Proceedings of The Institute of Acoustics

WHERE DO FEATURES INTERACT?

G.A. Foster

Lab. of Experimental Psychology, University of Sussex.

INTRODUCTION

The phenomenon of categorical perception has been interpreted by some authors as evidence that phoneme recognition is mediated by discrete decisions about the values of phonetic features present in the speech signal. There are others however, who hold that the categorical results obtained with speech continua have a psychoacoustic rather than a phonetic basis. The phonetic feature of voicing is cued by the relative timing of two events: the release burst and the onset of periodic excitation, (commonly referred to as VOT). The psychoacoustic view seems to be supported by the finding that nonspeech analogues of VOT continua yield categorical perception effects when Ss are asked to make a "simultaneous / successive" judgement (e.g. Miller et al., 1976; Pisoni, 1977). Pisoni has argued that the voiced and voiceless classes of stops fall on different sides of the simultaneity-successivity threshold, and that this psychoacoustic boundary underlies categorical perception of VOT continua. If this is indeed the case, then there is no need to postulate speech-specific (ie. phonetic) memory codes in order to explain the categorical phenomenon.

In their strongest forms, both the phonetic and the psychoacoustic accounts of categorical perception are undermined by the finding that features interact. That is, the acoustic correlates of any one feature may vary with the value of other features present in the signal. For example, the voicing boundary typically shifts to longer values of VOT as place of articulation moves back in the mouth (eg. Lisker and Abramson, 1970). This poses a problem for the phonetic hypothesis because the justification for a phonetic level of processing rests critically on the independent evaluation of features during auditory analysis. Clearly, if phonetic features are perceived categorically, then feature interactions must arise at a pre-phonetic (ie. auditory) level of analysis. The strong version of the psychoacoustic hypothesis is also difficult to reconcile with feature interactions, but for a different reason. This account requires that there be an auditory basis for shifts in the simultaneity-successivity threshold with place of articulation.

Two lines of evidence have been cited in support of an auditory explanation of place-voicing interactions. Firstly, there is evidence from studies using non-human subjects who obviously cannot be employing phonetic codes. Chinchillas have been found to yield similar identification functions to human subjects on VOT continua, and moreover, comparable place-contingent boundary shifts (Kuhl and Miller; 1977). Secondly, VOT boundary locations have been found to shift as a function of spectral characteristics, primarily in the region of F1, which typically vary with place (eg. Summerfield and Haggard 1977). However, the extent to which these findings can be said to support the psychoacoustic explanation of categorical perception has been challenged. It has been argued that although the data obtained with chinchillas certainly mirrors the behaviour of English subjects, it is very difficult to understand

in view of cross-language differences in the location of VOT boundaries (eg. Lisker and Abramson; 1970). Secondly, the spectral explanation of place-dependent shifts in boundary values of VOT is difficult to equate with evidence that French subjects are not sensitive to this dimension (Simon and Fourcin; 1978). Furthermore, Summerfield (1982) has shown that spectral manipulations analogous to variations in the first formant, (which shift the voicing boundary on a VOT continuum), do not cause systematic shifts in the simultaneity-successivity boundary on either a tone- or a noise-onset-time continuum.

In view of these conflicting claims, it has not been possible to attribute feature interactions either to the auditory, or to some phonetic level of processing. This has largely been due to the problem of separating these two stages experimentally. This has recently become possible using an audiovisual illusion which was first discovered by McGurk and MacDonald (1976). Here, dubbing particular speech sounds produced at one place of articulation over inappropriate lipmovements can yield percepts at a different place of articulation. For example, a bilabial stop consonant dubbed over nonlabial lipmovements will generally be perceived as an alveolar stop having the same voicing value as the auditory input. Thus, /b/ dubbed in this way will normally be perceived as /d/, while /p/ will yield /t/ percepts.

In this study, bilabial and alveolar voicing continua were synthesized. These were expected to yield voicing boundaries at different values of VOT (ie. the typical place-voice interaction). The bilabial series was dubbed over nonlabial sets of lipmovements in order to yield alveolar percepts. This audiovisual continuum could thus be described as having the place feature bilabial at the auditory level of processing, but the feature alveolar at the phonetic level. If the interaction between place of articulation and VOT boundary location arises exclusively at the auditory level of processing, then the voicing boundary obtained on the dubbed series would be expected to fall close to that obtained on the auditory bilabial continuum. In contrast, if feature interactions arise at a phonetic level, one would predict an audiovisual voicing boundary close to that obtained with the auditory alveolar series.

METHOD

A. Subjects. Eighteen graduate and undergraduate students from the University of Sussex were employed as subjects in this experiment. They were all native speakers of British English, and none of them had any speech, hearing, or visual disorder or impairment. They were each paid £4.50 for their participation.

B. Stimuli and design. Two types of stimuli were employed: auditory and dubbed items. The auditory stimuli were synthetic /ba-pa/ and /da-ta/ VOT continua, each varying in VOT in eleven equal steps. In order to avoid stimulus set effects the auditory stimuli were embedded within a larger set which included an 11-member /ga-ka/ series. In the dubbed condition, each member of the /ba-pa/ series was dubbed over lipmovements appropriate for a nonlabial stop. The dubbed stimuli were also presented as part of a larger stimulus pool. In this condition, each member of the 3 auditory continua was paired with lipmovements for /ba/, for /da/, and for /ga/, making 9 audio-visual voicing continua, (99 dubbed items in all). Three video-tapes were prepared, each comprising four instances of each of the 99 dubbed items, in three different random orders. These were each preceded by a practice

tape. The practice stimuli were the six endpoint tokens from the three VOT continua each dubbed over the three sets of lipmovements. The practice tapes consisted of a random sequence of two tokens of each of these 18 audio-visual pairs.

For each subject, the three sets of practise + experimental tape were used in separate identification sessions on three consecutive days. In the first session, Ss were presented with the auditory track of one of the sets of practise + experi mental tape. The practise session therefore consisted of six occurrences of each of the six continuum endpoint tokens, whilst the experimental session comprised a random sequence of twelve occurrences of each of the three eleven member voicing continua. The second and third sessions involved audio-visual presentation of the remaining two sets of tapes. Excluding practice then, the subjects were presented with eight instances of each member of the nine audio-visual voicing continua. The order of tape presentation was counterbalanced across subjects, (ie. each of the tape sets was assigned to the auditory condition for 6 of the 18 Ss).

C. Stimulus preparation. (i) Auditory components: The three eleven-member VOT continua were produced on a Klatt parallel/cascade-resonance synthesizer. Synthesis parameters were modelled on the author's (female) speech. Since the majority of speech synthesis work has been modelled on male rather than female speech, the synthesis parameters used in this experiment are described below in some detail. Some attempt has been made to differentiate between the parameter settings which were important for female speech synthesis, and for the quality of synthetic speech generally. The formant frequencies, the overall amplitude contour, and various durational measures (averaged over twelve natural utterances of each of the six stops) provided good first approximations of the values which were eventually chosen for the six continuum endpoint tokens. Where the appropriate values could not be obtained using LPC analysis, it was necessary to resort to a mixture of informed guesswork and trial and error methods. This was especially the case for F0, and for individual formant bandwidth and amplitude values. Speech-like quality was considerably enhanced by using (i) non linear formant transitions and F0 contours (ii) gradual onsets and offsets of the various energy sources, and (iii) bursts spliced from natural speech stimuli.

Each of the synthetic syllables was 440 msec long, and had a 10 msec natural burst spliced onto it. The vowel steady-state /a/ was characterized by formant frequencies set at 960, 1300, 3000, and 4100 Hz for F1, F2, F3 and F4 respectively. F1 onset frequency was 275 Hz for all three continua, and F4 was flat. The three continua varied in F2 and F3 onset frequencies which were set respectively at 1170 and 2680 Hz for the bilabial continuum, at 1900 and 3350 Hz for the alveolars, and at 1900 and 2000 Hz for the velar continuum. The formant transitions lasted for 60 msec, but their effective durations were probably shorter than this since the transitions were nonlinear, with most of the frequency shift occurring during the initial part of the transition period. 30% of the total transition shift was accomplished in the first 5 msec, 50% 10 msec into the transition period, and 75%, 80% and 95% at points 20, 30, and 45 msec from transition onset time respectively.

The amplitudes and bandwidths of the individual formants were crucial to stimulus quality, rather than to phonetic identity, and were constant both within and across continua. Their values were rather different from those typically used in synthetic stimuli modelled on male speech. The amplitudes of F1 and F2 were set

at 50dB, F3 at 40dB and F4 at 30dB. The bandwidth were 50 Hz for F1, 150 Hz for F2 and 250 Hz for F3 and F4. The fundamental frequency (FO) contour was also an important determinant of general speech-like quality, whilst the frequency range within which it was set was obviously important for achieving a female quality. For Stimulus 1 in each series (VOT = 10 msec), FO was generally characterized by an initial rise from 187 Hz to a peak value of 227 Hz, followed by a (more gradual) fall to 170 Hz. Precise details of the contour are given in Table 1. For the other members of each continuum, FO onset coincided with voice onset, such that FO onset frequency was higher for tokens with longer values of VOT.

For each of the three eleven member continua, VOT was varied in 5 msec steps from 10 to 60 msec. However, the effective values of VOT may have been longer than these nominal values, since the voicing source (AV) was turned on gradually. At the nominal voicing onset time, AV was set to 30 dB, rising to 39 dB over the next 5 msec, and further to 46 dB after 10 msec. Regardless of voice onset time, AV then rose linearly to 51 dB at 90 msec from burst onset, and then to 52 dB at 100 msec, at which level it remained for 80 msec. The amplitude of the voicing source then fell linearly to 50 dB over the next 70 msec, and thence to 40 dB over a further 100 msec. At this point, (330 msec from burst onset) AV was turned off, and the aspiration source (AH) was set at 37 dB. The amplitude of Ah then fell linearly to 27 dB over the last 100 msec of the syllable. This "breathy" sounding termination of the syllables produced a slight improvement in their perceived naturalness in pilot work. Except for stimulus 1 in each series, which had a VOT value corresponding to burst duration, the period between burst offset and voice onset was filled with aspiration noise. The aspiration source (AH) was set at 60 dB at onset, and fell linearly to 55 dB at voicing onset when it was switched off.

The three natural bursts (one for each place of articulation), were spliced from the syllables on which the synthesis parameters were based. Voiceless unaspirated bursts were selected which could be cut as close as possible to 10 msec without introducing any clicks in the waveform. These were spliced onto the synthetic syllables, with the burst amplitude being set separately for each continuum but remaining constant within continua. Relative amplitude levels were selected such that: (i) there were no obvious discontinuities in the overall amplitude contours at burst offset and either voice or aspiration onset; and (ii) the burst did not cause any of the tokens at the voiced end of the continua to be perceived as voiceless.

(ii) Visual components: The lipmovements were recorded in a professional television studio on a Sony U-matic colour video recorder. In each articulatory sequence the picture showed a bowed head which was raised to present a full-face view with the lips held slightly parted just prior to syllable production. The face almost filled the frame during articulation. Lighting was provided by two 1 Kwatt bulbs placed to minimize facial shadow. Lip and jaw movements were clearly visible, but tongue movements were difficult to see, the oral cavity being in shadow. Three sets of lipmovements appropriate for bilabial, alveolar and velar stops respectively were selected from a pool of 72, (12 tokens of each of the syllables /ba,da,ga,pa,ta,ka/). These had been recorded at the same time as the auditory syllables upon which the synthesis parameters were based and were therefore of similar durations to the auditory stimuli described above. The three sets were chosen such that the bilabial/nonlabial distinction was very clear, whilst the alveolar and velar tokens could not be distinguished (based

on a pilot study using a same/different task).

(iii) Dubbing procedure: A computerized dubbing procedure was used to ensure that burst onset of each dubbed syllable coincided with visual consonant release to within 0.1 msec. Audio-visual temporal alignment has been shown to affect place of articulation perception, (Foster; 1983). The three sets of lipmovements were spliced from the master film, and a high frequency tone was dubbed onto their auditory tracks just prior to (auditory) burst onset. Visual release was operationally defined as that point in the visual signal which corresponded to auditory burst onset on the original film. Burst onset (and thus visual release) was located relative to tone onset using a waveform editing facility on a PDP 12 computer. The three visual stimuli were then copied repeatedly at 5 sec onset intervals to form the visual tracts of the three sets of practice + experimental tapes. The auditory track of the master film (tone + original auditory signal) was simultaneously copied to audio track 1 of these new films. The tones were later used to trigger the dubbing procedure. The 33 auditory stimulus components, (natural burst + synthetic syllable), were transferred from the VAX to the PDP 12 computer. The wave files containing these stimuli were edited such that burst onset coincided with the beginning of each file. Each tone on track 1 triggered the output of one of the stored auditory components onto track 2 of the films, at a delay determined by the tone-burst interval on track 1.

D. Procedure. Each S was required to perform a closed-set identification task in first the auditory and then the dubbed condition. In the auditory condition, the response set consisted of the six stop consonants /b,d,g,p,t,k/. In the dubbed condition, in addition to these six stops, the response set also included any pair of the V+ or the V- stops. In both conditions, the Ss were instructed to report what they heard. In the dubbed condition, it was emphasized that they should nevertheless pay close attention to the lipmovements on the film. The Ss typed their responses into a terminal connected to a VAX computer, and their results were sorted and analysed on-line.

RESULTS

The results were analyzed in three stages. Firstly, the responses obtained on the auditory-alone continua were tested to ensure that (i) they had been perceived correctly according to place of articulation, and (ii) the voicing boundaries occurred at longer values of VOT for the alveolar than the bilabial continuum. In the second stage, response distributions obtained with the audio-visual stimuli were examined to see whether they had yielded above chance alveolar percepts. Finally, for the Ss who met the criteria outlined above, the /da-ta/ boundaries obtained with the audio-visual continuum were compared with those found on each of the auditory-alone continua.

Auditory conditions. There were very few place of articulation errors on either of the auditory continua. Averaged across both Ss and continuum points, the bilabial stimuli were correctly classified 99.1% of the time, and the alveolars 93.1% of the time. The data was normalized where place errors did occur: the V+ and V- responses were calculated as proportions of the total number of place-correct responses for each subject at each value of VOT on the two continua. Voicing boundaries were obtained from these data using the method of probits (Finney; 1971). The mean boundaries fell at 26.9 and 30.8 msec of VOT for the bilabial and alveolar continua respectively. These values were significantly different on a one-tailed analysis of variance test ($p < 0.01$). Examined

separately, 16 of the 18 Ss were found to conform to this pattern. The other two were dropped from the data pool.

Dubbed condition. The auditory-bilabial/visual-nonlabial stimuli were perceived as alveolar stops 78.8% of the time, averaged across the remaining 16 Ss and the 11 continuum points. However, on separate analyses, it was found that 4 of the Ss failed to achieve above chance (ie. 33.3%) alveolar percepts at every value of VOT. These Ss responses were therefore dropped from further analysis. The mean percent alveolar responses for the remaining 12 Ss was 90.5, ranging from 86% to 94% at different values of VOT. For each of these Ss, the proportions of "d" and "t" responses were calculated as a function of the total number of alveolar responses at each continuum point. Auditory-bilabial/phonetic alveolar voicing boundaries were obtained from these data.

Auditory versus dubbed conditions. The auditory-bilabial/phonetic-alveolar boundaries were compared with the auditory-bilabial, and the auditory-alveolar boundaries. The individual and mean boundaries obtained for the three continua are given in Table 2. Only the 12 Ss who had satisfied the criteria described above were included in these comparisons. The boundary VOT value for the auditory-alveolar continuum was significantly longer than that for the auditory-bilabial series ($p < 0.01$). The mean boundary obtained on the dubbed continuum was significantly longer than that found with the auditory-bilabial series ($p = 0.01$), but was not significantly different from that obtained with the auditory-alveolar series. This pattern of results is seen in Fig. 1 which shows the pooled normalized identification functions for the three continua.

DISCUSSION

The results support the hypothesis that interactions between place of articulation and VOT boundary location can arise at the phonetic level of processing. Feature interactions cannot therefore be explained solely in terms of general properties of the auditory perceptual system. The remarkable coincidence between the performances of Chinchillas and English-speaking human Ss on VOT continua therefore requires some other explanation. In the light of the data reported here, they can no longer be interpreted as evidence for the purely psychoacoustic account of the phenomenon of categorical perception. An alternative explanation could be that speech sound systems have evolved such that phonetically relevant contrasts tend to span regions of heightened auditory sensitivity (eg. the simultaneity-successivity threshold). This account has the advantage of providing scope for both (i) cross-language differences in the location of phonetic boundaries, and (ii) the phonetic feature interactions reported here.

Feature interactions at the phonetic level also throw doubt on the traditional view that categorical perception effects, and the related phenomenon of adaptation, reflect the operation of discrete rather than continuous decision mechanisms. These interactions imply that the phonetic stage has access to probabilistic information about features. The same conclusion has recently been drawn on the basis of evidence obtained from studies using two very different approaches. Firstly, it has been found that listeners can be trained to place their boundaries at arbitrary locations on speech continua. A new voicing contrast learnt on a continuum produced at one place of articulation will generalize to continua with other values of the place feature (McClaskey et al.; 1983). This suggests that Ss have access to within-category information.

WHERE DO FEATURES INTERACT?

A second source of evidence comes from Miller's experiments using the dichotic adaptation paradigm. Adaptation effects have generally been assumed to reflect the fatigue of feature detecting mechanisms. When Ss are asked to report one of a pair of dichotically presented stimuli, there will generally be some intrusions from the non-target ear. Adaptations with either voiced or voiceless adaptors increases the number of intrusions when the unadapted value of the voicing feature occurs in the non-target ear (Miller; 1975). In a subsequent experiment (Miller; 1977) voiceless adaptors were used which varied in VOT. The post-adaptation shift towards voiced responses increased with increasing values of VOT. The output of the feature evaluation mechanism could not therefore have been a simple binary decision. Traditional categorical perception effects do not reflect categorical mechanisms in the speech perception system.

References

1. Finney, D.J. Probit Analysis, Cambridge University Press: Cambridge, (1971)
2. Foster, G.A. 'The intergration of audio-visual speech stimuli as a function of temporal desynchronisation' , Proc.I.O.A. Autumn Conference 1983, Bournemouth, ppH3.1-H3.5
3. Kuhl, P.K. & Miller, J.D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. J.A.S.A., Vol. 63, no. 3, 905-917, 1978.
4. Lisker, L., & Abramson, A.S. The voicing dimension: Some experiments in comparative phonetics', Proc. 6th Int. Congr. Phonetic Sciences, 563-567,1967.
5. Mc Claskey, C.L., Pisoni, D.B. & Carrell, T.D. 'Transfer of training of a linguistic contrast in voicing", Perception & Psychophysics, 34(4), 323-330, 1983.
6. McGurk, H. & MacDonald, J. 'Hearing lips and seeing voices', Nature, 264, 746-748, 1976.
7. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly. W.J. & Dooling, R.J. 'Discrimination and labeling of noise-buzz sequences with varying noise lead time: An example of categorical perception', J.A.S.A., 60, 410-417, 1976.
8. Miller, J.L. 'Properties of feature detectors for speech: evidence for the effects of selective adaptation on dichotic listening', Percept. Psychophysics, 18, 389-397, 1975.
9. Miller, J.L. 'Properties of feature detectors for VOT: The voiceless channel of analysis', J.A.S.A., 62, 641-648, 1977.
10. Pisoni, D.B. 'Identification and discrimination of the relative onset time of two-component tones: Implications of voicing perception in stops, J.A.S.A. 61, 1352-1361, 1977.
11. Simon, C., & Forcin, A.J. 'Cross-language study of speech-pattern learning', J.A.S.A., 63, 925-935, 1978.
12. Summerfield, A.Q. & Haggard, M.P. 'On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants', J.A.S.A., 62, 435-448, 1977.

Table 1. Fo as a function of time, with linear interpolation between stated values (T = 0 at burst offset).

Time (msec)	0	10	25	45	60	70	85	100	120	180	200	220	240	340
Fo (Hz)	187	195	207	219	225	227	227	224	221	191	185	180	176	170

Table 2. Individual and mean boundaries obtained with auditory and dubbed continua.

Subject	Auditory Bilabial	Auditory Alveolar	Dubbed
1	23.23	24.57	24.61
2	28.85	34.90	33.66
3	26.57	36.07	36.60
4	29.55	32.96	29.95
5	28.67	29.16	27.86
6	29.88	38.26	44.19
7	26.16	29.48	28.33
8	32.86	40.32	43.33
9	26.70	29.85	27.43
10	27.79	34.35	29.06
11	24.65	31.08	27.81
12	23.27	24.85	26.70
Mean:	27.3	32.2	31.6

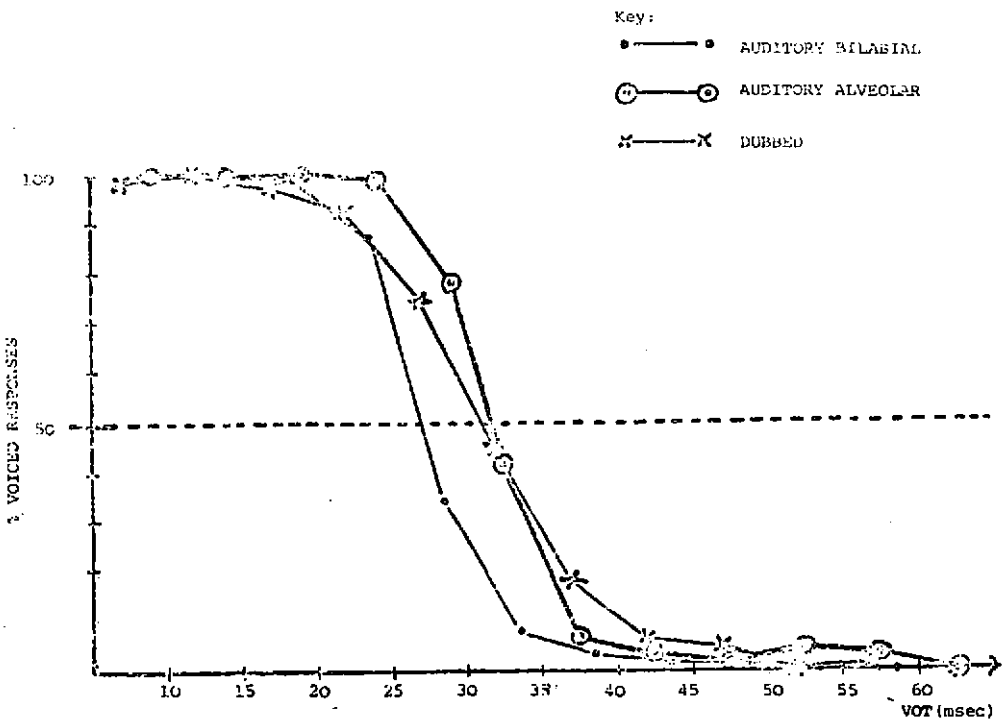

FIGURE 1: Pooled normalized identification functions obtained with auditory and dubbed voicing continua (identification functions fitted to normal ogives).