PERCEPTUAL MECHANISMS AT THE FIRST LEVEL OF SPEECH PROCESSING

ADRIAN FOURCIN

1. DESCRIPTION

Three sets of experiments will be briefly described. The first deals with a particular aspect of the perception of pitch and is intended to make a contribution to the understanding of the low level processing of intonation in speech. The second set of experiments deals with a feature of the perception of tone in Cantonese and is particularly concerned with the way in which the phonetic value of a tone is determined by its relation to a reference pattern. The last experiments involved the use of single formant whispered stimuli which are now interpreted in relation to a reference pattern which is a function of an inferred vocal tract. The results of the three lots of experiments are interpreted as showing that although reference to the mechanism of speech production may be of great value, very powerful perceptual mechanisms are necessarily available for the perceptual processing of speech in ways which are not related to its formation in the speaker's vocal tract, and, indeed, that our ability to produce speech may be in large measure dependent on our auditory ability to perceive its patterns and partly controlled by a process of auditory pattern feedback.

The voiced sounds of speech are always phonetically associated with a quasi-periodic larynx excitation of the vocal tract. This periodicity gives these sounds a regularity in frequency, as well as in time, and there is a wealth of experimental evidence which shows how a sensation of pitch may be associated with stimuli designed to be either periodic in time or periodic in frequency. The presence of only a part of a regular harmonic spectrum in the auditory stimuli used in an experiment can lead to a strong sensation of pitch which is highly correlated with its putative fundamental frequency. Similarly, the periodicity of a stimulus has a high inverse correlation with its pitch, but as there is always a strong relation between frequency spectrum periodicity and waveform periodicity, it is difficult to determine their relative contributions to the perception of pitch. This difficulty has been partly resolved in experiments with 'residue' pitch stimuli (Schouten, Ritsma and Cardozo, 1962) in which it has been found that in cases in which there is a small discrepancy

between line spectrum and stimulus periodicity the sensation follows periodicity. A similar situation was found to exist with smooth spectrum noise added or subtracted from itself delayed to produce an 'echo' pitch sensation (Fourcin, 1965). Here the pitch was again not an obvious function of spectral peak placing but was directly dependent on the delay, but now very large changes in spectrum could be associated with only second-order changes in pitch.

In both residue and echo pitch, however, there is an uneliminated possibility that a pitch mechanism might operate directly from spectral information by, for example, basing its estimate on the spacing of adjacent spectral peaks rather than on their absolute locations. This problem cannot be resolved with stimuli of normal composition, but there is one class of binaural stimuli in which the temporal organisation does not influence the spectral form, and it has been found possible using these to study a 'central' pitch (Fourcin, 1970) which can only owe its origin to time processing on the part of the nervous system, rather than frequency analysis in the inner ear. The perceptual range of the periodicities of central pitch corresponds to the range of periodicities of speech and to the range of residue pitch. It seems likely that normal place analysis combined with this temporal analysis may jointly provide the perceptual processes with a highly developed method of processing this aspect of speech, Figure 1.

The pitch experiments were made in order to explore some aspects of the way that voice pitch is so readily perceived. For normal speech the changing resonances of the vocal tract and the varying nature of its excitation make the reliable tracking of larynx frequency a difficult matter in the laboratory. At present the best solution to this problem (using Cepstral analysis) is still not completely satisfactory and is barely capable of operating in real time. This is a task which our perceptual processes appear to accomplish with ease for both normal speech and speech which is presented in adverse conditions of noise and distortion.

In order to study some of the ways in which larynx excitation itself is perceived, and once perceived, utilised, we have developed a means (Laryngograph, Fourcin and Abberton) of extracting a correlate of vocal fold movement. The approach that we have adopted was originated by Fabre but, although the basic principle of monitoring the electrical impedance across the speaker's throat by means of superficially applied electrodes is the same, our arrangement differs in several important respects. The electrodes, on either side of the Adam's apple, each have a guard ring at earth potential and operate with screened connections to low impedance circuits which employ an automatic adjustment to compensate for the gross impedance differences which exist between different speakers.

We have made a cine-stroboscopic comparison of the glottal area function with the output of the laryngograph (Fourcin, Roach and Donovan, unpublished) and an example of the form of the relationship is shown in Figure 2. The most important feature of the laryngograph waveform *(Lx)* which is confirmed by this analysis, is that there is a good — although not perfect — correlation between the positive

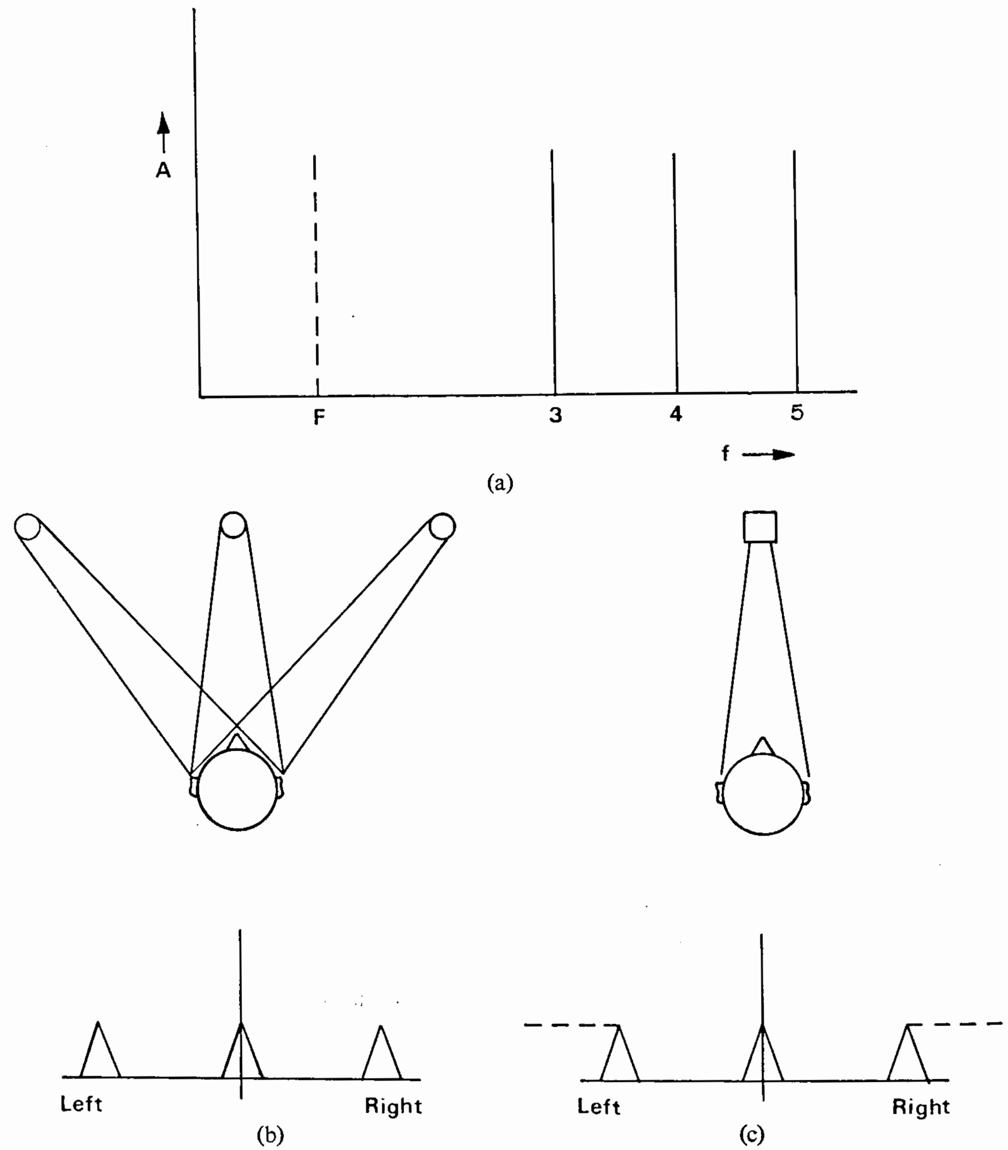

Fig. 1. Pitch Pattern Inference. 1(a) illustrates an effect basic to the residue pitch phenomenon. Three pure tone components, whose frequencies are in the ratios 3 : 4 : 5 and fall in the middle range of audition, can give rise to an auditory pitch which corresponds to the frequency of their highest common factor. This frequency, F in the diagram, need have no physical existence, its presence is effectively inferred by the auditory mechanism. For all periodic sounds, the inference of the fundamental frequency, F, provides a key to the structure of the whole sound. 1(b) illustrates how another key-feature of an acoustic stimulus can provide pitch information. Three regularly spaced independent white noise sources can be perceived in terms of three independent lateralization images by a listener. In 1(c) a periodic stimulus also produces a series of lateralization images. If, for 1(b), the left and right images are produced by delays which are greater than 1 mS, then a pitch sensation can result which is directly related to that of the periodic source in 1(c) when operating to give the interaural delays.

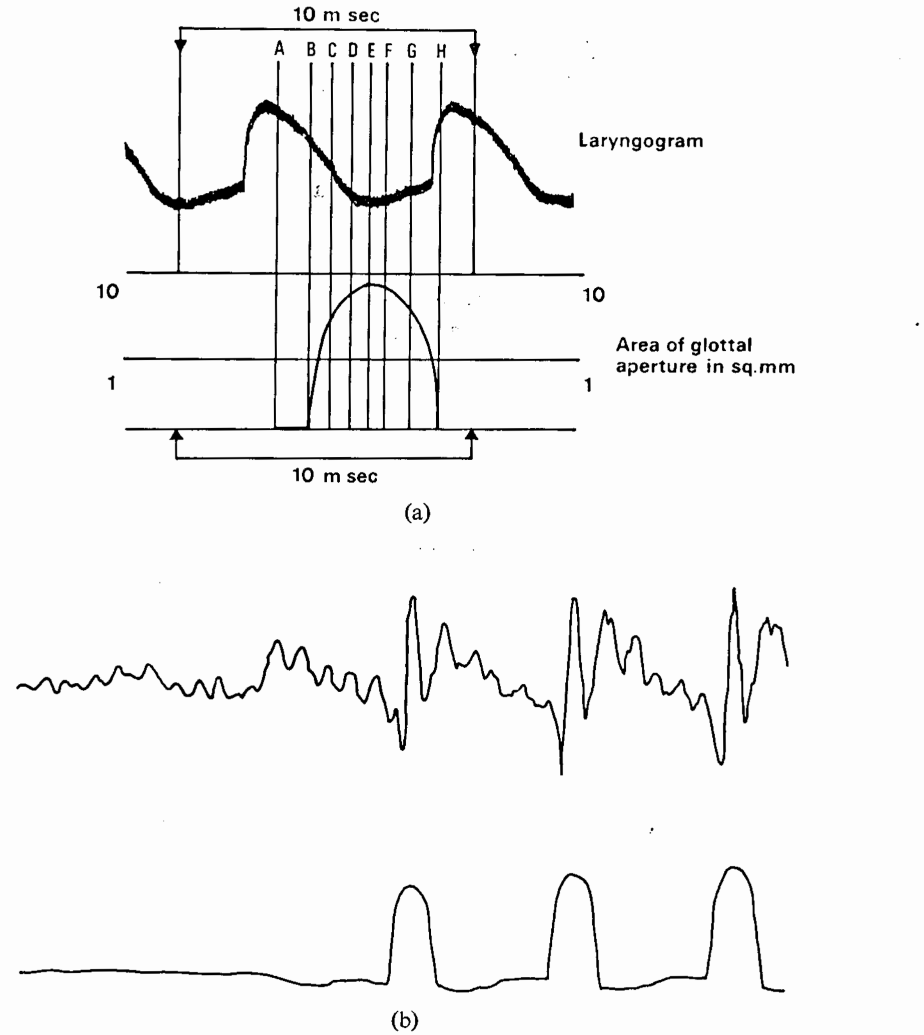

Fig. 2. Laryngograph Waveform Comparisons. 2(a) relates the output of the laryngograph (Lx), above, to the associated glottal area function, below, for one particular voice quality (falsetto) of a male speaker. The laryngograph output is only associated with vocal fold closure. Figure 2(b) shows the speech pressure onset in breathy voice above and the corresponding Lx waveform below. Once vocal fold contact is established the periodicities of the two waveforms are identical.

(low impedance) peaks of its output and the closed phase of vocal fold movement. This appears to be potentially useful in the diagnosis of abnormal vocal fold vibration and provides a basis for the simple determination of its frequency (Fourcin and Abberton). It also provides a basis for a range of perceptual experiments since the laryngograph output can also be used directly as a source of sound. In this way, stimuli can be prepared which, whilst depending only indirectly on the vocal tract, contain very clear fundamental frequency *(Fx)* information, give a good indication of duration, and some indication of the spectral form of the larynx excitation.

In one particular series of experiments (Angela Chan and Fourcin, 1971) stimuli of this type have been used to investigate the response of Cantonese listeners to lexical tones. The tests were made in Hong Kong and involved the use of five main types of stimulus, all based on single syllable utterances: normal speech; the laryngograph output corresponding to the normal speech utterances; synthesized *Lx*, with a constant pulse shape for all speakers but original *Lx* duration and *Fx* contour; synthetic *Lx* with constant duration and original *Fx*; synthetic *Lx* with constant duration and linear (on a linear frequency scale) *Fx* variation. Cantonese has essentially a six tone system, plus three glottalised tones, and all nine tones were used and were spoken by a man and a woman with partially overlapping fundamental frequency ranges. In the main set of experiments, the subjects were required to label the stimuli in terms of a reference set of Chinese characters whose values had been defined at the beginning of the response session by the recorded natural utterances of the two speakers involved. The initial instructions were given in English. In a subsidiary set of experiments, the subjects were once more required to label the stimuli in terms of the same set of ideograms, but the characters were not associated with any speech sounds could serve as a reference set.

In brief, the results of the main experimental sessions, for which the stimuli from the two overlapping sources were presented in random order, showed that although all stimuli could be fairly readily recognised, there appeared to be a definite pattern in the distribution of errors. The large number of subjects employed in the experiments, however, made a detailed analysis of their errors meaningful, in spite of the generally low error probability. With few exceptions it was found that all errors could be interpreted in terms of the effect of the previous context of the misjudged sound and on the employment by the listeners of a decision sequence based on the use of a reference pattern derived from the precursive examples given by the two model speakers. If the stimulus set was changed so that it did not fit in with this pattern, then the subject's responses were modified in a way which depended on the closest interpretation which could be applied, in terms of the reference frame, to the new stimulus.

If for the simplified *Fx*, equal duration, stimuli no precursive speech models were provided, then the responses were completely different. As the experiment proceeded the subjects gradually inferred a reference system and applied it to the stimuli as though they were all derived from the same, unreliable, source.

It does not appear to be necessary, however, to give a complete set of precursors for a reliable pattern to be inferred by the listeners. In another preliminary series of experiments (R. Chan and Fourcin), the same simplified *Fx* stimuli were presented with only one precursor — the rising tone, which covers the whole range — to define the range. This precursor was a normal voiced utterance, and occurred before each stimulus from the corresponding set of syntheses. The same high labelling scores

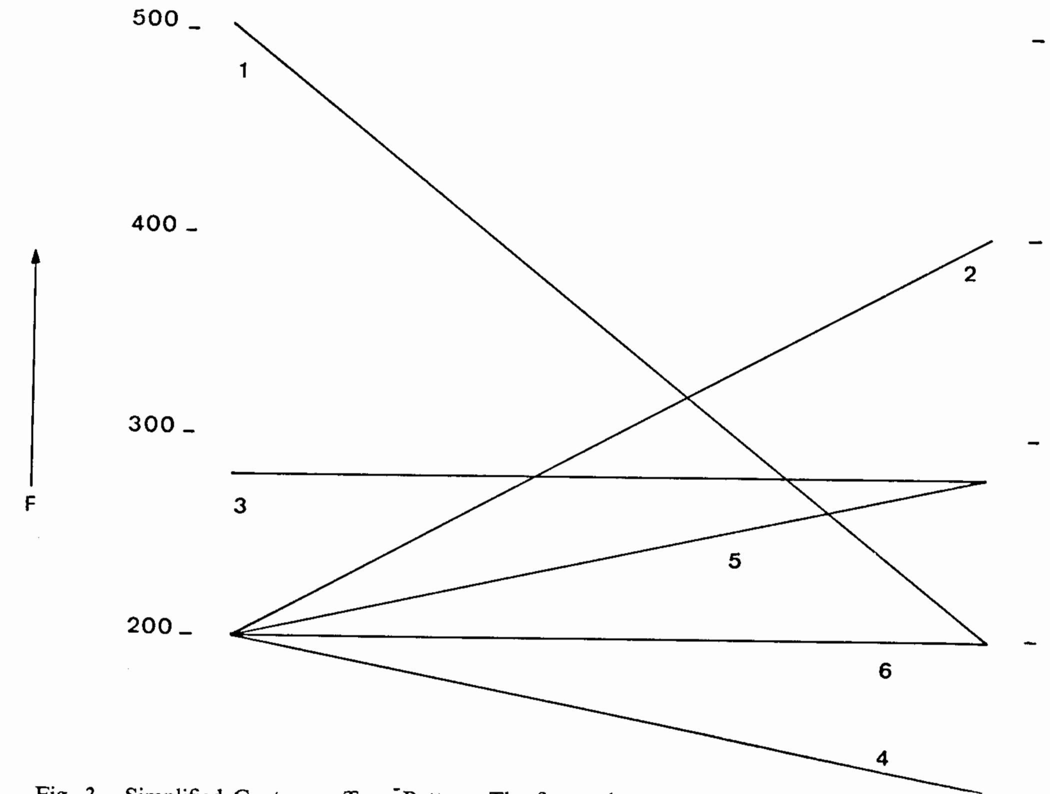

Fig. 3. Simplified Cantonese Tone Pattern. The figure shows a set of tone contours which more than cover the combined speech fundamental frequency range for a normal man and woman combined. The contours were used in an experiment in which three formant generators were set in frequency, as in Figure 1(a), to be at the third, fourth and fifth multiplies of a common frequency. When the generators were excited with random noise, ramp controlled in amplitude for the tone duration, clearly recognizable tones were heard by Cantonese listeners who were able to infer the tone system and give the correct lexical labels to its members. The physical nature of the stimulus was quite incompatible with that for normal speech, only its pitch pattern was similar.

were obtained using these single precursors as for the full sets. In addition, left and right single ear experiments with these precursor-stimulus pairs showed a significant right ear advantage. These results show that these stimuli are being treated as speech sounds, rather than psycho-acoustic stimuli, and that identification depends on their perception relative to a pattern which can be readily displaced and readily inferred.

An outline of the decisions used in referring to this pattern is shown in Figure 4.

When perceptual processing of ultimate phonetic importance is applied to vocal tract — rather than excitation — features of speech, a similar situation can exist. Just as for the tone experiments, it is necessary to evaluate the sound source so that its individual output sequences can be placed within a pattern. The process of tuning in to a new speech source must take place at every possible level of speech processing. The resulting process of adaptation is complicated because not only are there different levels of abstraction at which adjustment must be made but these levels can

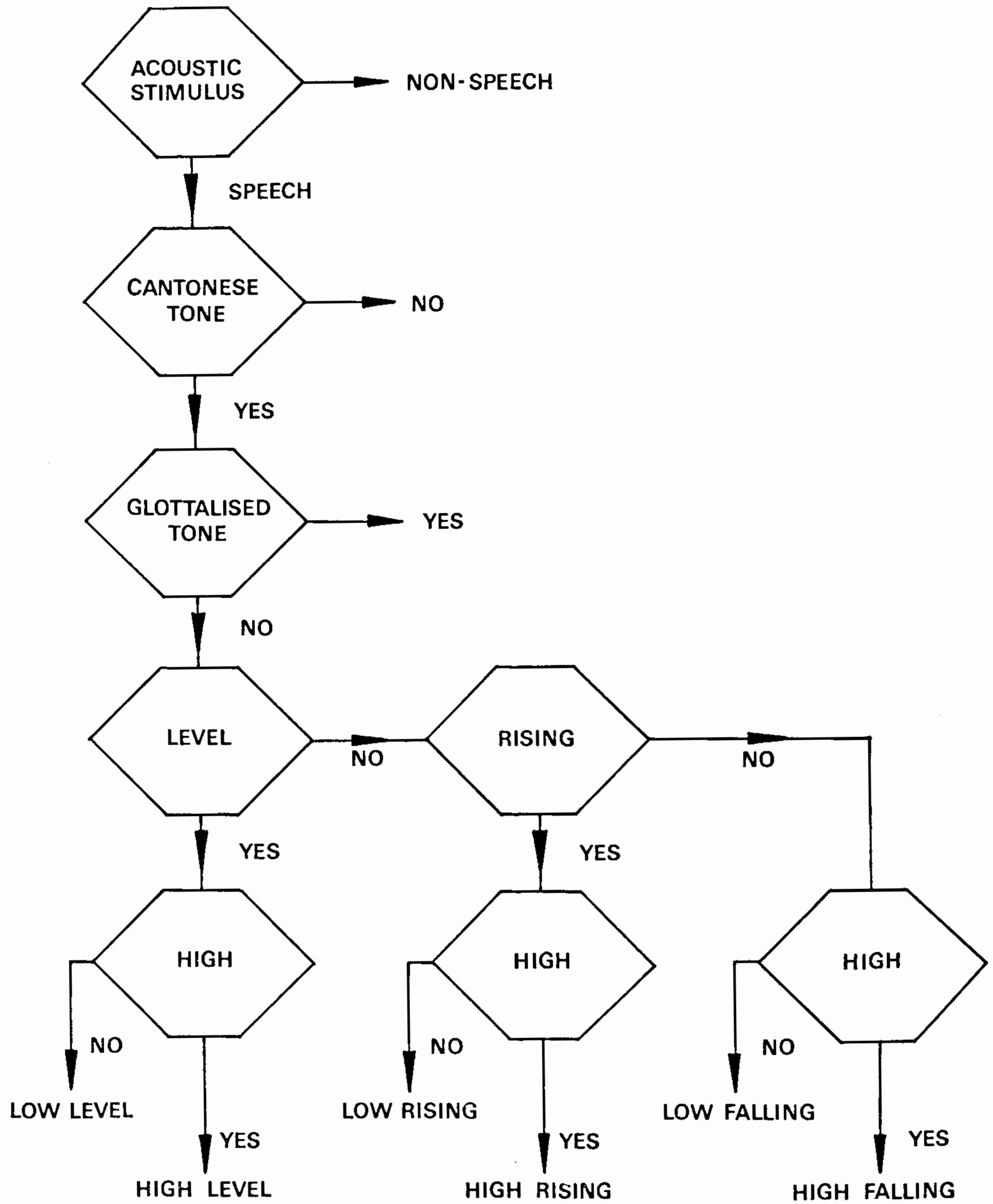

Fig. 4. Possible decision sequence in the perception of Cantonese Tones. The response errors made by Cantonese listeners in Dr. A. Chan's experiments corresponded to the application of the hierarchy of decisions which is shown in the figure. This decision sequence is an important factor in defining the nature of the pattern of tones employed and explains how the stimuli of figure 3 can be processed as speech sounds although their physical form is quite different from that of speech.

interact with each other and may interact between speaker and listener. In order to study only the very first level of processing, at which acoustic information is transformed into phonetic material, it is necessary to eliminate the contributions from the other levels and control the information available for tuning in at this level. The use of recorded material, in a single dialect, with only phonetic significance is familiar. The final control can be achieved by making the stimuli so that they contain only just enough information for a speech pattern to be discernible and by preceding their presentation with a source-identifying precursor of defined form.

Important source information can be derived, in principle at least, from the absolute values of a sound's formant frequencies, their relative spacing and from the spectrum of the larynx excitation. In a particular experiment in which this general approach was adopted (Fourcin, 1968), these source clues were largely eliminated from the stimuli by the use of single formant whispered sounds. These were synthesized in a range of frequency which straddled the combined $F2$ ranges for two speakers, a man and a six-year old child. Both speakers (father and daughter) had the same English dialect. Five synthetic formant frequencies were used and each formant had an initial transition which came from one of eleven different locus frequencies. The locus frequencies were also chosen so that they more than covered the locus frequency range for [b] and [d] in the speech of the particular man and the child. The $F2$ range of frequencies was used because the second formant is the most important and most apparent carrier of speech sound pattern information in whispered speech. It is consequently to be expected that listeners will have a tendency to treat whispered single formant sounds as though they were at an $F2$ frequency.

The precursor, containing only source information, was the single word 'hallo', voiced and spoken with a fall rise. This word was chosen primarily because it is acceptable as a precursor and within small compass gives a good indication of the acoustic possibilities of its source; and because it contains no explicit [b] or [d] locus information.

Forced choice labelling experiments were made, presenting these single formant whispered stimuli in sets of eleven. Each set was preceded by one of the two, man or child, recorded 'hallo' utterances. All the possible initial transitions were used in random order, but only one vowel formant frequency was employed in each set. The listeners' task was merely to write 'b' or 'd' for each of the stimuli. The response for a group of listeners was a labelling function with high scores for [b] at low locus frequencies and high for [d] at high locus frequencies. This function was characterised by interpolation to determine the point along the locus frequency axis at which the 50% level of judgment occurred. From this, one locus frequency value was obtained for each response set.

Each set of responses was associated with one of the two precursors and one of the five formant frequencies: the results obtained from this interpolation of the whole experiment are shown in Figure 5. There is a fairly clear pattern of response.

First, the same stimuli can evoke different responses dependent on the precursor with which they are associated, and on their formant frequency. If the formant frequency of a stimulus is common to the $F2$ ranges employed by both precursive

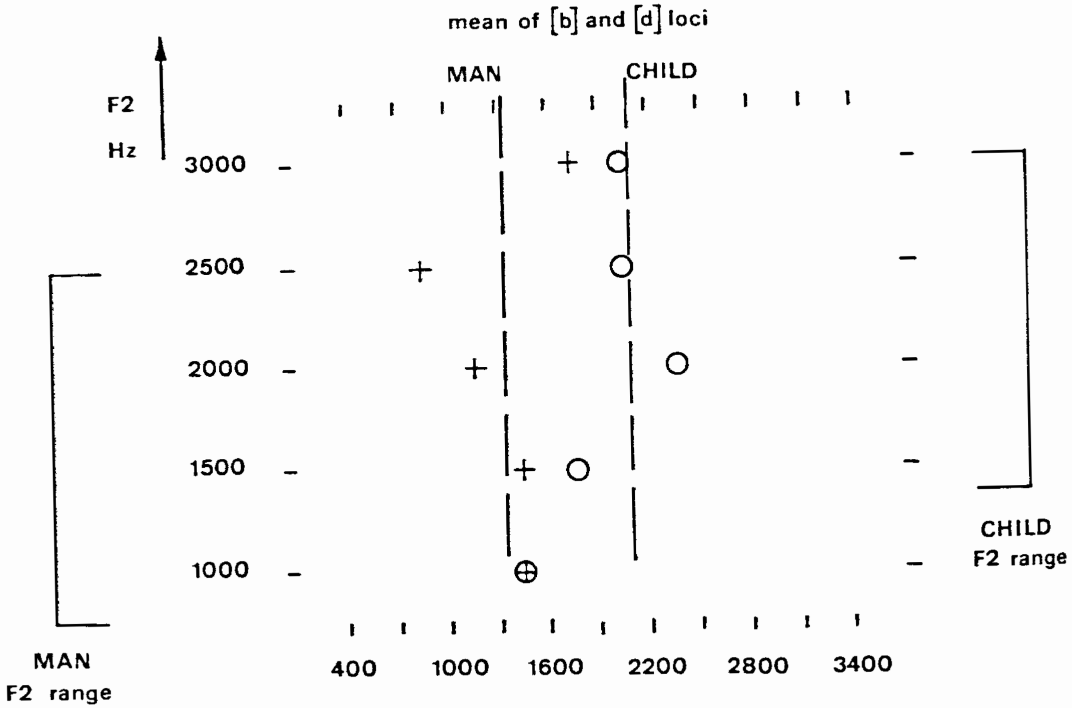

Fig. 5. Vocal Tract Inference. The results of an experiment in which whispered single formant stimuli are capable of different interpretation dependent on the reference speech with which they are associated. The effects observed are not dependent on dialect differences. They seem best explained in terms of the listener's appraisal of the sounds in terms of their acoustic speech patterns.

speakers then a stimulus is very likely to be interpreted in terms of its precursor's vocal tract.

Second, when the formant frequency of a stimulus is not in the range which could possibly be associated with its precursor, then the listeners tend to interpret the stimulus without regard to the precursor at all. The source clue given by the particular value of *F*2 determines the assessment of the transition, in terms of the sound pattern of the other precursive speaker.

Third, the responses to the stimuli made by the listeners correspond to the locus frequencies actually employed by the speakers when producing [b] and [d] in initial position. Measurements were made of these [b] and [d] loci for each speaker for all vowels in their dialect, and the mean of the linear differences is shown for the man and for the child as vertical dashed lines on the figure (Figure 5). There is, in spite of the approximate nature of this comparison, a meaningful correspondence between what the listeners have inferred from the 'hallo' precursors and what the speakers really produce, although the precursive words contain no locus information.

This experiment gave large and highly significant changes in subject response with change of precursor. This was partly because the precursors were rich in source information, and partly because there is a marked difference between the physical characteristics of phonetically identical plosives produced by adults and children. Other consonant features than second formant transition could be expected to show just the same effect to a degree directly related to their variation in speech, and in particular, the same type of effect is to be expected for vowel sounds. Vowels, however, differ physically, between adult and child, less than these plosive consonant transitions.

Similar source inference experiments have been made (Margaret Robertson, 1971) using synthetic whispered single formant vowel stimuli in the *F2* range with and without both natural and synthetic voiced and whispered precursors ('what's this', spoken with a falling intonation). As before, both the precursive speakers and the listeners came from the same dialect area. In brief, her results showed that:

(1) In the absence of precursors, or source type instructions, the stimuli were themselves used to provide source information and the listeners fitted their responses to cover the stimulus range; rather as though a speaker were involved with the low *F2* values of a man and the high *F2* values of a child, with a cross-over region afflicted with some uncertainty.
(2) Good quality voiced precursors had a significant effect on the responses to the stimuli. The listeners tended to label the sounds according to the *F2* values that would have been employed by the particular precursive speaker even when this information was not explicit and only available to them by inference.
(3) When the greatest similarity between precursor and stimulus existed (*F2* only synthesis), then the effect of the precursor was greatest, although here the precursor was less rich in source information than when a normal voiced utterance was employed.

2. DISCUSSION

The three sets of experiments described are only concerned with a very small part of the complex whole of the perceptual processing of speech. They do, however, give a particular view of some of its salient features, which are possibly of greater consequence. The pitch experiments show how the operation of complex neural processing can enable quite different aspects of a stimulus to lead to a common percept. Separate mechanisms of processing the physical features of frequency, period and lateral disposition can each independently lead to a pitch percept which is in good accord with what would have resulted from their joint operation. The detailed operations here are dependent on innate processing, and although this may not be so for the listener's responses in the tone and vocal tract experiments, the same type of behaviour is found, and a correct inference of the whole is made on the basis of the presentation of part of the pattern.

The speech-like stimuli require a further operation, however, since there is not an absolute relation between the parts of the *Fx* and formant patterns in the way that there is for frequency, period and lateralisation. On the basis of detailed knowledge of the rules governing the interrelation between its parts, and on the basis of a hypothesis about at least how one presented part fits in, the overall form of a new speech source has to be inferred. This process of inference has to occur before any speech processing can take place in the experiments which have been described, and it must be the first operation to be performed in any normal speech processing situation, since the perceptual mechanism must be of finite capacity. In abnormal situations the process is of even greater importance. Speech presented at a rate greater than any normal means of production allows, or with its formants beyond the normal possible range, can be understood by normal listeners. And children learning to speak, as well as spastics congenitally unable to produce any of the sounds of speech, can make perceptual distinctions which they are unable to produce only by pattern inference.

These operations of pattern inference must then precede operations of detailed analysis, but since this inference process exists and is evidently of considerable power, it would be wrong to consider that it has no part to play in subsequent processing. The pattern processing necessary for tuning in to a source could be employed to categorise its individual sequences, without any reference to the mechanisms employed by the source in its process of production. Indeed, given the existence of this ability to perceive patterns, our ability to produce speech may be largely mediated by our ability to perceive the patterning of our output acoustic sequences.

This would involve the use of an overall auditory-to-production pattern feedback process, in which the details of the process of production were of minor significance, once the rules governing the form of acceptable outputs had been auditorily established and auditorily learnt. However, once the motor sub-routines had been set up, production on a short-term basis could proceed with little reference to auditory control.

Before this overall long-term auditory control can take place, the patterning which it has to employ must be established. This cannot be achieved without reference to the process of production. The sound sequences of speech are basically determined by the mechanism which produces them. In consequence, from an evolutionary point of view, production must be given precedence, but once the system has been established and the rules governing its auditory patterns have been defined, it is most practical to suppose that the primary load both in perception and production is borne by the perceptual processes, and it is most economic to suppose that any reference to the details of production in the process of decoding speech may be of secondary rather than of primary importance.

Department of Phonetics
University College London

REFERENCES

Chan, Angela
1971 "A Perceptual Study of Tones in Cantonese", Ph.D. thesis in the University of London.
Fourcin, A.J.
1965 "The Pitch of Noise with Periodic Spectral Peaks", 5e *Congrès International d'Acoustique*, Liège, Vol. 1a, B52.
1968 "Speech Source Inference", *TAIEEE* 16, 65-67.
1970 "Central Pitch and Auditory Lateralization", *Frequency Analysis and Periodicity Detection in Hearing* (R. Plomp and G.F. Smoorenberg, eds., Sijthoff).
Fourcin, A.J. and E. Abberton
1971 "First Applications of a New Laryngograph", *Med. Biol. Ill.* 21, 172-182.
Robertson, Margaret
1971 "Some effects of Source Inference on Speech Perception", Ph.D. thesis in the University of London.
Schouten, J.F., R.J. Ritsma, and B.L. Cardozo
1962 "Pitch of the Residue", *J. Acoust. Soc. Amer.* 34, 1418-1424.

DISCUSSION

SANTERRE (Montréal)

J'ai relu à plusieurs reprises et avec le plus vif intérêt la communication du professeur Fourcin; je l'en remercie et le félicite chaleureusement.

Les considérations et les quelques questions que je veux lui soumettre ne visent aucunement à infirmer ses conclusions, mais à satisfaire ma curiosité sur les données de ses expériences.

1. Vous avez dit que le présentatif (precursor) *hallo* ne contenait pas d'information explicite sur les locus de [b] ou de [d]. Il me semble, au contraire, que les transitions à partir de [l] et de [d] sont très ressemblantes, aussi bien sur *F2* que sur *F1*. Les *TR2* remontent ou descendent selon que la langue remonte ou descend sous le palais dur après la rupture de l'occlusion apicale: à partir de [b], toutes les transitions sont remontantes. Comme les auditeurs n'avaient le choix d'opter qu'entre [b] et [d], les stimuli qui se rapprochaient de [l] de *hallo* pouvaient être interprétés comme [d], et les autres, comme [b]. C'est pourquoi, à mon avis, le présentatif ne renseignait pas seulement sur la personne (homme ou enfant) à qui les stimuli pouvaient être attribués, mais aussi sur le mouvement des transitions de [d] opposées à celles de [b], au moins pour les stimuli dont les fréquences pouvaient être rapprochées des *F2* des voyelles plutôt ouvertes.

2. A propos de 'locus', j'aurais aimé connaître les fréquences de départ pour toutes les voyelles après [b] et [d], chez l'homme et chez l'enfant, plutôt que les seules moyennes de 1300 Hz (homme) et 2100 Hz (enfant) entre les deux consonnes; ces moyennes me masquent une partie de l'expérience, parce que je ne crois pas à l'existence d'un 'locus' pour chaque consonne indépendamment des voyelles. La fréquence de départ des transitions varie considérablement selon les voyelles; il y a à vrai dire autant de 'locus' que de voyelles qui précèdent et qui suivent la con-

sonne; à tout moment, les fréquences de résonance du canal buccal sont déterminées par les positions articulatoires, qui varient beaucoup plus en raison des différentes voyelles qu'en raison des modes d'articulation, occlusif ou constrictif, par exemple aux dents ou aux alvéoles. Les transitions à partir de [d] sont montantes ou descendantes selon les *F2* de la voyelle qui précède la consonne et de celle qui la suit. Toutes les transitions à partir de [b] sont montantes. Le 'locus' 1300 Hz ne me semble pas le point de séparation entre [b] et [d], ni nécessairement entre les transitions montantes et les transitions descendantes à partir de [d]. C'est pourquoi j'aurais aimé voir toutes les transitions et non seulement la moyenne pour chaque locuteur.

3. Il serait de plus intéressant de savoir comment ont été interprétées par les auditeurs les transitions droites, i.e., quand *TR2* et *F2* sont à la même fréquence de 1000 Hz, soit dans la zone qui ne pouvait être attribuée qu'à la voix d'homme, d'une part, et d'autre part, à 2500 Hz, dans la zone commune à la voix d'homme et à la voix d'enfant.

Je suppose, mais je voudrais savoir si je me trompe, qu'elles ont été entendues comme [d] dans tous les cas, puisque les bilabiales abaissent les fréquences de résonance du canal buccal, tandis que le seul fait, par lui-même, de poser l'apex sur les dents ou les alvéoles n'affecte pas sensiblement la fréquence de *F2*, cette fréquence étant déterminée à tout moment par la hauteur de la langue sous le palais dur et par la largeur du passage pharyngal.

Or, le point $TR2 = 1000$ Hz et $F2 = 1000$ Hz est dans la zone de [b] sur le tableau des résultats des tests; quand il n'y a pas de transition, comme c'est ici le cas, il est normal que le stimulus soit senti comme [d] par opposition à [b], et c'est ce qui arrive d'après le tableau pour l'homme et pour l'enfant; quand le point de départ de la transition était assez haut et la transition montante, j'aimerais savoir où s'est fait le partage entre [b] et [d].

Vous ne pouviez, bien sûr, nous fournir ici tous ces détails. D'ailleurs, sans les connaître, on ne peut qu'être d'accord avec les conclusions que vous tirez de vos expériences.

FOURCIN

1. Le locus de [d] ne correspond pas à celui de [1]. Quand vous attendez que le précurseur renseigne au minimum, l'effet décrit et *F2* pour l'homme sont au maximum.
2. J'ai utilisé les données que vous demandez en calculant les moyennes indiquées; je vous les montrerai.
3. Il suffit de construire la ligne $F2 = FL$ sur la Figure 5, afin de décomposer les stimuli entre transitions montantes, situées au-dessus de la ligne, et descendantes, situées en dessous.

PICKETT (Washington)
I would like to know whether, in your vowel experiment, the preceding *F1* pattern

for *hello* might have served to cue the listener to normalize the test vowel for vocal tract length and I wonder if, by using two or more formants in the test vowel, you could get results that would indicate whether the auditory processing has an articulatory reference or a purely 'acoustic' normalization.

FOURCIN
Vocal tract length is not by itself sufficient to enable a listener to tune in to a speech source. This is because the listener must have information about what the speaker really does, not what he might be capable of doing. The construction of a vocal tract provides limits within which the speaker is constrained to operate but he has the possibility of choosing from an indefinitely large number of articulatory settings within these limits. The listener must estimate the relationships between the corresponding sounds and he could, in principle, do this on the basis of purely acoustic or purely vocal tract normalisation. Theoretically these are merely transforms of each other but there does appear to be an advantage of simplicity in supposing that our perceptual processing makes use of both types of information with a much greater bias in favour of acoustic pattern manipulation. I have only given an indication of the possibility in my paper. I think that rather more convincing evidence will accumulate when more experiments of the type that I have described are performed with stimuli which could not physically be produced by a human vocal tract. In addition it will be important to employ listeners who have never been able to speak and are in consequence denied a direct knowledge of vocal tract-acoustic pattern relations.

FROMKIN (Los Angeles)
Did you run the pitch perception experiment with non-Cantonese speakers? It would appear that the labelling of the six tones is less specifically related to speech perception in this experiment than to general auditory pattern recognition.

FOURCIN
In the experiments I have described, the subjects were required to respond with lexical labels. This involves a specific speech response and is not dependent on the prior psycho-physical training which would be needed for non-Cantonese speakers.

GAGE (Washington)
I do not know how you know that no fundamental is present once you get past the middle ear, when you supply only the third, fourth, and fifth harmonics.

FOURCIN
Licklider has shown for coherent stimuli of this type, and it is easy to verify, that

low frequency masking noise can obliterate any possible products of distortion at fundamental frequency yet leave the low pitch percept unimpaired. For the tonal stimuli which were synthesized from three harmonically related incoherent noise peaks, it is not possible to have a physical component at the common fundamental frequency.

WAJSKOP (Bruxelles)
A la suite des résultats obtenus au cours des deux dernières expériences citées, quelle liaison faites-vous entre eux et ceux obtenus par Broadbent et Ladefoged (1957-1961)?

L'influence des stimuli précurseurs sur les réponses de vos sujets semble justifier l'exactitude du phénomène de constance perceptive ou de la théorie du niveau d'adaptation proposée par Helson (1948).

FOURCIN
I think that the results of the experiments which were reported by Broadbent and Ladefoged can best be interpreted as being due to the combined operation of two types of response behaviour on the part of the listener. In the first, the listener assesses the stimuli with reference to his gauging of the speaker's range of sound patterns. In the second, the listener assesses the stimuli with reference to his estimation of the inter-relations involved in the speaker's dialect. These are quite different levels of operation.

The experiments which I have discussed employed a single dialect situation and the listeners responded within the context of their assessment of a particular speaker's characteristics. A number of complex perceptual processes are involved but they are all closely tied to the sensory input. When the characteristics of a new dialect have to be determined the listener operates at a further level of abstraction and without proper controls the factors contributing to his response cannot be determined.

Adaptation occurs partly as a function of the actual mechanism of the sensory mediation of a stimulus. It is not primarily a function of his acquired knowledge. In consequence, although it comes within the first level of operation with which I am concerned, I do not believe that it contributes to the particular results which I have described.