

A goodness optimization method for investigating phonetic categorization.

Paul Iverson and Bronwen G. Evans

Department of Phonetics and Linguistics, University College London, London, U.K.

E-mail: paul@phon.ucl.ac.uk, bron@phon.ucl.ac.uk

ABSTRACT

Existing methods for examining phonetic categorization (i.e., identification or goodness judgments) typically require listeners to give responses on every member of a stimulus set. This article describes a new method that is more efficient for higher-dimensional stimulus sets (i.e., more phonetic detail) in which the number of perceptually distinct stimuli is too large to be played within an experimental session. The method uses a traditional goodness-rating task (i.e., subjects hear individual synthesized stimuli and give goodness ratings on a continuous scale). However, the stimulus selection is determined by an iterative computer algorithm that is designed to find the minimum value of a function within a multidimensional variable space. The method has been applied to map best exemplars of English vowels in a 5-dimensional space including formant movement and duration; best exemplars can be found within the set of 100,700 possible vowels after playing subjects only 35 trials per vowel category.

1. INTRODUCTION

The recognition of monophthongal English vowels is influenced by relatively fine-grained phonetic variation, related to formant movement and duration, in addition to steady-state F1 and F2 target frequencies; talkers increase formant movement and duration differences when speaking clearly [1], vowels are more accurately recognized when they include formant movement and duration differences [2, 3], and vowels can be recognized even when the steady-state portions have been removed [4]. This emphasis on fine-grained phonetic details parallels work on episodic memory and talker differences, which suggests that such details are an important contributor to speech understanding, rather than a nuisance that must be normalized or removed [5].

However, it is difficult to include such fine-grained variation when mapping vowel spaces within perceptual experiments. Previous experiments have generally restricted vowels to a 2-dimensional F1/F2 space. With such low-dimensional spaces, it is possible to play subjects a set of stimuli that span the space in order to map identification boundaries [6], and to give subjects a grid of stimuli from which to choose best exemplars [7]. Increasing phonetic detail increases the number of

stimulus dimensions (i.e., adding formant movement and duration), such that the number of stimuli required to adequately span this space becomes very large. With such high-dimensional stimulus sets it is no longer feasible to play the full set of stimuli to subjects.

This paper describes a method that allows for best exemplar locations to be found within high-dimensional stimulus sets. The method involves playing subjects individual stimuli, having them rate the category goodness of each stimulus, and using a computational procedure based on standard function minimization algorithms [8] to iteratively find the stimuli within the set that are best exemplars of different vowel categories. To demonstrate this technique, this procedure was applied to a 5-dimensional stimulus space of hVd vowels embedded in English sentences, with the dimensions being F1 onset, F1 offset, F2 onset, F2 offset, and duration. Although there are thousands of distinguishable stimuli that can be synthesized within the constraints of this 5-dimensional space, the minimization algorithm is able to find reliable estimates of best exemplar locations after only 35 trials for each vowel category.

2. STIMULI

The stimuli consisted of hVd syllables embedded in recordings of the carrier sentence *Say " " again*. The carrier sentence was produced in a Standard Southern British English accent by a male speaker. Initial and final words, plus the burst of the /d/, were edited from the natural recording, and the hVd syllables were synthesized to match the vocal timbre, pitch, and higher formant frequencies of the talker. Each syllable had static formant frequencies during the /h/ that matched the onset of the vowel, the F1 and F2 formant frequencies could change linearly from the onset to the offset of the vowel, and F1 fell at the end of the vowel to simulate the /d/ closure. The durations of /h/ and the /d/ closure were fixed, and the duration of the vowel was allowed to vary from 148 to 403 ms. F1 frequency was restricted so that it had a lower limit of 5 ERB and an upper limit of 15 ERB. F2 frequency was restricted so that it had a lower limit of 10 ERB, was always at least 1 ERB higher than F1, and had an upper limit defined by the equation $F2 = 26 - 5 * F1 / 10$. The stimuli were synthesized in advance with a 1-ERB spacing of the vowel space, and with 7 levels of log-spaced duration values, for a total of 100,700 individual stimuli.

2. PROCEDURE

On each trial, subjects heard one sentence and rated on a continuous scale whether the hVd was close to being a good exemplar of a word that was displayed on the computer screen. They gave ratings for 13 words, *heed* (/i/), *hid* (/i/), *hayed* (/eɪ/), *head* (/ɛ/), *had* (/a/), *hard* (/ɑ/), *hod* (/ɒ/), *hoard*, (/ɔ/), *hud* (/ʌ/), *heard* (/ɜ/), *hoed* (/əʊ/), *hood* (/u/), and *who'd* (/u/). The goodness optimization procedure involved searching along individual vectors through the stimulus space (i.e., 1-dimensional straight-line paths), and finding the best exemplar on each vector; there were a total of 7 search vectors and 5 trials per vector for each vowel.

For Vector 1, goodness was optimized along a straight-line path that passed through two points: (1) the F1 and F2 formant frequencies of the natural productions of the target word, and (2) a neutral stimulus in the middle of the vowel space (F1 = 500 Hz and F2 = 1500 Hz, at both the onset and offset); duration was not varied along Vector 1. The endpoints of the vector were constrained by the boundaries of the vowel space. For example, the first search vector for *heed* crossed diagonally across the vowel space, starting from the high-front boundary of the space (i.e., low F1 and high F2, near /i/), passing through the middle of the space, and ending at the low-back boundary of the space (i.e., high F1 and low F2, near /ɑ/).

On the first two trials for the search vector, subjects heard the most extreme stimuli that it was possible to synthesize along the vector (e.g., in the case of *heed*, they heard extreme high-front and low-back vowels, with the order of these two trials randomized). The selection of stimuli on the remaining trials was based on the subjects' judgments, using formulas that were designed to find stimuli along the path that would be perceived as better exemplars. On the 3rd trial, subjects heard a stimulus that was selected by a weighted average of the first two stimuli, according to the equation

$$c = a * \frac{f(b)}{f(a) + f(b)} + b * \frac{f(a)}{f(a) + f(b)}, \quad (2)$$

where a and b are the positions on the search path for the first two trials, $f(a)$ and $f(b)$ are the goodness ratings for the stimuli on those trials (the goodness responses of close to far away were scaled from 0 to 1), and c is the new path position selected for the 3rd trial. On the 4th and 5th trials, the stimuli were selected by finding the minimum of a parabola that was defined by the equation

$$\min = \frac{b \square 0.5 * \{ [b \square a]^2 * [f(b) \square f(c)] \square [b \square c]^2 * [f(b) \square f(a)] \}}{[b \square a] * [f(b) \square f(c)] \square [b \square c] * [f(b) \square f(a)]}, \quad (3)$$

where b was the path position of the best stimulus found thus far; a and c were most recently tested positions on either side of b ; and $f(a)$, $f(b)$, and $f(c)$ were the goodness ratings of those stimuli. At the completion of the 5th trial, subjects were allowed to repeat the search if it had

produced a poor exemplar. If the best exemplar was correct, the parameters of the best stimulus found thus far were passed onto the next stage of the search algorithm.

The same 5-trial search algorithm was used for the other search vectors. Vector 2 varied duration, keeping formant frequencies fixed. Vector 3 varied the onset F1 and F2 formant frequencies (i.e., duration and offset formant frequencies were fixed) along the same basic path as the first vector (i.e., through a straight-line path including a neutral vowel and the onset formant frequencies of the natural production). Vector 4 was orthogonal to Vector 3 in the F1/F2 onset space. Vectors 5 and 6 were analogous to Vectors 3 and 4, except that the offset F1 and F2 frequencies were varied. Vector 7 varied all dimensions, passing through the best values found thus far on all dimensions and the neutral vowel.

The search paths were chosen so that Vector 1 would likely get close to a best exemplar quickly, Vector 2 would adjust duration, Vectors 3-6 would adjust the onset and offset formant frequencies, and Vector 7 would fine-tune the selection by allowing subjects to make their best exemplar more or less extreme in the vowel space.

2. RESULTS AND DISCUSSION

Best Exemplar Locations

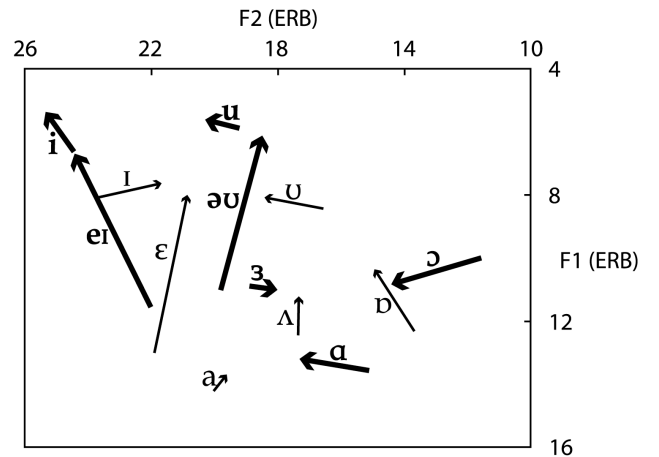


Figure 1: Average best exemplar locations for 9 speakers of Standard Southern British English. Lines go from the formant frequencies at the onset to the formant frequencies at the offset. Vowels with long durations (>250 ms) are in bold with thick lines; vowels with short durations (<250 ms) are displayed with thin lines.

Results were collected from 9 speakers of Standard Southern British English (Figure 1). The best exemplars of different vowels varied both in the magnitude and direction of their formant movement. Vowels such as /a/, /ʌ/, /ɜ/, and /u/ had no statistically significant formant movement. The other monophthongal vowels were preferred to have a median of 2 ERBs of formant movement; /ɛ/ had as much formant movement as the diphthongs /eɪ/ and /əʊ/, despite the fact that it had a

shorter duration. Although there was a general tendency for all formant movement to lower F1 and raise F2 (i.e., move toward the /i/ corner of the vowel space), there were significant differences in direction; vowels differed in whether their formant movement predominantly changed F1 or F2, and /ɪ/ moved toward a centralized position. The duration of the preferred vowels also varied significantly; listeners preferred markedly shorter durations for /ɪ/, /ɛ/, /a/, /ɒ/, /ʌ/, and /ʊ/; than for /i/, /eɪ/, /ɑ/, /ɔ/, /ɜ/, /əʊ/, and /u/.

Productions of Modeled Speaker

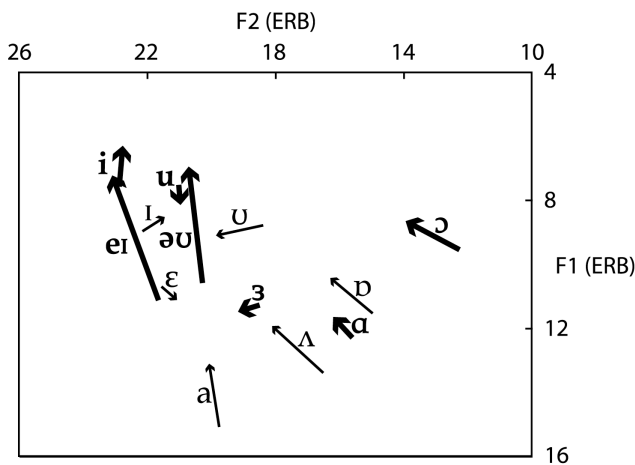


Figure 2: Formant frequencies for the natural productions of the modeled speaker. Lines go from the formant frequencies at the onset (20% of the duration) to the formant frequencies at the offset (80% of the duration). Vowels with long durations (>250 ms) are in bold with thick lines; vowels with short durations (<250 ms) are displayed with thin lines.

The overall patterns of preferred formant movement and duration generally corresponded with the acoustics of the natural productions of the modeled talker (Figure 2). One difference is that the listeners preferred much more formant movement for /ɛ/. The preferred vowels were hyperarticulated compared to the productions. This is particularly noticeable in the separation between the front and back vowels; the best exemplars of the traditional back vowels were less fronted than were the productions.

One advantage of this type of perceptual adjustment procedure is that it can be used to examine individual differences among talkers more easily than can production measurements; production measurements need to take into account physiological differences between talkers (e.g., overall differences in formant frequencies due to vocal tract lengths) but the best exemplar locations found using this procedure are all mapped onto the same talker. For example, Figure 3 displays the best exemplars for a female native German speaker who is highly fluent in English. Her preferred English vowels are generally in native-like locations in the vowel space, other than /ʊ/ and /u/ in which more traditional back vowels are preferred.

However, her patterns of formant movement do not match native speakers. For example, /i/, /ɪ/, /eɪ/, and /ɛ/ vary mostly in vowel height and duration, having similar magnitudes and directions of formant movement, rather than having the more complicated patterns of formant movement displayed by native speakers. Such a procedure thus may be useful for revealing more subtle differences in phonetic categorization due to language experience than can be found using more traditional measures that vary static F1 and F2 frequencies.

English Best Exemplar Locations for Native German Speaker

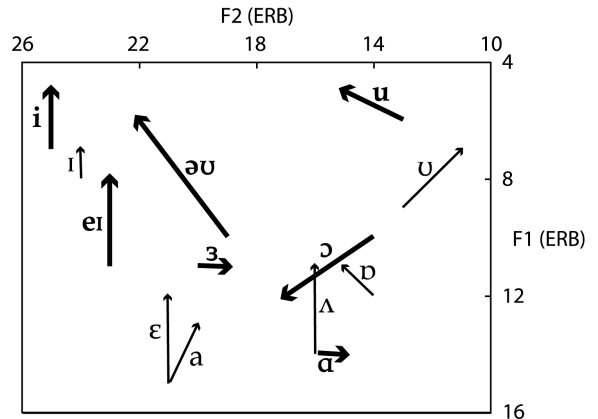


Figure 3: Average best exemplar locations of English vowels for one native German speaker. Lines go from the formant frequencies at the onset to the formant frequencies at the offset. Vowels with long durations (>250 ms) are in bold with thick lines; vowels with short durations (<250 ms) are displayed with thin lines.

4. SUMMARY

The results demonstrate that this goodness optimization method is able to make useful estimates of best exemplar locations within a 100,700 vowel stimulus set after 35 trials. This basic method can be used, in principle, with a wide variety of phonetic contrasts. For example, recent research suggests that the difficulty that Japanese adults have in learning the English /r/-/l/ distinction is due to their relatively high sensitivity to acoustic cues that are not used by American English listeners to categorize these phonemes [9]. Examining phonetic categorization within larger cue spaces, rather than just focusing on the one or two most important cues for native listeners, may thus be useful for further examining effects of language experience.

ACKNOWLEDGMENTS

This research was supported by an EPSRC Doctoral Training Award. We are grateful to C. A. Smith for her assistance with the experiments.

REFERENCES

[1] Ferguson, S. H., and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for

normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 112, 259-271.

- [2] Hillenbrand, J., and Nearey, T. (1999) Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, 105, 3509–3523.
- [3] Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (2000). Some effects of duration on vowel recognition, *Journal of the Acoustical Society of America*, 108, 3013–3022.
- [4] Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels, *Journal of the Acoustical Society of America*, 74, 695–705.
- [5] Pisoni, D. B. (1997). Some Thoughts on "Normalization" in Speech Perception. In K. Johnson & J. W. Mullenix (Eds.), *Talker Variability in Speech Processing* (pp. 9-32). Academic Press: San Diego.
- [6] Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of Acoustical Society of America*, 85, 2088-2113.
- [7] Johnson, K., Flemming, E., and Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated, *Language*, 69, 505–528.
- [8] Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. H. (1992). *Numerical recipes in C: the art of scientific computing*. Cambridge; New York: Cambridge University Press, 2nd ed.
- [9] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes, *Cognition*, 87, B47-B57.