Asymmetries between speech perception and production reveal phonological structure

Boersma, P.; Chládková, K.

Publication date
2011

Document Version
Final published version

Published in
The 17th International Congress of Phonetic Sciences: Hong Kong, China, August 17-21, 2011: congress proceedings

ASYMMETRIES BETWEEN SPEECH PERCEPTION AND PRODUCTION REVEAL PHONOLOGICAL STRUCTURE

Paul Boersma & Kateřina Chládková

University of Amsterdam, the Netherlands
paul.boersma@uva.nl; k.chladkova@uva.nl

ABSTRACT

It has been observed that in production, the boundary between the vowels /i/ and /e/ is diagonal, i.e. it involves both F1 and F2; in perception, by contrast, the boundary has been observed to be horizontal, i.e. listeners do not use F2 as a cue for distinguishing the two vowels. The same is true of the /u–o/ boundary. With computer simulations of virtual language learners we show that this perception-production discrepancy can be explained if vowels are structured as bundles of phonetically based phonological features.

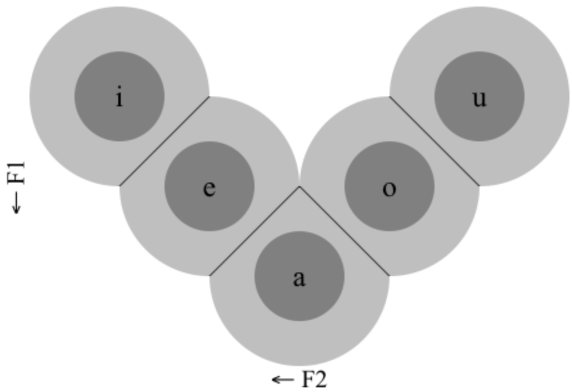
Keywords: phoneme boundaries, vowel systems, phonetically based features, cue constraints

1. OBSERVED VOWEL PRODUCTION AND PERCEPTION

Previous research on vowel production in many languages shows that the realizations of a vowel exhibit large variation both in F1 and in F2. This causes the distributions of neighbouring vowel categories to overlap both in F1 and in F2, as illustrated schematically in Fig. 1. As a result, the *production boundary* between e.g. /e/ and /i/ is diagonal; this boundary is defined as those F1–F2 combinations that the speaker must have equally likely intended as /e/ and as /i/. Acoustic analyses of vowel productions confirm that these boundaries are indeed diagonal in American English [8, 14], Dutch [1], French [16], German [16], Portuguese [7], and Czech [5].

Language users participate both in production and in perception, and if communication is to be successful we should expect symmetry between these two directions of phonetic processing. An *optimal perception* strategy of a listener confronted with the production environment of Fig. 1 would be to perceive every F1–F2 pair as the vowel category that was most likely intended by the speaker; in this way, the listener could minimize her perception errors. With this optimal perception strategy, the category boundaries in perception (i.e., the tokens that have an equal chance of being perceived as either of the two neighbouring vowel categories) should correspond to the category boundaries in production, that is, the *perceptual boundaries* should be diagonal, just as the production boundaries in Fig. 1.

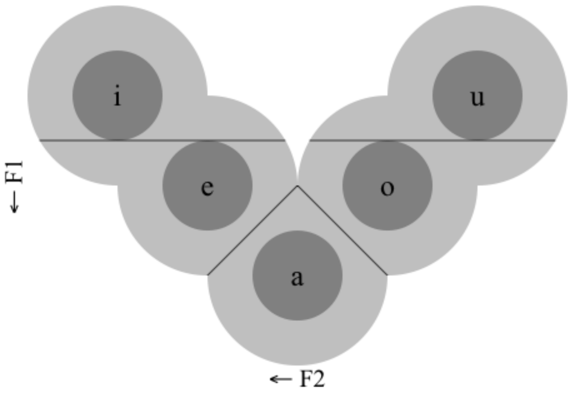
Figure 1: Stylized summary of the literature on production distributions in five-vowel systems. Dark grey disks could denote one standard deviation, light grey disks two standard deviations. The lines denote the production boundaries between pairs of vowels along the front or back edges of the vowel space.



However, this correspondence between production and perception boundaries is not what is observed in humans. In the results of vowel perception studies (Swedish [4]; Czech, Spanish, Polish, Italian, German, Dutch, Finnish [15]) we see that while the perception boundaries in the low-vowel region can indeed be diagonal, the perception boundaries between high vowels and their corresponding (high-)mid vowels are typically horizontal (as was noted by [4]); this situation is shown schematically in Fig. 2.

In other words, for the distinction between high and mid vowels listeners seem to ignore the F2 cue, although this cue is utilized in their language environment. This discrepancy between perception and production, which seems not to have been noticed before, calls for an explanation. In this paper, we propose an explanation in terms of phonetically based phonological features, supported by computer simulations with artificial language users.

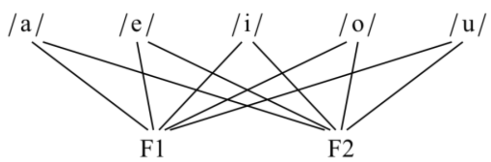
Figure 2: Stylized summary of the literature on vowel perception in five-vowel systems. The lines denote the perceptual boundaries between pairs of adjacent vowels (the disks show the production distributions of Fig. 1, for reference).



2. COMPUTATIONAL MODELLING

As summarized above, the production boundaries attested across languages look like the diagonal ones in Fig. 1, whereas the attested perception boundaries look like the horizontal ones in Fig. 2. Here we will derive this asymmetry within the linguistically oriented computational frameworks of Optimality Theory (OT) and Harmonic Grammar (HG). In these frameworks we represent the language user's knowledge of phonetic perception and production as a set of *connections* between phonological elements (e.g. vowel phonemes) and auditory cues (F1 and F2 values), as illustrated in Fig. 3.

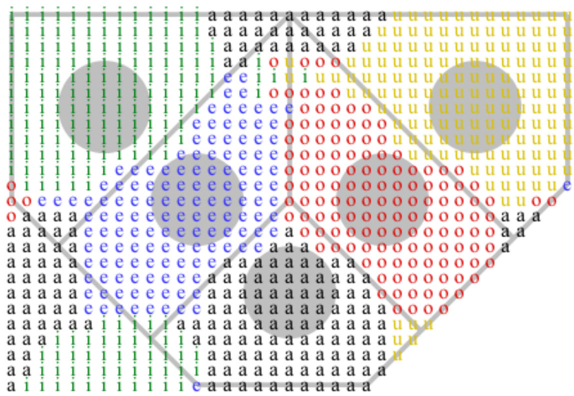
Figure 3: The phonetics-phonology interface when the phonological elements are phonemes.



2.1. Modelling with phonemes

Boersma and Escudero [3] modelled the perception of vowel distributions like those in Fig. 1 within Stochastic OT [2]. The connections of Fig. 3 were *cue constraints* such as "an F1 value of [*x*] is not the phonological vowel category /e/" and "an F2 value of [*y*] is not the phonological vowel category /i/". These cue constraints existed for all possible values of F1 and F2, and for all five vowel categories. Before learning began, all cue constraints were ranked at the same height; the virtual baby was then fed combinations of F1, F2 and the correct vowel category, and a simulated error-driven perceptual learning procedure [2] caused the cue constraints to become ranked in an optimal way, i.e. minimizing the probability of misperception. Figure 4 shows the ultimate perceptual behaviour of one typical virtual learner for the distributions of our Fig. 1 (100,000 pieces of data drawn from the five distributions of Fig. 1 with equal probability; evaluation noise 2.0; plasticity 0.01): all perceptual boundaries have become diagonal.

Figure 4: The perceptual behaviour of a simulated 'phonemic' Stochastic OT learner, computed by running F1–F2 pairs through the final simulated perception grammar. The thick grey lines stylize the perceptual boundaries.



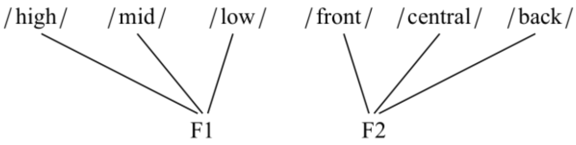
Boersma and Escudero showed that if the learners were modelled with Noisy HG instead of Stochastic OT, the exact same result applies.

The diagonal boundaries seen in Fig. 4 do correspond to the production boundaries of Fig. 1 and therefore represent *optimal perception*, but as the result is different from the behaviour of human listeners (Fig. 2) we conclude that Boersma and Escudero's phonemic cue model of Fig. 3 does not suffice to explain how real human listeners behave.

2.2. Modelling with features

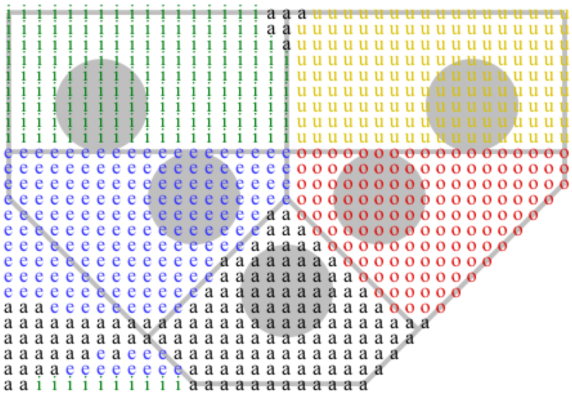
To improve the link with human behaviour, we now model the vowels as combinations of six *features* instead of in terms of unanalysed vowel phonemes: /a/ is the feature combination /low, central/, /e/ is /mid, front/, /i/ is /high, front/, /o/ is /mid, back/, and /u/ is /high, back/. The phonetics-phonology interface then comes to look like Fig. 5 instead of Fig. 3.

Figure 5: The phonetics-phonology interface when the phonological elements are features.



The vowel perception process is then modelled with six families of featural cue constraints, such as “an F1 of [*x*] is not /high/” and “an F2 of [*y*] is not /back/”. A typical simulated OT or HG listener now ends up with the perceptual behaviour of Fig. 6, where several boundaries are horizontal.

Figure 6: The perceptual behaviour of a simulated ‘featural’ Stochastic OT learner (“Greek” type).



The horizontal boundaries can be understood as follows. Between neighbouring vowels that differ in two features, the boundary is diagonal (there is *cue trading* of F1 and F2); this happens between /a/ and /e/ and between /a/ and /o/. Between neighbouring vowels that differ in only one feature, only F1 *or* F2 can be a distinguishing cue, and therefore the boundary has to be horizontal (as between /e/ and /i/ and between /o/ and /u/) or vertical (as between /e/ and /o/ and between /i/ and /u/).

The horizontal boundaries in this simulation correspond nicely with what the humans of Fig. 2 did. A crucial assumption needed to achieve this result was that the phonological features in Fig. 5 are *phonetically based*, i.e., F1 is linked only to the three height features and F2 only to the three backness features. If we redo the simulation with more arbitrary relations between the auditory level and the phonological level, i.e. with both F1 and F2 being cues for all six features, the result will be similar to the phoneme-based learning of Fig. 4.

2.3. Differences between five-vowel systems

Figure 1 is too much of an idealization. In reality, all five-vowel systems are slightly different. With different featural representations of the vowels we obtain Figs. 7 through 10.

Figure 7: As Fig. 6, but with /a/ being /back/ instead of /central/ (“Hebrew” type).

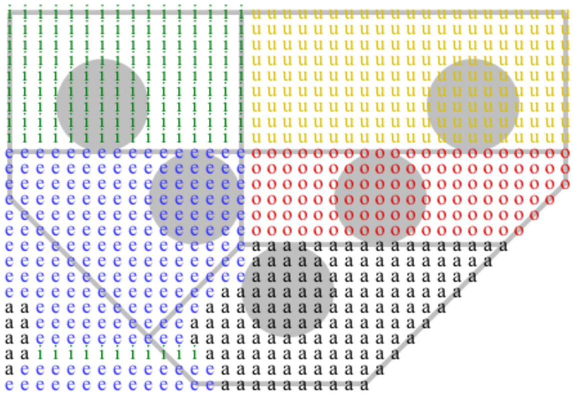


Figure 8: As Fig. 7, but with /e/ being /low/ instead of /mid/ (“Czech” type).

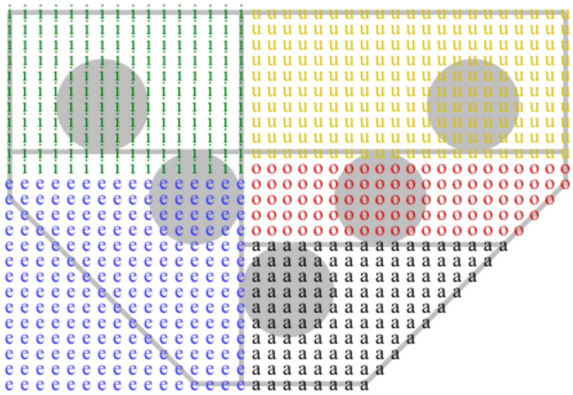


Figure 9: As Fig. 6, but with /e/ having a separate (fourth) place feature (“Spanish” type).

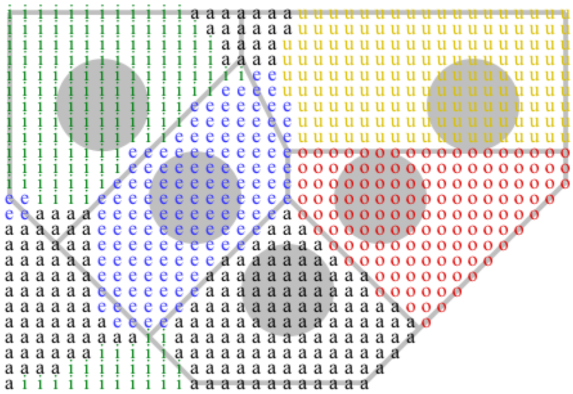
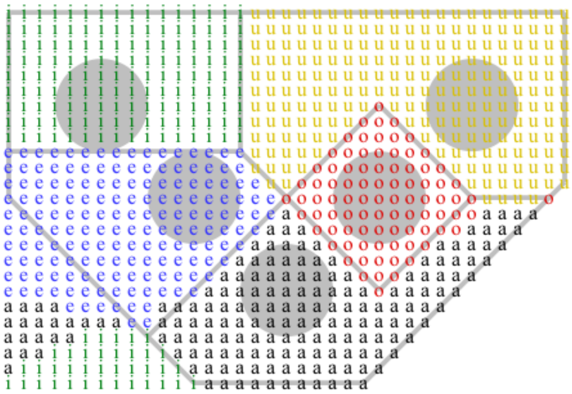


Figure 10: As Fig. 6, but with /u/ being /central/ rather than /back/ ("Japanese" type).



In Figs. 6-10, the perceptions are no longer centred around the tops of the input distributions. This means that if the learners use the same constraint rankings later in *production*, they will shift the distributions towards the centres of the perceptual spaces (i.e. away from the boundaries in the figures). Thus, the productions of the next generation will look similar to the vowel systems of Greek, Hebrew, Czech, Spanish and Japanese speakers, respectively [5, 6, 9, 10, 13].

3. CONCLUSION

Our simulated OT or HG listeners turn out to exhibit the same perception behaviour as humans (Fig. 2), i.e. typically with some horizontal boundaries that do not occur in production (Figs. 6–10). The crucial condition for this to work is our assumption that cue constraints refer to *phonetically based features* (i.e. with F1 connected only to /high, mid, low/, and F2 connected only to /front, central, back/) rather than to *phonemes* (i.e. with both F1 and F2 connected to all of /a, e, i, o, u/) or to *arbitrary features* (i.e. with both F1 and F2 connected to all of /high, mid, low, front, central, back/). Results of experiments on feature generalization with humans indeed suggest that listeners can attend to features, such as vowel height and backness [11, 12].

Furthermore, we related differences between seemingly similar vowel systems to the idea that the "same" vowel can be represented by different feature bundles in different languages.

In representing the production distributions (Fig. 1) we have been simplifying: the true distributions are whatever the listener cannot normalize away; if the listener can normalize for between-speaker variation but not for vowel reduction (Jan-Willem v. Leussen, p.c.), the clouds might not be circular but might instead be ellipses whose long axis is radial in the vowel space. In this way, we might obtain nearly horizontal boundaries, even if vowels are represented as phonemes. This possibility must remain an object of further study.

In general, we have provided a method for detecting phonological structure from asymmetries between phonetic perception and production. This principle can in the future be applied to other cases than five-vowel systems.

4. REFERENCES

[1] Adank, P., van Hout, R., Smits, R. 2004. An acoustic description of the vowels of Northern and Southern standard Dutch. *J. Acoust. Soc. Am.* 116, 1729-1738.

[2] Boersma, P. 1997. How we learn variation, optionality, and probability. *IFA Proceedings* 21, 43-58.

[3] Boersma, P., Escudero, P. 2008. Learning to perceive a smaller L2 vowel inventory: An Optimality Theory account. In Avery, P., Dresher, E., Rice, K. (eds.), *Contrast in Phonology: Theory, Perception, Acquisition*. Berlin & New York: Mouton de Gruyter, 271-301.

[4] Chistovich, L., Fant, G., de Serpa Leitao, A. 1966. Mimicking and perception of synthetic vowels, part II. *STL-QPSR* 7(3), 1-3.

[5] Chládková, K., Boersma, P., Podlipský, V.J. 2009. On-line formant shifting as a function of F0. *Proc. of Interspeech 2009*, 464-467.

[6] Chládková, K., Escudero, P., Boersma, P. In Press. Context-specific acoustic differences between Peruvian and Iberian Spanish vowels. *J. Acoust Soc. Am.* 130.

[7] Escudero, P., Boersma, P., Rauber, A.S., Bion, R.A.H. 2009. A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *J. Acoust. Soc. Am.* 126, 1379-1393.

[8] Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099-3111.

[9] Jongman, A., Fourakis, M., Sereno, J.A. 1989. The acoustic vowel space of Modern Greek and German. *Language and Speech* 32, 221-248.

[10] Keating, P.A., Huffman, M.K. 1984. Vowel variation in Japanese. *Phonetica* 41, 191-207.

[11] Kingston, J. 2003. Learning foreign vowels. *Language and Speech* 46, 295-349.

[12] Maye, J., Aslin, R.N., Tanenhaus, M.K. 2008. The weckud wetch of the wast: lexical adaptation to a novel accent. *Cognitive Science* 32, 543-562.

[13] Most, T., Amir, O., Tobin, Y. 2000. The Hebrew vowel system: Raw and normalized acoustic data. *Language and Speech* 43, 295-308.

[14] Peterson, G.E., Barney, H.L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175-184.

[15] Savela, J. 2009. *Role of Selected Spectral Attributes in the Perception of Synthetic Vowels*. Ph.D. dissertation, University of Turku.

[16] Strange, W., Weber, A., Levy, E.S., Shafiro, V., Hisagi, M., Nishi, K. 2007. Acoustic variability within and across German, French, and American English vowels: phonetic context effects. *J. Acoust. Soc. Am.* 122, 1111-1129.