IS CONSONANT PERCEPTION LINKED TO WITHIN-CATEGORY DISPERSION OR ACROSS-CATEGORY DISTANCE?

Valerie Hazan & Rachel Baker

Department of Speech, Hearing and Phonetic Sciences, University College London (UCL), UK
v.hazan@ucl.ac.uk; rachelbaker81@googlemail.com

ABSTRACT

This study investigated the relation between the internal structure of phonetic categories and consonant intelligibility. For two phonetic contrasts (/s/-/ʃ/ and /b/-/p/), 32 iterations per category were elicited for each of 40 talkers from a same accent group and age range, and measures of cross-category distance and within-category dispersion were obtained. These measures varied substantially across talkers but were not correlated across both contrasts suggesting that degree of cross-category distance or within-category dispersion is not consistent within-speaker. For each contrast, consonant identification tests in mild babble noise, that presented the complete set of iterations for eight talkers showing extreme values in these two measures, revealed some talker effects on reaction time. However, these did not appear to be correlated with either cross-category distance or within-category dispersion for those talkers.

Keywords: talker variability, consonant production, consonant perception

1. INTRODUCTION

The acoustic characteristics of utterances vary significantly between speakers and across speaking styles; to what degree does between- and within-speaker variability impact on the listener's comprehension what is being said, or ease with which speech is processed? Studies that have investigated the relation between speaker intelligibility and acoustic-phonetic characteristics of the speech have found significant but weak correlations with a number of acoustic-phonetic characteristics, such as amount of energy in the mid-frequency range, size of vowel space, fundamental frequency range and speech rate [2, 5]. The fact that correlations are rather weak and variable across studies could be due to individual speakers using different strategies to achieve greater clarity, e.g. [5]. It could also be the case that speech clarity may not necessarily be associated with more extreme values in an acoustic-phonetic 'space'; intelligibility might be linked to features linked to the internal structure of phonetic categories. One contender is the degree of within-category dispersion in acoustic-phonetic patterns: speakers showing little within-category dispersion may be clearer than speakers who are less consistent in their productions for a given phonetic category. A related but not identical factor is the degree of between-category distance and/or overlap in acoustic-phonetic patterns. Some support for these two options comes from [7] who examined the degree of overlap in the main energy distribution of two fricatives /s/ and /ʃ/ in different speakers. Greater within-category dispersion and cross-category overlap were associated with slower response times in identification tests. The impact of within-category dispersion and cross-category distance requires further investigation.

The first research aim was to explore, for two phonetic contrasts (/s/-/ʃ/ and /b/-/p/) how the factors of cross-category distance and within-category dispersion varied across a range of 40 talkers. The study investigated whether, for a given talker, these factors were correlated across two different phonetic contrasts and therefore more likely to represent a general characteristic of this talker's speech, or whether within-category dispersion and cross-category distance were contrast-specific. The second aim was to investigate whether these measures were correlated with consonant intelligibility. Tokens from a subset of 8 talkers showing extreme values in these two measures were included in the perception test. Findings of these two studies should inform our understanding of the impact of within-speaker variability on speech perception.

2. STUDY 1: PRODUCTION

2.1. Participants

Speech materials were recorded from forty native talkers of Southern British English (20 M, 20 F; 19-29 yrs old), who were students or staff from the University of London.

2.2. Speech materials

The speech materials for this task were recorded as part of a large corpus of spontaneous and read casual and clear speech, the LUCID corpus [7]. To collect multiple iterations of a set of word tokens, participants carried out a picture naming task. An easily-recognizable picture was found for each of the 36 keywords (18 near-minimal pairs) containing the phonemes /p,b,s,ʃ/ in initial position. 30 of these keywords were represented by a picture of a noun (e.g., ‘ball’), and 6 were represented by a picture of a verb (e.g., ‘push’).

2.3. Speech recordings

The picture elicitation task was run with the stimuli presented via DMDX software [3], and participants wearing Beyerdynamic DT297PV microphone headsets. In the recording session, a picture appeared on the screen and participants were instructed to name each picture using one of two frame sentences: ‘I can see a (noun)’ or ‘the verb is to (verb)’. The 36 pictures were each presented 8 times in a pseudo-randomized order (nouns and verbs were presented in separate blocks).The speech recorded for each utterance at a sampling rate of 22050Hz was automatically saved by DMDX into a separate file in wav format.

2.4. Acoustic-phonetic analyses

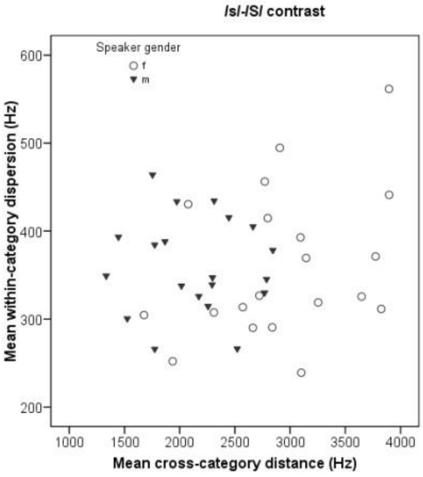
For the dispersion analysis, all token iterations for a subset of the minimal pairs were analysed: beach-peach, bee-pea, bill-pill, bin-pin, sea-sheep, seat-sheet, cell-shell and sack-shack, giving 32 tokens per talker for each of the four phonetic categories (1280 tokens in total per category). All tokens were annotated in Praat [1]. For the words with initial /s/-/ʃ/, the start and end of the initial fricative segment (excluding mixed excitation) were marked, and a Praat script was used to calculate spectral centre of gravity (CoG), a measure reflecting the spectral distribution of the frication. For the words with initial /p,b/, Voice Onset Time (VOT) was marked from burst release to the onset of the first voiced period, and a Praat script was used to calculate VOT duration for each of the iterations. For both contrasts, two further measures were derived to quantify the characteristics of the within- and across-category distributions for each talker: a measure of within-category dispersion, calculated per talker as the mean standard deviation averaged across both categories in the contrast, and a measure of cross-category distance, calculated at the difference between the mean (for /p/-/b/) or median values (for /s/-/ʃ/) for the two categories in the contrast.

2.5. Results

2.5.1. Measures of within-category dispersion and cross-category distance for /s/-/ʃ/

There is substantial cross-talker variability in within-category dispersion and cross-category distance in CoG for the fricative segments for /s/-/ʃ/ (Figure 1). Z-scores were calculated for these measures separately for male and female speakers to take account of gender-based differences in fricative CoG. Using z-scores in the calculations, within-category dispersion and cross-category measures were found not to be significantly correlated, as can clearly be seen from Fig. 1.

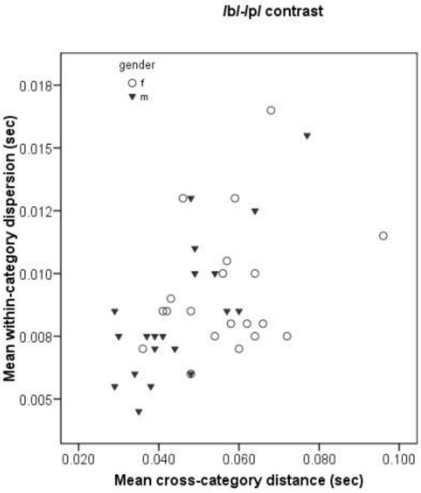
Figure 1: Cross-plot of the mean cross-category distance and within-category dispersion in fricative Centre of Gravity (CoG) for the /s/-/ʃ/ contrast.



2.5.2. Measures of within-category dispersion and cross-category distance for /p/-/b/

VOT measures for the /p/-/b/ contrast again show substantial variance in both within-category dispersion and across-category distance across talkers (Figure 2). However, here the correlation between these two measures was significant (r=0.536; p<0.001): there was a tendency for larger cross-category distances to be associated with a greater degree of within-category dispersion (especially for the /p/ category, as might be expected). As illustration, Figure 3 shows token distributions for male talkers showing the extremes values for within-category dispersion.

Figure 2: Cross-plot of the mean cross-category distance and within-category dispersion in VOT for the /p/-/b/ contrast.



2.5.3 Correlation across contrasts

If it is the case that speakers are more or less consistent or more or less extreme in their articulations, then a positive and significant correlation would be expected for measures of within-category dispersion or cross-category distance across the two phonetic contrasts under investigation. Values were first converted to z-scores within the male and female speaker groups. Pearson's product-moment correlations showed that neither measures of cross-category distance (r=-.205; p>0.05) or measures of within-category dispersion (r=0.063; p>0.05) were correlated across contrasts, thus refuting the hypothesis that these may be more general talker characteristics rather than contrast-specific.

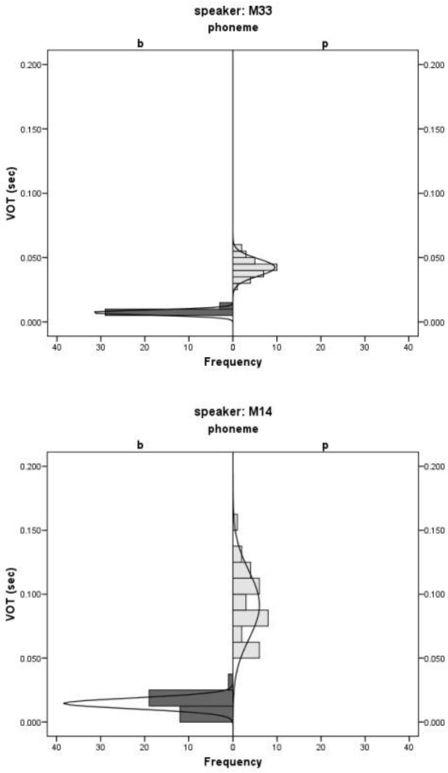
3. STUDY 2: PERCEPTION

The perception study investigated whether tokens produced by talkers showing greater within-category dispersion and/or smaller cross-category distance would be less easily perceived (as shown by a slower reaction time) than tokens produced by talkers who were more consistent and/or showed greater cross-category distance. The approach used was similar to that used in [7] except that our study presented tokens in a mild degree of background noise, that it included two contrasts rather than only /s/-/ʃ/, and 8 talkers rather than 2 per experiment. A similar number of iterations per talker was used in both studies.

3.1. Participants

The listener group included 32 right-handed monolingual participants (7 M, 25 F, age range: 19-30 yrs), from the same accent group as the speakers, divided into two groups of 16 listeners. They were screened for normal hearing thresholds. Each listener group carried out the experiment for one of the contrasts.

Figure 3: Distributions of VOT measures for /p/-/b/ for male talkers showing small (M33) and large (M14) within-category dispersion and distance.



3.2. Materials

For the /s/-/ʃ/ contrast, the perception test included all 64 tokens (32*2 consonants) for each of eight male talkers showing extreme values in terms of cross-category distance or within-category spread. Two were chosen from each quadrant in Fig. 1: high dispersion-high distance (HH), high-dispersion-low distance (HL), low dispersion-high distance (LH), low dispersion-low distance (LL). For /p/-/b/, given that distance and dispersion were correlated, the eight male talkers were selected at equal intervals across the range. Again, all 64 tokens for each talker were included. All speech files were normalized to a fixed intensity level then mixed with 8-talker babble noise at a signal to noise ratio of 0 dB, using a matlab script.

3.3. Method

Participants carried out the study in a sound-treated booth, with the tokens presented via headphones at a comfortable listening level, and randomized across talkers and words. Listeners were instructed to pay attention to the initial segment of the word and to press one of two keys on a keyboard corresponding to the initial consonant as quickly and accurately as possible but only after the whole word had been produced. Response keys were counterbalanced across participants to minimize any handedness effects.

3.4. Results

For /s/-/ʃ/, for each listener, a median RT per talker was calculated over all correct tokens after outlier RTs (>2 SDs of mean) had been removed. Repeated-measures ANOVA with group and talker as within-subject factors showed a significant effect of group [F(3, 45) = 20.7, p<.001], with a slower RT for the low distance, low dispersion group. However, as the talker by group interaction was also significant, the two talkers in each group did not show similar trends. For the LL group, RT was 267 ms for Talker 1 but only 238 ms for talker 2. Also, listeners were not slower to respond to tokens from the low distance-high dispersion group which would be likely to be even more confusable.

Table 1: Median reaction times (RT) in ms for the four talker groups.

Talker group	RT	Talker group	RT
HH	217.9	LH	219.2
HL	228.6	LL	253.0

For the /b/-/p/ data, for each listener, a median RT per speaker was calculated over all correct tokens after outlier RTs had been removed. Repeated-measures ANOVA with talker as within-subject factor revealed a significant effect of talker [F(7, 119) = 6.8, p<.001]. However, the talkers for which a shorter RT was obtained were not those showing extreme values of cross-category distance or within-category dispersion.

Table 2: Median reaction times (RT) in ms for the eight talkers listed in order to increasing distance/dispersion.

Talker	RT	Talker	RT
M33	249 (77)	M07	234 (73)
M10	275 (75)	M13	235 (75)
M41	248 (72)	M08	255 (71)
M17	249 (73)	M14	225 (67)

4. DISCUSSION

The study by Newman et al. [7] suggested that there were perceptual consequences to talker variability, with greater consistency of production leading to an easier and faster classification of initial consonants. Our acoustic-phonetic analyses of multiple iterations of tokens with /b/-/p/ and /s/-/ʃ/ initial consonants also found substantial variability in within-category dispersion and cross-category distance across talkers consistent in age and regional accent. However, talkers who had a high or low degree of within-category dispersion or cross-category distance for one phonetic contrast did not necessarily do so for the other contrast suggesting that these are not general talker characteristics per se. Further, in our perception study involving a larger number of talkers than in [7] and testing two phonetic contrasts, cross-talker effects on reaction time were found but the talkers whose consonants were easier to classify were not those showing a small degree of within-category dispersion or high degree of cross-category distance, so the talker effects do not appear directly linked to internal category structure. The fact that more talkers were used in the test might have made it harder for listeners to map internal phonetic category structure for individual talkers. These data suggest that the conclusions of Newman et al. [7] might have been premature.

5. ACKNOWLEDGMENTS

This project funded by the UK Economics and Social Research Council (RES-062-23-0681).

6. REFERENCES

[1] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 9(10), 341-345.

[2] Bradlow, A.R., Torretta, G.M., Pisoni, D.B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Com.* 20, 255-272.

[3] Forster, K.I., Forster, J.C. 2003. DMDX: A windows display program with millisecond accuracy. *Behavior Res. Methods, Instruments, & Computers* 35, 116-124.

[4] Hazan, V., Baker, R. Conditionally accepted. Acoustic-phonetic characteristics of clear speech produced with and without communicative intent. *J. Acoust. Soc. Am.*

[5] Hazan, V., Markham, D. 2004. Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acoust. Soc. Am.* 116, 3108-3118.

[6] Newman, R.S., Clouse, S.A., Burnham, J. 2001. The perceptual consequences of acoustic variability in fricative production within and across talkers. *J. Acoust. Soc. Am.* 109, 1181-1196.