on the maturity of speech motor control skill. Specifically, older children can achieve a minimal but perceptually adequate spectral contrast between onset and offset in shorter time compared to younger age speakers. This ability may reach in its peak performance by age 14 or 15 years old.

Per the relative importance of onset, offset, and transitional segments for the perceptual definition of a diphthong, it is difficult to draw any general conclusions from this analysis. Clearly the onset of the diphthong is the most robust feature for speech classification experiments and is closer in the acoustic space to nearby monophthongs [see Fig. 8(a)]. Also when splitting diphthongs using the mid-point landmark, the first portion defined relative to the onset dominates in terms of duration for all diphthongs with the possible exception of /OY/ (see Fig. 9). Thus, there are strong indications that diphthong onsets are perceptually very important. There is not a clear winner, however, when comparing the relative perceptual importance of offset and transitional segments of diphthongs. In terms of classification results, offset features achieve better performance than transition rates when combined with onset features. Also diphthong offsets show clear age-dependent trends and seem to developmentally co-evolve with onsets (see Fig. 5), while the developmental trends for the average transition rates (see Fig. 6) are less consistent especially for F1 and F2. Also the role of the transitional segment on diphthong perception may be more important than our classification experiments indicate. Based on the evidence in this study, it is safe to say that the [onset + offset] definition of diphthongs has phonetic and potentially also phonological relevance.

Note that the development behavior of /OY/ seems different along various acoustic correlates. First the duration of /OY/ is significantly higher (almost double) than that of other diphthongs (see Fig. 2). Second, /OY/ onset and offset move far away from nearby monophthongs with age [see Fig. 8(b)]. These developmental differences can be also observed also in Fig. 7 comparing ages 5 and 14 years old and the position of the /OY/ solenoid relative to the rest of the diphthongs. It is unclear if these differences imply also a different perceptual definition of /OY/. Based on the evidence on the co-evolution of onset-offset away from monophthong targets, it seems that the [onset + offset] definition is the most fitting for /OY/.

A final remark has to do with the F0 contrast in fundamental frequency between diphthong onset and offset for ages 5 and 6 years old, as shown in Fig. 3. The contrast decreases with age and quickly disappears for older children. F0 onset-offset contrast could be a correlate in the acquisition of diphthongs that eventually (as the diphthongs mature developmentally) looses its relevance and disappears. Another possible explanation is the elicitation method for (some of) the 5 and 6 year olds (isolated words vs carrier sentence). Specifically, softer voice (i.e., relatively lower volume velocity) at the end portion of diphthong production, augmented with longer duration, could be a possible explanation for the contrast, as the children imitated carefully produced diphthong sounds by a female speech pathologist. Note that differences in duration for ages 5 and 6 years old could be also partially attributed to the different elicitation method.

VI. CONCLUDING REMARKS

In this study, we analyzed durational and spectral patterns of five diphthongs in American English as a function of age and gender. Together with our previous study on monophthongs,[1] our results are consistent with the literature[17–20] and confirm the reduction of magnitude and group variability of surface speech acoustic parameters as a function of age. As part of this study we also provide a set of basic acoustic parameters that have been measured for subjects of age 5 through 18 years old, as well as adult speakers. This is a valuable resource, however, some inherent limitations in the current (and previous vowel) study should be underscored: (1) the data are cross-sectional (rather than longitudinal) and, therefore, the results should be interpreted as group behaviors as a function of age, (2) the contexts in which vowels are embedded are somewhat limited, and (3) despite our best efforts to remove outliers, it is possible that there exist some erroneous data points due to measurement difficulties (e.g., formant estimation). Overall, the data set and analysis method adequately represents averaged behaviors of vowel acoustic parameters as a function of age and gender. Future studies should collect and analyze longitudinal data with rich vowel contexts to overcome the limitations of the current study in order to investigate further the developmental aspects of speech production mechanism of vowel sounds.

Results of the positional comparison of diphthong onset and offset positions against monophthongs suggest that diphthongs develop their own onset and offset positions as speakers grow older that are often farther in the acoustic space compared to nearby monophthongs. However, diphthongs and monophthongs do tend to co-evolve and co-develop in the acoustic space, with the exception of /OY/ where the distance between onset/offset and nearby monophthongs increases significantly with age. In general, /OY/ stands out also in respect to its longer duration and in having a midpoint landmark significantly earlier than other diphthongs. These observations may imply that initially younger age speakers use monophthongs to anchor or determine diphthong onset and offset positions, but as speakers grow older, they may develop diphthong-specific onset and offset positions in the acoustic vowel space. In summary, based on those observations, as well as the results of diphthong classification experiments the [onset + offset] definition of diphthongs appears to have phonetic as well as phonological relevance.

The reduction of magnitude and variability of surface speech acoustic parameters as a function of age is attributed

to two main factors: Vocal tract growth and speech motor control skill maturation. However, age is just a crude correlate of the physical and physiological development factors associated with speech production. In order to infer the true relationships among them using acoustic measures, it is desirable to examine surface speech acoustics with explicit articulatory measurements (i.e., measurements on both the vocal tract anatomy and the vocal tract shaping associated with speech sound), not just with the age variable alone. With recent advances in non-invasive technologies for collecting articulatory speech production data, such as real-time magnetic resonance imaging[21] (MRI) and electromagnetic articulography[22] (EMA), it is now possible to gather sagittal and three-dimensional articulatory data from a large number of speakers with a wide age range of lower age limit down to 4 or 5 years old. Future developmental studies, especially longitudinal ones, that utilize such vocal tract data acquisition technologies should contribute into gaining detailed insight on the developmental aspects of human speech production mechanisms. Such studies will also provide useful information for advancing speech-science driven speech technologies such as articulatory speech synthesis and speech recognition.

ACKNOWLEDGMENTS

The Children's speech database analyzed in the study has been made available by the courtesy of the Central Institute for the Deaf at St. Louis, Missouri and Southwestern Bell Technology Resources, Inc. (now a subsidiary of AT&T). This work was partially supported by the National Science Foundation and the National Institutes of Health. The duration, fundamental frequency and formant statistics shown in Tables II–V may be downloaded from http://sail.usc.edu/children-speech/.

[1]S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468 (1999).
[2]A. Holbrook and G. Fairbanks, "Diphthong formants and their movements," J. Speech Hear. Res. **5**, 38–58 (1962).
[3]I. Lehiste and G. E. Peterson, "Transition, glides, and diphthongs," J. Acoust. Soc. Am. **33**, 268–277 (1961).
[4]T. Gay, "Effects of speaking rate on diphthong formant movements," J. Acoust. Soc. Am. **44**, 1570–1573 (1968).
[5]M. Gottfried, J. D. Miller, and D. J. Meyer, "Three approaches to the classification of American English diphthongs," J. Phonetics **21**, 205–229 (1993).
[6]T. Gay, "A perceptual study of American English diphthongs," Language Speech **13**, 65–88 (1970).
[7]F. Sánchez-Miret, "Some reflections on the notion of diphthong," Pap. Stud. Contrastive Linguistics **34**, 27–51 (1998).
[8]G. Hare, "Development at 2 years," in *Phonological Development in Children: 18–72 Months*, edited by J. V. Irwin and S. P. Wong (Southern Illinois University Press, Carbondale, IL, 1983), pp. 55–88.
[9]E. M. Prather, D. L. Hedrick, and C. A. Kern, "Articulation developments in children in ages two to four," Natl. Student Speech Language Hear. Assoc. J. **18**, 96–102 (1991).
[10]L. Paschall, "Development at 18 months," in *Phonological Development in Children: 18–72 Months*, edited by J. V. Irwin and S. P. Wong (Southern Illinois University Press, Carbondale, IL, 1983), pp. 27–54.
[11]R. D. Kent and H. K. Vorperian, "Anatomic development of the craniofacial-oral-laryngeal systems: A review," J. Med. Speech-Language Pathol. **3**, 145–190 (1995).
[12]H. K. Vorperian, S. Wang, M. K. Chung, E. M. Schimek, R. B. Durtschi, R. D. Kent, A. J. Ziegert, and L. R. Gentry, "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**, 1666–1678 (2009).
[13]J. D. Miller, S. Lee, R. M. Uchanski, A. F. Heidbreder, B. B. Richman, and J. Tadlock, "Creation of two children's speech databases," in *Proceedings of ICASSP* (Atlanta, GA), pp. 849–852 (1996).
[14]P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.1.1) [computer program]," available at htpp://www.praat.org (Last viewed April 7, 2014).
[15]Y. Wada and M. Kawato, "A via-point time optimization algorithm for complex sequential trajectory formation," Neural Networks **17**, 353–364 (2004).
[16]J. Liljencrants and B. Lindblom, "Numerical simulations of vowel quality systems: The role of perceptual contrast," Language **48**, 839–862 (1972).
[17]S. Eguchi and I. J. Hirsh, "Development of speech sounds in children," Acta. Otolaryng. Suppl. **257**, 1–51 (1969).
[18]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).
[19]R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Tutorial," J. Speech Hear. Res. **19**, 421–447 (1976).
[20]R. E. Turner, T. C. Walters, J. J. Monaghan, and R. D. Patterson, "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," J. Acoust. Soc. Am. **125**, 2374–2386 (2009).
[21]S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776 (2004).
[22]S. Lee, J. Kim, and S. S. Narayanan, "On the interactions among speech parameters across emotions and speakers in emotional speech production," in *Proceedings of the International Seminar on Speech Production (ISSP)* (Cologne, Germany, 2014).