In most studies, the monophthong vs. diphthong status of a given vowel is simply assumed a priori to hold across the entire sample of speakers in one’s study (Clopper, Pisoni, & de Jong, 2005). However, the conventional assumptions can occasionally be wrong, e.g., monophthongized /o/ in certain Finnish-heritage communities (Rankinen & Albin, 2015). Moreover, blindly assuming one or the other option is uninformative in contexts where the monophthong vs. diphthong distinction is the very phenomenon of interest, e.g., for studies of the diphthongization of [æɪ] in Southern U.S. English (Labov & Ash, 2014, 513-514). The present study’s research question is: What cutoffs for what acoustic measurements lead to the highest accuracy in splitting the vowels conventionally assumed to be monophthongs and diphthongs?

To explore this question, analyses were conducted on a corpus of 14,543 stressed vowels extracted from a reading passage task administered to 133 speakers of American English in Michigan’s Upper Peninsula (Rankinen, 2018). A total of 13 vowels were represented: [æ e ə o u] and [i ɪ e ə o ɔ ʌ ø ø u]; refer to Figure 1 for precise vowel qualities. From each vowel token (regardless of vowel category), F1 and F2 tracks were extracted at 1 ms intervals, and various statistical models were fit to the raw (non-normalized) formant data. From these fitted models, a total of 31 potentially relevant measurements (including Euclidean distances) were extracted for every vowel. To factor out the imbalance in the number of tokens, each measurement was averaged for each combination of speaker and vowel. Thus, for the final analysis, there were 1,728 data points for every measurement: (133 speakers x 13 vowels) minus 1 cell with unusable data. This dataset was analyzed by exhaustively exploring the space of possible binary decision rules, e.g., If duration ≥ 100ms, predicted to be a diphthong; otherwise, monophthong.

Many of the best-performing measurements came from a simple linear regression model, fit separately to the F1 and F2 tracks of each token. Among the most strongly predictive was the absolute value of the difference (in Hertz) between the beginning and end of the line fit to the F2 track, referred to here as $F_2$ Deflection; refer to Figure 3 for an illustration of this measurement. In particular, the decision rule $F_2$ Deflection > 537.9 Hz picks out [eɪ] and [æ] as diphthongs with an error rate of just 3.0%, i.e., 52 misclassifications out of 1,728 speaker+vowel combinations; refer to Figure 2 for a visualization of these results. This F2-based decision rule, however, misses [ɔʊ], whose formant movement is primarily in F1. The parallel decision rule for F1 is $F_1$ Deflection > 162.4 Hz, which picks out [æ] and [æ] as diphthongs. The error rate for this decision rule is higher, however, at 11.3% (195/1,728), due to partial monophthongization of [æ] in this community. Overall, F1 and F2 deflection show the most promise in empirically determining the monophthong vs. diphthong distinction. Future studies are needed to determine generalizability.
Figure 1: Overall vowel space for community (in normalized formant values)

Figure 2: Results of applying the decision rule $F2 \text{ Deflection} > 538 \text{ Hz}$

(a) Diphthong [ui]  
(b) Monophthong [a]

Figure 3: Illustration of how deflections are calculated. Green dashes represent regression lines. For (a), the absolute value of the difference between the F2 line’s beginning and end is 1899 Hz. For (b), it is only 0.15 Hz.

References


