CUE-TRADING RELATIONS FOR INITIAL STOP VOICING CONTRAST AT DIFFERENT LINGUISTIC LEVELS

U THEIN-TUN

School of Communication Disorders
Lincoln Institute
Melbourne, Australia 3053

ABSTRACT

The speech mode of processing is a special mode of information processing in the sense that the cue-trading relationship of multiple cues signifying one phonemic contrast is most effective in the speech mode but not in the sense that cue-trading does not exist at non-linguistic levels.

INTRODUCTION

Fitch, Halwes, Erickson and Liberman [8] claimed that virtually every phonetic contrast is cued by several distinct acoustic properties of speech signal and within limits set by the relative perceptual weights and by the ranges of effectiveness of these cues, a change in the setting of one cue can be offset by an opposed change in the setting of another cue so as to maintain phonetic percept. This phenomenon of of perception is generally known as phonetic cue-trading relation. Using the results of new works by Bailey, Summerfield and Dorman [1] , Best, Morrongiello and Robson [2] and Repp [12] , Repp [13] argues that the cue-trading relation operates only in the phonemic mode of perception but not in the auditory mode of perception. On the basis of this point, he claims that speech perception is a special mode of perception different from the mode of perceiving non-speech sounds.

Best et al. [2] investigated the cue -trading relation between silence gap and F1 onset frequency for the "say"/"Stay" continuum as their test stimuli. They found that the cue-trading relation was evident only in the group which was instructed to treat the sinewave analogues as speech sounds but not in the group which was told that the stimuli were non-speech computer sounds. However, their findings provoke further questions about the nature of cue-trading and its relevance to the nature of speech perception. All their test stimuli were confined to the word level. Thus it is necessary to ascertain how the cue-trading relation as a phenomenon behaves at non-linguistic and linguistic levels other than words i.e. sentence, syllable, phonetic and auditory levels, so that the finding may shed more light on the controversy of "speech specificity". The present study was conducted in order to answer the following research questions:

(i) along the five linguistic levels mentioned above, do individuals' (both normal hearing and hearing impaired) cue-trading relations at one linguistic level differ from those of the other levels, with initial voicing contrast as an example?
(ii) if the cue-trading relations differ from one level to another, what is the interlinguistic level pattern of cue-trading?
(iii) how do the answers to the above questions fit into the present controversy of special v. non-special mode of speech perception?

METHOD

Selection of Segments and Creation of Linguistic Levels

The stop consonant type selected for initial voicing contrast was alveolar because alveolar stops have the most confined range of initial F2 and F3 frequencies (see [7] :123). The vowel /a/ was chosen for the syllable level continuum. The diphthong /ai/ (forming the words "dye"/"tie") was chosen for the word level continuum. The same /dai/-/tai/ continuum was chosen for the sentence level as well. In order to create the sentence level processing, the individual steps of the /dai/-/tai/ continuum were placed at the end of PL carrier sentences. The phonetic and auditory level stimuli were the sinewave analogues of the syllable level /da/-/ta/ ten VOT steps.

Syllable Level Stimuli

Using the 12 parameter serial analogue speech synthesiser designed by Clark [3] and [4] , step one of the syllable level continuum, one with 0 VOT i.e. the good /da/ of the /da/-/ta/ continuum, was created first. The frequency values of the three formant patterns of the good /da/ were set after the averaged measurements of the five /da/ spectrograms of five general to broad Australian male native speakers of English. The duration was 300 msec. The fundamental frequency was constant at 125 Hz over the first 85 msec and fell linearly to 90 Hz. The initial formant transitions were stepwise linear and 45 msec in duration. F1 rose from 285 to 770 Hz, F2 fell from 1540 to 1233 Hz and F3 fell from 3019 to 2520 Hz. The duration of the synthesis time

frame was 5 msec i.e. the synthesis data was updated at every 5 msec. Then the VOT continuum for the remaining nine 5-msec steps was created by replacing the periodic voiced (V) excitation with noise and simultaneously increasing the bandwidth of F1 transition to its maximum and hence virtually eliminating the existence of F1 transition . The first /da/-/ta/ continuum created in such a way produced a good/da/ on one end and a good /ta/ on the other. The amplitude levels of the noise and vowel portions (though different in actual measurements) in this continuum were given the nominal 0dB each. Therefore the first /da/-/ta/ continuum can be described as bearing the nominal amplitude pattern of 0dB A (aspiration noise) and 0dB V (vowel portion). Eight more /da/-/ta/ continua were created by increasing and decreasing both the 0dB A and 0dB V amplitudes by 6dB as described below.

A amplitude		V amplitude
+6dB	orthogonally	+6dB
0dB	combined	0dB
-6dB		-6dB

The formulation of these syllable level continua was almost the replica of Repp [11]. Through such an arrangement it was expected that the stimuli with the A=+6dB and V=-6dB pattern would produce more /t/ responses and those with A=-6dB and V=+6dB would produce more /d/ responses. For normal hearing listeners, an increase or decrease in amplitude by 6dB is appropriate to make the stimulus noticeably louder and fainter respectively. Every step from each continuum served as one stimulus. On the test tape all the ten VOT steps of the nine continua (i.e. 90 stimuli) were randomised four times with an interstimulus interval of 3.5 msec and these four sets of randomisation served as four blocks of test stimuli for the syllable level.

Phonetic and Auditory Level Stimuli

The phonetic and auditory level stimuli were the same four blocks of 90 stimuli each from the syllable level. The only difference was that the phonetic and auditory level stimuli were the sinewave analogues of the syllable level stimuli.

Word Level Stimuli

The word level stimuli were in principle the same as the four blocks of 90 stimuli each from the syllable Level. The difference was that the word level stimuli were from the nine /dai/-/tai/ continua instead of the nine/da/-/ta/ continua of the syllable level. With the exception of the initial transition duration, the durations and frequency values of the formant trajectories in the /dai-tai/ continua were set after the corresponding averaged values of the five /dai/ spectrograms of five (general to broad) Australian male native speakers of English. As with the good /da/ stimulus at the syllable level, the duration of the initial formant transitions was 45 msec long in the good /dai/ stimulus. Such an arrangement was necessary in order to maintain the uniformity of VOT steps along the the different linguistic levels. It was the duration of the initial formant transitions which was progressively replaced by noise in order to create VOT steps in this experiment.

Sentence Level Stimuli

The sentence level stimuli were the same four blocks of the word level /dai/-/tai/ ("dye"-"tie" as words) stimuli. In order to create the sentence level processing for the subjects, the "dye"-"tie" test stimuli were presented in the context of the sentences whose semantic, syntactic, vocabulary and phonetic variations were controlled. Every carrier sentence consisted of seven syllables including the stimulus word at the end. On the sentence level test tape, the 90 sentences used were synthesised according to the synthesis by rule system of Australian English by Clark [5] and [6]. The average amplitude of the carrier sentences was maintained at the same value of the nominal 0dB of the V amplitude in the stimulus word. Fourteen spectrograms of the seven basic sentences (each carrying a good "dye" and a good "tie") spoken by Professor Clark were used as norms in synthesising the sentences.

Subjects

Thirteen male and 15 female normal hearing listeners and 7 male and 2female (sensorineurally) impaired listeners took part in the listening tests. They were all native speakers of Australian English and naive listeners of synthesised speech, between 20 and 40 years of age.

Test Procedure

The listening test was conducted in the acoustically treated speech perception laboratory of the Macquarie Speech, Hearing and Language Research Centre (SHLRC). All the listeners wore the Telephonics TDH 49P audiometric headphones with circumaural seals and each listener sat at a test booth. The test tapes were played on a Revox B77 MKII stereo tape recorder. The output level of the tape was controlled by an HH professional power amplifier TPA 25-D with calibration control panel and level meter attached. The tape output level was adjusted in such a way that the amplitude of the loudest sound on the tape was approximately 80dB SPL. The stimuli with the best /d/ and the best /t/ (i.e the A=-6dB/V=+6dB and A=+6/V=-6 amplitude patterns respectively) were used in the anchoring procedure for all the levels. At the sentence level the anchoring was conducted with the original seven basic carrier sentences. For the anchoring at the auditory level the subjects were told to treat one sound (the best/da/ analogue) as sound one and the other (the best /ta/ analogue) as sound two. At the phonetic level, the subjects were told that the sounds were whistled imitations of the /da/ and /ta/ speech sounds. There was a time lapse of at least two weeks between the tests of different levels.

At the four lower levels the task of the subjects was to identify every stimulus (sound one v. sound two, whistled /da/ v. whistled /ta/, /da/ v.

/ta/ speech sounds, and "dye" v. "tie" respectively) and tick their decisions on the sheets provided. At the sentence level the task of the subjects was to write down the whole sentence as they heard it.

METHOD OF ANALYSIS

The /d/ (sound one) responses were counted at every level. The initial data consisted of ten /d/ responses (at the ten VOT steps) for each of the nine continua at every level. Since the role of the +6dB A and V amplitudes as traded cues along the five linguistic levels was more important than the role of the VOT duration, the data were reorganised for every subject in two frameworks i.e. the framework of the role of A and V amplitude levels and the framework of categoricality distance. The following formula was used to reorganise the data for the first framework:

$$\frac{\Sigma \text{/d/ (or sound one) at each continuum}}{\text{number of trials for each stimulus (4)}}$$

If the cue-trading relation was operating as could be expected from the data of previous works (e.g. Repp 1979, Pisoni 1977, Miller et al. 1976 etc.), the continua with the A=-6dB/V=+6dB should attract more /d/ (sound one) responses and those with A=+6dB/V=-6dB should attract less /d/ responses. For the second framework, the framework of categoricality distance, the following formula of data reorganisation was followed.

$$\frac{4 - \Sigma \text{/d/ of the first five steps} + \Sigma \text{/d/ of the last five steps}}{\text{total number of VOT steps (10)}}$$

If the responses were strictly categorical, the number of /d/ (sound one) responses for each of the first five steps would be 100% i.e. 4 and that for each of the last five steps would be 0. If the cue-trading is operating, the responses can be expected to be less categorical when the V amplitude decreases and the A amplitude increases. The strength of the first framework of data organisation lies in the comparison of the roles of A and V amplitude levels. The strength of the latter lies in the interlinguistic level comparison of the roles of A and V amplitudes. The reason for this is that the variations in the former were not tied to any fixed value common to all the levels whereas those in the latter were tied to the idealised categoricality of responses.

The analysis of variance with planned contrasts was conducted. The contrasts was planned to ascertain the interlevel comparisons (a. lower two levels v. upper three, b. auditory v. phonetic levels, c. syllable level v. two upper levels and d. word v. sentence levels).

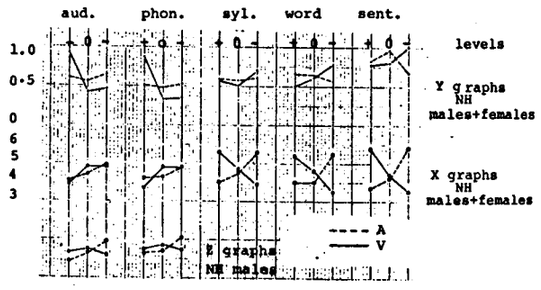

Figure 1 NH Pooled means

SUMMARY RESULTS AND DISCUSSION

In the first analysis the normal hearing (NH) and hearing impaired (HI) group difference was significant at alpha level=.025 (1,33 F.025=5.52 8.57). In the second analysis the NH and HI group difference was significant at alpha level =0.5 (1,33 F.05=4.14 4.88).
In the first analysis, over the grand total of 37 subjects, none of the four contrasts ("a" to "d") was significant. However, interlevel contrast was not the strength of the first analysis. In the second analysis over the grand total of 37 subjects, contrast "b" was not significant but "a", "c" and "d" were (all at alpha level=.025).
This implies that there was no difference in cue-trading relation between auditory and phonetic levels while the cue-trading relations across the upper three levels were different. The pooled means of the A and V amplitude effects for the NH group across the five levels in both the analyses are summarised in figure 1. In the Y graphs of the second analysis, the higher the means move away from the zero line the less categorical are the identifications. In addition, in the Y graphs of the second analysis, the steeper the unbroken lines moves diagonally up from left to right, the more effective is the cue-trading along the V factor and the reverse (left to right downward) is the case with the broken line for the A factor. In the X graphs of the first analysis, the further the means move away from the bar of value 5 (up or down) the stronger is the cue-trading while the steepness of the unbroken and broken lines represents the effect of V and A respectively. In both the analyses, from the syllable level onward, the higher the level, the more linear is the V factor and less linear and more quadratic is the A factor. This suggests that the Va factor exerted a stronger influence than the A factor at the higher linguistic levels. In the overall analysis it seems that there was no cue-trading relationship at the two lowest levels. However, there was some evidence of cue-trading at the two lowest levels in the HN male group though the extent of such lower level non-linguistic cue-trading is rather insignificant compared with that at the linguistic levels. See Z graphs in figure 1.

CONCLUSION

The results of the experiment indicate that from

the syllable level onward, the higher the linguistic level the stronger the cue-trading and the less categorical were the identifications for both the NH and HI listeners. There was a certain degree of cue-trading at the auditory and phonetic levels though it is not as strong as that at the linguistic levels. The speech mode of perception is special and speech specific in the sense that cue-trading in the speech mode is significantly stronger than that in the non-speech mode, but not in the sense that cue-trading does not exist in the non-speech mode of processing. *

REFERENCES

[1] BAILEY, J.P., SUMMERFIELD,Q.& DORMAN,M. (1977) "On the identification of sinewave analogues of certain speech sounds", Haskins Lab. Status Reports on Speech Research, SR-51/52, 1-25.

[2] BEST, C.T., MORRONGIELLO,B.,& ROBSON,R.(1981) "Perceptual equivalence of acoustic cues in speech and non-speech perception", Percept. & Psycho., 29, 191-221.

[3] CLARK, J.E. (1075) "A 12 parameter serial formant speech synthesiser" , Working Papers of the Speech and Lang. Res. Centre, Macquarie University, January 51-71.

[4] CLARK,J.E., (1976) "Specifications for a 12 parameter formant speech synthesiser", Occasional Papers of the Speech and Lang. Res. Centre, Macquarie University, April

[5] CLARK, J.E. (1979)"Synthesis-by-rule system for Australian English speech", (A doctoral thesis, Macquarie University).

[6] CLARK, J.E,(1981) " A low-level speech-synthesis-by-rule system", Jour. of Phon., 9, 541-497.

[7] FANT, C.G.M., (1973) "Speech sounds and features", (MIT).

[8] FITCH,H.L., HALWES,T., ERICKSON,D.M.& LIBERMAN,A.M. (1980) Perceptual equivalence of two acoustic cues for stop consonant manner", Percept. & Psycho., 27, 343-350.

[9] MILLER, J.D., WIER,C.C., PASTORE,R., KELLY,W. J. & DOOLING,R.J. (1976) "Discrimination and labelling of noise-buzz sequences with varying noise lead times: An example of catergorical perception", Jour. Acoust. Soc. Amer. 60, 410-417.

[10] PISONI ,D.B. (1977) "Identificaiton and discrimination of the relative onset time on two component tones: Implications for the perception of voicing in stops", Jour Acoust. Soc. Amer.,61, 1352-1361.

[11] REPP, B.H. (1979) "Relative amplitude of aspiration noise as voicing cue for syllable initial stop consonants", Lang. & Speech 22,137-189.

[12] REPP, B.H. (1981a) "Auditory and phonetic trading relations between acoustic cues in speech perception: Preliminary results", Haskins Lab. Status Reports on Speech Research, SR-67/68, 165-189.

[13] REPP, B.H. (1981b)"Phonetic trading relations and context effects: New experimental evidence for a speech model pf perception", Haskins Lab. Status Reports on Speech Research, SR-67/68, 1-40

* This research was conducted under the supervision of Professor John Clark, School of English and Linguistics, Macquarie University, Sydney.

The preliminary version of this paper was circulated at the first Australian Conference of Speech Science held at the Australian National University, Canberra, November 1986.