INDIVIDUAL VARIABILITY IN THE PERCEPTION OF CUES TO AN INITIAL BA-PA VOICING CONTRAST

V. Hazan and B. Shi

Dept of Phonetics & Linguistics, University College London, U.K.

ABSTRACT

This study investigates whether individual variability in the categorisation of a voiced-voiceless speech contrast is related to the stimulus type used in the perceptual experiment. Continua constructed using copy-syntheses, computer-edited natural tokens and stylised syntheses were used. Categorisation of reduced-cue continua was also examined for the copy-synthesised and natural-edited ranges. Greater variability was generally found in the labelling of copy-synthesised continua.

1. INTRODUCTION

An initial study on the perception of initial stop place contrasts [1] has shown that individuals may vary greatly in the extent to which they are affected by the neutralisation of specific cues to these contrasts. Greater individual variability was found in the labelling of stops in an /ei/ environment than in an /a/ vowel environment and in the perception of complex syntheses, copied from a natural utterance, than that of more stylised syntheses.

The aim of the present study was to assess whether the variability was related to stimulus type, by controlling vowel environment. Stimulus types used included computer-edited natural speech, high-quality copy-synthesis based on the same natural tokens and a highly stylised synthesised continuum created at the Haskins Laboratories [2].

2. STIMULI

The natural-edited and copy-synthesis versions were presented in three conditions:

a. *full-cue* (VOT and F1 cutback): change in VOT from -20ms to +70 ms in nine steps *and* change in the F1 onset frequency.

b. *Ba/VOT*: same change in VOT with, at vowel onset, formant frequencies characteristic of [ba] (rising F1 onset throughout).

c. *Pa/VOT*: same change in VOT, with, at vowel onset, formant frequencies characteristic of [p^ha] (flat F1 onset throughout).

2.1. Natural edited stimulus continua

Recording were made of tokens of /pa/ and /ba/ produced by an adult English male speaker. Two tokens were chosen which were characterised by clear formant patterns and a regular fundamental frequency trace of around 100Hz to facilitate editing in 10ms steps.

The creation of natural-edited stimulus continua was done on a mini-computer using a "cut and paste" technique. For the Ba/VOT continuum, the vowel portion from the [ba] token was appended to the burst and aspiration portion from the [p^ha] token. For the creation of stimuli with VOT between 70 ms and 5 ms, the aspiration was progressively deleted from the vowel end

in 10 ms steps. For stimuli with negative VOTs, the prevoicing portion was cut out from the [ba] stimulus and appended to the initial burst. The same process was carried out for the Pa/VOT continuum, except that, in this case, both the burst/aspiration and vowel portions were taken from the [p^ha] token. In the 'full-cue' continuum, for steps with positive VOTs, initial cycles of the vowel portion were deleted as VOT increased to create formant cutbacks in voiceless tokens.

2.2 Synthetic stimuli

The natural tokens used as a base for the natural-edited continua were analysed using a ten pole closed-phase LPC analysis to derive the formant frequencies. Amplitude control parameters were obtained using an FFT analysis [3]. A first resynthesis through a 4 kHz bandwidth, software parallel formant synthesiser was performed. Further modifications to the syntheses were then made on the basis of comparisons between the natural and synthetic spectra on a Kay digital spectrograph until a close match was obtained. Analogous conditions to the ones created for the natural-edited speech were prepared. For more details on stimulus preparation, see [4].

The stylised synthetic Haskins continuum was presented in the full-cue condition only. The VOT range used was the same as above.

3. SUBJECTS

Subjects were 18 paid volunteers with normal hearing as defined by average thresholds of 10 dB HL or better, from .25 to 8 kHz. The listeners ranged in age from 18 to 29 years (mean: 20.7 years) and had no previous listening experience of synthetic speech.

4. TEST PROCEDURE

Stimuli were presented in the form of two-alternative forced-choice identification tests over four sessions. At each session, seven tests were presented. Each consisted of 10 tokens repeated randomly eight times. Stimuli were presented at a comfortable listening level through headphones.

5. RESULTS

A statistical approach based on generalized linear models (GLMs) fit by maximum likelihood estimation was used to determine the extent to which performance varied across different test conditions. This technique, analogous to Analysis of Variance, was used as it is especially tailored to the analysis of multi-variate data involving binary responses (for a more detailed description, see [1]).

Using GLM, phoneme boundary and gradient measures were derived from the best fit cumulative normal to the four repetitions of each test condition for each of the 18 subjects (Fig. 1). A mean VOT phoneme boundary value was then derived for each of the three "full cue" conditions. The mean boundaries obtained were 13.5 ms (s.e. 5.1) for the natural edited condition, 13.3 ms (s.e. 6.2) for the stylised synthesis condition and 22.4 ms (s.e. 5.0) for the copy-synthesis condition. The mean gradient values obtained were -2.492 (s.e. 1.865) for the natural edited condition, -2.604 (s.e. 1.997) for the stylised synthesis condition, and -1.289 (s.e. 0.784) for the copy-synthesis condition. Highly similar phoneme boundary and gradient values were therefore obtained for the stylised syntheses and natural-edited stimuli. The copy-synthesis condition was less sharply labelled and showed a shift in boundary.

The next step of the analysis was to investigate difference in labelling between conditions for individual subjects. For each subject, the condition deviances, which are quantitative, statistically interpretable, measures of the extent to which subjects change their labelling behaviour across conditions (see [1]) were calculated. Labelling of the stylised synthesis condition was

compared with labelling of the other full-cue conditions. 83 % of subjects showed significant deviances at the 0.001 level (deviances greater than 26.1) between the copy-synthesised and stylised synthesis continua. Significant deviances were only found for 44% of subjects when the natural edited and stylised synthesis stimuli were compared and the range of deviances obtained (8.9 to 61.9) was generally smaller than in the first comparison (22.6 to 174.6).

Next, the effects of cue reduction on phoneme boundary and gradient for copy-synthesised and natural-edited continua were examined. For the natural edited stimuli, the mean phoneme boundary increased from a value of 13.52 ms for the full-cue condition, to 16.89 ms (s.e 6.14) for the Ba/VOT condition and decreased to 0.47 ms (s.e. 11.73) for the Pa/VOT condition. For the copy-synthesised stimuli, the shift was from 22.41 ms for the full-cue to 25.04 ms (s.e. 5.53) for the Ba/VOT and 10.1 ms (s.e. 15.4) for the Pa/VOT condition.

Condition deviances were again calculated to compare labelling for the full-cue condition and the two reduced-cue conditions for individual listeners. For the natural edited range, very few listeners (11%) showed a significant deviance ($p<0.001$) between the full-cue and Ba/VOT condition. For the copy-synthesised stimuli, a greater number of listeners (33%) showed such an effect. Greater differences in labelling were found between the full-cue and Pa/VOT conditions. Generally greater individual variability in the labelling of this reduced cue condition was obtained, showing that some listeners were more greatly affected by changes in the spectral characteristics than others (Fig. 2). With the natural edited stimuli, all listeners showed a significant deviance between the two conditions with condition deviances ranging from 58.7 to 201.4, while, with the copy-synthesised stimuli, only 72% showed such an effect (deviances ranging from 9.4 to 240.2).

6. DISCUSSION

When full-cue ranges were presented, more similar results were obtained for natural-edited and highly stylised Haskins synthetic continua than for a copy-synthesised continuum based on parameters measured from the same natural tokens. One explanation might be that, in the Haskins continuum, the unnaturalness of the highly stylised stimuli is compensated by the clear enhancement of the cues which are present. With the copy-synthesised stimuli, listeners are having to deal with a complex set of patterns which may also contain slight inaccuracies in terms of formant bandwidth values and intensity relations for example. Certain listeners, especially in reduced-cue conditions, may be more sensitive to these inaccuracies and as a result, show greater variability in categorisation.

When looking at the effect of cue reduction, it was found that the lack of an appropriate F1 onset with short VOT (Pa/VOT condition), generally led to a smaller number of "voiced" responses, showing the importance of spectral cues to the voicing contrast. For both stimulus types, individual listeners varied in the extent to which they were affected by the spectral cue to the voicing contrast as shown by large differences in condition deviance measures obtained. However, more homogeneous results were obtained with natural-edited stimuli than with copy-synthesised stimuli.

7. REFERENCES

[1] HAZAN, V. and ROSEN, S. (1991) Individual variability in the perception of cues to place contrasts in initial stops. *Perception and Psychophysics*, vol.49, 2.

[2] LISKER, L and ABRAMSON, A. (1970). The voicing dimension: some experiments in comparative phonetics. *Proc. of the 6th ICPhS*, Prague, 1967 (Academia, Prague), 563-567.

[3] HOLMES, W. (1989) Copy synthesis of female speech using the JSRU parallel formant synthesiser. *Proc. of Eurospeech '89* (Paris), 513-516.

[4] SHI, B. and HAZAN, V. (in press) Effect of stimulus type on the labelling of a /ba/-/pa/ voicing contrast. *Speech, Hearing and Language, UCL Work in progress*, vol.5.

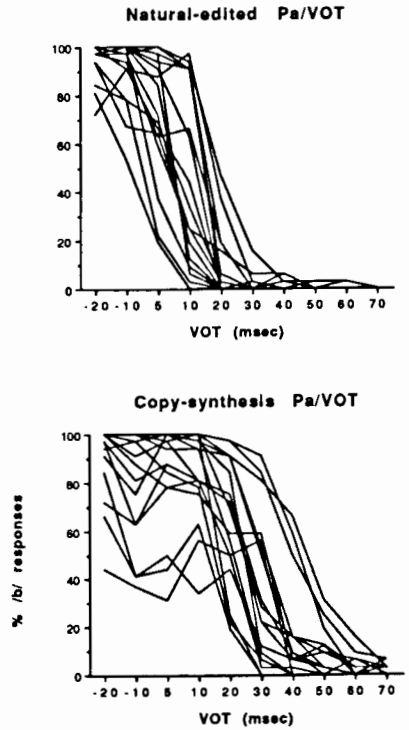

Figure 2: Individual labelling functions for the two Pa/VOT conditions.

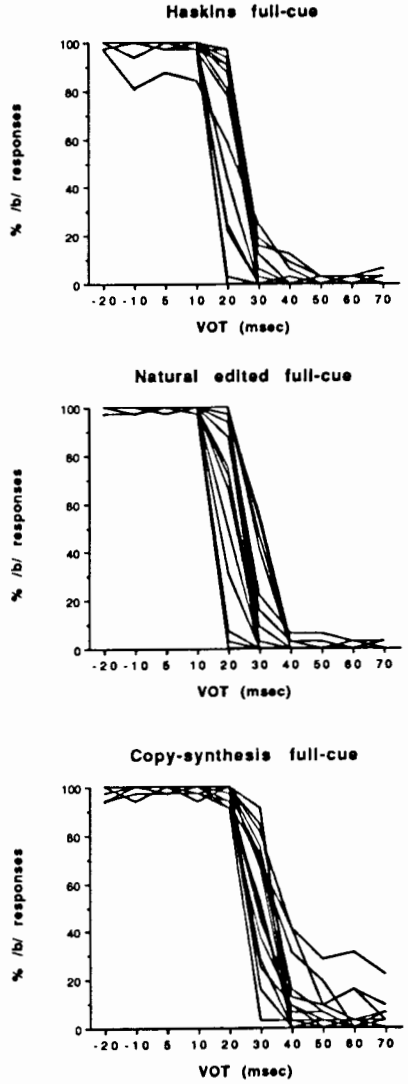

Figure 1: Individual labelling functions for the three full-cue conditions.