# Perception of Diphthongized Vowels in Rhode Island English

**Harriet S. Magen**
Haskins Laboratories, New Haven, CT, USA
Rhode Island College, Providence, RI , USA
E-mail: hmagen@ric.edu

## ABSTRACT

Dialects of American English vary in extent of diphthongization. The Rhode Island dialect includes the usual mid-height /eɪ/ and /oʊ/ and also diphthongizes /ɪ/ and /ɔ/. To examine whether one portion of the diphthong is perceptually dominant, we compared production and perception of 10 RI speakers on vowel pairs (/ɪ/ /eɪ/) and (/oʊ/ /ɔ/). Vowel formants were measured at three points and interpolated, yielding two pairs of trajectories roughly parallel in F1. In a perceptual study, listeners identified the best exemplar from a range of steady-state synthetic vowels. Perceptions varied: closer to the onset for /eɪ/, to the midpoint for /oʊ/ and /ɪ/; and to the glide for /ɔ/. Results indicate that for diphthongized vowels the perceptually dominant portion of the diphthong is variable. The heavier weighting of the offglide for /ɔ/ places it in a part of the vowel space more typical of other dialects.

## 1. INTRODUCTION

Vowels can be fairly well perceived from two-formant, steady-state synthetic versions [1]. Vowel formants are often compared only for single measurement points in time. Peterson and Barney [2] and Hillenbrand et al [3] did not measure /oʊ/ and /eɪ/ for just this reason. Especially for diphthongs, though, single measurement points can obscure the way vowels are actually distinguished. While listeners have been found to rely more heavily on a specific portion of the diphthong [4,5], listeners have also been found to rely on information on the rate of spectral change [6,7,8]

Dialects of American English vary with respect to extent of diphthongization. The Rhode Island dialect includes the usual mid-height /eɪ/ and /oʊ/ but also diphthongizes /ɪ/ and /ɔ/. The existence of phonemic /ɔ/ and its contrast to /ɑ/ has been noted as one of the factors differentiating the Providence dialect from the Boston dialects [9,10] as well as establishing it as one of the more conservative American English dialects [11]. Measurements at more than one point may be revealing in capturing the vowel space for a given dialect.

The aim of the present study is to examine the dynamic separation of some pairs of Rhode Island dialect vowels that are very close acoustically at a given point in production. A second aim is to examine perception and production results for intrinsically dynamic vowels to determine whether, in matching steady state synthetic vowel productions to their own diphthongal productions, listeners predictably use nucleus, glide, or some average of the trajectory.

## 2. EXPERIMENT

### 2.1 Participants

Participants were five female and five male students at Rhode Island College, ages 18-23, paid for their participation. Four of the five females had completed a semester course in phonetics while the fifth female and one male had had some practice in phonetic transcription. The remaining four males had no knowledge of phonetics. Participants were linguistically homogeneous; all ten had lived only in the Providence, Rhode Island area.

### 2.2 Speech Materials (production portion)

The speech materials for the production portion of the experiment were keywords containing 11 English vowels, “heed, hid, aid, head, had, odd, awed, hud, owed, hood, who’d” in the carrier phrase, “say ____ again.”

### 2.3 Stimuli (perception portion)

Synthetic vowel stimuli were generated by a software synthesizer, KLSYN88 [12] as implemented in the Sensimetrics program SenSyn. There were 298 steady-state F1/F2 combinations, with fifteen values of F1 ranging from 250 to 1000 Hz and twenty-two values of F2 ranging from 800 to 2900 Hz. The step size for F1 and F2 for all stimuli was about 4/10 of an auditory critical band. F3 was generated by separate regression formulas for front and back vowels, and F4 was 3500 Hz or 300 Hz more than F3, whichever was greater. Both F0 and duration were generated by formula [13]. Bandwidths were held constant: 75 Hz for F1, 100 Hz for F2, 150 Hz for F3, and 200 Hz for F4. Judgments were made using the same list of keywords containing the vowels of English that was used in the production portion.

### 2.4 Procedure

All participants participated in both the production and perception portions of this experiment. In the production


ISBN 1-876346-48-5 

portion, participants were asked to read five repetitions of the list of English words in the carrier phrase "Say _____ again", presented in five different randomized orders. They were given a chance to practice. Recordings were made using a Nakamichi microphone and Sony DAT recorder.

The perception portion of the experiment was run on-line with a Dell lap-top. Participants were seated at the CRT screen and provided with headphones. They saw a word on the screen, from among those they had just recorded, and were presented with a 15 x 22 grid, the squares of which corresponded to the F1/F2 combinations described above. Participants were told that their task was to select the best example of the vowel contained in the word that they would see presented visually on the screen, and to make as many attempts as necessary. They were instructed to use a mouse to select a particular square and to click on it in order to hear the vowel associated with it. To prevent participants from learning the locations of the stimuli, there were four orientations of the acoustic vowel space and the orientation was changed from trial to trial. Stimuli were presented binaurally over Sennheiser 430 headphones. Each keyword was judged 10 times, resulting in a total of 110 judgments. Participants were presented with a status bar so that they could monitor their progress through the experiment [14].

## 3. RESULTS

### 3.1 Production results

The vowel productions were measured at a single point from a narrow-band spectrum, the centroid of three harmonics, one-third of the way through the vocalic segment. Figure 1a shows the averaged raw formant values of this measurement point for the male participants for a subset of the eleven vowels, those produced in the front of the mouth. Figure 1b shows the averaged raw formant values of this measurement point for these same speakers for an additional subset of the vowels, those produced in the back of the mouth. Plots of these formant measurements showed that /ɔ/ overlaps with /oʊ/, although /oʊ/ is generally considered higher than and distinct from /ɔ/. /eɪ/ is higher than /ɪ/, although most plots of phonological vowel spaces across dialects show /ɪ/ to be the higher vowel [2].

To examine the effects of diphthongization, we made measurements at the 1/12 and 2/3 points for the two pairs of vowels. Figure 2a shows the additional measurement points for the front vowel pair /ɪ/ and /eɪ/ of a representative male speaker (with reference vowels /i/ and /ɛ/), and Figure 2b shows the additional measurement points for the vowels /oʊ/ and /ɔ/ from the same speaker

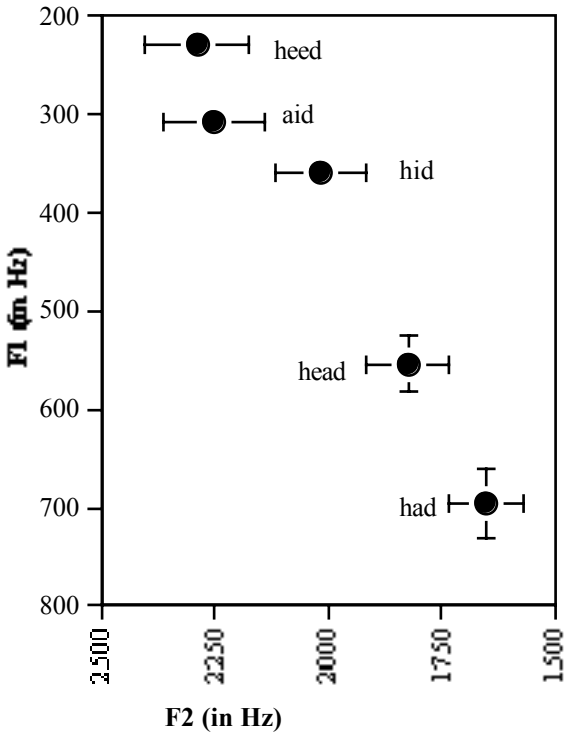


Figure 1a: Formant values for male speakers, front vowels.

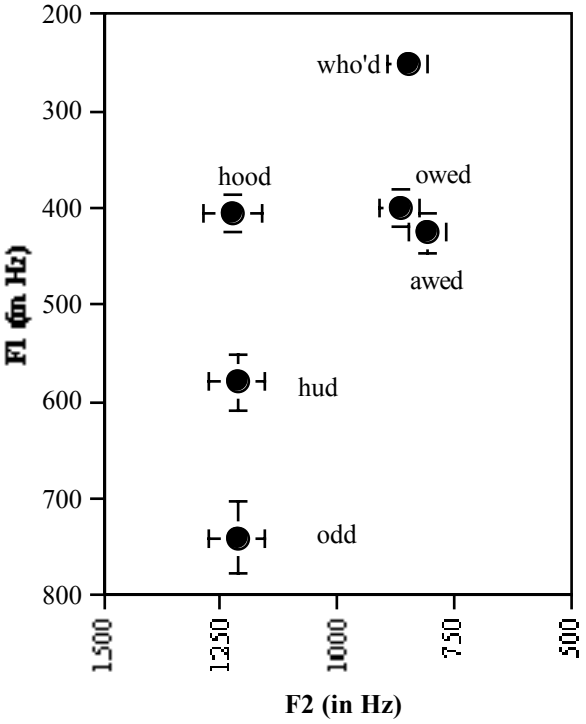


Figure 1b: Formant values for male speakers, back vowels.

(as well as reference vowels /u/ and /ɑ/). These values are shown in normalized space (see below). All vowels show movement. The expected nucleus followed by an upglide was found for /eɪ/ and /oʊ/. For /ɪ/ and /ɔ/, the nucleus is followed by an inglide. The two sets of trajectories appear

to run roughly parallel to one another, covering more or less the same area in the F1 dimension. The F2 of the member of each pair usually described as a diphthong is higher.

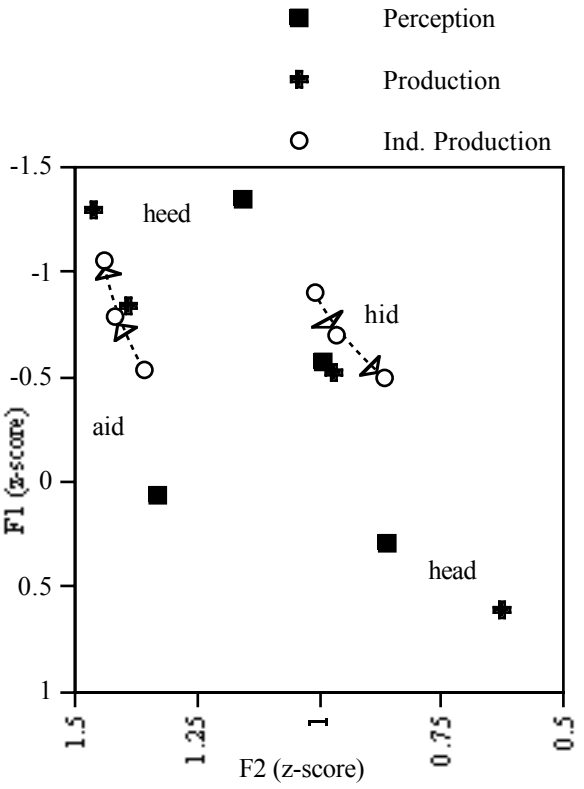


Figure 2a: Perception and production, front vowels. "Ind. production" shows values for a representative speaker at 1/12, 1/3 and 2/3 of total duration (arrows show direction).

**3.2 Perception results**

In our study, we compared the perception of a synthetic male voice to productions from several speakers, thereby introducing an often discussed issue in studies of vowel perception, that is, that listeners are able to perceive a vowel produced by different speakers as the same vowel, despite the differences among the speakers' vocal tracts and resulting formant patterns. To address this issue and to duplicate the removal of speaker variation that is automatically performed by the human perceptual system, various normalization procedures have been proposed. (For discussion, see e.g. [15-17].) One method shown to be successful in normalizing acoustic data is the z-score transformation [18] and we have chosen to use that method on our data.

Even though female and male perception results showed no significant differences [14], in considering the perception results, we concentrated on the male participants, since the synthetic voice was male. For the vowels [oʊ] and [ɪ], perception and production results were congruent, lining up with the midpoint of the vowel. For the vowel [eɪ] perception seems to coincide with the nucleus, whereas for [ɔ] it corresponds to the glide portion of the diphthong. In both of the latter cases, listeners report perceiving a vowel that is lower than the one they produce.

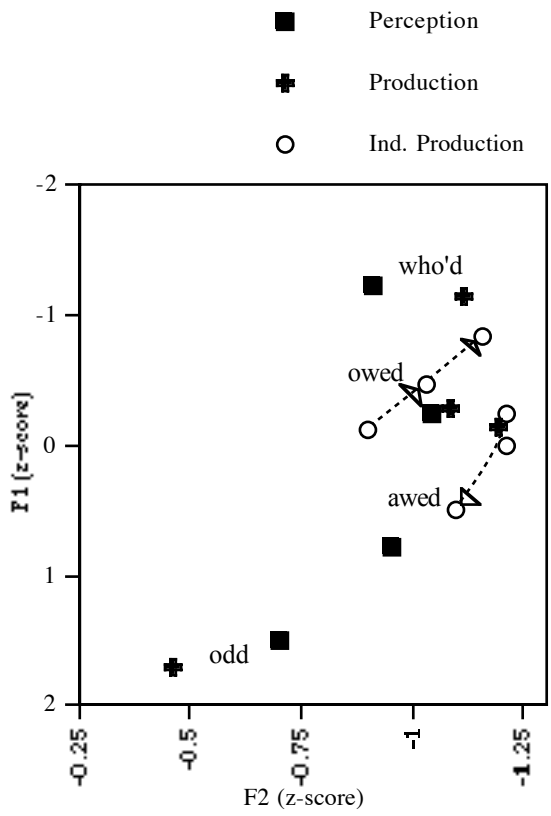


Figure 2b: Perception and production, selected back vowels.

## 4. CONCLUSIONS

Dynamic specification is needed for two pairs of Rhode Island dialect vowels whose trajectories show a substantial amount of overlap. While one member of each pair is typically considered a diphthong, one member is not. This dialect is considered conservative in that its speakers maintain a very clear distinction between /ɔ/ and /ɑ/; it may be that the maintenance of this distinction results in a production of /ɔ/ that moves into the /oʊ/ space. It is interesting to note also that the higher than normal productions of both /ɔ/ and /eɪ/ create a symmetrical gap in the usage of F1 for vowels.

From the point of view of perception of diphthongs, it is noteworthy that perception does not coincide with any measured time point for the vowels [eɪ] and [ɔ] but does coincide with the midpoint for the vowels [ɪ] and [ɔ]. For the vowels [eɪ] and [ɔ] it appears that if perception is showing some congruence with production, it corresponds


ISBN 1-876346-48-5 © 2003 UAB

to the nucleus for [eɪ] but to the glide for [ɔ]. Interestingly, speakers of this dialect report perceptions that are closer to the productions of a more standard dialect than are their own productions. Overall, since listeners appear not to be basing their perceptions consistently on the nucleus, the glide or some averaged location, the results point to a lack of predictability, arguing for dynamic specification.

## ACKNOWLEDGEMENTS

This work was supported by NIH grant DC-02717 to Haskins Laboratories and a Rhode Island College Faculty Research grant. Thanks go to Matthew Richardson and D. H. Whalen for help with this paper.

## REFERENCES

[1] P. C. Delattre, A. M. Liberman and F. S. Cooper. "Voyelles synthétiques à deux formants et voyelles cardinales," *Le Maître Phonetique,* vol. 96, pp. 30-36, 1951.

[2] G.E. Peterson and H.L. Barney, "Control methods used a study of the vowels," *Journal of the Acoustical Society of America,* vol. 24, 175-184, 1952.

[3] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America,* vol. 97, 3099-3111, 1995.

[4] C.I. Watson and J. Harrington, "Acoustic evidence for dynamic formant trajectories in Australian English vowels," *Journal of the Acoustical Society of America*, vol. 106, 458-468, 1999.

[5] A. Bladon, "Diphthongs: A case study of dynamic auditory processing," *Speech Communication,* vol. 4, 145-154, 1985

[6] T. Gay, "A perceptual study of American English diphthongs," *Language and Speech*, vol. 13, 65-88, 1970.

[7] Z. Bond, "Experiments with synthetic diphthongs," *Journal of Phonetics*, vol. 10, 259-264, 1982.

[8] C.B. Huang, "Modelling human vowel identification using aspects of formant trajectory and context," Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds., *Speech Perception, Production and Linguistic Structure*, pp. 43-61. Amsterdam: IOS Press, 1992.

[9] W.G. Moulton, "Structural Dialectology," *Language,* vol. 44, 451-466, 1968.

[10] R.I. McDavid, "Low-back vowels in Providence: A note on structural dialectology," *Journal of English Linguistics*, vol. 15, 21-29, 1981.

[11] W. Labov, *Atlas of North American English*, http://www.ling.upenn.edu/phonoatlas/.2003.

[12] D. H. Klatt. "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America,* vol. 67, pp. 971-995, 1980.

[13] K. Johnson, E. Flemming, and R. Wright, "The hyperspace effect: Phonetic targets are hyperarticulated," *Language*, vol. 69, 505-528, 1993.

[14] D. H. Whalen and H. S. Magen, "The 'hyperspace' effect in vowel perception," *Journal of the Acoustical Society of America,* vol. 109, p. 2291, 2001.

[15] D. Hindle, "Approaches to vowel normalization in the study of natural speech," in: *Linguistic Variation: Models and Methods*, D. Sankoff, Ed., pp. 161-172. New York: Academic Press, 1978.

[16] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *Journal of the Acoustical Society of America*, vol. 85, 2088-2113, 1989.

[17] J. Hillenbrand and R.T. Gayvert, "Identification of steady-state vowels synthesized from the Peterson and Barney measurements," *Journal of the Acoustical Society of America,* vol. 94, 668-674, 1993.

[18] P. Adank, R. van Hout, R. Smits, "A comparison between human vowel normalization strategies and acoustic vowel transformation techniques," *Proceedings Eurospeech 2001*, pp. 481-484, 2001.