AUDITORY SPEAKER RECOGNITION:

A THEORETICAL AND

EXPERIMENTAL

STUDY

ROGER BROWN

Thesis submitted for the degree of Ph.D.
University of Edinburgh
1980

This thesis is dedicated to my father,
who died on 28th June, 1979,
before its completion

A C K N O W L E D G M E N T S

My sincere thanks are extended to the many people who have given me help of various kinds during the preparation of this thesis:

- above all, to John Laver, of the Department of Linguistics, University of Edinburgh, who first suggested the field of speaker recognition to me, and, as both supervisor and friend, kindled my interest and supplied me with encouragement and much sound advice over the years.

- to Dr. P. Fisk, of the Department of Statistics, University of Edinburgh, for his enthusiasm in providing statistical help in the design and analysis of the experiments reported in Chapter 6. I am also grateful for the willing statistical assistance of Irene Macleod, of the Department of Linguistics, University of Edinburgh.

- to many other members of the University of Edinburgh for their constructive comments on previous drafts and seminar papers relating to this thesis; especially Ellen Gurman Bard, Gill Brown, Keith Brown, Karen Currie, Jody (Williamson) Higgs and Keith Mitchell.

- to the people, far too numerous to name, who willingly acted as speakers and listeners in the experiments reported here.

- for technical assistance, to the members of the Phonetics Laboratory, University of Edinburgh: David Cruickshank, John MacRae, Ron Motherwell,

Stewart Smith, and especially Jeff Dodds whose methodical approach and ever-willing service helped to alleviate many technical headaches.

- to the Department of Education and Science, the Social Science Research Council and the University of Edinburgh for financial support.

- to my typist, Grace Young, whose efficiency relieved me of many editorial worries.

To all these people I am very grateful. However, I must naturally take responsibility for all aspects of the final version of this thesis.

Several published articles have appeared in relation to this thesis. These are cited in the References section as Brown (1978a,b, 1979a,b). A reprint of Brown (1979a) is included here as Appendix 4.

D E C L A R A T I O N

This thesis is original work of my own execution and authorship.

Roger Brown

TABLE OF CONTENTS

CHAPTER 3

CHAPTER 4

N.B. A tape accompanies this thesis, containing

(i) recordings of nine speakers, whose voice quality is analysed in Figure 3.7.

(ii) the synthetic control voice and the 16 synthetic stimulus voices used in Experiment 2 (section 6.3).

LIST OF FIGURES AND TABLES

APPENDIX 1

A B S T R A C T

Speaker recognition is defined as the ability to recognise a speaker's identity on the basis of hearing a sample of his speech. Previous approaches to the subject have concentrated on the experimental manipulation in isolation of acoustic features of the speech signal.

The theoretical approach adopted here attempts to provide a conceptual framework for speaker recognition, in which emphasis is laid on auditory speaker recognition (as opposed to speaker recognition by machine or by the visual examination of spectrograms ("voiceprints")). The everyday use of speaker recognition is discussed in contrast to the possible artificialities of experimental formats. The nature and utilisation of phonetic speaker-characterising features of voice are examined within the context of

(i) other levels of features (syntactic, semantic, lexical, etc.)

(ii) other forms of indexical information (sex, age, regional origin, social status, etc.) and

(iii) other identity characteristics (names, physical appearance, etc.).

Attention is also focussed on two variables in the speaker recognition process which have been relatively neglected by previous writers and researchers:

(i) The nature and implications of differences in the tasks which listeners perform. The culmination of this discussion is a model in Boolean logic of the decision-processes involved in speaker recognition.

(ii) The possible effects caused by differences in the number, background and training of listeners.

The experimental approach adopted exploits the simultaneous manipulation of parameters made possible by the use of synthetic speech. The relative weighting rather than the absolute potentiality of parameters as speaker-characterising features can thus be examined. Results from voice similarity judgment experiments employing a factorial design indicate that:

(i) the parameters of mean pitch, mean formant position and formant bandwidth are important for speaker recognition, and

(ii) despite overall performance differences in judgments of similarity and difference, the responses of the individual listeners show comparable reactions to factorial changes.

CHAPTER 1

INTRODUCTION

CHAPTER 1

I N T R O D U C T I O N

1.1 INTRODUCTION

This thesis deals with speaker recognition - the ability to recognise who a speaker is from hearing his voice. In addition to being an investigation worthwhile in itself, study of the field is relevant to many other areas in phonetics and wider disciplines (see section 1.7). Most importantly, speaker recognition must be seen as a necessary and integral part of speech recognition, whether as a theoretical or practically implementable form of modelling. Analysis of the linguistically criterial features of speech cannot proceed (except in a rudimentary one-speaker model) unless normalisation by the elimination of the linguistically unimportant features of the signal has taken place. The attitude expressed in this thesis is that the range of features usable for successful speaker recognition is large; that the decision processes employed to achieve successful speaker recognition from this wealth of potential information are complex; and that, consequently, we are still a long way from being able to answer adequately even the most basic questions about the speaker recognition ability.

The scope of this thesis is limited in three respects, although it should not be thought that study of aspects of the speaker recognition field outside these three restrictions is considered in any way secondary. Instead, the narrowed scope reflects partly my interests as a phonetician and partly an attempt to redress the imbalance in the consideration afforded so far to the various aspects of the field.

(i) The speaker recognition field is traditionally divided into three separate aspects:

(a) human auditory speaker recognition

(b) human speaker recognition by the visual examination of spectrograms ("voiceprints") (see section 5.3)

(c) automatic speaker recognition by machine (see section 5.4)

The discussion in this thesis will be relevant only indirectly to the last two aspects, and this will be extended only as far as it serves to clarify the similarities and differences between these and human auditory speaker recognition.

Misinterpretation of terminology is possible here. Human speaker recognition should not be taken to mean the recognition of human speakers, but the recognition of speakers by human listeners.

(ii) Research into speaker recognition has concentrated on listeners' performance in experiments. Admittedly, this is the soundest method of investigation; however, it is argued that experimental findings are not necessarily very revealing with regard to the way in which listeners recognise speakers in everyday life. Certain artificialities are introduced by the experimental framework (see chapter 4), which restrict the projectability of findings onto real world situations. Equally importantly, little thought has been given towards describing what speaker recognition situations occur in everyday life. For these reasons, this thesis tries to relate the argumentation, wherever and as much as possible, to everyday life.

(iii) A further limitation is that claims are made explicitly only for the English language and native English speakers and listeners. This is not a great restriction, however, since the instances where the discussion might not be expected to be applicable equally well to other languages are few (such as the discussion of rhythm; sections 3.4.2.2.4 and 3.4.4.4).

As mentioned above, the whole speaker recognition field is relatively unexplored, and to attempt a more exhaustive coverage would be unwise and probably prove fruitless in such a thesis.

1.2 DEFINITION OF SPEAKER RECOGNITION

Hecker (1971) defines speaker recognition very broadly as 'any decision-making process that uses the speaker-dependent features of the speech signal' (Hecker, 1971:2), and this may be adopted as a working definition for this thesis. These speaker-dependent features are referred to as speaker information by Bricker & Pruzansky (1976) (see section 2.2), who amplify the term as 'those attributes that vary distinctively from speaker to speaker' (Bricker & Pruzansky, 1976:297). However, this expansion might be worded more strictly as 'those attributes which can vary distinctively from speaker to speaker'. This rewording allows us to avoid the implication that if two speakers' voices are identical in relation to a certain speech parameter, then that parameter cannot constitute speaker information (either at all or for those two speakers). The fact that two speakers' voices are identical in relation to a parameter may be just as characteristic a feature as if they were quite different. Of course, this similarity may make impossible the discrimination of these two voices by that particular parameter, but, by the same measure, the narrowing-down of a set of reference possibilities in an identification task to these two may be obstructed if this confusion is not maintained.

There is an associated fundamental distinction to be drawn between the probably quite large number of parameters which are potentially usable in speaker recognition and the subset of these parameters which human listeners habitually use to recognise speakers, or which have priority in the everyday situation. This distinction is discussed further in section 1.8.3.

To judge from the frameworks of speaker recognition experiments, there is implicit agreement among researchers in the field

that the speaker recognition process must take place in the following manner (and this assumption will be adopted in this thesis, with the possible exception of simultaneous presentation tasks; see section 4.3.1). Reference voice patterns are stored in the listener's memory. These patterns are composed of speaker-characterising features which have been extracted during the listener's previous exposure to the speaker's voice. This exposure may result from everyday contact or may be much shorter, as in some experimental tasks where the duration of the reference sample may be only a couple of seconds. These reference patterns are compared with the pattern composed of features extracted from the stimulus voice sample (the new utterance heard), and a decision is reached on the basis of the similarity of the two patterns.

1.3 SPEAKER-CHARACTERISING FEATURES, INDEXICAL INFORMATION AND IDIOSYNCRASIES

The speaker-characterising features are therefore present in the speech signal along with the linguistic signal which conveys what the speaker is saying. The speaker characteristics are in this sense an extra message (Peters, 1954), and are prerequisites for auditory speaker recognition. The term index (in one of the senses described by Peirce, 1940; Hartshorne & Weiss, 1931-5; Feibleman, 1946; Greenlee, 1973; Laver & Trudgill, 1979) has been generally adopted to refer to any such feature which allows the listener to infer information about the speaker (i.e. which is informative in Lyons' (1977) terms; see section 1.8.1).

Abercrombie (1967:7) and Laver (1968:48) classify the different kinds of indices which are present in speech - Abercrombie into

(i) those that indicate membership of a group (regional and status indices),

(ii) those that characterise the individual (idiosyncratic indices), and

(iii) those that reveal changing states of the speaker (affective indices);

Laver into

(i) biological information (size, physique, sex, age and medical state),

(ii) psychological information (personality), and

(iii) social information (mainly accent information of regional origin, social status, etc.).

It is intuitive to suppose that, of Abercrombie's categories, idiosyncratic indices are the most important for speaker recognition. However, a particular speaker's voice may be characterised by its reflection of membership of regional or social groups, or of changing affective states of which the listener is aware. However, it seems unlikely that correct speaker recognition is achieved by knowing that the indexical features of sex, age, regional and social class, etc., inferred from the stimulus voice sample, apply correctly to one of the reference speakers. That is, a stimulus voice is not matched with a reference speaker by virtue of the fact that it provokes judgments of sex, age, etc., compatible with known reality. This is not to say that those features of the speech signal by which indexical judgments are made are not the same as, or similar in nature to, those by which speaker recognition is achieved, but that speaker recognition is performed on a more holistic and less componential basis.

Lyons (1972:71) describes these three-way classifications as 'different, though not necessarily conflicting', and indeed one can equate Laver's social information with (or at least consider it the major component of) Abercrombie's group category, and Laver's medical and psychological information with Abercrombie's affective indices. However, it is not clear how the elements of Laver's categorisation relate to Abercrombie's idiosyncratic indices. This

problem arises partly from the difficulty in the definition of the term idiosyncratic. Any definition must make appeal to the notion that an idiosyncrasy is characteristic of an individual. However, it is not necessarily a characteristic which belongs to one person alone. It may be considered an idiosyncrasy of a speaker that he uses a 'burr' (a voiced uvular fricative) as a realisation of the phoneme /r/ (to quote Abercrombie's (1967:9) example). (This does not apply if he is a speaker from the area around Durham, for whom a burr is a feature of the regional accent.) However, to call this an idiosyncrasy does not necessarily mean that he is the only person to use it; there are many speakers with accents other than that of the Durham area who use a burr. Instead it merely means that its occurrence is not caused by any motivation, intentional or otherwise. It follows that the occurrence of a burr in non-Durham speech cannot be predicted on the basis of observations of other speakers in the person's speech community. On the contrary, since a burr is not considered a regular feature of the phonology of any non-Durham accent (because it is statistically rare and only occurs in certain individuals), one would expect any such speaker not to use it. At the other extreme, it is possible for an idiosyncrasy to belong to one individual alone, although it would need to be a very abnormal feature - certainly much less common than a burr. This argument applies equally well to features unexpected on sex, age, social or psychological grounds etc. as it does to regionally unmotivated features.

To sum up therefore, any apparently unmotivated variation may be considered idiosyncratic. In short, an idiosyncratic feature is one which cannot be correlated with group factors such as sex, age, regional origin, social status, health, etc. Someone who does everything (and only everything) which is typical for a person of his sex, age, regional origin, social status, health, etc. cannot be said to have any idiosyncrasies. A consequence of this is that an idiosyncratic category must be added to Laver's

classification since idiosyncratic features are indexical but, by definition, do not indicate memebership of any biological, psychological or social group.

1.4 THE ABILITY TO NAME SPEAKERS

The term most widely used to refer to the process is speaker recognition. However, this term is misleading in that it might be taken to imply that the process is completed only when the identity of the speaker has been revealed, for example by giving his name. That the ability to name a speaker is unimportant to auditory speaker recognition in general terms is illustrated by a not uncommon real world situation. Often on hearing a voice, one realises quickly that one knows the voice (i.e. that it corresponds to an already stored reference voice pattern) but cannot say at the same time who that speaker is. The situation can be represented by Figure 1.1.

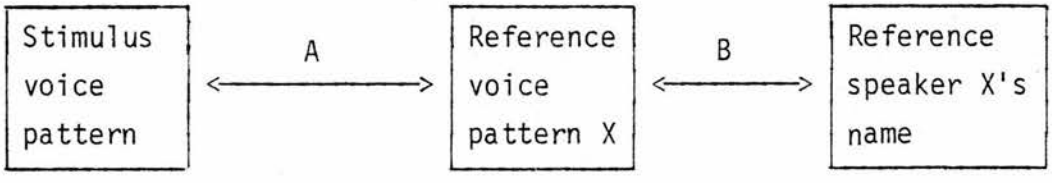


Figure 1.1 Speaker recognition and speaker naming

The listener is able to match the stimulus voice pattern with the stored reference voice pattern (to perform comparison A), but not to match the stored reference pattern with the stored name of that reference speaker (to perform comparison B). Comparison A is the domain of auditory speaker recognition, while comparison B is not. That is, speaker recognition refers to the recognition of voices (sound-patterns) not of speakers (people). It is in this

sense that speaker recognition is a misnomer; voice recognition would be more accurate. However, since speaker recognition is the most widely used term, I shall continue to use it.

This model of the matching processes involved in speaker recognition may be extended if we consider visual and other representations to be further forms of speaker identity. A more complex but still schematic version of Figure 1.1 may then be proposed (Figure 1.2).

In broad psychological terms, the representation stored in an individual's memory of people whom he knows, directly or indirectly, may be said to consist of various kinds of characteristic - at least the following four:

(i) a description such as the reference person's position in life relative to the individual ("the milkman; the man who owns the grocer's shop on the corner; the lady I met on the bus this morning").

(ii) the reference person's name ("Mr. Smith; John Smith; John") (names have characteristically different properties from descriptions; Searle, 1969).

(iii) his physical form or appearance, most importantly his face, but also his overall build, habitual gestures, the way he walks, etc.

(iv) his voice, a specification of distinctive acoustic/perceptual characteristics.

Other kinds of characteristics no doubt exist in this overall representation, although these four are probably the most important. It will not necessarily be the case that the specification for each reference person contains all these four kinds of characteristic. Thus an individual will have no information about the physical characteristics of a person with whom he only ever talks over the telephone (although he may form expectations about this; Laver & Trudgill, 1979). There may be many people whom he does not know by name. I may refer to Pierre Trudeau, the Prime Minister of

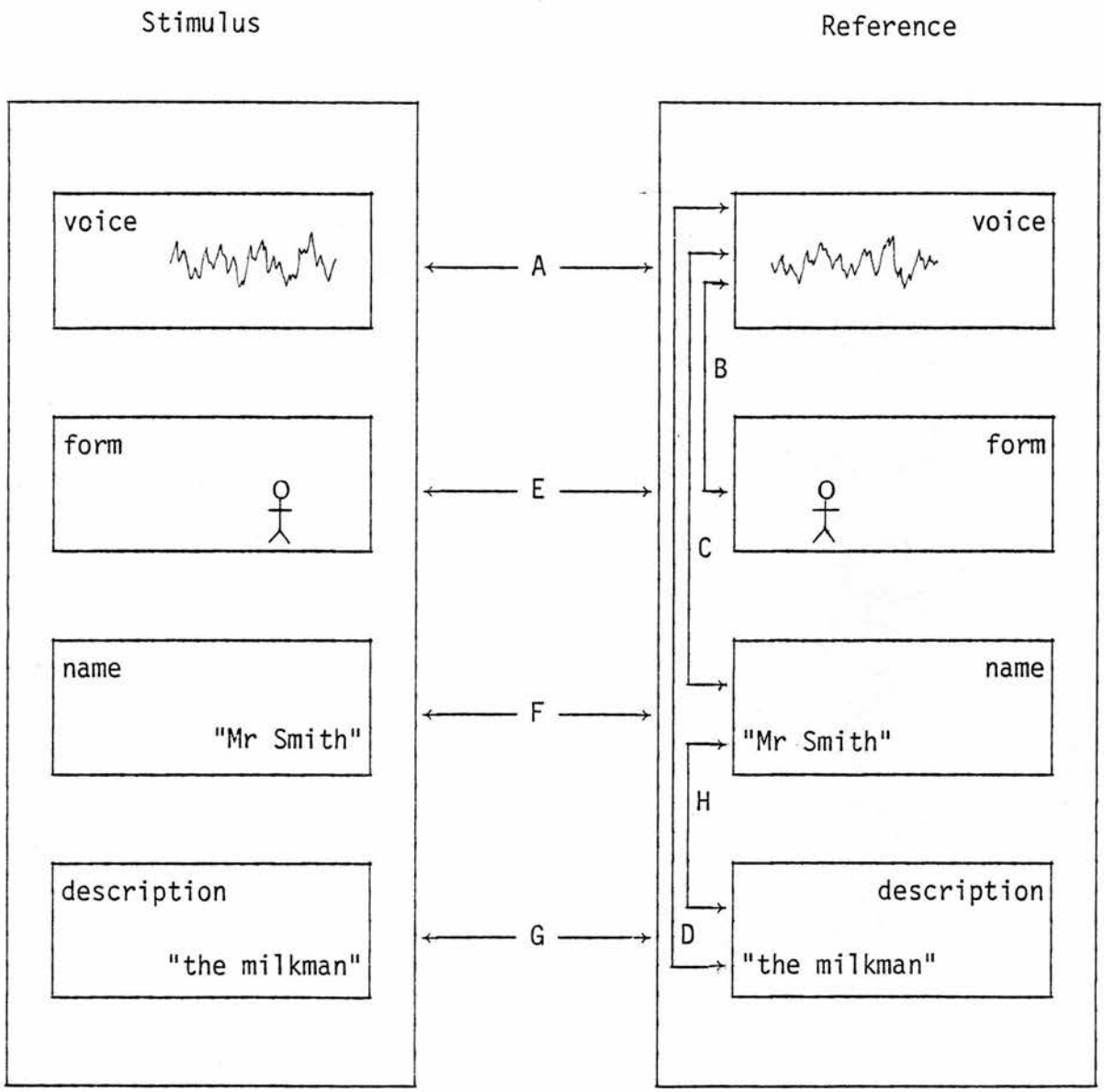


Figure 1.2 Reference to identity characteristics

Canada, without knowing what his voice sounds like. However, we may consider descriptions to be basic in that they are always present even if they have only been devised expressly for the purpose of classification of that reference person in the individual's memory store. That is, every reference person relates to the individual, whether remotely or closely, and it is a representation of this relation which is stored.

This four-way categorisation of reference characteristics may now be related to the question of speaker recognition. The voice representation (reference voice pattern) must be present as one of the characteristics, for otherwise auditory speaker recognition is of course impossible. It has just been argued that some form of description must be present, although the other two (form and name) need not be. It will be assumed for the following examples that they are present. This model then allows us to distinguish various kinds of recognition failure, which are expressed in the following plausible verbal responses to auditory speaker recognition tasks. In each case, it should be taken that the listener does in fact know the stimulus voice.

1. "I don't know that voice." Failure to match stimulus and reference voice patterns (process A in Figure 1.2).

2. "I know that voice but I can't put a face to it." Successful matching of stimulus and reference voice patterns (process A), but failure to connect voice pattern with physical form representation (process B).

3. "I know that voice but I can't put a name to it." Successful matching of stimulus and reference voice patterns (process A), but failure to connect voice pattern with reference's name (process C).

4. "I know that voice but I can't think in what context." Successful matching of stimulus and reference voice patterns (process A), but failure to connect voice pattern with description (process D).

Other more complex kinds of failure may be represented as combinations of the above four types. An important observation is that in examples 2, 3 and 4 above, auditory speaker recognition has been successfully completed, even though the stimulus speaker has not been fully recognised in terms of the other reference characteristics.

The model may be easily extended to visual person recognition where the stimulus would consist of a token of the person's physical form. This might be to actually see the person, as in normal everyday situations or in an identity parade, or to be presented with a less direct stimulus such as a photograph or an Identikit picture. The matching process would then be process E in Figure 1.2. It has been reported that patients compensate for prosopagnosia (inability to recognise faces) by means of voice recognition (Bornstein & Kidron, 1959; Beyn & Knyazeva, 1962). The two are therefore distinct processes relying on different media, although both contribute to speech perception (McGurk & MacDonald, 1976; Binnie et al., 1974).

It is difficult to think of similar experimental tasks involving the matching of names or descriptions. However, this corresponds to the everyday process of reference (Lyons, 1977), as when someone tells you "Miss Ryan wants to see you" or "My wife has left me" (processes F and G respectively). Instances where a reference-stimulus matching of descriptions and a reference description - reference name connection are required are quiz instructions such as "Name the inventor of the phonograph" (processes G and H).

The above examples show that a person's voice is merely one form of characterising feature. Exactly what constitutes a person's identity is a psychological or philosophical question beyond the scope of this thesis. Similarly, the connecting of a person's reference voice pattern with his reference form, name or description is a psychological process outside the field of auditory speaker

recognition. Therefore, in this thesis, speaker recognition refers only to the matching of voice patterns (process A). In experiments where more than one reference voice is used, investigators can only test listeners' abilities to recognise voices by requiring them to respond with reference speakers' names, or some such labels (processes A and C in Figure 1.2). However, these are two separable processes, and in the real world success in one is not always accompanied by success in the other (as is shown in example 3 above).

1.5 EVERYDAY OCCURRENCES OF SPEAKER RECOGNITION

Some writers demonstrate the relevance of the study of speaker recognition by including one or two examples of situations in everyday life where the ability to recognise speakers plays an important part in communication.

> 'Most listeners have little difficulty in identifying the voices of familiar speakers over the telephone or on the radio. Recognition of the voices of familiar speakers in the darkness or when the speaker is out of sight of the listener is also a common occurrence.'
>
> (Compton, 1963:1748)

> 'Each of us has had the experience of recognising a friend's voice and being able to name the talker without seeing him.'
>
> (Bricker & Pruzansky, 1966:1441)

> 'An old friend telephones you unexpectedly and you know who it is before he declares himself. After watching a TV talk show for a few minutes, you know which panel member is speaking before the camera picks her up.'
>
> (Bricker & Pruzansky, 1976:295)

The view represented in these quotations is a very narrow one, in that the situations contained in the examples are limited in

three important respects. Firstly, the situations quoted as plausible everyday events all deal with the initial recognition of speakers. The voice to be identified either has not been heard by the listener in the recent past, or does not correspond to the voice most recently heard. The novelty of the stimulus causes a reasonably conscious decision to be made by the listener. He may frown, turn his head in concentration, say "Who's that?" either internally to himself or out loud. All these are signs that he is making a conscious effort to arrive at the correct identity of the speaker. The assumption that speaker recognition deals with this initial identification of voices is a narrow standpoint to adopt. It has obvious attractions since initial identification corresponds to a task very commonly set experimentally. However, the process which takes place in initial recognition has strong similarities to one which often takes place during the course of conversations. If you are holding a telephone conversation with someone, you need to be sure, each time that you hear a voice on the line, that it is the same voice, i.e. that you are still talking with the same person and have not been interrupted. More is said about this confirmatory function of the speaker recognition process in section 4.5.2, where it is argued that the process involved is of the same nature as one kind of experimental task, but that it is performed in a much less conscious manner; the listener does not show any outward signs of conscious effort.

Secondly, the situations quoted presuppose that the listener is deprived of visual clues as to the speaker's identity. Therefore the listener is described making a phone-call, listening to the radio, with his back turned, in the dark, etc. For initial identification in face-to-face situations, it is reasonable to assume that the visual stimulus of actually seeing the speaker plays a greater part than the acoustic stimulus of hearing his voice. It is difficult to imagine any possible situation in which an acoustic stimulus contradicts a visual one (as in the format employed by McGurk & Macdonald, 1976). Knowing this, the listener may justifiably

subordinate the speaker-characterising information contained in the acoustic stimulus, preferring to rely on that contained in the visual. However, it is not reasonable to assume that the acoustic signal plays no part in the speaker recognition process when visual clues are present, since the possibility of misinterpretation of visual clues exists.

Thirdly, exemplification of real world instances of speaker recognition by reference to stimuli occurring on the television, radio or telephone suggests that localisation clues play no part in confirmatory (and possibly initial) speaker recognition decisions. The voice which acts as the stimulus in these situations is not live in the sense that the speaker is in the vicinity of the listener, but is presented via the loudspeaker of the television, etc. Localisation clues, which in the live situation allow the listener to track any movement (or indeed stationary position) of the speaker, therefore do not apply. Visual and localisation clues are not usually made available in experimental formats and this may be a reason why such a limited view of speaker recognition is often presented. For writers to start with a description of one or two illustrations of the everyday importance of speaker recognition gives the reader a false impression that there exists a theoretical framework relating the processes involved in experimental tasks to real world possibilities. No such explicit theory has been found in the literature which refers to considerations such as the above and those contained in Chapter 4.

1.6 NON-PHONETIC SPEAKER-CHARACTERISING FEATURES

While the above discussion has dealt with the acoustic/perceptual properties of the speech signal, it should not be forgotten that higher-order properties (syntactic, semantic, lexical) may also be speaker-characterising. A person's choice of words and of grammatical constructions is very idiosyncratic. When writing some

extended piece such as a letter or a thesis, it is often difficult to avoid using the same limited repertoire of words and constructions again and again. One can usually identify an unknown passage by Henry James, for example, from its interrupted syntax. Impersonators exploit a speaker's semantic and syntactic idiosyncrasies as much as phonetic features of his speech. Although the characterisation afforded by these features is much weaker (see section 3.3) than that of phonetic features, they should not be totally ignored and may play a part in speaker recognition in exceptional circumstances. In this thesis, however, attention will be concentrated on the recognition of speakers by the use of phonetic features.

1.7 THE RELEVANCE OF SPEAKER RECOGNITION

Why should one want to study the field of speaker recognition? There are two answers to this question. Firstly, there is the practical answer relating to the applications to which an automatic speaker recognising device may be put in the real world; these are dealt with in the next section. Secondly, there is the theoretical answer, which relates study of speaker recognition to study in other fields, in an overall attempt to increase our knowledge of the human perceptual mechanism.

1.7.1 Theoretical Applications

Speaker recognition relates closely to the field of speech recognition. It is a major criticism of many present-day automatic speech recognition devices that they will operate successfully with input speech from only a few specific speakers, and sometimes only one. With a realistically large population of speakers, any successful speech recognition model, whether it be a theoretical representation of the process, or a physical device which replicates the human ability, requires a speaker recognition model as one of its component parts. Thus a speech signal must preliminarily be normalised so that inter-speaker differences of pronunciation are diminished or eliminated

before the process starts which attempts to recognise the linguistic units of the signal (allophones, phonemes, words, etc.).

Just as speech recognition is dependent upon preliminary speaker recognition, as has just been argued, so speaker recognition is dependent upon speech recognition. To recognise a speaker, one needs to be able to recognise and discount the linguistic features of the speech signal in order to derive the speaker-characterising component which constitutes the inter-speaker variation by which speaker recognition is achieved.

Inter-speaker variation is a major factor accounting for the difference in success rates achieved so far in automatic speech recognition systems on the one hand and speech synthesis devices on the other. As was shown above, the problem of inter-speaker variation must be solved for any but the simplest speech recognition device. However, the criterion for success of a speech synthesis system is that the output speech is intelligible and reasonably natural, not that it has to sound like one particular speaker. In other words, while inter-speaker variation is a major stumbling-block for many present-day speech recognition systems, this problem is bypassed in speech synthesis. A fairer comparison could be made between speech recognition and speech synthesis systems (although it would have limited practical purpose) if the speech synthesis system were required to produce an output which sounded like one of a population of specified speakers, i.e. if the factor of inter-speaker variation were introduced.

A second field where the relevance of speaker recognition is great is child language acquisition. A child gradually learns to speak his native language by imitating the adults around him. In general, the adult with whom he has greatest contact is his mother, and, other things being equal, a child normally grows up to speak with an accent similar to that of his mother, at least before school age. However, it is unreasonable to suppose that speech from other adults does not have any influence on this learning

process. That the child manages to learn his native language despite the fact that he hears speech from a variety of differing voices is evidence for the argument that the child must possess some form of normalisation process based on speaker recognition principles. There is evidence (Friedlander, 1968) that speaker recognition ability is acquired very early in the child (at roughly 12 months). The child is thus able to take into account inter-speaker differences, and to abstract and concentrate on the linguistic component, which is the common denominator of all the acoustic signals. His success in imitating precisely what he hears is limited by intrinsic factors (see section 1.8.1), anatomical and physiological in nature. He has a shorter vocal tract that the adults he imitates, and the proportional relationship of the various articulators and cavities also differs.

The ability to recognise human voices is not only a human faculty. In the 1890's an English artist, Francis Barraud, painted a picture of his dog Nipper recognising his master's voice being played back on a gramophone. The Gramophone Company adopted the symbol in 1900, and we are all familiar with it today as the trade-mark of the His Master's Voice record label. Study of the ability of animals other than humans to recognise human voices may shed limited light on the human ability.

Speaker recognition is relevant not only to speech recognition but also to the wider field of communication theory. The place of speaker recognition in communication in general is examined in section 4.3.1.

1.7.2 Practical Applications

There are many applications to which a device which can recognise the voice of an individual or a small group of individuals can be put, some of them obvious and some less immediately apparent. Obvious examples are situations where access to information or areas

is limited to certain authorised personnel. This might involve information stored in computer data files, as in modern banking, or confined areas, as in prisons. A criterial feature of all these situations is the requirement of optimal security. Success of the speaker recognition device can be measured as the minimisation (or ultimately elimination) of incorrect acceptance decisions. There are two limitations governing the practical use of present-day speaker recognition devices. Firstly, in practical situations, the device has to be capable of not being misled by mimicry by impostors, a factor which has not been investigated in depth by researchers. Secondly, the level of security required in banking, prison, etc. is very high. Present-day speaker recognition devices cannot provide this security with realistic populations of speakers and they are therefore used only in conjunction with other (e.g. fingerprint) security systems.

Practical applications which are perhaps less obvious include instruments for people whose main means of practical everyday interaction is speech, e.g. the physically handicapped. If a person is so disabled that it is a struggle for him to perform a relatively simple operation like turning the key in his front door, then any device which can produce the same effect of opening the door without physical requirements and with a comparable level of security is an invaluable aid. Again, the level of security required and the likelihood of impostors places severe criteria on the sophistication of the device and on the feasibility of the whole project.

A more technical application relates to vocoder devices. These analyse speech into an optimally small set of parameters, which may then be transmitted by cable or radio-waves, to be re-assembled into intelligible speech by a synthesiser system. The advantages for telecommunications of an optimally small set of parameters are obvious. However, such devices do not preserve all the speaker-characterising features of the signal, which is tolerable

for some purposes, but less so for others. Considerations relevant to speaker recognition will probably also have significance for this latter set of situations.

Finally, identification of criminals or suspects in court is generally performed auditorily by human listeners, sometimes supported by evidence from the visual examination of spectrograms by trained experts (Tosi, 1979). The reliability of the human methods in isolation is thought generally not to be high enough for legal applications (Stevens et al., 1968; Bolt et al., 1969, 1970, 1973). The reliability required of an automatic device would again be so high in this case that its practical application might be limited to providing supportive evidence for the human expert.

1.8 PRELIMINARY DISTINCTIONS

Three distinctions will now be discussed. Since the distinctions are of basic importance to human speaker recognition and are a prerequisite for the categorisations which follow in the thesis, the preliminary examination given to them in this section will be extended to some length.

1.8.1 Intrinsic and Extrinsic Factors

Although the writers who have talked about the first distinction have treated it as a strictly binary one, it will be shown that, at least for the purposes of speaker recognition, certain flexibility and further explanation are necessary before such a categorical standpoint can be adopted.

In short, the distinction lies between those production factors over which the speaker has some control, and those over which he has no control. The problem concerning the definition of this distinction is that various other distinctions have been treated as though they stood in an unambiguous one-to-one relationship with it.

In this section these closely related distinctions are examined and the degree to which they correspond to that outlined above is assessed. I shall therefore take the above criterion of controllability as the basic one and see how the others diverge from this.

The drawback of many of the criteria which writers have offered as defining features of this distinction is that they rely heavily on the interpretation of certain key-words. The implication is that these key-words are unambiguous and therefore do not need further elaboration. In this way, the definition, which ought to be an explicit statement of unambiguous interpretation, relies upon the implicit (and probably slightly differing) interpretations given to it by individual readers.

The criterion which I am taking as basic, namely that of controllability, is a good example of the above point. Writers define this criterion by reference to the term control. It is simple to treat this as a binary feature; the distinction is, as stated above, between those factors over which the speaker has no control whatsoever, and those over which he has at least partial control. However, one can see that there is a large difference in the degree of control, and in the nature of that control, which the speaker can exert over factors subsumed in the "partial control" category. The greatness of this difference may lead to differences of interpretation of this criterion.

A notion which is not implied in the distinction being discussed but which is so closely allied to it that it is a major possible source of ambiguity is that of skill of control. Because of this possible ambiguity, many writers have referred to potential controllability. The degree to which this skill of control is present for any one speaker will depend upon two factors:

(a) from a productive point of view, his ability to use auditory and proprioceptive forms of feedback in the performance of articulatory adjustments, and

(b) his perceptual ability to use auditory forms of information to detect sound differences, in order to establish target values for the production entailed in (a).

In this way, (a) is dependent upon (b), although both are entailed in the notion of skill of control. In short, a speaker cannot consistently perform an articulatory adjustment if he is not aware of a difference between his target articulation and his actual articulation.

Those factors over which the speaker has no control are taken to be attributable solely to features of the organic structure of his vocal apparatus.

> 'Examples of such uncontrollable features are the maximum and minimum possible length of a given speaker's vocal tract, the maximum possible area of velo-pharyngeal opening, the length of the vocal cords, factors of mechanical inertia of muscles of a given mass, and so on.'
>
> (Laver, 1975:26)

Also subsumed under this category are features which one would call structural "defects" such as cleft palates or loss of teeth, although such features can nowadays be easily eradicated or minimised by artificial means such as surgery or false teeth. Controllable factors will therefore not derive from organic constraints, and will include

> 'not only the choices of segmental articulation, but also the choices of manipulation of the voice dynamic strand of speech, and all the potentially controllable habitual muscular settings which characterise the manipulable component of his voice quality.'
>
> (Laver, 1975:26)
>
> (see sections 3.4.2.1 and 2 for a description of voice dynamic and voice quality features)

Thus organicness can be taken as a binary criterion. However, further explanation is necessary to account for scalar variation within each category. Laver hints at this by defining uncontrollable features as being 'dependent solely on the normally invariant physical foundation of the speaker's vocal equipment' (1975:324) and as 'the permanent and ephemeral organic foundation of the speaker's anatomy and physiology' (1967:523) (my underlining in both cases). The fact that uncontrollable features may accompany a variation in this physical foundation is mentioned by Abercrombie (1967). However, these variations are always quite temporary.

> 'They may arise, for example, from such causes as adenoids, tonsillitis, laryngitis, pharyngitis, or a common cold. These and other infections involve inflammation of the tissues of the vocal tract at various points, which will usually result in modifying the quality of the sound which the vocal tract conveys.'
>
> (Abercrombie, 1967:92)

Uncontrollable features may result from factors of an even shorter-term duration. A whispery or breathy phonation quality may be attributable to the fact that the speaker has just run up a flight of stairs. Of course, this uncontrollable feature will last for a much shorter time than one caused by a common cold, and this will in turn last for a much shorter time than any permanent organic feature. All these kinds of features would fall under the "uncontrollable" category, but for the purposes of categorising speaker-characterising features, it is useful to draw a distinction between long-term (permanent) and short-term features. The distinction may be characterised by the criterion of irreversibility; after illnesses such as laryngitis or very short-term effects such as that produced by running up stairs, the tissues and muscles of the vocal apparatus will eventually return to their previous unaffected permanent state. However, this will not be the case for long-term (permanent) effects such as those caused by puberty, excision of vocal cords, etc. Factors such as the fitting of false teeth or

the surgical repair of a cleft palate will involve an adjustment of the normality categorisation of a speaker's permanent organic structure. The relevance of this distinction to a discussion of speaker-characterising features is that long-term uncontrollable features will be strongly speaker-characterising, while short-term uncontrollable features will be of limited characterising strength (see section 3.3).

It follows that if certain features of a speaker's vocal output derive from permanent, uncontrollable, organic factors, then those features will pervade all the susceptible segments of his speech, i.e. they will be omnipresent. Certain phoneticians, including Abercrombie, Ladefoged and Laver, have concentrated on this aspect and restated the distinction as one between phonetic and non-phonetic features, on the grounds that phonetics does not deal with the permanent, invariable organic structure of a speaker's vocal apparatus but with the way in which these organs are manipulated for the intentional conveying of indexical and signalling information (see section 3.2). If a feature is not manipulable, then it cannot be used for the purpose of conveying any linguistic or paralinguistic information. This is not to say that such features do not convey indexical information about the speaker's sex, age, health, etc. (Laver, 1968). This is referred to by Lyons (1977) as the distinction between communicative and informative properties of signals.

> 'A signal is communicative ... if it is intended by the sender to make the receiver aware of something of which he was not previously aware. Whether a signal is communicative or not rests, then, upon the possibility of choice, or selection, on the part of the sender. If the sender cannot but behave in a certain way (i.e. if he cannot choose between alternative kinds of behaviour), then he obviously cannot communicate anything by behaving in that way. ... "Communicative" means "meaningful for the sender". A signal is informative if (regardless of the intentions of the sender) it makes the receiver aware of something of

> which he was not previously aware. "Informative" therefore means "meaningful to the receiver". ... Sender's meaning involves the notion of intention and receiver's meaning the notion of value, or significance.'
>
> (Lyons, 1977:33)

Both intrinsic and extrinsic factors therefore can be informative, although only extrinsic ones can be communicative. The disadvantage of the use of the terms phonetic/non-phonetic to refer to this distinction is that it implies that such uncontrollable, organic features are of no immediate interest to the phonetician.

> 'The general "quality" of the individual's voice ... is chiefly determined by the individual anatomical characteristics of the larynx and is of no linguistic interest whatever.'
>
> (Sapir, 1921:47, FN)

Although the quality caused by anatomical factors performs no distinctive, linguistic function, the phonetician must take account of these factors in order to establish norms with reference to which descriptive systems can be set up. In other words, organic factors are very relevant to the phonetician as a background factor in answering questions such as "How can a descriptive system, such as the International Phonetic Association alphabet, be applied to the speech of a cleft palate patient?", or "If a speaker has had his vocal cords excised, can his speech be said to contain the productive or perceptual features of voicing?", and so on.

If a feature is manipulable, then the speaker has a choice as to the manner in which it is used for the conveying of signals or as an index, usually social or psychological in nature. Where the speaker has a choice, certain writers have described the manipulation of these features as habits, learned by imitation. There are two drawbacks to using this as a criterial feature for the distinction:

(a) it may be interpreted as implying that this manipulation is the result of conscious learning by the imitation of easily definable exemplars. That is, it does not allow for the possibility that any such manipulation may be purely idiosyncratic, perhaps even as a dissimilatory process or without definable reason.

(b) it implies that the manipulation, having been "learned", can be easily "forgotten". However, as Abercrombie (1967) notes,

> 'though acquired by learning, the habit of such muscular tensions can, once acquired, be so deeply rooted as to seem as much an unalterable part of a person as his anatomical characteristics.'
>
> (Abercrombie, 1967:93)

If such transference from one category to another is accepted, then the distinction cannot be held to be binary. This is obviously closely related to the concepts of controllability and skill of control.

Another distinction which is related to that of controllability is between habitual parameters and extreme parameters. Laver (1968) summarises this distinction:

> 'The anatomy and physiology of a speaker determines the width of the potential range of operation for any one voice quality feature, and the long-term habitual settings of the larynx and the vocal tract restrict this feature to a more limited range of operation. For example, a man's voice may be physically capable of spanning a wide pitch range; in normal speech, however, he habitually selects a more restricted range within the total possibilities. Basic anatomy and physiology thus determine the possible extremes, and voluntary muscular settings determine habitual ranges between these extremes.'
>
> (Laver, 1968:44)

The crux of the problem is that many writers use the terms "normal speech" or "habitual range" without further definition. Since this is a very basic question for speaker recognition, I shall spend a little time investigating the way in which writers have treated this problem. I shall take as the parameter for illustration habitual pitch range since writers have discussed it at some length, corresponding as it does to the stave widely used for the transcription of intonation curves.

The following are quotations from writers defining the meaning of the two-line stave:

> 'two parallel lines, representing the upper and lower limits of the normal voice range'
>
> (Schubiger, 1958:4)

> 'two parallel lines representing the upper and lower limits of the normal voice register'
>
> (MacCarthy, 1944:162-3)

> 'two horizontal lines representing the normal high and low limits of the voice'
>
> (O'Connor & Arnold, 1961:6)

> 'double lines, representing the upper and lower limits of the speaking voice'
>
> (Lee, 1960:8)

> 'the top line stands for a relatively high and the bottom line for a relatively low pitch'
>
> (Palmer & Blandford, 1924:13)

It can be seen that each of these definitions describes the meaning of the two-line stave in terms of the undefined notions of "normality", "the speaking voice" or "relativeness". In other words, these descriptions may be didactically helpful as a means of explanation (the quotations come from intonation handbooks for foreign learners) but they do not constitute definitions.

People's estimates of the physical distance between the upper and lower limits implied by the two-line stave for an average speaker vary widely. The narrowest is about half an octave (Kaplan, 1960); Delattre (1965) puts the figure at between 7 and 10 tones; Christophersen (1956) estimates that one octave is normal; Schubiger (1958) and Van Riper & Irwin (1958) reckon one and a half octaves; Gimson (1962) and Jones (1964) have similar estimations of up to just over two octaves; O'Connor (1973) is non-committal at between one and two and a half octaves. The habitual, normal repertoire is therefore typically a very restricted range in comparison to the range of physical possibilities. However, the great variation between these estimates does not mean that they represent contradictory views, but is rather because the definition of habitual pitch range, for which each writer provides an estimation, differ; the variation reflects a difference in the flexibility of the writer's viewpoint, as to how much or how little should be included in the calculation of habitual pitch range.

Jassem (1952) describes the two-line stave somewhat differently as

> 'two parallel lines to indicate the bottom pitch and a probable top speaking pitch, and by placing the strokes between the lines in such a way that they should show the relations of the intonations of the particular units to each other and to the bottom pitch, the intonation patterns of those tone-groups can be represented...'
>
> (Jassem, 1952:53)

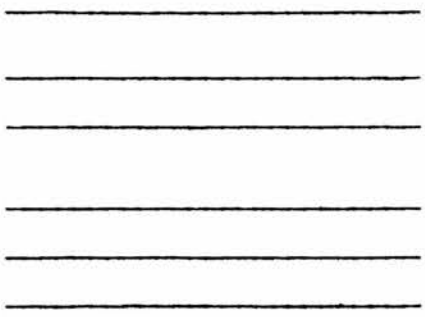
Jassem appreciates the fact that it is the top line which is the less stable of the two and which affects people's estimates of pitch range. The bottom line is relatively fixed, being restricted more severely by physiological constraints. This is reflected in the examples of intonation contours given by Jassem, which rarely rise above midway in the stave but frequently fall as low as the bottom line. This is also perhaps the reason why Ward

(1929) and Armstrong & Ward (1931) resort at times to the use of a single line (the bottom line) without any upper extreme to the stave. When the pitch range is altered, because the speaker is excited for example, it is intuitive to suppose that the extension at the top of the normal range is greater than that at the bottom. If this is so, we might ask why, given that most writers claim to be using the stave as a representation of the limits of the normal speaking voice, do few writers allow excursion outside the stave in instances where what is being represented constitutes an abnormal, exaggerated form of speaking. The reason is probably that it would confuse the reader of the book, being written as most of them are as text-books of the phonetics of English for foreign learners. From a theoretical point of view, however, there is no reason why, if the stave represents a normal range, abnormal styles of speech should not extend outside the stave. Abercrombie (1969) touches on this when defining tessitura (which corresponds to the pitch range phenomenon being discussed here):

> 'If we disregard the occasional extremes, a speaker has a characteristic range of notes, or compass, within which the pitch fluctuation of his voice falls during normal circumstances.'
>
> (Abercrombie, 1967:99)
>
> (my underlining)

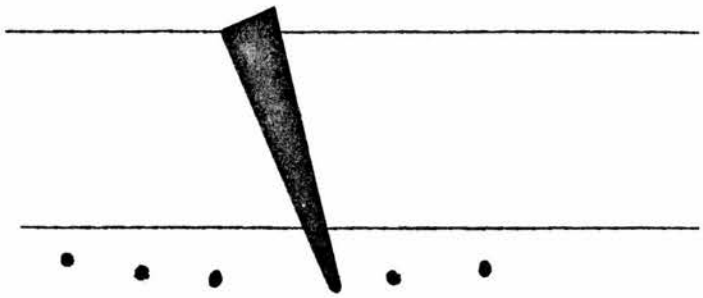
Excursion outside the stave is witnessed in the works of Kingdon (1958 a,b). The different staves which Kingdon distinguishes relate directly to the problem stated above of deciding how much is to be included in the normal speech category, and how abnormal forms are to be treated.

> 'The range which is in common use in speech may be divided into three portions, a central one which is used in normal unemphatic delivery, and above and below it narrower bands which are brought into use when emphatic syllables are being uttered. The full range of pitches which can be used by the human voice for speaking purposes can therefore be represented somewhat as follows:



> These ranges, even in the case of an individual speaker, are not fixed, either absolutely or relatively to one another. They may, according to circumstances, be shifted slightly up or down, or expanded or contracted to a moderate degree. Any expansion of the normal and emphatic ranges will, of course, be at the expense of the emotional ranges, the outer limits of which are physiologically fixed at the frequencies at which the individual's vocal cords refuse to vibrate, at the upper end of the scale because they cannot be placed under greater tension, and at the lower because they have reached the limit of possible relaxation.'
>
> (Kingdon, 1958 a:2)

Although he does not define it as such, Kingdon's emotional range may therefore be taken as equivalent to Laver's extreme pitch range, i.e. an uncontrollable, organically determined feature. Kingdon thus allows explicitly more flexibility in the range represented by staves than other writers probably do implicitly. Kingdon's estimate of the physical distance of the normal unemphatic range is one octave, and one and a half octaves for the emphatic range. Although this is more explicit, it still differs quite considerably from some writers' estimates quoted above. For transcription purposes, Kingdon uses only the normal unemphatic stave, and allows the transcription of emphatic utterances to extend beyond the limits of this stave - an excursion which few other writers allow; for example



It was in - cre - di - ble

(Kingdon, 1958 b:32)

A similar system is proposed by Brown (1977:127), who distinguishes between normal speaking range, "squeak" range and "growl" range. Excursion into the "squeak" or "growl" ranges is taken as the phonetic realisation of exposed emotion or personal attitude. Another possible correlate of excursion outside the normal range is a change in phonation type. An excursion into the "squeak" range may well be accompanied by a change (perhaps only momentary) to a tense, falsetto voice quality (see section 3.4.2.1). Similarly, a concomitant feature of the use of the "growl" range might be a change to a lax, breathy or creaky voice quality.

> '[Creaky voice] frequently accompanies very low pitches in intonation. ... Falsetto is not uncommon in reaching higher than normal pitches for expressive purposes, so an extra wide fall on Wonderful! may drop from high falsetto to very low creaky voice.'
>
> (O'Connor, 1973:267)

Naturally, these features can only be adopted as clues to such excursion if the voice quality components concerned are not already features of the speaker's normal voice quality.

The above discussion has been extended to some length to emphasise the fact that although writers are generally happy to use key-words such as those underlined above (control, skill of control, organic, omnipresent, phonetic, manipulable, learned, choice, habitual), these terms are not easy to define rigorously, and even less so to quantify.

At the beginning of this section, controllability was adopted as the basic criterion for the distinction. While keeping it as the criterion, I shall adopt Laver's (1975) usage of the terms intrinsic and extrinsic as labels for the two parts of the distinction. Although they are less mnemonically transparent, they carry none of the possibly misleading implications of all the other proposed terms. Thus intrinsic factors are the uncontrollable, organic ones, while extrinsic factors are the manipulable, learned, habitual ones.

Intrinsic factors are likely to be more important than extrinsic ones from the point of view of speaker characterisation because they are omnipresent and unchangeable. The great problem for speaker recognition (and for voice quality and voice dynamic theory; see sections 3.4.2.1 and 2) is that intrinsic and extrinsic factors may produce identical effects on the speaker's vocal output. This is well illustrated by the following three examples.

(1) Abercrombie (1967) mentions the possible diachronic influence of intrinsic conditions on extrinsic adjustments.

> 'A striking example of this is afforded by some urban slum communities where adenoids, due doubtless to malnutrition and lack of sunlight, are prevalent, with their consequent effect on [intrinsic] voice quality, but where people can be found with adenoidal voice quality who do not have adenoids - they have learnt the quality from the large number who do have them, so that they conform to what, for that community,

> has become the norm. (Continuing velic closure, together with velarisation, are the principal components needed for counterfeiting adenoidal voice quality). The accent of Liverpool seems to have had its origin in such circumstances.'
>
> (Abercrombie, 1967:94-5)

This is not a universally accepted explanation. For example, it does not account for the fact that adenoidal voice quality is characteristic of the speech of Liverpool but not of all such urban areas. Whilst it is not accepted fact, it is nevertheless a realistic possibility.

(2) Intrinsic factors cannot be manipulated by the speaker for any linguistic or paralinguistic purpose. However, since intrinsic and extrinsic factors may produce perceptually identical effects on the speaker's vocal output, there is the danger of misinterpretation on the part of the listener of the function of a particular feature of an utterance.

> 'For example, laryngitis or a heavy cold often result in a phonation type that sounds very similar to whispery voice. Most of us have probably had the experience of suffering from this condition, and having listeners reply to one's intrinsically whispery voice in whispery voice or whisper themselves, mistaking the physical index for an affective one, joining in the conspiratorially confidential interaction they thought was being signalled. Thus it may well also be the case that conclusions drawn by listeners about a speaker's personality can be influenced to some degree by the auditory similarity of features in his habitual voice quality to those of paraphonology.'
>
> (Laver, 1975:313; see also Laver & Trudgill, 1979)

In this case, how would a speaker with an intrinsically whispery phonation signal the paralinguistic markers of secrecy? One possibility is the substitution of an equivalent marker.

> '[It is interesting to consider] what substitution an individual might make in expounding a particular prosodic or paralinguistic feature if in fact his voice set was one which made habitual use of a parameter normally a key characteristic in the articulation of the feature in question.'
>
> (Crystal & Quirk, 1964:30)

A more realistic possibility is the reinforcement or exaggeration of the marker. For certain events, such as a cough, there may be great acoustic/auditory similarity between the physiologically caused intrinsic event on the one hand and the controllable extrinsic paralinguistic marker of scepticism on the other. It is perhaps for this reason that the events have to be exaggerated, and are therefore usually interpreted as somewhat jocular, in order to convey the paralinguistic intention.

Therefore, exaggeration or substitution must be employed for the conveying of the paralinguistic effect of a feature which also belongs to the extralinguistic component of a person's habitual speech.

> 'Whatever the signalling capacity any component of a voice set may have outside that voice set, this is neutralised for its owner, who may therefore have to exaggerate the component or substitute another in the given conventional function.'
>
> (Crystal & Quirk, 1964:11)

It is a problem for English speakers that the two distinct kinds of events described above are referred to by the same words. Lexemes such as <u>cough</u>, <u>swallow</u> and <u>sniff</u> may therefore refer either to intrinsically caused events, or to extrinsic paralinguistic markers of scepticism, amazement and disdain respectively. If the need arises, as it sometimes does in literature, the two interpretations can only be disambiguated by circumlocution (as in <u>He cleared</u>

his throat) or by the use of adverbials, as in the following examples.

(1) (intrinsic) He coughed convulsively.
(extrinsic) He coughed sceptically.

(2) (intrinsic) He swallowed involuntarily.
(extrinsic) He swallowed in amazement.

(3) (intrinsic) He sniffed to relieve his blocked-up nose.
(extrinsic) He sniffed disdainfully.

This seems to be true of other verbs in English which can have an intrinsic and an extrinsic sense (such unsavoury examples as gasp, snort, belch, etc.).

A problem for the definition of the distinction given above of controllability is that, although events such as extralinguistic coughs have their origin in uncontrollable, physiological factors, there is nevertheless a degree of control available to the speaker; he will typically be able to postpone the cough for a few seconds and will do this when he wants to finish what he is saying before breaking off (to avoid being interrupted for example). However, since the event is initiated by uncontrollable, physiological factors, this should be taken as the criterial feature for this intrinsically caused event.

(3) The acoustic/auditory similarity between certain intrinsically determined features and extrinsically determined features is exploited by professional impersonators, who use extrinsic voice features in order to produce a perceptually similar output to the impersonated person's extrinsic and intrinsic voice quality. This involves two kinds of process:

(i) the imitation of the impersonated person's voice quality - usually in addition, of course, to ("intrinsic" and "extrinsic") visual features of the speaker, and

(ii) the masking of the impersonator's own voice quality, except, of course, where it corresponds to that of the impersonated person. Abercrombie (1967) notes that:

> 'it is even possible to neutralise, by means of muscular adjustments, the components in voice quality which are anatomically derived - at least to some extent, and perhaps even, given enough skill, entirely. ... The extreme of virtuosity, probably, was reached by a certain music-hall performer, a large middle-aged man, who had learnt to produce, completely convincingly, the voice-quality of a seven-year-old girl, showing that it is possible to compensate, by muscular adjustments, for extreme anatomical differences.'
>
> (Abercrombie, 1967:94)

This is not to say that the impersonator's intrinsic voice quality is removed, but that its effect is minimised.

Although intrinsic and extrinsic factors may produce identical effects on a speaker's vocal output, it is desirable theoretically to treat them as two distinct categories; but it may be impossible from a practical point of view, such as that which must be adopted for automatic speaker recognition, to separate the two kinds of effect.

1.8.2 Inter-speaker and Intra-speaker Variability

The second distinction concerns the variability of the features which compose the speech signal. When a speaker repeats the same word, phrase or sentence, the second utterance will never be exactly the same in articulatory or acoustic terms as the first. The truth of such a statement made from the perceptual point of view

would depend somewhat on the sophistication of the individual listener's perceptual processes. It can be seen that the statement, even from the articulatory and acoustic points of view, presupposes a pseudo-procedure; that is, a method

> 'so arduous and time-consuming as a way of conducting an investigation that no one in their senses would ever set out to use it. If they did, they would certainly never carry it through.'
>
> (Abercrombie, 1963:9)

In short, no-one has measured a sufficiently large corpus of repetitions for this statement to be made as a categorical fact. However, it is a universally held opinion that such a procedure would reveal constant variation - a speaker never utters a phrase identically on two occasions. This difference is magnified if the two occasions in question involve a difference of style, or physical or emotional health of the speaker.

The second half of the distinction refers to the fact (again, an accepted fact based on the results of hypothetical exhaustive investigation) that no two speakers ever say a phrase in exactly the same way, articulatorily or acoustically. This is taken to be true even in cases of identical twins, professional impersonators, etc. The first type of variation is referred to as intra-speaker variation, the second as inter-speaker variation.

Both kinds of variation may derive from various sources, from extrinsic as well as intrinsic factors. Stevens (1972) reports that it is thought that the larynx is responsible for the greatest proportion of the intra- and inter-speaker variation in speech, although the significance of features elsewhere in the vocal apparatus is not small. The problem with such statements is that it is difficult to quantify this variability, and thereby to produce measures of the contributions of the different vocal organs.

There are various fields of research which exploit intra- and inter-speaker variability. Features with large intra-speaker variability are used in the determination of aspects of a speaker's physiological or emotional state from measurements on his speech. It is a criterion for a feature to be useful in speaker recognition, that it shows small intra-speaker but large inter-speaker variation, or at least that the inter-speaker variation far exceeds the intra-speaker (see sections 2.4 and 4.2). Thus if a feature varies substantially within one person's speech, and is not relatively distinctive between speakers, then it is improbable that that feature figures in the human speaker recognition process, or that it will be usable for automatic speaker recognition.

1.8.3 Potentially Usable and Habitually Used Parameters

It has just been argued that features with small intra-speaker and large inter-speaker variation differentiate optimally between speakers and are therefore very relevant to both the human and the machine speaker recognition processes. However, one cannot state on the above grounds that such features are the most relevant or important, since practical considerations play a role in this argument. For automatic speaker recognition, a parameter should ideally be easy to measure, unaffected by reasonable background noise, etc. (see section 2.4). It should likewise be remembered for the human process that, in most everyday circumstances, recognition is required in a minimal length of time for communication not to become seriously disrupted. An equally relevant criterion therefore is whether the value of a parameter for a particular speaker can be calculated from a short stretch of speech (see section 3.3). Practical considerations such as these therefore limit the importance of the criterion of inter- and intra-speaker variation.

Many of the experiments summarised in the review of the literature in Chapter 5 have similar formats: listeners are presented

with stimuli which have been modified acoustically so that one acoustic feature acts as the independent variable. Results generally show that this modification has a significant effect on listener performance, and the conclusion is drawn that the parameter therefore is a highly relevant one for human speaker recognition. However, such experiments ignore the fact that practical considerations such as those outlined above affect the relevance of parameters for human speaker recognition in everyday situations.

There is thus a basic distinction to be drawn between the probably quite large number of parameters which are potentially usable in speaker recognition and the subset of these parameters which human listeners habitually use to recognise speakers, or which have priority in the everyday situation. (This use of the term habitual, referring to parameters which listeners use in everyday situations, is quite different from its use as in habitual pitch range (section 1.8.1), referring to features which speakers use in normal unemphatic speech). This distinction is restated by Clarke et al. (1966) as a criticism which may be levelled against much of the speaker recognition experimentation.

> 'To show that human observers can use a particular type of information in discriminating among talkers does not demonstrate that they do use this type of information in typical situations.'
>
> (Clarke et al., 1966:42)

They extend this distinction to formulate a hypothesis regarding the flexibility and adaptability of the human faculty of speaker recognition -

> 'that the human observer uses a large number of characteristics of the waveform, each rather inefficiently, in distinguishing among talkers. It would appear that, as certain relevant

> information is eliminated from the signal, the listener selectively attends to remaining cues that are of value.'
>
> (Clarke *et al.*, 1966:54)

The point of view expressed in this hypothesis has been adopted for this thesis. The experiments referred to above only investigate the former category of potentially usable parameters. It is hoped that this thesis sheds some light on the nature of the latter category of habitually used parameters. This is possible only if the practical considerations are examined by which the habitually used parameters are selected from the potentially usable parameters. Much of the thesis is therefore devoted to a necessarily tentative discussion of these considerations. In section 3.5 an approximate theoretically based specification is presented of the most important habitually used parameters (*first-order* parameters). The research reported in Chapter 5 constitutes an experimental approach to the specification of first-order parameters.

CHAPTER 2

THEORETICAL FRAMEWORK

CHAPTER 2

THEORETICAL FRAMEWORK

2.1 INTRODUCTION

This chapter contains not a review of the literature, but rather a review of the categorisations of experimental paradigms and theoretical considerations for speaker recognition, contained in the literature. The review of the literature itself - that is, a description of experiments performed and of their findings - follows in Chapter 5. It is delayed until then since the categorisations contained in this chapter and Chapters 3 and 4 serve to highlight the inadequacies in several areas of the literature, especially in regard to human auditory speaker recognition.

It ought to be pointed out straight away that relatively few categorisations of this sort have been presented in the literature. It is assumed that investigators have given some consideration to the theoretical issues underlying speaker recognition experimentation although the number of writers who have put such theorising on the experimental framework into print is very small. Therefore, credit is due to those writers who are quoted in this chapter, for having made their theoretical work explicit.

2.2 BRICKER & PRUZANSKY'S ANALYSIS OF EXPERIMENTAL SPEAKER RECOGNITION VARIABLES

Bricker & Pruzansky's (1976) summary of the literature (hereafter simply B & P) contains the most thorough categorisations yet proposed of the speaker recognition process, with the possible exception of Hecker (1971). Figure 2.1 contains their representation of the various features which can act as variables in experimental speaker recognition tasks, and how these relate to the various components of the speaker recognition process chain.

The top line represents an analysis of the speaker recognition process into its component stages. This line corresponds closely to the traditional analysis of the speech chain and may be divided simply into the three broad categories of phonetic description - articulatory, acoustic and auditory/ perceptual.

Speaker information (section 1.2) is present at each stage of the speaker recognition process and is transformed before it is passed on to the next stage. The speaker information contained in B & P's diagram belongs to the category of potentially usable, not habitually used, parameters (see section 1.8.3).

The second line of Figure 2.1 shows the form which this speaker information takes at each stage.

> 'Speaker information is latent in the speaker in the form of anatomical features and neurally stored habit patterns (Garvin & Ladefoged, 1963). It is converted to activity of the speaking apparatus, which we shall call speech gestures, as the talker forms an utterance. This activity results in an acoustic signal in which speaker information is encoded as an extra message (Peters, 1954) that eventually reaches the ear of the listener along with the speech intelligence. The listener's sensory apparatus converts the acoustic signal to neural ... messages which, in turn, serve as input to a perceptual processor. The processor is specialised for converting sensory speaker information into perceptual data for use by a decision-making process. ... In this form, speaker information is used by some decision process to arrive at a response for every stimulus.'
>
> (p.297)

In the third line of Figure 2.1 are shown those operational elements which are manipulable in the experimental speaker recognition situation, and the close correlation which holds between them and the

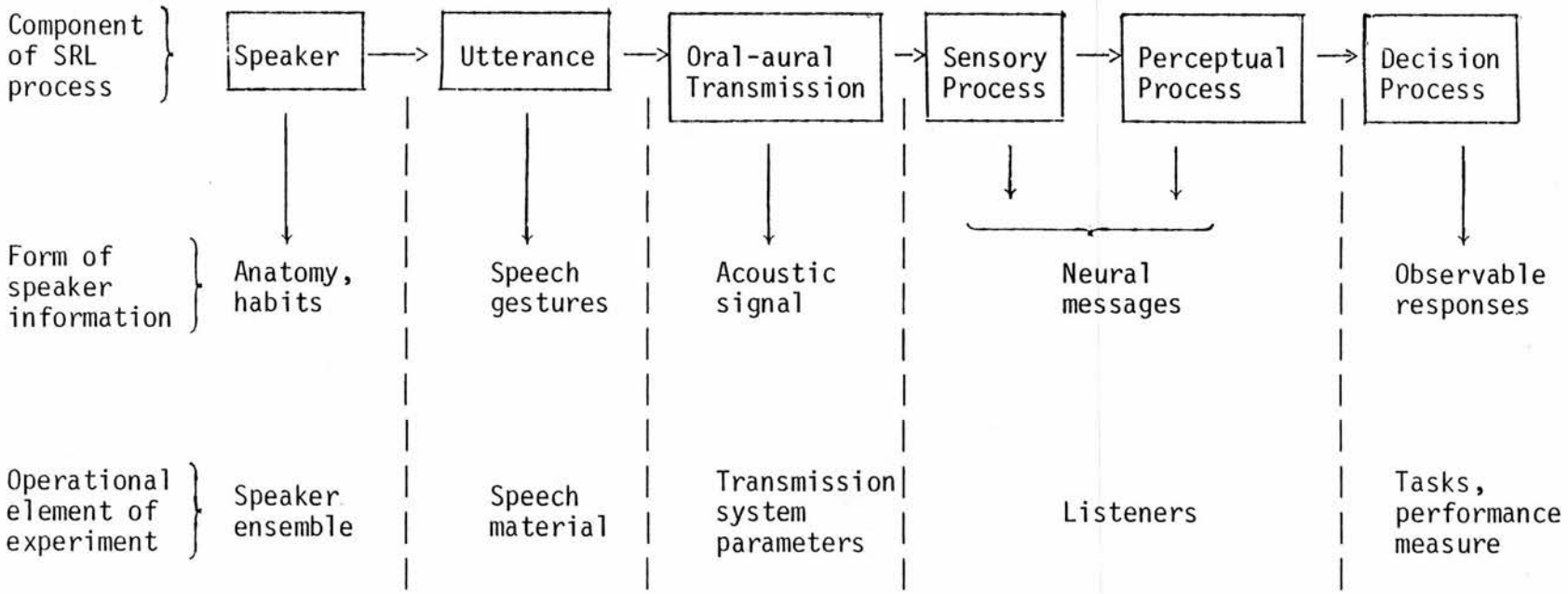


Figure 2.1 Schematic representation of the process of speaker recognition by listening, with related forms of speaker information and experimental operations (from Bricker & Pruzansky, 1976:298).

component forms of speaker information given in the second line. The speaker recognition experimenter must make conscious decisions in regard to all these operational elements for an experiment, and not merely to those which are being exploited as the variables under investigation.

B & P note that although such an analysis produces a regular, symmetrical representation of the process, it does not reflect the bias which exists as far as accessibility and consequent exemplification in the literature are concerned.

> 'The speech signal and the responses can be observed and measured directly. In contrast, the perceptual process is preceded by sensory processing and followed by a decision process, each of which performs its own transformation. Very little of the work we shall review has in fact dealt directly with the perceptual process.'
>
> (p.298)

I am in agreement with all parts of B & P's categorisation of their conceptual framework for speaker recognition experimentation.

2.3 BRICKER & PRUZANSKY'S TAXONOMY OF EXPERIMENTAL SPEAKER RECOGNITION TASKS

The tasks involved in speaker recognition experiments are the subject of a similar taxonomy contained in B & P. Criticisms of this taxonomy will be presented in this chapter; however, a more logical place for the revised version of B & P's taxonomy, which I propose, is in Chapter 4 (section 4.3), where it forms an integral part of the description of the formal, logical model of the speaker recognition process which takes place in the listener in the experimental situation. It is argued there that any speaker recognition model must take into account the distinctions which form the basis of that categorisation.

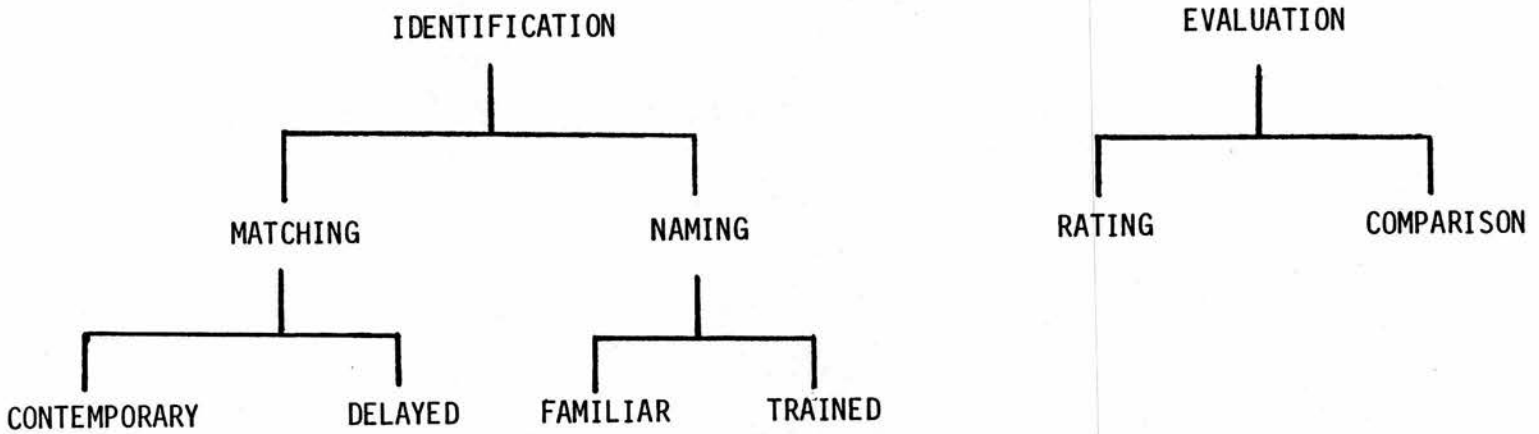


Figure 2.2 A taxonomy of speaker recognition tasks (from Bricker & Pruzansky, 1976:302).

The major criteria for the distinctions of B & P's taxonomy are stated in terms of two characteristics:

(i) the type of judgment required, and
(ii) the manner of presenting voice samples.

Their hierarchical taxonomy is represented by Figure 2.2.

The first distinction made by B & P is between tasks of identification and of evaluation. This distinction is made by reference to the nature of the response given by the listener.

> 'We classify as an identification task any in which some or all of the available responses denote an individual speaker. The term evaluation is applied to tasks that require the listener to judge the value of the stimulus-voice on some attribute, dimension, or characteristic. The accuracy criterion intrinsic to identification tasks (i.e. the scorability of identifications as to correctness) generally does not obtain for evaluation tasks.'
>
> (p.301)

B & P define matching tasks as ones

> 'in which the comparison stimuli that define the response categories are presented during the trial on which the test stimulus is judged.'
>
> (p.302).

I shall use the term reference voice sample to refer to what B & P describe above as 'stimuli defining the response categories', and stimulus voice sample for their 'test stimulus'.

> 'In the naming task, the listener's experience with voice samples corresponding to the response alternatives has been acquired prior to the judgment trial.'
>
> (p.302)

B & P consider matching tasks to entail short-term memory while naming tasks involve long-term memory.

Matching tasks are of two kinds:

> 'The delayed testing branch under matching ... is exemplified only in the work of McGehee (1937,1944), who presented the comparison stimuli after delays ranging from one day to five months. Contemporary matching experiments offer the listener one or more comparison stimuli during the trial on which the test stimulus is presented.'
>
> (p.303)

Naming tasks are also subdivided into two categories:

> 'In the variant used by most investigators, ... listeners had become familiar with the voices of the talker set prior to the time of the experiment through normal business and/or social contact. We shall refer to this task as familiar speaker naming. In the version of the naming task used by Williams (1964), learning of the talker set was under the control of the experimenter, who used a paired-associate task to train the listeners.'
>
> (p.302)

B & P subcategorise evaluation into rating and comparison tasks. The former involve 'a single voice sample at a time and one or more attribute scales' (p.302). These scales range from the objectively verifiable, such as sex and age, to the purely subjective, typically using the semantic differential technique of Osgood et al. (1957). This is composed of bipolar adjective scales, such as beautiful - ugly, along whose axes listeners are required to rate voice samples (see, for example, Voiers, 1964; Holmgren, 1967). B & P note that

> 'the literature contains no examples of comparative evaluations of two or more samples, although such procedures (e.g. pair-wise similarity judgments) have been used in the closely related area of speech transmission system quality (McDermott, 1969).'
>
> (p.302)

Few arguments can be levelled against the evaluation branch of B & P's taxonomy. Evaluation tasks are so different in nature from identification tasks that they do not warrant being treated as hyponymous elements. They are not of central interest to the argument of this thesis and, to this extent, the treatment afforded them in B & P's taxonomy may be taken as satisfactory. However, certain comments are appropriate.

The fact that no examples of comparative evaluation tasks exist in the literature does not mean that they do not merit a place in this taxonomy. But it should not be thought that comparative evaluation tasks constitute a totally independent category from rating evaluation tasks. They both involve much the same process of judging voice samples on attributional scales. For rating tasks, the judgments are of an absolute nature, whereas for comparison tasks they are relative.

In addition, one can derive comparative measures from rating judgments. Thus, if speaker A's voice is rated as "very beautiful" on the beautiful - ugly dimension, while speaker B's is rated as only "slightly beautiful", it is reasonable to assume that speaker A's voice would be judged "more beautiful" than speaker B's in a comparative evaluation task. I should like to avoid the essentially semantic argument as to whether the converse relationship holds, and simply conclude that we may derive comparative evaluations from absolute rating evaluations, but not necessarily vice versa.

The most reliable criterion for the evaluation/identification distinction is the main definition given by B & P, that an identification response denotes an individual speaker - in short, identification tasks involve reference voice patterns; evaluation tasks do not. Their secondary definition (that it is inappropriate to talk of the accuracy of an evaluation response) holds only for

subjective measures of the beautiful - ugly kind. It does not hold for objective measures such as sex and age, especially the former since it is not even a scalar dimension but a binary factor.

An additional characteristic, which may be used as an identifying feature for this distinction, is that the time factor (in the sense of exposure to the voice samples) is not crucial to evaluation tasks, but is very often a factor exploited by experimenters in identification tasks.

The evaluation branch of B & P's taxonomy will not be discussed further here, for reasons given above. More serious criticism, however, may be brought against their categorisation of identification tasks.

Firstly, it is not clear what B & P mean by the term trial. It is certainly being used in a different way from that of experimental psychologists. The psychologists' definition of the term entails whatever preliminary operations and stimulus presentations lead to a single listener response. The definition does not involve the notion of time-delay; thus all the preliminary operations, etc. constitute the same single trial irrespective of any time-delay. This is obviously different from the way in which B & P use the term. Interpreting the term in the sense described above, B & P's definition of naming tasks becomes self-contradictory, and of matching and contemporary tasks becomes tautologous.

B & P's definitions of matching tasks and of the subsumed category of contemporary tasks seem to be identical. Matching tasks are defined as ones 'in which the comparison stimuli that define the response categories are presented during the trial on which the test stimulus is judged' (p.302). Contemporary tasks 'offer the listener one or more comparison stimuli during the trial on which the test stimulus is presented' (p.303).

The major criticism against B & P's taxonomy arises from an incompatibility of categories. Matching tasks are considered to involve short-term memory. However, B & P note that

> 'the delayed testing branch under matching ... is exemplified only in the work of McGehee (1937,1944), who presented the comparison stimuli after delays ranging from one day to five months.'
>
> (p.303)

Psychologists' estimates of the time-span of short-term memory vary widely, from a few seconds to a few minutes. Few would agree with B & P, though, that a task entailing a delay of five months can involve short-term memory alone.

In summary, there are two main criticisms of the identification part of B & P's taxonomy:

(i) McGehee's delayed testing experiments cannot be categorised as short-term memory tasks, and

(ii) the term trial is not defined, but is not being used in a way consistent with its widely accepted meaning.

These arguments will be discussed again in section 4.3, where a revised version of B & P's taxonomy is presented, which avoids these inconsistencies.

2.4 WOLF'S CRITERIA FOR EFFICIENT ACOUSTIC PARAMETERS FOR AUTOMATIC SPEAKER RECOGNITION

The final categorisation to be considered in this chapter is Wolf's (1972) description of criterial attributes for efficient acoustic parameters for automatic speaker recognition. Other writers on automatic speaker recognition have proposed similar criteria (e.g. Floyd, 1964); Wolf's treatment is chosen for discussion here because it is the fullest of these classifications. It should be pointed out that these criterial attributes refer

specifically to automatic speaker recognition and not to the human speaker recognition process which is the main concern of this thesis. However, it is enlightening to relate the principles underlying these criteria to similar considerations underlying features of the human speaker recognition process - if only to emphasise the difference between these two sets of principles and thereby the fact that the results from automatic speaker recognition experiments are of little relevance to investigations of the human speaker recognition process. The discussions of this section are taken up again, therefore, in section 4.2 where the discussion of considerations underlying the human speaker recognition process forms a part of the framework of the model of the speaker recognition process described there. In relation to the human process, it is inappropriate to refer to "criterial" attributes except if one interprets criterial attributes as those characteristics which are habitually attended to (see section 1.8.3).

Wolf (1972) specifies six attributes which he considers criterial for efficient parameters for automatic speaker recognition:

> 'Ideally, the speech characteristics measured should;
>
> (i) occur naturally and frequently in normal speech,
>
> (ii) be easily measurable,
>
> (iii) vary as much as possible among speakers, but be as consistent as possible for each speaker,
>
> (iv) not change over time or be affected by the speaker's health,
>
> (v) not be affected by reasonable background noise nor depend on specific transmission characteristics, and
>
> (vi) not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice.'
>
> (p.2044-5)

The above criterial attributes are ordered in this way for one reason which Wolf gives, but another reason, which he at least does not make explicit, can be proposed for the ordering. The reason which Wolf gives is determined by practical considerations. He realistically admits that

> 'In practice, the simultaneous fulfilment of all these criteria is probably beyond the present state of the art. Partial or complete relaxation of some of these standards is reasonable for some research purposes and for limited practical speaker recognition. Specifically, in the research described here, the last three factors were not investigated, but were controlled.'
>
> (p.2045)

The unsatisfactory nature of this situation is rather played down by Wolf. The experimental situation, where factors may be controlled, is quite different from the practical situation, such as the use of an automatic speaker recognition device in a security system. In the latter situation, the above factors cannot necessarily be controlled and the last three of Wolf's attributes may have as important an effect on performance as the first three.

The other reason for the ordering, which I hypothesise, results from the fact that all the attributes are of a scalar rather than a binary nature. Therefore, for example, given that two parameters fulfil the other conditions equally well, the one which occurs more frequently in normal speech is preferable to the other, which occurs less frequently. The ordering of the attributes can then be taken to imply a weighting of their relative importance. To give another example, all other things being equal, a parameter which occurs frequently in normal speech is probably preferable to one which occurs less frequently but which is more easily measured.

There is an interaction between these two explanations, the one explicit and the other hypothesised. The hypothesised

explanation states that, other things being equal, attributes fulfilling criteria high in the list carry greatest weighting. However, the explicit reason states that those conditions low in the list are the most difficult to fulfil in practice. In other words, the presupposition of the first explanation (that other things are equal) will rarely hold. From Wolf's practical point of view, however, the latter is best taken as the reasoning behind the order.

Let us consider each of the six attributes in turn. It will be seen that they are not as consistent or homogeneous with each other from a theoretical point of view as might be supposed at first sight.

The first of Wolf's attributes - that a characteristic should occur naturally and frequently in normal speech - is an obvious prerequisite for any realistic form of speaker recognition process. However, precisely what is meant by normal speech is not discussed. This is not a criticism aimed specifically at Wolf because no writer, practical or theoretical, has tackled this problem by giving a satisfactory discussion of what normal speech refers to. The scope generally implied by normal speech in the context of speaker recognition (and other phonetic fields) includes that the speaker is in his usual state of emotional and physical health, that he is not trying to disguise his voice or imitate another speaker's, etc. These are exactly the situations treated in Wolf's later conditions.

One might similarly wonder whether it is reasonable to assume that a speaker <u>can</u> speak normally or naturally in the practical, automatic speaker recognition situation. In such a situation, the speaker knows that he is speaking into a microphone, and that what he says (or, more strictly, the way in which he says what he says) will be analysed. His knowledge of this will inevitably

have some effect on the way in which he says the utterance. In short, the easiest way to prevent someone from speaking naturally is to ask him to speak naturally. This is the same problem as occurs in most experimental situations and is one which is difficult to avoid. For automatic speaker recognition it is probably less of a problem than it is for phonetic research into normal, natural forms of speech. As long as a person's forced natural speech shows as much inter-speaker variation and intra-speaker consistency as his unforced natural style, then the practical difference will be minimal.

The second and fifth attributes are included as a consequence of the fact that these criteria refer specifically to automatic speaker recognition which is inevitably subject to the limitation imposed by practical requirements. These conditions are logically necessary; if the acoustic parameter to be considered cannot be registered by the machine because of interference or other, mechanical limitations, the speaker recognition process has no input and so cannot proceed.

The third condition - that the characteristic should vary as much as possible between speakers, but be as consistent as possible for each speaker - corresponds to the distinction discussed in section 1.8.2 of inter-speaker and intra-speaker variability. Naturally, the efficiency of this criterion depends largely on the threshold of variability chosen. It should be remembered that the concepts of inter- and intra-speaker variability refer specifically to how speakers <u>can</u> be recognised (which is therefore the relevant factor for automatic speaker recognition) rather than how they habitually <u>are</u> recognised.

The fourth condition states that characteristics should not change over time nor be affected by the speaker's health. Although the two parts of the fourth condition may appear at first to be homogeneous, it can be seen that those features which are likely to

fulfil the first part of the condition will be different from those likely to fulfil the second. Intrinsically determined features remain relatively stable over time whereas it is extrinsically determined features which are least affected by changes in physical health.

Referring to Wolf's conditions as a whole, there are various assumptions underlying these criteria. These will be summarised now because they constitute presuppositions about the kinds of parameter desirable for automatic speaker recognition. In section 4.2, these presuppositions are compared with those hypothesised for speaker recognition by humans, and it is shown that a radically different approach must be adopted.

The first of these assumptions is that the number of acoustic parameters which need to be considered for successful automatic speaker recognition is very small. This assumption is not stated explicitly in Wolf, but is evidenced by most of the experiments described in the literature, where experimenters have found that the analysis of sometimes only one acoustic parameter or the cross-correlation of a very small number of usually related parameters is required for the successful recognition of a limited population of speakers. An extended version of this assumption is that it is not accepted that an acoustic parameter which is not distinctive between two speakers can be used as a parameter for automatic speaker recognition (cf. Bricker & Pruzansky's (1976) definition of speaker information quoted in section 1.2).

Secondly, it is assumed that, as an objective measure (as is required for automatic speaker recognition), the acoustic parameter selected should be virtually constant and stable. That is, it should not vary over a moderate length of time, be affected by changes in the speaker's health, or be susceptible to disguise or imitation, etc. This is obviously a very stringent criterion

and it may be found that no parameter exhibits total stability. Adaptive systems, which are capable of modifying their acceptance criteria on the basis of acquired experience, may therefore be needed.

Thirdly, there is the presupposition that the system which is being aimed at is in some sense perfect. For automatic speaker recognition, perfection does not necessarily mean that all speakers' voices subjected to the system are correctly recognised in 100% of cases. Instead, in the practical situation of the use of automatic speaker recognition devices in security systems, the criterial factor of success is that the device does not allow access to impostors, i.e. that there is a minimal percentage of incorrect acceptances of stimulus voices (see section 4.4.2). In this situation, a "no decision" response is an alternative to the "speaker A" etc. and "stimulus does not correspond to any of the reference voices" responses, and a system with a low incorrect acceptance but high "no decision" score is preferable to one with a low "no decision" but high incorrect acceptance score.

Fourthly, there are the logically necessary practical considerations contained in the second and fifth attributes.

Lastly, there is the implicit suggestion that a consideration of such theoretical phonetic notions as the intrinsic/extrinsic distinction is not necessary. From a practical point of view and in the context of automatic speaker recognition (which is Wolf's standpoint), this is perhaps acceptable. But from the theoretical phonetic point of view and in the context of speaker recognition by humans (which is the standpoint of this thesis), it is not.

CHAPTER 3

SPEAKER-DEPENDENT FACTORS

CHAPTER 3

S P E A K E R - D E P E N D E N T F A C T O R S

3.1 INTRODUCTION

In this chapter discussion will be focussed on two related aspects of the speaker recognition process: firstly, on those elements of a speaker's total vocal output which contain information which is potentially speaker-characterising, and secondly on the acoustic parameters (and their perceptual equivalents) of this output. These parameters convey the information contained in the above elements. Also examined are the various ways in which the information conveyed may be considered strongly or weakly characteristic, so that the practical importance of all these potentially usable features in everyday speaker recognition may be specified in broad terms.

3.2 SIGNAL AND INDEX

A distinction must be made for most phonetic purposes between the different kinds of information from a semiotic point of view which are conveyed by different aspects of the stream of speech. These fall into two categories.

Firstly, when a speaker articulates an utterance in an interactional situation, he is using this utterance to convey meaning-bearing units to the listener. The reason for the interaction may be the transfer of information or may be much more phatic; but, whatever the reason, linguistic units are transmitted by the speaker and perceived by the listener. In this sense the utterance acts as a signal.

Secondly, when the speaker articulates the utterance, the listener also infers information about the speaker - his identity, sex, age, health, social status, etc. This kind of information conveyed by the utterance does not fulfil any linguistic function but acts as an index (Peirce, 1940; Abercrombie, 1967; Laver, 1968; see section 1.3). Thus, on hearing the utterance, the listener performs two processes - he extracts the criterial features which make up the meaning-bearing linguistic units, and also the speaker-characterising features which influence his perception of the physical, social, psychological and idiosyncratic characteristics of the speaker. From the phonetic standpoint of this thesis, the two forms of information are conveyed by theoretically distinct strands of the stream of speech, although their effects may not necessarily be physically separable. The strand which conveys the signal is of limited interest to the field of speaker recognition (see section 1.6).

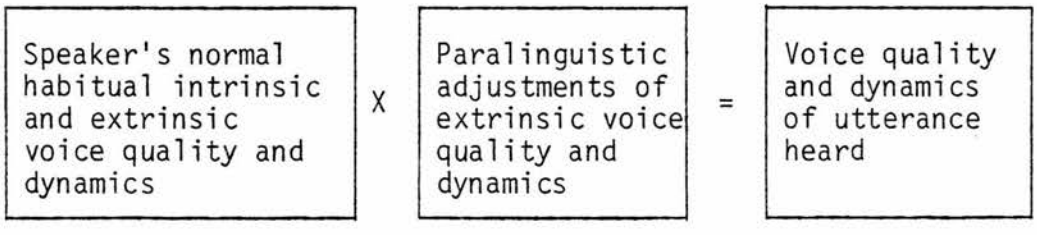
With regard to speaker recognition, it is useful to subdivide the indexical half of this distinction further. The two subdivisions are perhaps best explained by an example.

Speaker X says an utterance in an irritated manner, and uses a harsh phonation type with a clenched teeth setting and narrow loudness deviation. This utterance will then convey two indexical kinds of speaker-characterising information. Both will enable the listener to recognise the speaker correctly as X as against speaker Y, but with differing degrees of importance.

Firstly, the listener may know the intrinsic and extrinsic voice qualities and dynamics characteristic of speakers X and Y, i.e. have reference voice patterns for these two speakers, stored in his brain. The listener will find that the parameter values extracted from the utterance under consideration correspond more closely to those contained in the reference voice pattern for speaker X than for speaker Y. We may refer to this kind of index as parametric information.

Secondly, the listener may know that when speaker X is irritated, he typically uses the setting contained in the utterance heard (a harsh phonation type with a clenched teeth setting and narrow loudness deviation), whereas this is not the case for speaker Y. Thus the correct recognition of the speaker as X as against Y is facilitated by the fact that the utterance is in a voice quality and dynamics which is typical for X in that particular situation, but not for Y. We may call this second category information of frequency-of-occurrence. The criterion of frequency-of-occurrence may be interpreted in different ways (see section 3.3).

There is a problem inherent in this explanation of parametric and frequency-of-occurrence information. It has been assumed that the listener is able to realise that the feature of irritation is being unambiguously conveyed by paralinguistic markers. However, it is worth examining how the listener may be able to deduce this. The relationship which exists between the utterance heard and the paralinguistic markers of irritation may be expressed by the following equation.



The voice quality and dynamics of the utterance heard is expressed as a product rather than as a sum, to imply that there is an interactive effect between the preceding two factors and that their effects may not be easily separable. The problem derives from the fact that, for the listener, the first two of these factors may be

unknown - only the third factor, the voice quality and dynamics of the utterance heard, is necessarily immediately deduceable by the listener from the utterance. Since there are more than one unknown factors, guesswork based on expectation may be required on the part of the listener. Three possible situations may obtain in typical circumstances.

(i) the listener may know both the speaker's normal habitual voice quality and dynamics and the paralinguistic adjustments which are typical for him. Since the product of these two factors corresponds to the voice quality and dynamics of the utterance heard, the listener can assume with a measure of certainty that a paralinguistic feature is being conveyed. In this case, the listener will have little difficulty in recognising the speaker.

(ii) the listener may know the speaker's normal habitual voice quality and dynamics but not the paralinguistic markers of irritation which are typical for him. The listener will again have little difficulty in recognising the speaker since paralinguistic adjustments rarely alter the speaker's voice quality and dynamics radically ("beyond recognition"). Also, one's expectations of the adjustments entailed in particular paralinguistic effects are quite well specified, since these adjustments are governed largely by cultural conventions.

(iii) the listener may know neither the speaker's normal habitual voice quality and dynamics nor his typical paralinguistic adjustments. In this case, recognition on parametric information is, of course, impossible since the listener has no appropriate reference voice pattern stored in memory for the speaker.

In usual circumstances, the fourth situation will be impossible, i.e. that a listener can know the paralinguistic adjustments

which are typical for a speaker without also knowing that speaker's normal voice quality and dynamics.

If the listener does not have an appropriate reference pattern either for the speaker's normal habitual voice quality and dynamics or for his paralinguistic adjustments, it is interesting to consider how these are acquired from exposure to a sample of the speaker's irritated speech. Two sequentially ordered processes are performed. Firstly, the listener has to infer whether a paralinguistic adjustment is being used or not. He can do this from several sources of information; most of them, including semantic and syntactic ones, will derive from the physical and spoken context. These may be strengthened by phonetic expectations based on culturally determined norms. If harshness is the norm as a marker of irritation and is not the norm for normal voice quality in the particular speech community, then the listener may expect an instance of harsh voice quality in the majority of cases to indicate that paralinguistic intention. Since it is possible for a harsh phonation type to be a feature of the speaker's normal voice quality, the listener runs the not-too-great risk of misinterpretation (see section 1.8.1 for a quoted example of this) although these instances will be much rarer than those of correct interpretation.

Secondly, having made a reasoned guess that a paralinguistic feature is being employed, the listener may subtract the effects of this from the utterance heard to arrive at a rough approximation of the speaker's normal voice quality and dynamics. This specification will be an approximation since the paralinguistic adjustments can only be specified precisely when the speaker's normal habitual voice quality and dynamics are known. In the above situation, the adjustments can only be specified in as much detail as is allowed by the cultural norms, by which it was inferred in the first place that a paralinguistic adjustment was involved.

Therefore a speaker's normal voice quality and dynamics can only be specified in one of two ways. Firstly, by hearing the speaker in a situation where it is assumed that no paralinguistic effects are involved, or secondly, and less reliably, by deducing the normal habitual voice quality as the common denominator of a series of instances where paralinguistic influences may be at play.

3.3 RELATIVE STRENGTHS OF SPEAKER-CHARACTERISING FEATURES

Having seen that the large number of potentially usable speaker-characterising features may be significantly divided into at least the two major categories above of parametric and frequency-of-occurrence information, the various ways may now be examined in which the characterisation afforded by these indexical features may be said to be strong or weak. This should not be taken to be an exhaustive typology, but rather a discussion of some of the major kinds. The factors discussed do not constitute discrete categories but are generally overlapping in scope and interrelated. These strength criteria relate to one of two things: either to the elements of the categorisation contained in the next section, or to the acoustic/perceptual parameters of section 3.4.2. The parameters constitute the manifestation of the former elements.

(1) The first of Wolf's (1972) criterial attributes for efficient parameters for automatic speaker recognition (see sections 2.4 and 4.2) states that characteristics to be used should occur naturally and frequently in normal speech. Applying this criterion to speaker recognition by humans, a first strength hierarchy may be expressed in terms of the frequency with which a feature conveying indexical information occurs in normal speech. Allusion was made to this notion in the previous section, but in a vague way, and it is worth examining it in greater detail here. It is possible to equate at least three situations with the criterion of frequency-of-occurrence.

(i) Characterising strength increases in relation to the frequency with which an element or parameter occurs in normal speech. For example, paralinguistic effects are rarer in speech than extralinguistic elements, many of which are omnipresent. The paralinguistic category is therefore weaker in this respect than the extralinguistic. Similarly, parameters of formant structure, which is present and deducible in all phonetically voiced sounds and therefore occurs in a large proportion of a speaker's vocal output, are stronger than parameters such as those concerned with the nature of the friction in an [s] articulation, since this segment occurs relatively infrequently in normal speech.

(ii) Characterising strength decreases in relation to the length of utterance required for the calculation of a parameter. By this criterion, formant structure parameters are again very strong since they are calculable from segment-length utterances; whilst good examples of parameters which are weak according to it are temporal parameters of tempo and rhythmicality, which require much longer stretches (see sections 3.4.2.2.3 and 4; Appendix 1).

(iii) Characterising strength is greater for parameters reflecting parametric information than for frequency-of-occurrence information, for reasons allied to the above two factors. Thus parametric information of, for example, formant structure, which is calculable from segment-length utterances, is stronger than frequency-of-occurrence information, such as continuity (the occurrence of pauses in speech).

(2) A further kind of strength may be defined in terms of the factors from which the parameters conveying the indexical information derive. Thus, intrinsic factors may be more important for speaker recognition than extrinsic ones since they are by definition omnipresent and not manipulable; this is an intuitive suggestion but one which is difficult to verify since the effects of intrinsic factors are rarely separable physically from those of

extrinsic. However, assuming for the present purposes that this is the case, then a strength relationship can be stated where parameters which were most reliable as indicators of intrinsic factors would be stronger than those which were not. Again, this criterion would be related to the first one above, of frequency. Intrinsic factors are present in all forms of a speaker's vocal output and therefore parameters reflecting intrinsic factors occur very frequently in speech.

It should be noted that "intrinsic factors" in the above description refers only to long-term intrinsic factors, not short-term (such as a head-cold; see section 1.8.1). In speaker recognition, the effects of short-term intrinsic factors are of limited value. They have no strength other than (a) a weak form of non-phonetic frequency-of-occurrence strength - for example, if the listener knows that speaker X often has headcolds, and (b) a strength enabled only by a shift of norm - for example, if the listener knows that speaker X has a headcold at the present time, and hears an utterance containing what are obviously the acoustic effects of a headcold. However, in this last example, the headcold can no longer be considered to be exhibiting the typical properties of a short-term factor since it has now, for the listener, become the norm of reference for the voice pattern of that speaker.

(3) The next kind of strength relationship is perhaps best illustrated by an example. Let us suppose that the occurrence of silences is a feature, the discrimination afforded by which is polarised in relation to speaker recognition, in the sense that there is only a small proportion of speakers who can be characterised as having vocal outputs containing a significant number of stretches of silence. For the remainder of speakers (the majority), the discrimination afforded by this feature is poor and therefore the occurrence of silences is unlikely to be useful as a parameter for the recognition of these speakers. However, it can be seen that

since only a few speakers are able to be discriminated by this parameter, the fact that any one speaker is amongst the small population who are discriminated by this parameter will be a strong form of characterisation for that particular speaker. This kind of strength will not be discussed further because native speakers, although understanding the theory of this criterion, do not seem to have strong intuitions as to which parameters are characterised by this polarised discrimination. The occurrence of silences is selected above merely as a plausible but equally arguable candidate.

The above kind of strength dealt with the case where a parameter allowed a polarised discrimination. Let us now consider the case of parameters where the discrimination afforded is not so polarised, but instead is quite evenly distributed. If the discrimination is even, but small, one would expect the intra-speaker variability also to be small, i.e. the value of a reference speaker's speech in relation to that parameter would be relatively stable. If the discrimination is even, but large, one would expect the intra-speaker variability to be similarly large, i.e. the value is relatively unstable. It is for this reason that the reference values are considered in section 4.6.2 to be composed of two parts: (i) a value proper, being an absolute, standard measure for a particular reference speaker's speech in relation to a parameter, and (ii) a threshold, which is a measure of the variation possible for that speaker in relation to the parameter. With regard to the present discussion, parameters allowing a large range of values (and therefore greater discrimination between those values) for reference speakers will probably also entail similarly large thresholds. And, at the other extreme, parameters allowing a small range of values (and discriminability) for reference speakers will involve correspondingly small thresholds. Because of this, neither kind of parameter can be said to be stronger than the other.

(4) The final strength criterion which I shall discuss relates to the hypothesis contained in Brown (forthcoming; reproduced as Appendix 1). This states that the temporal parameters of tempo, rhythmicality and continuity are governed by the speaker's cognitive processes, as opposed to the physiological determination of the other parameters of pitch, loudness, etc. Cognitively governed parameters might be expected to be relatively stable and in this sense stronger, it is argued, since cognitive adjustment is more difficult than adjustment by physiological manipulation.

It can be seen that certain of these strength criteria oppose each other, in that parameters strong by one criterion are weak by another. For example, tempo parameters are strong by the cognitive/physiological distinction just described, but are weak in that the stretch of sample required for their calculation is relatively long. It is therefore desirable to rank the relative strength of these strength criteria which depends to a large extent on whether one is dealing with human or automatic speaker recognition. In section 4.2 the differences are examined between two processes, and between the kinds of parameters which are desirable for automatic speaker recognition and those which are probable in the human process. The basic difference is that stability and reliability are important in machine parameters while human speaker recognition parameters need to be immediately calculable. The various kinds of frequency strength ([1] above) therefore rank highly for the human process while those parameters which reflect relatively stable factors (in [2] and [4] above) will be more important for automatic speaker recognition.

3.4 THE SPEECH SIGNAL

3.4.1 Elements

3.4.1.1 Linguistic, paralinguistic and extralinguistic Elements

It is useful in considering speaker recognition to divide the speaker's vocal output according to two categorisations. The first of these is a categorisation of function into linguistic, paralinguistic and extralinguistic. Although it is to most people a traditional categorisation, there are still no universally agreed definitions for the parts of the trichotomy. The main disagreement lies in the definition of paralinguistic elements.

By linguistic elements is meant all those elements of a speaker's output which serve to convey information encoded in an arbitrary, conventionally determined (i.e. grammatical) way. Linguistic elements are therefore produced as the manifestation of those features which constitute the terminal nodes of the speaker's productive grammar.

Extralinguistic elements comprise all those parts of a speaker's output which, whether manipulable or not, are not being used by the speaker intentionally to convey any information.

Paralinguistic elements are more problematical and have been defined in different ways by different writers. Trager (1958, 1964) categorises paralanguage as consisting of vocalisations ('variegated ... noises, not having the structure of language' (1958:4)) and voice qualities, which modify both vocalisations and language sounds. Vocalisations consist of vocal characterisers (such as laughing and crying), vocal qualifiers (involving the raising or lowering in the ranges of intensity, pitch height and extent) and vocal segregates (segmental articulations such as the uh-uh of negation or the uh-huh of affirmation). The voice qualities are

composed of pitch range, vocal lip control, glottis control, pitch control, articulation control, rhythm control, resonance and tempo. Most writers agree that the above features are included in paralanguage (Crystal, 1963, 1964; Crystal & Quirk, 1964). Abercrombie (1968a), however, uses the term to refer to both audible vocal and visible nonvocal features. He defines paralinguistic activities as ones which 'must (a) communicate, and (b) be part of a conversational interaction' (1968a:56). The second criterion is introduced to exclude the act of taking one's hat off, or a "wolf whistle" (since these are not considered to enter into conversation). However, many writers would consider such events to belong to paralanguage. The validity of the second criterion may also be questioned since it typically follows on from the first. Thus it is usually nonsensical to say that one is communicating something unless one is engaged in a conversational interaction - that is, unless one has an addressee to communicate with. The rejection of this criterion for the present argument may be justified in that, for speaker recognition, we are dealing with vocal rather than nonvocal features, and it is towards the nonvocal features that this second criterion is directed. For the present purposes, therefore, we may adopt the definition proposed by Laver (1976). This defines paralanguage in a rather negative way as being neither linguistic nor extralinguistic. It is distinguished from extralinguistic features by being communicative (in Lyons' (1977) use of the term; see section 1.8.1). It is distinguished from linguistic features by lacking duality of structure.

> 'Every language has a system of significant units of sound (phonemes) and also, on another level, a system of significant units of form (morphemes), which consist of meaningful arrangements of phonemes. The structure of language is thus dual.'
>
> (Hall, 1964:6)

The structure of paralinguistic units is not dual in this sense. Laver's definition therefore includes as paralinguistic those activities excluded by Abercrombie above.

3.4.1.2 Segmental, suprasegmental and silence elements

The second three-way distinction derives from Sapir (1927). Firstly there are those features which refer to the articulation of specific vowels and consonants, i.e. segmental features. This category will be extended here to include paralinguistic and extra-linguistic events, as well as features produced by the speaker's phonology. Secondly, there are features of voice quality which refer to the intrinsically and extrinsically determined states of the vocal apparatus, which are best described in anatomical and physiological terms. The third part of the distinction refers to those features dealing with the way in which the vocal apparatus is manipulated (voice dynamics). It might be taken as a criterial feature for the distinction between voice quality and voice dynamics that voice quality is present segmentally, whereas the domain of voice dynamics parameters is generally greater than the segment. At least, for speaker recognition, the information derived from voice quality parameters in segmental articulations is much stronger than that from voice dynamics.

The features of voice dynamics and voice quality are generally referred to together as suprasegmentals. Suprasegmental features cannot stand in isolation but are in a sense imposed upon, and thereby dependent upon, the isolatable, segmental events. Whilst the categorisation into segmental and suprasegmental features is a well-established dichotomy, I should like for our purposes to add the category silence. Silence may be defined as the inaudible product of the speaker's vocal apparatus; that is, it is the residue of the speaker's vocal output when the segmental and suprasegmental elements described above are subtracted.

Having described the above two trichotomies, I shall now exemplify the intersecting categories. I shall not, however, attempt to correlate silence with the linguistic/paralinguistic/extralinguistic categories for two reasons.

(i) A definition of what would be understood by the categories linguistic silence, paralinguistic silence and extralinguistic silence would be problematical and ultimately arbitrary. It is difficult to imagine what would be meant, for example, by a paralinguistic silence - whether the duration of the silence constituted a criterial feature, and, if so, how this could be quantified; whether visual paralinguistic tokens would be a necessary concomitant feature, etc.

(ii) For the present purposes, there is no need to subdivide the category further. It may be left as a superordinate term since the characterisation applied to it below refers to all forms of silence (see Figure 3.2). Thus silence has characteristically different properties from segmental or suprasegmental features, but the different forms of silence are homogeneous in respect to this characterisation.

Linguistic segmental elements comprise the familiar phonetic elements. Suprasegmental linguistic elements are of two kinds: (a) variations of quality, such as the linguistic use of breathy phonation in many African languages, and (b) variations of dynamics features, such as the linguistic specification of pitch contours.

Paralinguistic segmental elements are events such as the use of a cough in English as an indicator of scepticism. Suprasegmental paralinguistic elements comprise (a) changes in quality such as the use of whispery phonation in English to indicate secrecy, and (b) fluctuations in dynamics parameters such as a high rise in pitch, probably with accompanying increases in loudness and length, to indicate surprise.

Abercrombie (1968a) categorises visual paralinguistic elements into

> 'those which can be independent of the verbal elements of conversation, and those which must be dependent on them. A participant in a conversation may nod his head, for example, at the same time as he says the word "yes"; or he may nod but say nothing - the nod will still communicate. This, therefore, is an independent paralinguistic element - it can occur alone, though it does not have to. Manual gestures of emphasis, on the other hand, must always accompany spoken words, and communicate nothing without them. These therefore are dependent paralinguistic elements.'
>
> (Abercrombie, 1968a: 56-7)

This categorisation seems to relate closely to the segmental/suprasegmental distinction discussed above, in that segmental paralinguistic events such as the cough of scepticism communicate even in isolation of other vocal activity, whereas suprasegmental elements such as fluctuations of pitch are necessarily dependent upon phonetic elements. However, the two distinctions are not totally equivalent. Vocal suprasegmental elements depend upon phonetic elements not only for their paralinguistic interpretation, but also for their realisation. It is possible, however, to make a manual gesture of emphasis independently of any vocal activity, although it will not have any paralinguistic effect. This results from the fact that visual and vocal activities employ different media, and in this respect visual paralinguistic elements can hardly be said to be suprasegmental to vocal activity.

Extralinguistic segmental elements are events such as involuntary coughs which interrupt the stream of speech. Suprasegmental extralinguistic elements are all those elements of quality and dynamics which do not fulfil any linguistic or paralinguistic function.

Although the above categories have been set up on reasonably well-defined grounds, it should not be thought that they reflect clear-cut distinctions, and differences of opinion will exist between individuals. A good example of the possible variation in the assignment of elements to categories are pause fillers. These are the noises used in order to avoid embarrassingly long silences in interactional situations in English, usually represented orthographically as er, um, ah, etc. The embarrassment is culturally determined; for example, Japanese speakers feel no obligation to avoid such silences in conversations. One might therefore consider the utterance of pause fillers as fulfilling a paralinguistic function, intentionally indicating to the other interactants that the speaker is thinking about what he is going to say. However, cultural conventions such as this tend to become so well-established for a native speaker that he will insert them subconsciously into his speech. The problem that this poses for the present definitions lies in deciding whether the criterion for paralinguistic status is the conscious insertion of features, as opposed to both conscious and subconscious.

Having discussed the above distinctions, the relation of the elements contained in them to the intrinsic/extrinsic distinction of section 1.8.1 can be examined (figure 3.1). The question is that of whether an element from the above categorisation reflects the presence of intrinsic factors or results from the manipulation of extrinsic factors (or possibly both).

The main generalisation to be drawn is that, while all the kinds of element may result from the manipulation of extrinsic factors, only extralinguistic elements can derive from the presence of intrinsic factors. This, however, is an obvious statement since it follows from the definition of intrinsic factors. Intrinsic factors are not under the potential volitional control of the speaker,

and therefore cannot be manipulated for the production of any system-governed, linguistic or paralinguistic element.

Figure 3.2 shows the relationships which hold between the elements of the linguistic/paralinguistic/extralinguistic and segmental/suprasegmental/silence categorisations and the categorisation of the forms of information conveyed by the speech signal (section 3.2).

The justification for the presentation of silence as a separate category can now be seen. It is the only kind of element which conveys no parametric indexical information. It may carry frequency-of-occurrence indexical information or act as a signal, although in both these cases the information conveyed will be weak. An example of frequency-of-occurrence information conveyed by silence is a speaker whose output contains an abnormally large number of pauses and other stretches of silence. Silence may convey a signal if we accept the undefined existence of the linguistic and paralinguistic use of silence.

It can also be seen that extralinguistic elements cannot carry a signal. This is self-evident since extralinguistic elements are defined as ones which cannot contain any conventional, system governed information.

Linguistic elements cannot carry any frequency-of-occurrence indexical information. This kind of information presupposes a choice on the part of the speaker - a choice which is precluded in the linguistically determined specification of phonetic elements. The restricted active vocabulary of a speaker may be a highly characterising feature, but of the semantic or syntactic kind discussed in section 1.6, not from the phonetic viewpoint of this thesis.

3.4.1.3 Relative strengths of elements

After the categorisation of elements of the speech signal, the discussion may now be related more specifically towards those elements which will be the most important for human speaker recognition

	INTRINSIC FACTORS	EXTRINSIC FACTORS
Linguistic segmental		*
Linguistic suprasegmental		*
Paralinguistic segmental		*
Paralinguistic suprasegmental		*
Extralinguistic segmental	*	(*) (1)
Extralinguistic suprasegmental	*	*
Silence	(*)	* (2)

Figure 3.1 The correlation of intrinsic/extrinsic with linguistic/paralinguistic/extralinguistic and segmental/suprasegmental/silence.

Notes

(1) Extralinguistic segmental elements may derive from either intrinsic or extrinsic sources. Thus a cough, for example, may be caused primarily by a build-up of mucus on the vocal cords (i.e. intrinsically), or may not. In the latter case, provided that the cough is not fulfilling a paralinguistic function (see this section), such an element must be categorised as originating from extrinsic sources - perhaps as a nervous habit. This extrinsic extralinguistic cough is probably much rarer than any other kind, and is unlikely to be useful as a primary source of speaker-characterising information.

(2) It is odd, and rather fruitless, to discuss whether silence may derive from either intrinsic or extrinsic sources. However, silence cannot be categorised as resulting from long-term intrinsic factors, since this would equal dumbness. It thus originates either extrinsically (choosing to remain silent by the manipulation - or, in this case, the non-manipulation - of extrinsic factors) or from short-term intrinsic factors (e.g. temporary breathlessness).

	Signal	Index	
		Parametric	Frequency-of-occurrence
Linguistic segmental	*	*	
Linguistic suprasegmental	*	*	
Paralinguistic segmental	*	*	*
Paralinguistic supra-segmental	*	*	*
Extralinguistic segmental		*	*
Extralinguistic supra-segmental		*	* (1)
Silence	*		*

Figure 3.2 The correlation of signal/index with linguistic/paralinguistic/extralinguistic and segmental/suprasegmental/silence.

Notes

(1) It may not be obvious that segmental and suprasegmental extralinguistic elements may convey frequency-of-occurrence indexical information. The intrinsic background quality of a speaker's voice is necessarily present in all forms of a speaker's vocal output. There is thus no choice as to their inclusion, and therefore no frequency-of-occurrence indexical information. However, certain intrinsically caused conditions may have an intermittent effect, which will carry frequency-of-occurrence information. For example, a long-term intrinsic condition may cause a speaker's output to be interspersed with frequent coughs, or to be uttered with sporadic whispery phonation. The frequency of the segmental cough or the suprasegmental whisperiness will constitute an admittedly weak form of indexical information.

in everyday situations. Those strength criteria (section 3.3) which refer to these elements are 1(i) and 2. The first of these deals with the frequency with which an element occurs in normal speech, and the second with whether the element derives from intrinsic or extrinsic factors. With regard to the first of these, paralinguistic elements are weak in that they are relatively rare. Extralinguistic segmental elements (coughs, etc.) are also rare. Silence is again a problematical category. Whether it is categorised as occurring frequently or not probably depends largely on exactly how the category is finally defined. With regard to the second criterion, only extralinguistic elements can derive from intrinsic factors (see Figure 3.1). In total, therefore, only suprasegmental extralinguistic elements are strong in relation to both criteria and may therefore be considered to be the most important of these categories for speaker recognition. The above argument is summarised in Figure 3.3.

	1(i) relative frequency/ infrequency	2 intrinsic/ extrinsic
Linguistic segmental	1	0
Linguistic suprasegmental	1	0
Paralinguistic segmental	0	0
Paralinguistic suprasegmental	0	0
Extralinguistic segmental	0	1
Extralinguistic suprasegmental	1	1
Silence	?	0

Figure 3.3 Relative strengths of elements.
1 = relative strength; 0 = relative weakness.

3.4.2 Parameters

In the following sections will be examined the kinds of parameters which are important from a phonetic point of view for speaker recognition. Whilst this is not intended as an exhaustive account of all the possible ways in which speakers' voices differ and by which they may be recognised, it is suggested that it covers the whole field of phonetic parameters and touches on the major ones in more detail. Preceding this examination is a discussion of the various criteria by which the parameters selected may be called major. Again, the criteria are not put forward as definitive but serve at least to highlight the difficulties involved in even the apparently simple choice of parameters.

The categories of silence and segmental features will not be discussed here. Silence contains no parametric information (Figure 3.2). On the other hand, the parametric information contained in segmental features (e.g. the distribution of friction in a speaker's [s] articulation) is so varied, resulting not only from the wealth of possible parameters but also from the large repertoire of segments eligible for selection, that it would be impossible to provide any kind of exhaustive typology. This is not to say that such segmental features are unimportant for speaker recognition; they are probably very idiosyncratic and therefore highly speaker-characteristic.

The first of these criteria relates to the distinction described in section 1.8.1 between habitual and extreme parameters. The example given in that section was of pitch range. Extreme pitch range is the distance between the highest pitch and the lowest pitch which a speaker is physically capable of producing, whilst his habitual pitch range is that span which he uses in normal unemphatic speech. It necessarily falls within his extreme pitch range and is usually a very restricted span relatively within the total possibilities.

As a speaker-characterising feature, extreme pitch range might be thought to be reliable since it reflects intrinsic, anatomical factors in a very direct fashion. However, extreme pitch range can be greatly affected by short-term intrinsic influences such as the momentary state of mucus on the vocal cords, differences caused by whether one has just been speaking, a head-cold, etc. Also, as a parameter to be used for speaker recognition, extreme pitch range is totally impractical. Under no normal circumstances does one ever hear the full pitch range capabilities of a speaker's voice. It might be usable, if not particularly convenient, as a parameter for automatic speaker recognition.

Habitual pitch range is therefore the relevant parameter for practical speaker recognition. Just as this argument applies to the parameter of pitch range, so it applies in exactly the same form to all the other parameters discussed in this section. Therefore, although they will not be explicitly described as such, the parameters discussed should be interpreted as implying parameters relating to a habitual form of speech.

Another criterion by which the following parameters can be called major is that of whether the parameter is remote and abstract or whether it is a surface one. For the purposes of automatic speaker recognition, there is no principled reason for preferring one kind to the other, except that by using less remote parameters economies are made in their mechanical extraction. For speaker recognition by humans, however, it is possible that the remoter a parameter is, the less likely it is to be useful in either the experimental or the real life situation. Although this is no more than intuitive speculation, it should be mentioned that for the present discussion it is not of paramount importance. What will be said about the less remote parameters to be discussed in this section should also be applicable to the more remote, derived parameters along the same dimension. That is, statements made about, for example, pitch deviation from the mean will apply, perhaps

somewhat less rigorously, to a remoter parameter such as rate of acceleration of pitch change. The latter is a more derived parameter of the same measure as that used for the calculation of pitch deviation from the mean.

To show that listeners are consciously aware of the parameters described in this chapter, a lay verbal description has been given, wherever possible, which a listener might use to refer to a particular parameter in relation to a speaker's voice. However, there is possible ambiguity inherent in such descriptions, and the implications of this are discussed in Brown (forthcoming; Appendix 1).

The parameters which will be considered in this section, therefore, will not be remote or of any great statistical sophistication. Instead, they will be defined in terms of the simple statistical functions of mean, range and deviation. The mean value of a parameter is the simple average, calculated by adding parametric values for as many discrete time sections as is required, and dividing the sum by the number of time sections. The range is the distance between the highest and lowest value reached during normal speech for that parameter. The deviation is a measure of the frequency and degree to which the value for a speaker's voice deviates from the mean within the range for that parameter, during speech.

The statistical functions of mean, range and deviation are applicable not only to the investigation of person-characterising features in speech, but also in other forms of human behaviour. Stuart & Godfrey (1970) discuss the relative richness of speaker-characterising voice features in comparison to those from another function of the human body - walking. The main point of their argument is that the variation in the organs used for walking is far smaller than that in the vocal apparatus.

> 'For example, there are people with big and with little feet - but there must be millions of people in the world who wear a size 10D shoe.

> There must be millions of men who are within a fourth of an inch, plus or minus, of being six feet tall. There must be millions of men who walk on the balls of their feet with even distribution of weight, with a stride of 18-23 inches, with a 1-5 inch lateral displacement of the footprints, and with no consistent toeing in or out. A personal characteristic or a complex of personal characteristics may be too general to be useful for individual identification.'
>
> (Stuart & Godfrey, 1970:105)

Whilst agreeing that the variation in the action of the vocal apparatus is probably greater than that for the organs used in walking, I disagree that the variation in walking styles is as poor as implied above. The fact that millions of people wear a size 10D shoe does not mean that their feet are all of the same size and proportions. It does not even mean that their left foot is identical to their right. It merely means that their feet fit into that pigeon-hole in the classification imposed by shoemakers - a classification which necessarily, for the sake of standardisation, overlooks some of the functionally less important distinctions between feet. The amount of variation (discriminability) possible within the stride length range of 18-23 inches is not small, and is made all the larger when multiplied by the variation possible in the lateral displacement range of 1-5 inches. The fact that a walker does not consistently toe in or out is just as person-characterising potentially as if he <u>did</u> toe in or out. If the statistical functions described above are introduced, we can calculate parameters such as the mean, range and deviation of the length of stride, time taken for a stride, weight displacement and so on. These parameters would generate indices which would be person-characterising and would certainly allow a great amount of walker-discriminability. Exactly the same arguments about richness as person-characterising features apply to the parameters of speech, discussed in the following sections.

3.4.2.1 Voice Quality

The most exhaustive categorisation of features of voice quality is that proposed by Laver (1975, 1979, 1980). His study draws together into a unified system work done by phoneticians, physiologists and acousticians. The categories of the taxonomy are defined in terms of manipulable adjustments in extrinsically determined factors from a neutral setting of the vocal apparatus. It therefore does not deal directly with anatomical and physiological differences between speakers, although the elements of the categorisation may well be helpful in referring to such differences, since these intrinsic differences may produce acoustically and auditorily similar effects to those of voice quality settings.

Voice quality settings may have differing time-domains (Laver, 1979). When used for linguistic function, a setting is relatively short-term. Paralinguistic usage generally involves somewhat longer-term settings, while extralinguistic settings, which are the most important for speaker recognition (Figure 3.3), operate on a quasi-permanent basis. Although the settings by definition operate on a stretch of speech larger than the segment, they do not necessarily affect every segment of the stretch in question. For example, because of articulatory considerations, nasality has no effect upon either nasal-stops or oral-stops. Laver applies the term non-susceptible to such segments. The criteria for susceptibility will therefore be specific to each setting and generally statable as being the consequence of articulatory constraints. The settings are most easily specified in articulatory terms and are described by reference to a baseline (the neutral setting) where

' - the supralaryngeal vocal tract is most nearly in equal cross-section along its full length

- the larynx is neither raised nor lowered

- the lips are not protruded

- front oral articulations are performed by the blade of the tongue
- the tongue-root is neither advanced nor retracted
- the faucal pillars do not constrict the vocal tract
- the jaw is neither closed nor unduly open
- the use of the velopharyngeal system causes audible nasality only where necessary for linguistic purposes
- the vibration of the true vocal folds is regularly periodic, efficient and without audible friction
- overall muscular tension throughout the vocal system is neither high nor low.'

(Laver 1979:34)

The settings are thus described in contrast with at least one of these requirements. The settings are grouped together as (i) supralaryngeal settings, (ii) laryngeal settings, and (iii) overall muscular tension settings.

Supralaryngeal settings (Honikman, 1964) are divided into longitudinal settings, latitudinal settings and velopharyngeal settings.

Longitudinal settings modify the longitudinal axis of the vocal tract by increasing or decreasing its overall length. This may be brought about either by raising or lowering the larynx, or by protruding the lips.

Latitudinal settings affect the cross-section of the vocal tract by constriction or expansion at various points. They are conveniently described by reference to the articulator performing this constriction:

Labial settings are categorised in terms of constrictions or expansions of the neutral labial opening in the horizontal and vertical axes of the lips. Ignoring scalar classifications, Laver's taxonomy of labial settings, if coupled with the longitudinal protruded and non-protruded settings, allows for the description of 18 different settings.

Lingual settings may be divided into two broad categories. Firstly, shifts in the body of the tongue on a long-term basis; Laver distinguishes eight of these, although this subdivision of the continuum may be too fine, and a grosser categorisation in terms of the three basic movements of tongue-fronting, tongue-retracting and tongue-raising may be more helpful for practical purposes. Secondly, settings of the tip and blade of the tongue, which may be enabled by adjustments of the tongue-body. Tip and retroflex articulations are thus distinguished from the neutral blade articulation.

Faucal settings refer to constriction of the pharynx by the faucal pillars. The pharynx may also be constricted by the action of the pharyngeal muscles.

Jaw settings have an interactive relationship with the other latitudinal settings. Laver distinguishes three settings - close jaw, open jaw and protruded jaw.

The third category of latitudinal settings refers to the action of the velopharyngeal mechanism. Nasal voice (nasality above what is required for linguistic purposes) and denasal voice (nasality less than this requirement) are distinguished from the neutral position.

Laryngeal settings describe the mode of phonation of the vocal cords (Catford, 1964). The articulatory, acoustic and auditory

descriptions of the six laryngeal components in Laver's categorisation are reasonably well-established from the large amount of study into the functioning of the laryngeal mechanism. The six components are modal voice, falsetto, whisper, creak, harshness and breathiness. They are grouped into three pairs on the basis of their ability to combine with each other; certain combinations are impossible owing to physiological incompatibility.

Overall muscular tension settings refer to the tendency to increase or decrease the muscular tension throughout the vocal apparatus, resulting in tense or lax voice. Although a single tendency, this feature is manifested as increases or decreases in tension in each of the local settings.

The above has been limited to a very concise description of the analysis of voice quality. Each of the settings has articulatory and acoustic specifications, most of which are agreed by workers in the field (for a fuller description of the system of analysis, and the correlates of its component categories, see Laver 1975, 1980). The above description of the system has concentrated on its categories and ignored the fact that for the practical analysis of voice quality, scalar degrees need to be included. For most components, three degrees seem to afford a usable amount of sophistication (Esling, 1978; see section 3.6).

It was stated above that features of voice quality are best described in anatomical and physiological terms, and Laver's description of the system concentrates on these articulatory aspects. However, these long-term articulatory adjustments have acoustic and perceptual correlates, which are mostly very complex and have been the subject of study for many experimenters over the years. Since the precise correlates of all the voice quality settings are not fully known, any specification of the criterial parameters involved

must necessarily be tentative. It is obvious, though, that the most profound effect which settings produce acoustically is in relation to formant structure, and in the experiments reported in Chapter 6 the parameters used to simulate long-term voice quality changes are formant position, formant range and formant bandwidth. Modifications along these parameters in synthetic voices produced convincing differences of voice quality although one cannot infer a direct relationship from this. Alterations of formant structure affect consonants less than vowels, especially of course in the case of voiceless segments. Formant structure is also largely irrelevant in the case of the tongue tip/blade setting, where the range of susceptible segments is also small; that is, only consonants articulated in the alveolar region are affected by the tip/blade/retroflex distinction, and the characteristic acoustic feature will be mostly the nature of the burst release of the closure. For the above reasons, it is more fruitful to refer not to formant structure but to spectral structure as the criterial acoustic characteristic of voice quality settings. It would be conjecture to state which parameters within the overall category of spectral structure are criterial, at least at the present state of knowledge, and any attempt to do so will be avoided. Similarly, our knowledge does not allow us to state the perceptual correlates of these unspecific acoustic features, although these will be the relevant factors for any discussion of human speaker recognition.

3.4.2.2 Voice Dynamics

Parameters of voice dynamics may be subdivided into the dimensions of fundamental frequency, amplitude and duration. Fundamental frequency (or FØ) is defined as the basic rate of vibration of the vocal cords. Amplitude refers to the strength of the sound-wave generated (either by a periodic or aperiodic sound source), and duration relates to the temporal characteristics of sounds. These three are physical

parameters which may be measured by the use of machines. However, in relation to the human speaker recognition process, the relevant parameters are not these physical, measurable ones but rather parameters referring to the way in which these are perceived by the listener. Phoneticians try to maintain a strict distinction between the physical parameters of FØ, amplitude and duration, and their perceptual equivalents - pitch, loudness and length (or temporal parameters). The latter are dependent on the former, but do not stand in a one-to-one relationship to them. An indication of the complexity of this relationship can be obtained from Ladefoged (1962:ch.2). It is sufficient to note for the present discussion that since perceptual measures are closely dependent on the physical, writers have used evidence from mechanical analysis to substantiate claims about perceptual phenomena, with varying degrees of reservation. The experiments summarised by Ladefoged are at a fairly rudimentary level, using convenient idealisations such as (i) sine-waves, as opposed to the very complex sound-waves which constitute speech, and (ii) steady-state tones, as opposed to the fluctuations in these which occur throughout speech. Therefore, surprisingly little of any detail is known about such phenomena as, for example, the changes in FØ which determine perceptual pitch fluctuations (intonation). In the following discussions, the parameters under consideration will be the perceptual ones. However, as has just been mentioned, for both voice quality and voice dynamics, we still know very little about criterial perceptual features, and are far from a rigorous exhaustive specification of acoustic characteristics. What is controvertible in the following sections, therefore, is the precise nature of the criterial perceptual characteristics, not the fact that these characteristics exist.

Length (temporal) parameters may be subdivided into tempo and rhythmicality (sections 3.4.2.2.3 and 4). The subdivision of the parameters described above can be summarised as in Figure 3.4.

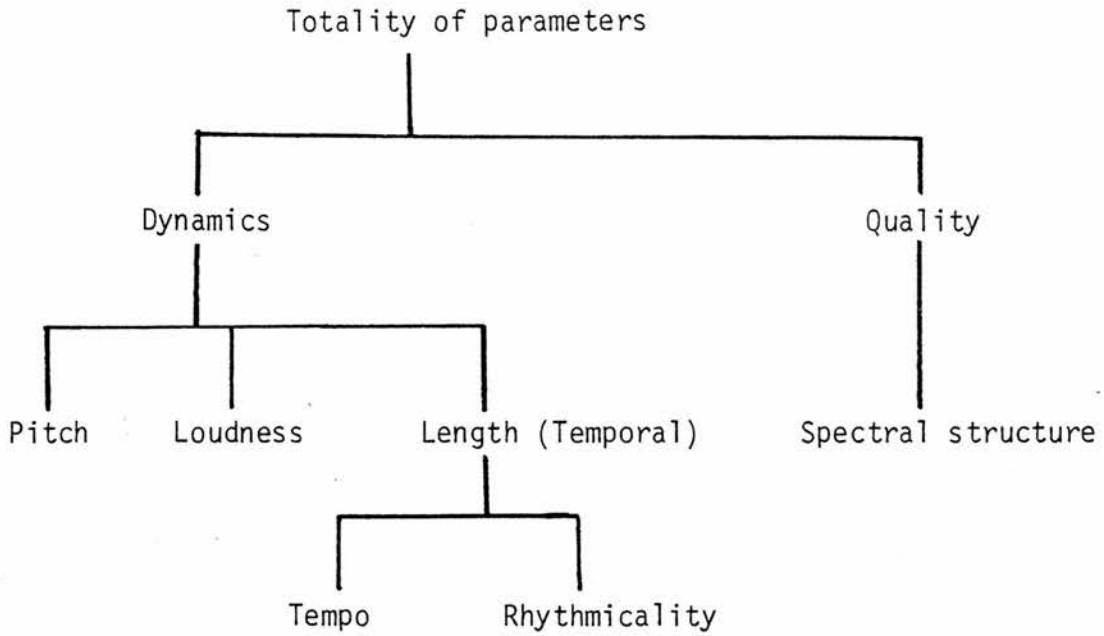


Figure 3.4 Subdivision of parameters.

The domain of these voice dynamic parameters is the utterance. The concept of the utterance need not be defined stringently for the present purpose of practical speaker recognition, but an indication of the approximate size of the stretch of speech implied by the term here may be given by stating that it is not smaller than one tone-group but may contain more than one. A criterial feature might be that it is bounded by pauses, although pauses of minimal length interrupting what may otherwise be perceived as an integral continuous unit should probably be excluded. Voice dynamic features which operate over stretches longer than the utterance are considered in section 3.4.4.

A concept which requires explanation before the dynamic parameters are discussed is that of susceptibility. (This use of the term susceptibility is slightly different from the technical meaning

assigned to it by Laver in relation to voice quality; see section 3.4.2.1). The data which is required for the calculation of certain parameters constitutes the whole of a speaker's vocal output in temporal segmental terms. For other parameters, however, certain segments are irrelevant for their calculation, and these may be called non-susceptible. The criteria for susceptibility will be specific to the particular parameter under investigation.

This form of susceptibility may be quite a problem for the experimenter in automatic speaker recognition, as in other machine-oriented phonetic fields. Consider even the widely used physical parameters of FØ and amplitude. Since the voiced/voiceless distinction is largely linguistically determined, there is justification for considering susceptible segments for FØ parameters to be any with greater than zero FØ (i.e. phonetically voiced segments), and for amplitude to be any with greater than zero amplitude (i.e. audible segments). An attractive feature of the adoption of these criteria is that they may be easily handled mechanically. However, there are certain speaker-characterising features of a person's speech which would be excluded by these criteria; for example, the phonetic devoicing in final positions of phonemically voiced consonants (usually oral stops and fricatives) in English. It is debatable whether values of aperiodic friction amplitude should be considered together with those of periodic voicing amplitude, as in most machine print-outs. These criteria are therefore generally adjusted by the experimenter to suit his ends and instrumentation.

Rather than discuss the ramifications of these questions for automatic speaker recognition, it ought instead to be pointed out that, when one considers the perceptual parameters of human speaker recognition, these problems concerning susceptibility disappear. This may be illustrated by taking as an example the pitch dimension. While voiceless sounds cannot be said to have a FØ , they may be said to have pitch.

> '... there can be no objective intonation [changes in FØ] when voiceless sounds are pronounced. The number of voiceless sounds occurring in connected speech is, however, small in comparison with the voiced sounds (about 20 per cent. of the sounds used in speaking a connected passage of English are voiceless), so that the intonation in any ordinary breath-group may be regarded as practically continuous. It is certainly subjectively continuous.'
>
> (my underlining) (Jones, 1964:275)

By 'practically continuous' Jones means 'continuous for practical purposes' rather than 'almost continuous' (Jones, 1950:144, note 1). Nevertheless, the fact remains that a FØ meter will register voicelessness by a return to the zero base-line, and that the automatic speaker recognition researcher must make some decision on the accommodation of these voiceless stretches. According to Jones, however, the human listener perceives the intonation of utterances as whole contours, not as a series of short bursts of pitch interspersed with breaks caused by voicelessness. He, as it were, "fills in" these breaks in FØ by internally hypothesising a plausible intonation curve between the preceding and following voiced stretches. The pitch contour is thus continuous and the problems of susceptibility concerning voiceless segments for the calculation of FØ parameters do not apply for pitch parameters. The only stretches which are non-susceptible for pitch contours are pauses and silences, when the speaker is not producing part of an utterance, and the hypothesising of a curve would not be possible.

It may seem that the above is merely a convenient and empirically unsubstantiable viewpoint, and that the unsubstantiability of the above arguments is mitigated by the issue of abstractness. The viewpoint is at present unsubstantiable because our knowledge of these forms of perception is still limited. However, the fact that there is a distinction between physical phenomena and our perception of these phenomena shows that it is not a mere convenience.

To illustrate that listeners are consciously aware of the parameters considered in this section, lay expressions referring to these parameters will be given when they are discussed. However, there are drawbacks inherent in the use of such expressions (or labels) for the substantiation of phonetic arguments. These are summarised by Brown (forthcoming; Appendix 1). The major criticism is that labels, by virtue of their very nature as expressions of lay usage, are vague or ambiguous, and thereby open to misinterpretation by the hearer. It is obvious that for indexical labels (Laver, 1974) there is much greater possibility of misinterpretation of the features of the voice which are being referred to and for this reason, descriptive labels will be given in the following sections to illustrate listeners' awareness of parameters. Laver (1974) contains a sizeable corpus of such labels, and I have taken most of my examples from that source. I have chosen those which seem to me to refer least ambiguously to the particular parameter under discussion. Interpersonal disagreement is inevitable, though, and the reader is referred to that corpus for further possibilities if the examples quoted seem unreasonable. This ambiguity is very great in the case of labels for features of voice quality. For this reason, no labels were given above in the discussion of voice quality parameters, although Laver's corpus contains a number of possible expressions.

3.4.2.2.1 Pitch

The susceptibility criterion for pitch parameters is that the segment forms part of an utterance, whether it is phonetically voiced or voiceless. In other words, those elements which are non-susceptible for these parameters are those stretches of silence and pauses, which cannot be said to constitute part of the speaker's utterance.

Pitch mean is the average of the pitch values for an utterance. It is usually to this parameter that listeners refer when they talk of 'a high voice' (but not necessarily 'a low voice', as Abercrombie (1967:90) notes).

It is tempting to assume that this pitch mean value corresponds to what Jones (1964) represents by his inclusion of a centre-line to the stave which he uses for the transcription of intonation. However, although it is an innovation on his previous practice (Jones, 1950), no explanation is apparently given for the inclusion of the extra line. He defines the centre-line as 'representing an intermediate pitch' (1964:276). The only other writer to use the centre-line is O'Connor (1973), who offers no definition. The traditional explanation (although, again, it has not been explicitly stated as such) is that it represents a value which is midway in pitch between the habitual maximum pitch value (the upper line of the stave) and the habitual minimum (the lower). Certainly this explanation is consistent with the equidistant relationship given to it in graphic representations. If this interpretation is correct, then this line does not represent any measure in relation to the speaker's intonation habits other than that it is the midpoint between his extremes of habitual pitch. It does not imply that the value represented by the line is a central value around which the speaker's intonation contours are symmetrical according to any of the statistical functions of mean, mode or median. (The mean of a series of values, as was explained above, is the average, calculated by adding together the values and dividing by the number of values; the mode is that value which occurs most frequently in the series; the median is that value such that half of the values are greater than it, and the other half are smaller). Whether it is helpful for transcription purposes to include a centre-line to the intonation stave or not, the line has no significance to speaker-characterising features and speaker recognition.

It is a largely statistical question whether the function of mean, mode or median is the most relevant to capture the speaker-characterising aspects of a speaker's intonation contours (or fluctuations in any of the other dimensions). All three can be thought to be relevant; perhaps the preferability of one over another may be determined by the dialect of the speaker under consideration and the intonation contour structures typical of that dialect. A concept relevant to this argument is that of an intonation contour being composed of a base-line with excursions from that base-line - itself a concept which may not be applicable to all dialects. In some dialects, this base-line is relatively level (Currie, 1979) while in others it includes a downdrift. For some dialects, the potential difference between the maximum and minimum pitch has been analysed as being greater at the beginning of the utterance than it is at the end, so that a "funnelling" effect is produced. Welmers (1959) has applied this kind of analysis to the intonation structures of African tone languages. These intonation effects may cause one statistical function to be preferable to another for a particular dialect. Since this is a highly speculative argument, however, I shall continue to refer to pitch mean, although the possibility of alternative functions being preferable should not be ignored.

Pitch range is the perceptual distance between the habitual maximum and minimum. The pitch range parameter thus describes the size of this range, not its height or lowness in the pitch scale (an indication of this is given by the pitch mean value). A large habitual pitch range is referred to by the expression 'a sing-song voice'. (Although this may have implications for pitch deviation as well, the main feature described by this label is the use of both high and low notes, the combination of which constitutes a large pitch range).

Pitch deviation is a measure of the frequency and degree to which the speaker's perceived pitch deviates from the mean. An

expression which refers to a low amount of pitch deviation is 'a monotonous voice'. The interdependence of the range and deviation functions is again evident in that this label may also (but not necessarily) imply a small pitch range.

3.4.2.2.2 Loudness

Labels which can be assigned with a reasonable degree of assurance exclusively to the loudness mean parameter are 'a loud voice' and 'a quiet voice'. The interdependence of parameters defined by the functions of range and deviation is again illustrated by the large number of labels which have implications for both. Examples include 'a booming/resonant/sonorous/piercing voice' and 'a rumbling/small/weak voice'.

So far, all the loudness parameters have been defined as referring to the overall loudness of a person's speech. However, the difference in loudness between stressed and unstressed syllables can be thought of as a very characteristic feature of a speaker. This relational aspect may be suggested as a preferable alternative to overall loudness values. Attention is thereby drawn to the upper limit, which seems the more salient perceptually than the lower limit. The upper limit is more extensible, as for pitch (see section 1.8.1) and it is a raising or lowering of this limit rather than the lower which is generally referred to by labels for loud and quiet voices.

Length (temporal) parameters

The length dimension may be divided into the characteristics of tempo, rhythmicality and continuity. All three are defined in terms of the temporal relationships between certain features. There is commonly an interrelationship between the three in normal speech, above that caused by their derivation from the same dimension. None of the three has received as much attention from phoneticians as, for example, pitch, but their complexities are at least partly understood.

The stretch of speech required for the calculation of continuity is greater than the utterance, and therefore the discussion of continuity is left until section 3.4.4.5.

While there is a separate term for the perceptual correlate of duration (length), there are no terms for the two length parameters discussed here (tempo and rhythmicality) to distinguish the perceptual from the physical. Speed and rate have been used to refer to tempo, but with no implications for any systematic terminological distinction. Rhythm or rhythmicality are the only terms used for that parameter. Therefore the terms are used for both the physical and the perceptual phenomena, although in the following discussion it is always the perceptual correlates which are under consideration.

3.4.2.2.3 Tempo

Tempo may be defined as the speed at which a speaker speaks, measurable in physical terms as the amount of speech output uttered per unit time (say, second). Syllables are the most appropriate units of speech output for English, although for other languages other units (segments, morphemes, words) may be more appropriate. For English, therefore, tempo may be expressed as the number of syllables uttered per second. A converse but equivalent way of stating this is as the length per syllable uttered. However, this second definition is mnemonically misleading in that the lower tempo limit would imply a small syllable length, i.e. a fast speed. For this reason, I shall continue to use the first definition in the following discussion, so that the lower tempo limit implies a small number of syllables uttered per second, i.e. a slow speed.

Since the definition of tempo is in terms of the number of syllables uttered per second, the criterial feature for susceptibility for tempo parameters includes all those stretches which constitute an utterance and thus excludes pauses and silence.

The tempo mean is therefore the average tempo measure for the utterance under consideration.

Tempo range is determined by the maximum and minimum number of syllables per second during the utterance (or, equivalently, by the minimum and maximum syllable length values). The extremes of tempo range may be considered to be constrained at the lower limit by the methodological definition of the notions of syllable and pause. The upper limit is governed by physiological constraints on articulator movement and by perceptual constraints on how long a part of an utterance must be to be perceivable as a syllable.

Tempo deviation is a measure of the frequency and degree to which the speaker's tempo deviates from the mean. If we restrict the domain of tempo to the utterance, it is unlikely that there will be any great range or deviation, or that these will be strong speaker-characterising features.

Brown (forthcoming; Appendix 1) deals with an observed regularity in the acceptability in English of certain syntactic constructions with certain labels for voices. In particular, labels describing temporal aspects of extralinguistic features of a speaker's voice are not acceptable in the frame 'a ____ voice', but instead use the 'a ____ speaker' or 'a ____ style' constructions. It is argued that this reflects a basic phonetic difference between temporal and non-temporal parameters in speech. Labels referring to mean tempo values are therefore 'a fast speaker' and 'a slow speaker', while 'a jerky speaker' seems to me to be the only label which carries strong implications for tempo range and deviation.

3.4.2.2.4 Rhythmicality

Rhythm has received a reasonable amount of attention from phoneticians, notably Abercrombie (1964a,b). His theory defines it as follows: the airstream emanating from the lungs for speech does not flow in a continuous manner but is pulse-like. Each pulse

constitutes a syllable. In addition to this system there is a second system of less frequent but more powerful pulses, each of which constitutes a stress. Rhythm is defined as the temporal regularity with which certain features recur. In syllable-timed languages, it is the syllables which occur more or less regularly, whereas in stress-timed, it is the stresses. Abercrombie claims that the stresses in English are isochronous (occur at regular intervals of time). However, it has been shown that feet in English (the time intervals from one stress up to, but not including, the next) are not of exactly equal durations (Uldall, 1971). This dichotomy may be resolved by considering various factors.

(i) The difference between syllable- and stress-timed languages is best handled not as a binary distinction but as a scalar one. Thus, the isochronous occurrence of foot boundaries in English can be stated as a tendency (Pike, 1945:34) rather than a rule regularly observable in its realisation. Various factors can be equated with the failure for this rule to be rigidly manifested, such as the syllable structure of feet and the style of utterance, as well as the idiosyncratic variation which forms the basis for the consideration of this parameter as a potential speaker-characterising feature.

(ii) The isochrony may be a perceptual phenomenon (Lehiste, 1973) which is not realised by total physical isochrony. Stress itself is above all a perceptual phenomenon, the physical correlates of which, in terms of changes in FØ, amplitude, duration and spectral structure, are extremely complex and still not fully understood (Fry, 1955, 1958, 1965). In interactional terms, stress may be considered as a perceptual phenomenon both auditorily and visually.

> 'If you watch an English speaker talking you will be able to see, without hearing what he is saying, where the stressed syllables are. All the big muscular movements that he makes

> are in time with the stressed syllables. When he waves his arms, nods his head, puts his foot down, raises his eyebrows, frowns, opens his jaw more widely, purses his lips; all this is done in time with the rhythm of speech.'
>
> (Brown, 1977:42)

Taking the isochrony of this stress also to be a perceptual phenomenon is therefore a viewpoint which is not unjustified and is advantageous for human speaker recognition, where the relevant parameters are perceptual ones. This is not to say, for example, that an English listener will mentally adjust what he hears to produce total perceptual isochrony in all cases, but that allowance will be made automatically for certain factors such as the syllable structure of feet, so that a measure of the relative underlying isochrony is achieved.

Further evidence for the perceptual nature of stresses is the occurrence of silent stresses - stresses perceived by the listener although they are not realised by any uttered syllable. Silent stresses occur, for instance, at the end of the first, second and fifth orthographic lines of limericks (Abercrombie, 1964a:23). Since such silent stresses are required for the calculation of rhythmicality, all forms of a speaker's vocal output are susceptible for rhythmicality parameters, including pauses (which was not the case for tempo parameters).

The parameters of rhythmicality can be defined as functions of the occurrence of foot lengths. A foot length is defined as the perceptual length from one stress up to, but not including, the next. The mean foot length is thus determined by averaging the lengths of the feet of the utterance under consideration. In English this would seem not to be a particularly relevant parameter. It does not represent any measure of the rhythmicality of a person's

speech. Instead, it might seem to be indicative of the tempo of the speech; however, it does not correlate with the mean tempo value since when a person's speech quickens up, it tends to drop certain stresses, and conversely when it slows down, extra stresses tend to be introduced. Thus the mean foot length value will tend not to vary overall as greatly as the mean tempo value.

More relevant parameters for speaker recognition are foot length range and deviation since they involve the comparison of foot lengths and are thus a measure of the rhythmicality of an utterance.

Labels referring to the parameters of rhythmicality are not so numerous as for the other parameters and usually carry implications for concomitant tempo values. Thus 'a jerky speaker' seems, to me at least, to involve large foot length range and deviation values and large tempo range and deviation values. Similarly, 'a tum-ti-tum style' indicates small foot length range and deviation values and also, probably, small tempo range and deviation values; 'a drawling/droning/monotonous/sombre style' all indicate, among other things, small foot length range and deviation values and small tempo mean, range and deviation values.

3.4.2.3 Relative strengths of parameters

Having categorised the various parameters which constitute the speech signal, the discussion may now be related more specifically to a consideration of which of these parameters will be the most important for human speaker recognition in everyday situations. The strength criteria of section 3.3 which refer to parameters of the speech signal are 1(ii), 1(iii) and 4. Criterion 1(ii) deals with the length of the stretch of speech required for the calculation of the value of a parameter. Criterion 1(iii) refers to whether the indexical information conveyed by the parameter is parametric in

nature, or whether it deals instead with the frequency-of-occurrence of certain features in speech. Criterion 4 concerns the distinction between the cognitively and physiologically determined nature of parameters.

It is the temporal parameters of tempo and rhythmicality which are weak in relation to both the above criteria. They require long stretches of utterance for their calculation, and refer to the temporal frequency and regularity of features of the speech signal (syllables and stresses in English). However, in relation to criterion 4, temporal parameters are cognitively governed (see Brown, forthcoming; Appendix 1) although the greater stability this suggests is relevant to automatic rather than human speaker recognition.

In summary, pitch, loudness and spectral structure parameters are the most important in human speaker recognition. The above argument is represented in Figure 3.5.

	1 (ii) short/ long stretch for calculation	1(iii) parametric/ frequency-of-occurrence	4 cognitive/ physiological
Pitch	1	1	0
Loudness	1	1	0
Tempo	0	0	1
Rhythmicality	0	0	1
Spectral structure	1	1	0

Figure 3.5 Relative strengths of parameters
1 = relative strength; 0 = relative weakness

In order to produce the above binary classification of strengths of parameters, it was necessary to adopt a Procrustean approach to the analysis. This was also true, but to a lesser extent, of the analysis of elements of the speech signal in section 3.4.1.3. The analysis therefore overlooks any scalar differences in characterising strength by imposing an arbitrarily selected threshold at some point along the continuum. One case where this results in the neutralisation of a significant difference is the strength of pitch and loudness parameters on the one hand and spectral structure parameters on the other. It has already been argued in this chapter that parameters of voice quality require shorter stretches of utterance for their calculation than parameters relating to pitch or loudness contours. However, all these parameters have been categorised as strong in relation to criterion 1(ii), since the difference in this respect between spectral structure and pitch/loudness is far less significant than that between pitch/loudness and temporal parameters on this continuum..

3.4.3 Correlation of Elements and Parameters

Figure 3.6 summarises the correlation of the voice quality and dynamic parameters just discussed with the elements of the linguistic/paralinguistic/extralinguistic and segmental/suprasegmental/silence categorisations given in sections 3.4.1.1 and 2. The issue under examination can be expressed by the question "Which of the elements which compose the speech signal interact to determine the value of the voice quality and dynamic parameters (for all speakers)?" An asterisk in Figure 3.6 indicates an element which influences the value of the particular parameter. A double asterisk indicates that the element has a great influence and may be the main determinant of the value of the parameter. A blank indicates that the element has no effect and should be considered non-susceptible for that parameter.

	Spectral structure	Pitch			Loudness			Length (temporal)					
								Tempo (1)			Rhythmicality		
		mean	range	deviation	mean	range	deviation	mean	range	deviation	mean	range	deviation
Linguistic segmental	*				*	*	*				*	*	*
Linguistic suprasegmental	*	*	*	*	*	*	*				*	*	*
Paralinguistic segmental		*	*	*			(2)				*	*	*
Paralinguistic supra-segmental	*	*	*	**	*	*	**	*	*	**	*	*	**
Extralinguistic segmental											*	*	*
Extralinguistic supra-segmental	*	*	**	*	*	**	*	*	**	*	*	**	*
Silence											*	*	*

Figure 3.6 The correlation of voice quality and dynamic parameters with linguistic/paralinguistic/extralinguistic and segmental/suprasegmental/silence.

Notes to Figure 3.6

(1) It is difficult to justify considering tempo parameters as being determined by anything other than (i) physiological and cognitive constraints, (ii) methodological criteria for the maximum duration of a syllable, and (iii) modifications for short-term paralinguistic effect (Brown, forthcoming; Appendix 1). Linguistic considerations might be considered to come into effect in "awkward" sequences of sounds, such as tongue-twisters, which generally require to be articulated more slowly than usual. However, this is probably best treated as a physiologically governed phenomenon, since the difficulty involved in tongue-twisters is easily accounted for articulatorily, as caused by the juxtaposition of a typically small set of sounds ([ʃ,s,θ] etc.).

(2) Although paralinguistic segmental elements are not silent, it is best to exclude them from any considerations of loudness parameters. Loudness may be produced by any articulators in the vocal tract, but for a paralinguistic effect they are, by no means always but at least very often, manipulated in a way which is not exploited for speech sounds. For example, a sniff of disdain will produce a loudness comparable to that of an egressive voiceless nasal. However, a sniff is an implausible speech sound for any language. Similarly, a paralinguistic cough of scepticism may produce an enormous loudness (since it has to be exaggerated; see section 1.8.1) but uses the vocal cords in a way not seen in speech.

It is now clear why rhythmicality is an unhappy parameter for speaker recognition. Although the probable main determinants of rhythmicality mean and deviation can be specified as extra-linguistic and paralinguistic suprasegmental elements respectively, the fact that all elements may contribute to the value of rhythmicality parameters is a possible reason why no totally satisfactory analysis of rhythm has been proposed at present.

Comparing the pitch and loudness columns, it can be seen that the effect which the various elements have on them is largely the same. From this, and from the discussion in Brown (forthcoming; Appendix 1), one can conclude that pitch and loudness have more in common with each other than temporal parameters have with either of them. They are certainly more easily measured objectively, and, for all the various reasons given above, it would seem that temporal parameters are not, in broad terms, useful for speaker recognition either by humans or by machine.

Silence is another unhappy category for speaker recognition, since its role is very limited. As can be seen from the diagram, it plays a part only in the determination of rhythmicality, and, even here, probably plays a much smaller part than non-silent segments.

3.4.4 Voice Dynamic Parameters in Discourse

The domain of all the voice dynamic parameters discussed in the previous section was stated as the utterance. The temporal parameters of tempo and rhythmicality are unlikely to be of great value when only such a short stretch is taken into consideration. This section examines some of the potential speaker-characterising features which come into play when one takes into account the role of the voice dynamics parameters in the larger realm of discourse. Comparatively little research has been done in this area, and therefore I shall simply list some features which have been shown to have certain validity.

By virtue of their nature as long-term effects, these features are weak according to the strength criterion which states that features to be measured should be calculable from a short stretch of speech. These features are therefore of limited importance to human speaker recognition, although they may be of use in automatic speaker recognition, where stability is the more relevant criterion.

3.4.4.1 Pitch

The pitch contours associated with tone-groups join together into larger structures. The function of these larger structures is to organise the discourse such that the units of information contained in each structure belong together. Thus when there is a substantial information break in the message conveyed by the discourse, a boundary is hypothesised between two of these larger structures. Brazil (1975) offers a definition of the realisation correlates of these larger structures, by making reference to Sweet (1906:70): 'each sentence or sentence-group has a general key or pitch of its own.' Brazil distinguishes three keys (high, mid, low), and his definition of high or low key is that it

> 'characteristically involves the raising or the lowering of the pitch of the whole tone-group relative to a pitch which can be established as the norm for the speaker concerned.'
>
> (1975:10)

Brazil thus defines 'a phonological unit of higher rank than the tone-group, having the structure:

$$(\text{High}_{1\ldots n})\quad \text{Mid}\quad (\text{Low}_{1\ldots n})\ '$$

(1975:10)

It is to this unit that Brown (1977:86) gives the name paratone (on analogy with the orthographic paragraph). Experimentation has

shown that this concept has a certain psychological validity (Rees & Urquhart, 1976), in which case the occurrence and structuring of paratones may be considered a potential speaker-characterising feature.

3.4.4.2 Loudness

It seems possible that the larger-scale variations in loudness will prove to be very similar to those of pitch. They may even be directly related to those of pitch as potential realisations of the division into large-scale perceptual structures. Thus Brazil suggests that

> 'It is likely that there will be predictable co-occurrence relationships [between pitch and] other variables such as loudness and speed. It may even be that ... the physical correlates of key are a set of phonetic features no single one of which is essentially present. Provisionally, however, I take pitch-level as defining.'
>
> (1975:10)

If loudness can function in the same way as pitch to organise discourse into paratones, then the variation in occurrence and structuring of large-scale loudness features may be considered a potential speaker-characterising feature.

3.4.4.3 Tempo

If only one utterance is considered for the calculation of tempo parameters, it is unlikely that any great variation in tempo will be found. This will result in small range and deviation values and will be attributable largely to the style and content of the utterance rather than to any idiosyncratic factors. Tempo mean is therefore the most speaker-characterising parameter in this respect. However, a measure of a more speaker-characterising nature is obtained

if the deviation in tempo between two or more utterances is considered. Again, the style and content of the discourse will exert an important influence although the tempo range and deviation values between utterances will be of some value. Tempo mean is nevertheless still the most useful parameter.

3.4.4.4 Rhythm

Just as tempo differences between utterances may be speaker-characterising, so may rhythm differences of this kind. It has been suggested (Abercrombie, 1968b) that persons taking over the role of speaker from one another tend to maintain the rhythm established by the previous speaker. Similarly, an effective way to interrupt or contradict a speaker is to break this established rhythm. Furthermore, the established rhythm of a speaker may be maintained during pauses and even through such performance errors as false starts, hesitations, tongue-slips, etc. However, these informal observations cannot be taken further until more is known about the perception of rhythm.

3.4.4.5 Continuity

Continuity is a feature of voice dynamics among those listed by Abercrombie (1967:95) but which requires for its calculation a stretch of discourse much larger than the utterance-length which has been considered so far.

> 'continuity refers to the incidence of pauses in the stream of speech - where they come, how frequent they are, and how long they are. The incidence of pauses, whether they are hesitations or whether they are deliberate cessations of talking for the purpose of taking breath, seems to be a highly idio-syncratic matter, and there is a lot of variation from speaker to speaker. Under the conditions of ordinary conversation nobody's speech is fluent, and it is probably true to say that the more thought there is behind what one is saying, the less fluent will be the speech. Pauses, for the most part, pass

> unnoticed by both speaker and hearer; yet they frequently occur at apparently unpredictable places, as is revealed by conversations which have been recorded. Pauses bear little relation to syntax, in spite of popular belief to the contrary.'
>
> (Abercrombie, 1967:96)

Brazil, quoted above, claims that changes in key, and thereby paratone boundaries, may be potentially realised by a set of phonetic features, rather than solely by pitch fluctuations. There is evidence (Brown, 1978b) that the incidence of, and the relative lengths of pauses may function as indicators of paratone boundaries.

Although labels such as 'jerky', 'interrupted' and 'fluent' refer to the parameter of continuity, no labels exist which one would assign with any confidence to discourse-scale features of pitch, loudness, tempo or rhythm. This may be taken as an indication that such phenomena are quite remote in nature and probably play a relatively small part in speaker recognition by humans.

3.5 FIRST-ORDER PARAMETERS

It is now possible to amalgamate the results of the strength categorisations of elements and parameters in order to be able to specify, in very broad general terms, those aspects of a speaker's vocal output which are the most important for human speaker recognition in everyday situations. It has just been seen that pitch, loudness and spectral structure parameters are the most relevant parameters, and their importance is all the greater if they occur as the manifestation of extralinguistic suprasegmental elements, which were shown in section 3.4.1.3 to be the most relevant elements of the speech signal. As was shown in Figure 3.6, extralinguistic suprasegmental elements provide a major contribution to the values of these parameters in speech. We might therefore refer to them as first-order parameters, meaning that they are the most important for

human speaker recognition, and in terms of the model presented in the next chapter, are those parameters in terms of which stimulus utterances are first analysed. By the use of the term here it is not intended to imply what is often meant by 'first-order', i.e. that the parameters are basic, or easily accessible or extractable, although this may well be the case.

It can be seen that there is a close relationship between the category of first-order parameters and the distinction discussed in section 1.8.3 between those parameters which listeners are potentially capable of using to recognise speakers, and the subset of these parameters which listeners habitually do use in everyday situations. That is, first-order parameters represent the latter subset.

We might also set up further orders of parameters (second-order, third-order, etc.) which have decreasing importance for the human speaker recognition process. Those parameters which listeners can use but do not habitually use will therefore figure in these lower orders. Their role in speaker recognition is considered to be supportive in that they only come into play if no decision is made on the basis of first-order parameters.

The experiments of Miller (1964) and Matsumoto et al. (1973) (see section 5.2.3.7), and those reported in Chapter 6 are all directed at determining the relative importance for speaker recognition of various parameters in speech; in short, they attempt to specify which parameters belong to the first order as opposed to lower orders. Our knowledge of the speaker recognition process would be greatly advanced if answers were obtained for the following questions.

(1) Which are the most important parameters for human speaker recognition in everyday situations?

(2) Is the relative importance of these parameters affected by differences of

(a) the individual speakers being recognised?

(b) the dialects of the speakers being recognised?

(c) the languages of the speakers being recognised?

(d) the content of the speech?

(e) the context of utterance?

(f) listeners?

etc.

At present we know little of very specific detail about the first of these questions, and much less about the second.

3.6 VOICE JUDGMENT PROTOCOL

The voice quality and voice dynamic categories proposed in the preceding sections are suggested as the most important parameters in the human auditory speaker recognition process. They may also be used for the practical description of voices. Figures 3.7 and 3.8 contain a protocol for this purpose. To establish its practical applicability, judgments in relation to the various categories for nine voices have been inserted. Recordings of these voices are included on the tape which accompanies this thesis. These recordings are taken, with permission, from a tape produced to accompany Laver (1980), and are copyright. The judgments represent the consensus of opinion among a panel of judges trained in Laver's (1980) classificatory system (see section 3.4.2.1).

The categories of the protocol correspond to those described in sections 3.4.2.1 and 2, with minor modifications. The voice quality categories therefore derive from Laver (1980), while the voice dynamics categories represent reasonably well-established divisions. The modifications made here serve to facilitate the applicability of the categories in the practical description of voices; certain of the categories described in sections 3.4.2.1 and 2

have been found to be too subtle for practical use, and therefore grosser categories are substituted. For convenience, settings which are rarely found, such as protruded jaw and faucalisation, have been omitted. Nevertheless, the sophistication afforded by these grosser categories is still considerable. This sophistication is increased by the incorporation of scalar degrees into each category. Esling (1978) found that three degrees provided a workable amount of sophistication. Thus, in Figure 3.7, the numbers 1, 2 and 3 represent slight, moderate and extreme versions of the category respectively.

Voice quality settings are defined by reference to excursions from the neutral setting (see section 3.4.2.1). In Figure 3.7 no category appears which forms part of the neutral setting; for example, blade articulation is not included in the front oral articulation category for this reason. This restriction does not hold for the phonation type category, since modal voice may be present either by itself (the neutral setting) or in conjunction with other components. Similarly, absence of an entry in any one category should be interpreted as representing the neutral setting for that category; for example, absence of an entry in the larynx position category implies the neutral setting, i.e. neither raised nor lowered. In relation to simple phonation types (where only one component is involved), the relevant component in Figure 3.7 is conventionally assigned the number 3, since it is nonsensical to refer, for example, to slight modal voice or extreme modal voice - the phonation type either is or is not modal voice. However, for compound phonation types (where two or more components are combined) the numbers assigned to each component represent the proportion of that component in the compound. Thus (modal voice 3) (whisper 1) indicates a compound composed predominantly of modal voice with a much smaller whisper component. In Figure 3.7, numbers in brackets indicate features whose influence is intermittent; in other words, they denote features which are characteristic but not totally pervasive in all susceptible segments of the person's speech. Certain of the voice

quality settings do not have strong acoustic/perceptual influences; for example, labial settings are obviously more prominent visually than auditorily, and the judgments in Figure 3.7 have accordingly been made on visual rather than auditory grounds.

The recordings on the tape are not examples of spontaneous conversation. Instead, all nine speakers are reading the same passage (the Rainbow passage; Appendix 2); this enables a more direct comparison between the voices to be made. Unfortunately, it also means that the speakers adopted the style appropriate for reading aloud. Although this may affect voice quality settings somewhat, it is likely to have a much greater influence on voice dynamic features. Thus the speech is slower, more rhythmical and more fluent than spontaneous speech, and exhibits a smaller set of intonation contours. In other words, it would give a distorted picture of the voice dynamics features of the speakers' normal conversational style to base these judgments on examples of deliberately read prose. For voice dynamics, therefore, the appropriate categories are given in Figure 3.8, although no corresponding judgments have been included because of this possible interference. Certain of these parameters may remain relatively unchanged, though. It is easily appreciated that speaker B has a characteristically lower mean pitch than speaker A (whether in reading aloud or in conversation). Three scalar degrees may provide a usable amount of sophistication for voice dynamic parameters, as for voice quality.

		Speaker								
		A	B	C	D	E	F	G	H	I
Larynx position	(a) raised	2				1	1			
	(b) lowered		2	2	1					
Labial Factors	(a) labiodentalisation	1								
	(b) lip protrusion	1		2					1	
	(c) open rounding	1		1					1	
	(d) close rounding									
	(e) lip spreading		(1)				1	1		
Front oral articulation	(a) tip									
	(b) retroflex									
Tongue body	(a) fronted						3	1		
	(b) retracted	1			1					
	(c) raised				1					
Mandibular factors	(a) open jaw		2							
	(b) close jaw									3
Velopharyngeal factors	(a) nasal	1		1		2	1		2	2
	(b) denasal				2					
Phonation type	(a) modal voice	3	2	3	3	3	3	2	3	3
	(b) falsetto									
	(c) creak(y)	1	3	1	1	1	2	3	1	2
	(d) whisper(y)	1	2		1			2		
	(e) breathiness		(1)	1						
	(f) harshness							1		
Overall tension	(a) tense	1					1		1	2
	(b) lax									

Figure 3.7 Voice quality judgment protocol (see accompanying tape for illustration of speakers' voices (A,B,C etc.)).

1 = slight version of the setting;

2 = moderate;

3 = extreme.

Pitch	(i) mean	(a) high							
		(b) low							
	(ii) range/ deviation	(a) wide							
		(b) narrow							
Loudness	(i) mean	(a) high							
		(b) low							
	(ii) range/ deviation	(a) wide							
		(b) narrow							
Temporal parameters	(i) tempo	(a) fast							
		(b) slow							
	(ii) rhythm	(a) rhythmical							
		(b) unrhythmical							
	(iii) continuity	(a) fluent							
		(b) interrupted							

Figure 3.8 Voice dynamics judgment protocol.

CHAPTER 4

A MODEL OF SPEAKER RECOGNITION

CHAPTER 4

A MODEL OF SPEAKER RECOGNITION

4.1 INTRODUCTION

The previous chapter has concentrated on speaker-characterising features. These features will be of the kind which constitute reference voice patterns. In this chapter the discussion turns away from speaker-characterising features and concentrates rather on the decision-making component of the speaker recognition process. A formal, logical model of this process is presented. It is preceded by a preliminary discussion of background issues, which takes up the debate started in section 2.4 about Wolf's (1972) criterial attributes for efficient acoustic parameters in automatic speaker recognition, and Bricker & Pruzansky's (1976) taxonomy of speaker recognition tasks (section 2.3). These are discussed here because the formal model presented (and any such model) of the speaker recognition process will require to take into account the considerations which form the basis of the categorisations.

4.2 WOLF'S CRITERIA (2.4) AND CRITERIA FOR AUDITORY PARAMETERS

The criterial attributes which Wolf (1972) specifies for efficient acoustic parameters for automatic speaker recognition were discussed in section 2.4. To summarise briefly, the six criterial attributes were that characteristics measured should

'(i) occur naturally and frequently in normal speech,

(ii) be easily measurable,

(iii) vary as much as possible among speakers, but be as consistent as possible for each speaker,

(iv) not change over time or be affected by the speaker's health,

(v) not be affected by reasonable background noise nor depend on specific transmission characteristics, and

(vi) not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice.'

(p.2044-5)

From these attributes were deduced several assumptions, which constitute presuppositions about the kinds of parameter desirable for automatic speaker recognition. There were five of these:

(i) the number of acoustic parameters which need to be considered for successful automatic speaker recognition is small,

(ii) the acoustic parameter(s) should be virtually constant and stable,

(iii) the system aimed at is, in some sense, perfect,

(iv) practical considerations (measurability, background noise, etc.) must not only be taken into account but also play an active part in the selection of one parameter over another,

(v) a consideration of theoretical phonetic notions (e.g. the intrinsic/extrinsic distinction) is not necessary.

Let us now consider each of these five presuppositions in turn and see how they relate to the kinds of parameter which are probable in the human speaker recognition process.

Whereas in the case of automatic speaker recognition it is assumed that a small number of acoustic parameters are required for

successful recognition, there is no reason to suppose that the set of parameters habitually used by humans is so severely limited. Indeed, all the experimental evidence suggests that human listeners are capable of calling on a wide variety of parameters as cues in successful speaker recognition. For example, it has been shown (Abberton, 1974) that listeners are capable of successfully recognising speakers on the basis of glottal source information (laryngograph waveform) alone. However, in complementary experiments, it has been found that listeners can equally successfully recognise speakers when inter-speaker differences in glottal source information are eliminated. This may be done by using a channel vocoder (Shearme & Holmes, 1959) or an electro-larynx (Coleman, 1973). In other words, in these experiments, recognition takes place on the basis of vocal tract information. (These experiments are described in greater detail in Chapter 5.) There is therefore a wide range of parameters which are potentially usable by listeners. Again it must be stressed, though, that the fact that human listeners <u>can</u> use a wide range of parameters does not necessarily imply that they habitually <u>do</u> use that full range in speaker recognition in the everyday situation (see section 1.8.3). However, this distinction does not exist in machine recognition.

Wolf's attributes suggest that the speech characteristics measured should be virtually constant and stable. However, it was noted in section 2.4 that it is unrealistic to expect any parameter to remain stable in all styles of normal speech, over a moderate length of time, and to be so distinctive for each speaker that it is unsusceptible to imitation or disguise. For this reason, an adaptive system which allows for a degree of flexibility of operation not only is preferable for automatic speaker recognition but also corresponds more probably to the manner in which the human process operates. It is difficult to prove this statement; however, the experiments just described hint at the flexibility of the human process. Such flexibility is also assumed for the model presented

later in this chapter of the human process, of which different types of feedback mechanism are an important feature.

The third presupposition was that the system at which automatic speaker recognition research is aimed is, in some sense, perfect. If this presupposition is correctly inferred, then the system at which automatic speaker recognition is aiming cannot correspond to the human process, which is in no sense perfect; no experimenter has yet found a 100% success rate in any but the most trivial tasks. However, at the other extreme, human capabilities ought not to be underestimated. Success rates of 90% or more have been achieved in the performance of even seemingly very difficult tasks.

It is usually quite easy to specify the aim of any automatic speaker recognition system. These aims fall into two main categories:

(i) a maximally high success (correct response) rate, and

(ii) a maximally low failure (incorrect response) rate.

The second of these is not only the easier to fulfil but is also the more commonly required aim of automatic systems in practical, security situations. On the other hand, it seems paradoxical to talk of the "aim", or "criteria" for parameters, of the human speaker recognition process. This is firstly because there is rarely any incentive for improvement; the situations are rare in which the difference between the successful and the unsuccessful recognition of a speaker has important consequences, although performance in a legal situation is a counterexample. A second reason is that it is difficult to imagine how improvement might be achieved. Improvement is not impossible; otherwise there would be no reason to assume that blind listeners are more successful at speaker recognition, and other

listening tasks, than sighted listeners. However, it would be conjecture to discuss:

(i) how this improvement in performance might come about, and

(ii) whether it can be deliberately achieved, or is merely due to practice through necessity.

Therefore, the "aim" of the human speaker recognition process might be stated as the limited but satisfactory success in recognising speakers in the everyday situation. The problem, then, for human speaker recognition experimenters is to quantify this limitation and to state what its main determinants are.

Fourthly, the characteristics considered by Wolf are subject to the limitation imposed by practical requirements. The human speaker recognition process is not so subject to practical restrictions, but performance may be affected to some degree. Recognition is sometimes made impossible by background noise or transmission characteristics, such as on a bad telephone line. However, the amount of signal degradation required for recognition to become a problem is quite great, and this may be attributed to the versatility of the human speaker recognition process. The fact that it may draw upon information from a wide variety of parameters means that recognition may continue even when data is restricted to only a few bits of information on any one parameter (see Clarke et al's (1966) view, quoted in section 1.8.3).

A consideration of theoretical issues was not taken to be of great importance for automatic speaker recognition. For the human speaker process, its importance is probably greater, but again not vital. It was shown in section 2.4 that a consideration of the distinction between intrinsically and extrinsically determined voice features helped to highlight a discrepancy in Wolf's criteria as homogeneous categorisations of desirable speech characteristics.

Since such speaker-characterising features will play the most important part in the human speaker recognition process, it is unlikely that the intrinsic/extrinsic distinction will be of no relevance.

There are thus differences between what is desirable for the automatic speaker recognition process, and what is probable in the human process. For this reason, only limited relevance can be assigned to the findings of automatic speaker recognition experimentation as regards its implications for the human process.

4.3 A REVISED VERSION OF BRICKER & PRUZANSKY'S TAXONOMY (2.3)

Bricker & Pruzansky's (1976) taxonomy (hereafter B & P) was described in detail in section 2.3. The taxonomy was summarised diagrammatically in Figure 2.2. Criticisms of this taxonomy were also given in that section, the most important being that:

(i) McGehee's (1937, 1944) experiments cannot be categorised as short-term memory tasks since they involve time delays of up to five months, and

(ii) the term trial is not defined, but is not being used in a way consistent with its widely accepted meaning, and this fact renders some of the criterial definitions unclear.

In this section is presented a revised version of that taxonomy, avoiding these inconsistencies. This revised version then plays a part in the description of the model of the human speaker recognition process presented later.

It should be pointed out at the beginning that this revised version deals only with the identification half of B & P's taxonomy. As was argued in section 2.3, evaluation tasks are so

different in nature from identification tasks that they do not merit being part of the same taxonomy. They are not a central interest of this thesis and, to this extent, the treatment afforded them in B & P's taxonomy may be taken as satisfactory.

The first revision is in regard to the superordinate term. As is explained in the next section, the term identification is often used technically with a more specific meaning than that intended by B & P. For this reason, it might be misleading to use it here. Instead, a more neutral term, such as designation, would be better. There are very few terms in this semantic field which have not been used by one writer or another with a specific technical meaning. Designation is still an unsatisfactory term but is the least bad alternative.

It was shown that B & P's distinction between matching and naming tasks is badly defined. In particular, a time-delay between the presentation of reference and stimulus voice samples is incompatible with their criterion for matching tasks; that is, McGehee's experiments are examples of long-term, not short-term, memory tasks. The distinction between short-term and long-term memory tasks can still be kept, though. The model of memory which I am using in this thesis is a categorisation (e.g. Norman, 1969) into three systems:

(i) a sensory image of events occurring in the immediate past,

(ii) a short-term memory containing limited information extracted from the rapidly decaying sensory image, and

(iii) a long-term memory, with a much greater capacity than the short-term.

In future I shall use the term short-term memory to refer to the second stage, and long-term memory for the third (and this seems consistent with B & P's usage of these terms). Whilst the first,

sensory image stage probably plays no significant part in speaker recognition by humans, it may be argued that there is one case where it is at least a relevant factor in the process. This case is discussed later.

Short-term memory tasks involve the stimulus voice sample being presented after a delay of at most a few minutes from the presentation of the reference voice sample. In long-term memory tasks, the stimulus voice sample is presented after a longer delay. In this case, the reference voice pattern will be stored in long-term memory either through rehearsal or because the reference voice pattern was already in long-term memory before the experimental session. Rehearsal is the process of, silently or overtly, "going over" information stored in short-term memory. In this way, the information can enter long-term memory. Otherwise, it is lost. By session I mean that period of time during which a number of stimulus voice samples are presented (as well as, possibly, some reference voice samples). A session usually lasts for not more than an hour, its time limits being governed by factors of listener performance. If a listener is required to perform a task for a period of time, his performance will first rise to an optimal level and later fall due to fatigue, boredom, etc. Results obtained after a listener has reached this latter stage will be statistically unreliable (unless one is performing a study of the fatigue factor). Similarly the experimenter normally tries to allow the listener to reach the optimal level before the test proper begins, by practice trials. An hour is a liberal estimate of the length of time before a listener's performance drops off; the practical duration of a session is usually much less than this.

Long-term memory tasks may be subdivided along similar lines to those of B & P. In tasks involving familiar reference voice patterns, these patterns have been acquired through social or business contact. However, the other half of this division should

consist of any other reference voice patterns stored in long-term memory. These will include

(i) patterns which have been presented to the listener at a much earlier stage in the experimental session, so that they have been stored in long-term memory through rehearsal (this is presumably what B & P mean by training), and

(ii) patterns which have been presented prior to the experimental session (as in McGehee's experiments).

I shall call both these kinds of reference voice patterns familiarised. The choice of familiarised patterns to be used is under the total control of the experimenter, whereas for familiar patterns he only has control over the selection of a subset from the total inventory (not over the identity of the total inventory, since this depends on the listener's social and business circles). An important point is that reference patterns of speakers which are familiarised assume the same status as those "familiarised" through social or business contact outside the experimental situation. This similarity is often overlooked, e.g. Tosi (1975):

> 'The long-term memory process is utilised when the voice to be identified is a familiar one to the listener. The short-term memory process is used when the unknown and known voices to be compared are available to the listeners through recordings.'
>
> (p.400)

Examples of experiments using familiar reference voice patterns are Pollack et al. (1954) and Abberton (1974), while Williams (1964) and Stevens et al. (1968) involve familiarisation of reference voices. Although it is not clear from Stevens et al. (1968) whether they consider their task to involve long-term memory, it must be categorised as such. Since the same eight reference voices are used throughout the experiment, a learning process (storage into long-term memory) must be involved.

Short-term memory tasks subsume what B & P categorise as contemporary tasks, in which the stimulus voice sample is presented after (often immediately after) the reference voice sample. However, there is another manner of presentation of voice samples, which has been ignored by experimenters: this is to present the two voice samples at the same time, i.e. one superimposed on the other, by double-tracking on a tape-recorder. I shall call this latter kind of task simultaneous presentation, and the former sequential presentation. Examples of sequential short-term memory tasks are Coleman (1973), Doehring & Ross (1972) and Shearme & Holmes (1959). I have not found any examples of simultaneous short-term memory speaker recognition tasks in the literature. Williamson (1961a) investigated the relative effects of sequential and simultaneous presentation of samples, and I have carried out experiments using a simultaneous presentation format. This format is discussed in greater detail in the next section.

Figure 4.1 summarises the revised version of the categorisation described above. The criterial feature for each division can be stated in terms of the time-span, not only between the presentation of reference and stimulus voice samples, but also, as a consequence, of the memory system involved. The terminal elements of the categorisation have been ordered to emphasise this temporal scale; the length of time for which the reference voice pattern has been stored in memory increases from the left to the right of the diagram.

By defining the divisions in terms of time-span, it is possible to avoid using the terms matching and naming, both of which are connotationally misleading. In a sense, all speaker designation tasks involve matching or comparison (of the reference and stimulus voice patterns). One might almost say that in evaluation tasks, stimulus

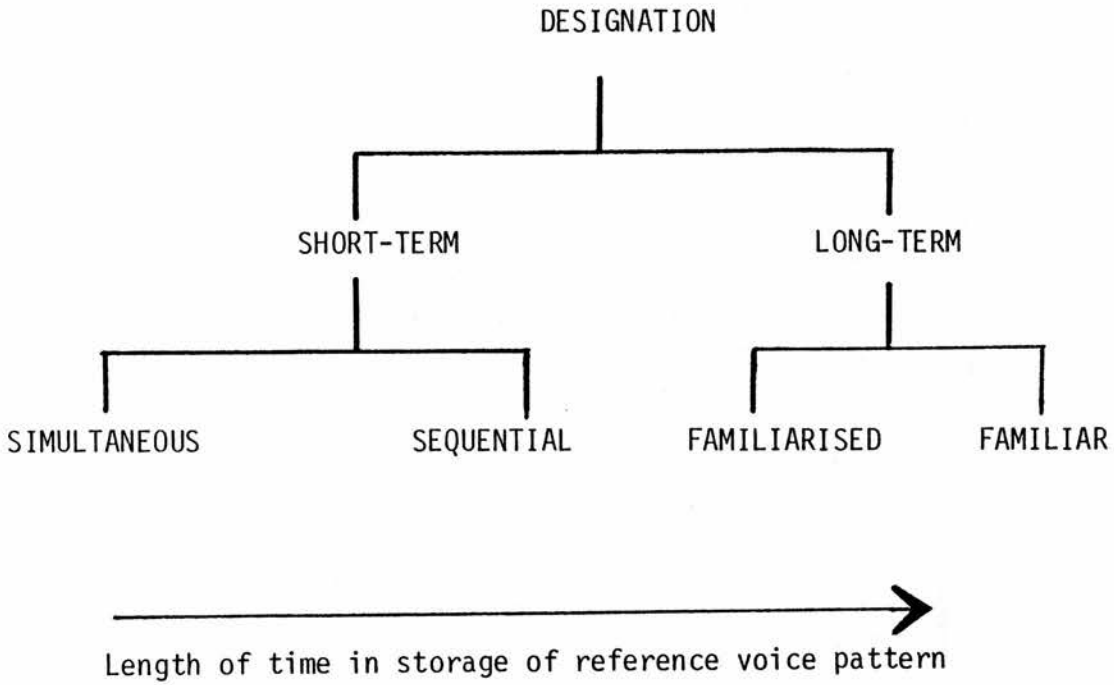


Figure 4.1 A categorisation of experimental speaker designation tasks (compare with Bricker & Pruzansky's taxonomy, Figure 2.2).

voices are matched with reference attributional norms. Similarly, naming is a misleading term in that it implies that there is a significance in knowing a reference speaker's name. As was explained in section 1.4, speaker recognition deals with the recognition of voices, not with the associated process of the connection of speaker identity characteristics (such as names) with those voices.

4.3.1 The Simultaneous Presentation Task

The simultaneous presentation format has its origins in the so-called cocktail party problem, discussed by Cherry (1957). Cherry defines the phenomenon as 'the ability to listen to, and follow, one speaker in the presence of others' (p.278). He rightly considers this to be a surprisingly efficient and important human faculty, and one which is usually taken for granted. In this respect, the term problem is therefore a misnomer since the human listener normally succeeds, albeit sometimes with some effort, in performing this discrimination. The situation only becomes a problem when one's normal ability is impaired by fatigue, hearing loss, overindulgence in cocktails, etc. Cherry's use of the term problem is directed more at the fact that no mechanical device has yet been produced which is capable of filtering out one voice from a number of others. Since this thesis is concerned with human rather than mechanical issues, I prefer to call the phenomenon the cocktail party situation.

To show that this discrimination is a normal human capability (although one requiring some mental effort), Cherry carried out a series of experiments (Cherry, 1953) in which listeners had to perform such discrimination tasks. The same format was used in all the experiments: a tape-recording was made of a reading of a passage, on top of which was then superimposed a reading of a different passage

by the same speaker. Cherry describes the resulting recording as 'a complete babel', which he then presented to the listener. The listener was required to separate out the two recordings and to dictate bits and pieces of phrases as he identified them. From his observations on the findings of these experiments, Cherry distinguishes four kinds of uncertainty, on which numerous inductive inferences must be carried out concurrently for communication to be achieved.

> '(1) Uncertainties of speech sounds, or acoustic patterning. Accents, tones, loudness may be varied; speakers may shout, sing, whisper, or talk with their mouths full.
>
> (2) Uncertainties of language and syntax. Sentence constructions differ; conversational language may be bound by few rules of syntax. Vocabularies vary; words may have near-synonyms, popular usages, special usages, et cetera.
>
> (3) Environmental uncertainties. Conversations are disturbed by street noises, by telephone bells, and background chatter.
>
> (4) Recognition uncertainties. Recognition depends upon the peculiar past experiences of the listener, upon his familiarity with the speaker's speech habits, knowledge of language, subject matter, et cetera.'
>
> (Cherry, 1957:277)

This categorisation and later series of experiments (Cherry, 1953; Cherry & Taylor, 1954) demonstrate that Cherry's major interest, as a communications theorist, is in the semantic and syntactic factors which have an influence on successful communication (see Clark & Clark (1977:216-219) for a summary of the selective listening experimentation). From a phonetic point of view, their interest lies in the fact that it is a prerequisite for the performance of the tasks set in the experiments (the correct dictation of passages) that the successful discrimination of the two acoustic signals has been

achieved; if the two signals cannot be distinguished, then the task cannot be performed.

That successful communication is normally achieved is a measure of the skill of the human perceptual process, which can draw on information from many sources. Thus, in the cocktail party situation, the clues whereby we are able to distinguish one speaker from all others and to follow what he is saying, include the following:

(1) visual clues - lip reading, accompanying gestures and facial expressions,

(2) localisation clues. The voices come from different directions and/or distances. This results in acoustic differences in intensity and phase (for a fuller description of these complex factors, see, for example, Stevens & Davis, 1938:ch.6),

(3) clues of speakers' personal production characteristics, both intrinsic and extrinsic, and

(4) semantic and syntactic transition probability clues.

In Cherry's series of experiments described above, the influence of all but the fourth of these factors was avoided by

(i) recording the readings on tape,

(ii) presenting them to listeners monophonically through headphones, and

(iii) using the same one speaker throughout.

Thus the fourth factor - semantic and syntactic transition probabilities - was investigated as the independent variable. This format can be easily modified, though, for the investigation of speaker recognition. Readings are again recorded on tape and presented monophonically to listeners. In the speaker recognition

experiment, the influence of the fourth factor may be avoided by

(i) having the speakers read the same passage, and

(ii) making the samples presented to listeners short enough for the recognition of phrases to be made impossible.

This leaves the third factor (speaker-characterising features) as the experimental variable.

Therefore, the basic framework for simultaneous presentation tasks has been with us since Cherry (1957). However, only one writer since then (Williamson, 1961a) has investigated the potential advantages of this format. Her study includes a comparison of sequential and simultaneous presentation formats, with striking results. These are summarised in Figure 4.2.

	Sequential stimulus presentation	Simultaneous stimulus presentation
A. Five-second stimulus duration	90%	63%
B. Ten-second stimulus duration	90%	70%

Figure 4.2 Mean percentages of correct judgments per test (from Williamson, 1961a:28)

For the sequential presentation format, a stimulus duration of five seconds is sufficient for listeners' performance to have reached an optimal level (90%), and an increase in the stimulus duration to ten seconds therefore does not produce an increase in performance. However, this is not the case for the simultaneous presentation format.

> 'When the superimposed items were of five-second duration, 63% were judged correctly. This is only slightly better than chance. When the speech sample duration was increased to ten seconds, the percentage of correct judgments increased to 70%. This difference in results is greater than could reasonably be attributed to chance. ($\chi^2 = 68.0$, $p < 0.1\%$). The duration of the speech sample, therefore, seems to affect speaker identification when the voices are presented simultaneously.'
>
> (Williamson, 1961a:32)

These results might lead us to the following hypothesis. The sequential presentation task requires the listener to calculate the voice patterns of the two voice samples and to perform a comparative decision on these. This is obviously not a very difficult task with samples of a reasonable length, as shown by Williamson's results. The simultaneous presentation task, on the other hand, requires not only this comparative decision but also the prior separating out of the two voice samples, which is a prerequisite for that decision. That this "compound" task is more difficult than the "single" task entailed in sequential presentation is witnessed by the much lower percentages of correct judgments. The above explanation is tentatively proposed because other explanations of this phenomenon are possible. One of these alternatives is based on observations on experiments I have carried out using a simultaneous presentation format (see Chapter 7). It is a stumbling block for the thinking underlying the above explanation that different strategies may be adopted by listeners when performing this task. The above explanation presupposes that the task is performed in a similar way to

other tasks; namely, that reference and stimulus voice patterns are extracted from the two voice samples and are then compared. The simultaneous presentation task performed in this way would involve a very short-term form of memory, and the sensory image stage of memory might also play a part. It can be seen, though, that it is somewhat arbitrary to refer to one of the simultaneously presented voices as the reference and the other as the stimulus. Instead, this kind of task may be performed not on a reference-stimulus basis but rather on a discrimination basis; that is, listeners may base their responses on their ability to separate the two voices. This seemed often to be the case in the experiments which I carried out. If the listener succeeded in separating out the two voices (in a 2.5 second sample duration in my experiments), he took this as an indication that the voices were "different", whereas, if he did not succeed in this separation, he would give a "same" response. In this case, the actual comparison of the two voice patterns may play hardly any part in the decision. This difference in strategy may lead to a difference in the response given. If the "separability" approach is adopted, inability to distinguish the two voices will result in a "same speaker" response. On the other hand, if the task is performed on a reference-stimulus basis, failure to separate out the voices will probably result in a "don't know" response.

Among the experiments reported in Pollack et al. (1954) is one using a simultaneous presentation format. However, there is an important difference between their tasks and the one described above: in their experiment, listeners were presented with simultaneous recordings of two stimulus samples corresponding to familiar reference voices, both of which were to be identified. In other words, their listeners were presented with two stimulus samples and no reference samples, as opposed to one stimulus and one reference as in (one interpretation of) the above task. It is interesting to read, though, that they also found differences in performance, dependent upon the strategy adopted by the listener.

> 'Wide individual differences among listeners were obtained in the proficiency of identification of combination of voices, especially at short durations. It appears that listeners, who are able to identify the combinations of voices as unique patterns, identify these combinations nearly as well as they identify single voices. Other listeners, who attempt to abstract each of the two voices, individually, identify the combinations more poorly than the individual voices.'
>
> (Pollack et al., 1954:405, FN)

The simultaneous presentation format is one of phonetic and psychological interest. This interest arises partly from the fact that the simultaneous presentation situation is one which is unnatural, in that it cannot occur in the real world with a choice of "same" or "different" speakers. In short, two simultaneously heard voices must belong to different speakers in the real world situation (if we disregard the possibility of tape-recorders, etc. since these belong to the experimental situation). This point is discussed further in section 4.5, where the categorisation of experimental speaker recognition tasks is related to real world possibilities.

4.4. DECISION PROCESSES IN EXPERIMENTAL SPEAKER RECOGNITION TASKS

As was mentioned above, the evaluation half of Bricker & Pruzansky's taxonomy will not be discussed further. In this section a categorisation is presented of a different aspect of the other half of that taxonomy (the identification half, which I prefer to call designation - the reason for this substitution will become clear in this section). The previous two sections contained a categorisation in terms of the memory systems involved in experimental speaker recognition tasks; the categorisation in this section is in terms of the decision process which takes place in the listener when

performing such experimental tasks. This categorisation is conveniently approached by considering the terms which experimenters have used to refer to these various tasks. It will be seen that significant correspondences can be stated between these task terms and the elements of the temporal categorisation of section 4.3 (Figure 4.1).

There are eight terms which I shall consider: speaker authentication, speaker differentiation, speaker discrimination, speaker identification, speaker recognition, speaker validation, speaker verification and same-different testing. This should not be taken to be an exhaustive list of the terms used to describe speaker recognition tasks. However, these eight are by far the most common terms used in the speaker recognition literature. This list covers the full range of tasks so far used by experimenters; any other terms which may be found in the literature will be comparable, roughly, with one of them.

These eight task terms can be classified into three groups:

1. Identification, recognition
2. Differentiation, discrimination, same-different
3. Verification, authentication, validation.

The terms within each group are, more or less, interchangeable; that is, they are used throughout the literature to refer to the same kind of task. It should be remembered that what I am discussing here are the kinds of task which experimenters have so far used, and the terms which they have used to refer to these tasks. I am not discussing the full range of tasks which it is logically possible for experimenters to give listeners to perform. In this respect, this categorisation differs from the temporal categorisation described above, which includes elements which are logically possible but which have not been employed in the literature (comparative

evaluation tasks). In addition, this categorisation is of the experimental tasks which listeners have been required to perform. The relationship between the elements of this categorisation and those categories underlying what may occur in the real world is examined in section 4.5.

Let me describe the typical meaning (excluding anomalous, idiosyncratic usages) associated with each term. I shall use the label identification in future to refer to the first group, differentiation for the second, and verification for the third.

Identification

The terms in this group are used in two distinct senses:

(i) a general sense, in which the terms are used to refer to the whole field. This is especially true of speaker recognition, but also, to a lesser degree, of speaker identification. In this sense, the terms subsume differentiation, verification and the second sense of identification tasks described below. The term speaker recognition was perhaps adopted to describe the whole field on analogy with the established term for the related field speech recognition.

(ii) a more specific sense, where they may apply to short-term or long-term memory tasks. The task in both cases requires the listener to decide "Which of the reference voices does the stimulus voice correspond to?"

Long-term memory identification tasks may involve familiarised or familiar reference voice patterns, potentially quite great in number because of the very large capacity of long-term memory. This kind of task is presumably what B & P intend by their naming category.

Short-term memory tasks for which the term identification has been used are of two kinds:

(a) the fixed-sequence method of presentation. One version of this is the so-called ABX format,

where the listener is presented with a first reference voice (A), then a second reference voice (B), and finally the stimulus voice (X) which corresponds to either A or B. The fixed-sequence method has also been used with more than two reference voices. For this, certain experimenters have presented the stimulus sample before the reference samples, although, as is shown below, this has a significantly different effect on listener performance from presenting the stimulus sample after the reference samples.

(b) the free-comparison system. Listeners are able to refer at will to reference voices and may return to the stimulus voice at any time for direct comparison. This method imposes complex mechanical requirements.

Differentiation

The terms in this group have only been applied to tasks involving short-term memory. A typical format is to present the listener with two voice samples and require him to decide "Are these samples spoken by the same or different speakers?" The reference voice pattern is therefore acquired on the basis of the brief exposure to the first of these voice samples.

Verification

The terms in this group are typically used in the context of automatic speaker recognition, where they correspond to the task involved in security situations. The machine is then being required to decide "Is the input speaker who he claims to be?"

In the context of speaker recognition by humans, the task is typically one using long-term memory. The task closely resembles the machine recognition task described above in that the listener is required to compare the stimulus voice with a specific reference voice pattern. This corresponds to the task usually involved in court situations.

The speaker recognition task terms may be correlated with the temporal categorisation of speaker recognition tasks given above (Figure 4.1). The relationships which hold, in the typical usages of the terms, are given in Figure 4.3. An asterisk indicates that that task term has been typically applied to that kind of task in respect of the temporal categorisation. A blank indicates that it has not (or would not).

	Short-term		Long-term	
	Simultaneous	Sequential	Familiarised	Familiar
1. Identification		(*)	*	*
2. Differentiation	*	*		
3. Verification			*	*

Figure 4.3 A classification of the realm typically implied by speaker recognition task terms, in respect of the temporal categorisation of Figure 4.1.

The reason why the intersection of identification tasks and sequential presentation has been assigned a bracketed asterisk is as follows. It has been assigned the asterisk because several writers have used the term identification to refer to sequential presentation tasks. The dubiety implied by the bracketing has been included

because it may be argued that the two forms of short-term memory identification tasks (fixed-sequence and free-comparison) fall within the category for which the term differentiation has generally been used. Let me consider the two forms of presentation separately.

(a) the fixed-sequence method. Hecker (1971) points out that

> 'in an ABX design, for example, some listeners may compare X individually with A and B, while other listeners may disregard A and only compare X with B. This may lead to large individual differences.'
> (Hecker, 1971:32; see also Clarke et al, 1966)

In other words, in the second case, the ABX format is being simplified into a BX format. Of course, this can only be done if X corresponds to either A or B (i.e. there are no other possible alternatives), and if the listener is aware of this. However, when this simplification is adopted, there is only one reference voice under consideration instead of two.

The other form of fixed-sequence presentation, where more than two reference voices are used, is less easily analysed. Certainly the involvement of several reference voice patterns imposes severe restrictions on the listener's performance. It has been shown (Clarke et al., 1966; Doehring & Ross, 1972) that the separation between the stimulus voice sample and the correct reference voice sample has a significant effect on listener performance. Clarke et al. found that when the stimulus voice sample followed the reference voice samples, correct "identifications" fell from nearly 80% when there was no intervening sample to below 50% when one sample intervened (see Figure 4.4). This effect was very different from the situation where the stimulus voice sample preceded the reference voice samples. Here there was only a slight degradation in performance.

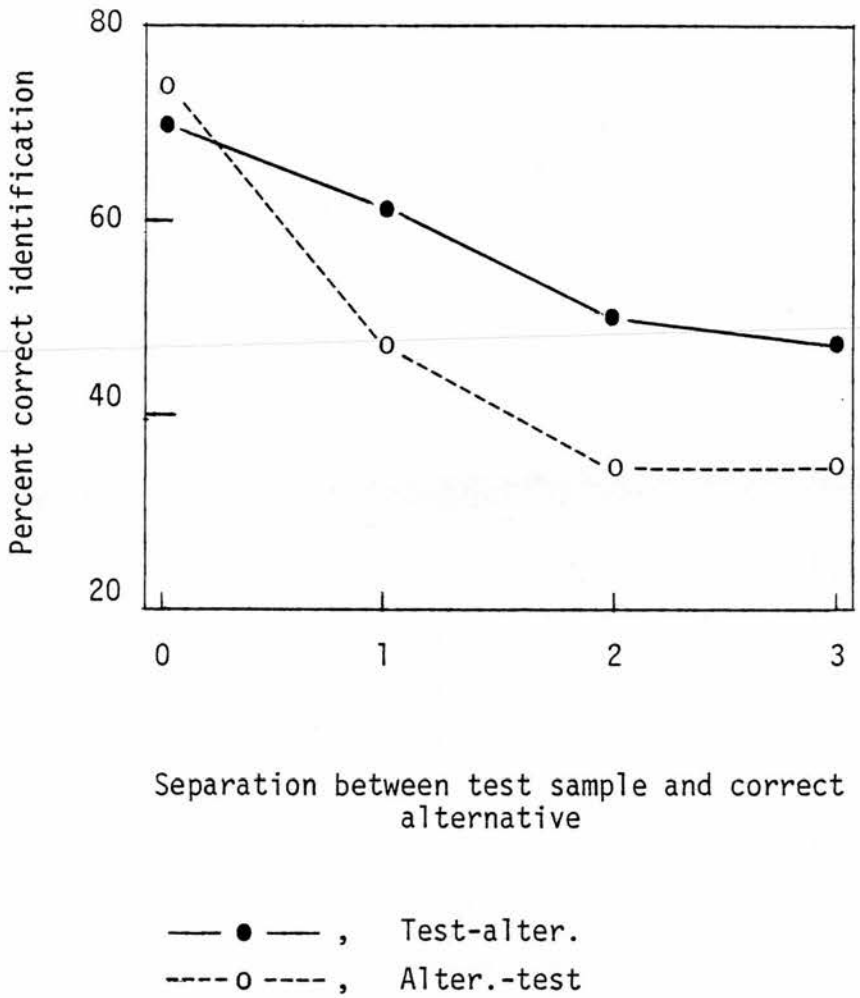


Figure 4.4 Effect on speaker identifiability of separation between test sample and correct alternative in four-choice identification test. Data are shown for both orders of presentation of test sample and alternatives. (Reproduced, with permission, from Clarke et al., 1966:11).

This whole argument hinges on the storage capacity of short-term memory, i.e. the physiological limitations upon implementation, about which relatively little is known. However, this evidence suggests

that short-term memory is incapable of reliably handling more than one reference voice sample (whilst the storage capacity of long-term memory is much larger and obviously capable of this). The case where the stimulus sample is presented after the reference samples necessitates several reference voice patterns being held in short-term memory. This is reflected in the low performance for this kind of task and is probably why alternative strategies are adopted. In the other situation, where the stimulus sample is presented before the reference samples, only one pattern has to be stored in short-term memory (the "stimulus" voice pattern), and this accounts for the better performance. In this case, it seems difficult to justify the use of the term "stimulus", as this sample comprises the pattern which is stored in memory, for which the term "reference" is normally used.

(b) the free-comparison method. This method permits direct sequential comparison between the stimulus voice sample and any reference voice sample. Although several reference voices may be used, only one need be held in short-term memory by the listener at any one time.

Therefore the free-comparison method of presentation and fixed-sequence presentation tasks of the ABX type may both become tasks in which the stimulus voice sample is directly compared with only one reference voice sample, and this is the kind of process for which the term differentiation has generally been reserved. The tasks under consideration in Clarke et al's experiments do not involve simplification, but their results show that listener performance is severely degraded when more than one reference voice needs to be taken into account, i.e. human listeners cannot perform this task with a success rate comparable to that achieved in other speaker recognition tasks.

Therefore, although the above tasks superficially contain more than one reference voice, if we consider the strategies adopted by listeners, and the decision processes underlying the tasks, they generally involve only one reference voice pattern at any time, and are thus what are generally referred to as differentiation tasks.

The major difference between all the tasks described in this section lies in the number of reference voice patterns under consideration. Apart from that difference, all the tasks involve much the same decision process; namely, the listener has to compare a stimulus voice pattern with one or more reference voice patterns. The fact that they all involve a matching or comparison was the reason why the term matching was considered a misleading label for a node in Bricker & Pruzansky's taxonomy (section 2.3). Where there is only one reference voice pattern, I shall call this decision a single comparison, and where there are more than one, a multiple comparison. This is a slight simplification: for example, the thresholds used in a single comparison decision will be somewhat different from those in a multiple comparison decision.

It is now possible to correlate the above three factors (memory system, decision process and task term) with each other, but only if one adopts a rather prescriptive position in relation to the task terms. The position is prescriptive, not in the more usual sense that one is proposing how people ought to use the terms, but in the sense that one is describing the most widely used meaning applied to the term. The meanings implied by the terms identification, differentiation and verification in the categorisation below are thus more literal than the flexible usages described previously. The categorisation is represented in Figure 4.5.

The lack of a term to describe short-term memory multiple comparisons is due to the incompatibility of the two factors, which was argued above.

	Single comparison	Multiple comparison
Short-term memory	Differentiation	
Long-term memory	Verification	Identification

Figure 4.5. A categorisation of the decision processes involved in speaker recognition tasks, using literal senses of the terms identification, differentiation and verification.

The above categorisations have all been stated as formally as possible, since a consideration of such issues is a prerequisite for any formal model of the speaker recognition process, such as the one presented in this chapter. However, for the sake of formalisation, certain problems and associated issues have been simplified or omitted. A limitation of minor importance is that the categorisation given above of the terms experimenters have used to refer to speaker recognition tasks overlooks the fact that within each of the three groups (identification, differentiation, verification), the various terms may connotationally imply different aspects of the same decision process. For example, in the title of Tosi's (1975) article ("The problem of speaker identification and elimination"), on the use of speaker recognition evidence in legal situations, it is easily appreciated that the term identification focusses on the acceptance of the stimulus speaker as one of the suspects, whereas the term elimination focusses on his rejection. However, the task implied by both terms involves the same decision process on the part of the listener and for this reason would be treated similarly in the categorisation. The term elimination was omitted from the list at

the beginning of this section because the only instance of its use which I have found is in the work of Tosi (1975, 1979).

There is one situation which occurs very commonly in the real world, but which has been ignored in the above discussions. Experimentally, this is the format where the stimulus utterance may have been produced by a speaker who does not correspond to any of the population of reference speakers. The listener is usually told this fact by the experimenter. This is therefore an identification task since it involves long-term memory and more than one reference voice; I shall refer to it as open identification (as opposed to the closed task discussed previously). The number of response alternatives available to the listener in the experimental situation is thus one greater than the number of reference speakers, the extra alternative corresponding to "the stimulus utterance was not produced by any of the reference speakers". Further discussion of this task is delayed until section 4.8, where the process underlying the task is explained more easily in terms of the speaker recognition model to be presented in section 4.6.

4.4.1 Response Alternatives

Since the decision processes and the number of reference voice patterns under consideration differ for the three kinds of task defined above (identification, differentiation, verification), there is also a difference in the number and nature of response alternatives available to the listener in the experimental situation (Figure 4.6).

The similarities and differences between the tasks can be stated quite simply. Open identification has one more response alternative than closed; the extra possibility is the "stimulus does not correspond to any of the references" response. In Figure 4.6 the label "speaker A/B/etc." should be interpreted as an abbreviation for "stimulus is speaker A", "stimulus is speaker B", etc. That is,

TASK		RESPONSES
1. Identification	(a) closed	(1) "Speaker A/B/etc." (2) "Don't know"
	(b) open	(1) "Speaker A/B/etc." (2) "Stimulus does not correspond to any of the references" (3) "Don't know"
2. Differentiation		(1) "Same speaker" (2) "Different speakers" (3) "Don't know"
3. Verification		(1) "Stimulus is speaker X" (= "same speaker") (2) "Stimulus is not speaker X" (= "different speakers") (3) "Don't know"

Figure 4.6 Response alternatives for experimental speaker recognition tasks.

this abbreviation represents as many response categories as there are reference speakers - they do not all belong to the same single response category. The number of reference speakers used is, of course, a matter of choice on the part of the experimenter. These response categories are clustered together by this abbreviatory convention in Figure 4.6 because the decision process underlying a response "stimulus is speaker A" is of the same kind as for a response "stimulus is speaker B", whilst a "don't know" response is totally different in nature. The two "don't know" responses have different statuses for the two tasks; the nature of, and the need for, these "don't know" responses are discussed below.

Since both kinds of identification involve the long-term memory system, reference speakers will always be known to the listeners, either by name, or by some other form of identity (letters, numbers, etc.). The distinction between reference speakers whom the listener knows personally and those he does not is the sole cause of the difference in the response categories for differentiation and for verification tasks. The response alternatives are identical in effect; they differ only in wording. Therefore, the response "stimulus is speaker X" in a verification task represents the same process and final decision as the "same speaker" response for differentiation.

The alternatives shown in Figure 4.6 represent the greatest range of responses available to the listener in an experimental task. In particular, a "don't know" response is shown as an alternative for each kind of task, as a means whereby the listener can express his extreme lack of confidence in any other response; that is, for him to give any other response would be little more than a best guess. However, not all experimenters allow listeners the luxury of a "don't know" alternative. Where this response is denied them, the task is referred to as a <u>forced choice</u> format; listeners must respond with

one of the designatory response categories. However, it is easily appreciated that if listeners are only allowed these responses, then they have no way of showing that some of their responses are made very confidently while they may have serious doubts as to the correctness of others. Similarly, the experimenter has no way of avoiding assigning equal weight to all responses when he comes to score them, and thus the difference in listener confidence is neutralised.

One way in which this difference may be preserved is to allow listeners to assign confidence ratings to their responses. They still have to give a designatory response but can now express whether they are dubious as to its correctness. A typical scheme is a three point confidence rating scale corresponding to "very sure"/ "fairly sure"/"unsure". Data obtained in this way can be used to plot Receiver Operating Characteristic (ROC) curves, which provide measures of the decision thresholds used by listeners in tasks (whether, for example, one listener in a differentiation task is more inclined to respond "same speaker" than another listener) and thereby of whether listeners, although using different thresholds, are equally capable of discriminating between speakers. The theory of ROC curves is put forward in Egan et al. (1959) and is summarised concisely in Hecker (1971) and Tosi (1979). The statistical details will not be discussed further here.

A second means for the listener to express extreme lack of confidence is the "don't know" alternative. The "don't know" responses in Figure 4.6 are not homogeneous in that they do not carry the same weighting. Let us suppose that, in an experimental task, listeners are allowed neither a "don't know" response nor a confidence rating. In other words, they have to make a response of one of the designatory categories. Let us further suppose that one such listener is very dubious about a decision:

(i) If the task is one of differentiation, there are different probabilities depending on the task format and the strategy adopted by the listener:

(a) if there is a simultaneous presentation format and the listener adopts a "separability" approach, he will give a "same speaker" response, as was argued in section 4.3.1.

(b) if there is either a simultaneous presentation format and he adopts a reference-stimulus approach, or a sequential presentation format, he is likely to give a "different speakers" response. In these situations, a "same speaker" response is conceptually considered a positive designation, whereas the "different speakers" response is taken as the negative, non-committal alternative. The theoretical justification for this strategy being adopted by the listener is dubious. However, in my experiments reported in Chapter 7, listeners who said that they had not adopted the "separability" strategy reported that they felt a "same speaker" response to be one of positive recognition of similarity, whereas lack of this recognition (and also moderate doubt about recognition) was taken as an indication that the voices were "different". In short, listeners expected "same speaker" responses to be made with some confidence.

(ii) The probable response in a verification task will correspond to the latter given above, for the same reasoning; namely "stimulus is not speaker X" (corresponding to "different speakers"). This is especially probable in a court situation, where a positive acceptance might have important consequences, and where a jury would treat an unconfident listener's acceptance response as evidence of dubious validity.

(iii) In an open identification task, he will probably respond that "the stimulus does not correspond to any of the references" since a "speaker A/B/etc." response will be considered a positive recognition in the same way as the "same speaker" response just discussed.

(iv) The listener in a closed identification task may find himself unable to respond. He will not want to give a positive "speaker A/B/etc." response for the reason just stated. This will be especially so if none of the reference alternatives seem at all similar to the stimulus and to give any positive response would be simply a guess. In this task, therefore, the listener has no means at all of expressing his doubt, whereas, in the other tasks, he at least has some means even if they are rather indirect and of dubious theoretical justification. The inclusion of a "don't know" response alternative is thus most needed in closed identification tasks. In the model of the speaker recognition process presented in this chapter, the "don't know" response is kept as a full and required alternative for closed identification tasks; it is not included as a feature of the other tasks because it is more of an optionally available response for them, and one which many experimenters choose to omit. Omission of the "don't know" alternative facilitates the process of scoring responses and errors (see the next section).

4.4.2 Error Possibilities

Just as there are different response alternatives for the tasks just described, so there are differences in the possibilities for response error. The terms used in Figure 4.7 to refer to the different error possibilities derive partly from Tosi et al. (1972), although it should be pointed out that their study deals with the process of speaker recognition by the visual examination of spectrograms.

The "don't know" responses contained in the previous figure have been disregarded for the purpose of this categorisation. As was explained above, the "don't know" alternative is important only for closed identification tasks. However, it might be argued that a "don't know" response is a weak form of error, in that incorrect responses and "don't know" responses both constitute failures to

TASK	ERROR	EXAMPLE
1. Identification (a) closed	(1) False identification	Stimulus was speaker A, but "speaker B" response was given
(b) open	(1) False match identification	Stimulus was speaker A, but "speaker B" response was given
	(2) False no-match identification	Stimulus did not correspond to any of the references, but "speaker B" response was given
	(3) False elimination	Stimulus was speaker A, but response "stimulus does not correspond to any of the references" was given
2. Differentiation	(1) False similarity	Stimulus was different from reference, but "same speaker" response was given
	(2) False dissimilarity	Stimulus was the same as reference, but "different speakers" response was given
3. Verification	(1) False acceptance	Stimulus was not speaker X, but "stimulus is speaker X" response was given
	(2) False rejection	Stimulus was speaker X, but "stimulus is not speaker X" response was given

Figure 4.7 Error possibilities for experimental speaker recognition tasks.

recognise a correspondence or lack of correspondence between the stimulus and reference(s). This view would lead to great difficulties in assigning weightings to the various kinds of error when scoring responses, and is not a generally held view. In closed identification tasks, therefore, only one kind of error is possible.

There are three forms of possible error in open identification tasks. In Figure 4.7, the term match denotes that the stimulus speaker in fact corresponds to one of the references; no-match means that there is no such correspondence. Listeners may

(i) identify the stimulus speaker as one of the references, when he in fact corresponds to another of the references,

(ii) identify the stimulus as one of the references, when he is not in fact represented among the reference population, or

(iii) reject the stimulus as any of the reference population, when he is in fact represented in it.

It is debatable exactly what relative weighting should be assigned to each of these three kinds of error. Such problems are compounded if confidence ratings are also used. For this reason, the open identification task has not been employed greatly in the literature, even though it corresponds to a very common real world situation (see the next section). Experimentally, this is a less attractive format to the experimenter than closed identification which, having only one error possibility, lends itself to being combined with confidence ratings.

The similarity seen previously between the response alternatives for differentiation and verification tasks is mirrored by the similarity in their possibilities for error. The difference is again one of wording, not of effect, and derives from the difference in memory system involved. Thus, the false acceptance category

corresponds to false similarity, and false rejection to false dissimilarity.

4.5 REAL WORLD TASKS

The above categorisations relate to the listener's performance in the experimental situation. However, the range of experimental possibilities is greater than the range of situations which usually occur in the real world. Consideration of the experimental possibilities enabled the categorisations given above to be more exhaustive than if they had related solely to the real world range of possibilities.

In relation to speaker recognition, a definition of the real world must include the recognition of live voices in everyday situations. Nowadays, use of such devices as the telephone, television and radio must also be included as belonging to everyday life, as must the use of audio- and video-tape recorders. However, the latter (and also, possibly, the former) may be employed in experimental tasks. Therefore, it is perhaps best simply to define the real world as the opposite of the experimental situation, i.e. as being characterised by the lack of selection of voices, controlled variables, pressure to produce a conscious response, etc.

The easiest way to discuss the real world possibilities is to examine the experimental tasks (Figures 4.1 and 4.5) and to decide whether they are still possible.

As was explained in section 4.3.1, the simultaneous presentation format is artificial in that it cannot exist in the real world while preserving the choice of "same"/"different" alternatives; that is, in the real world, two simultaneously heard voices must belong

to different speakers. If there is no choice of alternative, then there is no question of differentiation. This is not to say, however, that this situation is of no interest to speaker recognition. Two simultaneously heard voices may still need to be identified; the two voices then both act as stimuli, rather than one constituting the stimulus and the other the reference (as for a differentiation task). This task of identifying two simulultaneously heard voices corresponds to the experimental task set by Pollack et al. (1954), discussed in section 4.3.1.

The main characteristic of the distinction between familiar and familiarised reference voice patterns is that the choice of familiarised patterns to be used is under the total control of the experimenter. But since the control of the experimenter is precluded in the real world, this feature cannot be used as a defining characteristic. The distinction then rests on the time factor and it becomes a matter of deciding the length of time in storage for patterns to be familiar rather than familiarised. Any decision of this sort would be arbitrary and therefore any attempt to subdivide the category of reference voice patterns stored in long-term memory will be abandoned.

All three kinds of task (identification, differentiation, verification) are possible in the real world. An example of identification would be the situation where you are in a room with a number of other people. Having turned yourback, you hear somebody speak. The voice will therefore correspond to one of the people in the room. Naturally, in this case, speaker recognition will be helped by localisation clues (see section 1.5), but speaker recognition by voice alone will play an important part.

An example of differentiation is a situation where you telephone an office and are connected initially to the switchboard operator. Having asked to be put through to a certain extension, you then wait. The next voice to be heard may be either a different one (an unknown voice replying on the required extension) or the same

one (the switchboard operator, to tell you that she cannot get a reply on that extension).

Verification in the real world usually takes place in court situations, where a witness is required to recognise the voice of one specific suspect. If the suspect were required to produce an utterance in court, knowing that this might form the basis of his conviction, it is likely either that he would try to subtly disguise his voice, or that any utterance he produced would not represent his normal articulation. For this reason, the suspect's voice is normally presented as a recording made when the suspect was not aware of it. This use of the tape-recorder is experimental rather than everyday, and would be excluded by the definition of the real world given above. Similarly, it is odd to refer to speaker recognition "tasks" in the real world. The process is rarely conscious and rarely has important consequences - you are under little pressure to achieve success. However, this is not true for the usual kind of verification task described above, even if the suspect's voice is heard as a live utterance. This kind of task is therefore perhaps best omitted from any discussion of real world possibilities.

Other forms of verification do occur, though less frequently, in the real world. An example of this would be the situation where you telephone a friend and ask him to call back (e.g. because you are using somebody else's phone and do not want to increase their bill). When the telephone next rings, you are therefore expecting your friend and the immediate task is to verify the caller as your friend as against anyone else.

Verification tasks are rare in the real world, probably because they require that strong expectations be had as to the identity of the voice. If expectations are not strong, the task turns into one of identification because the population of probable speakers increases above one.

The categorisation of experimental tasks may therefore be revised for real world situations as summarised in Figure 4.8.

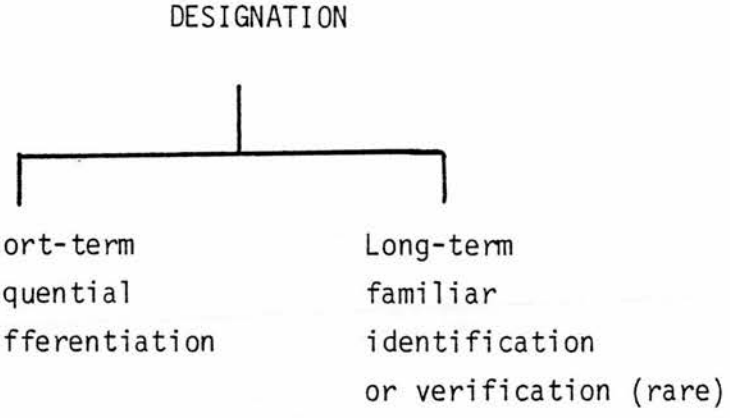


Figure 4.8 A categorisation of real world speaker designation tasks

4.5.1 Expectations for Probable Reference Candidates

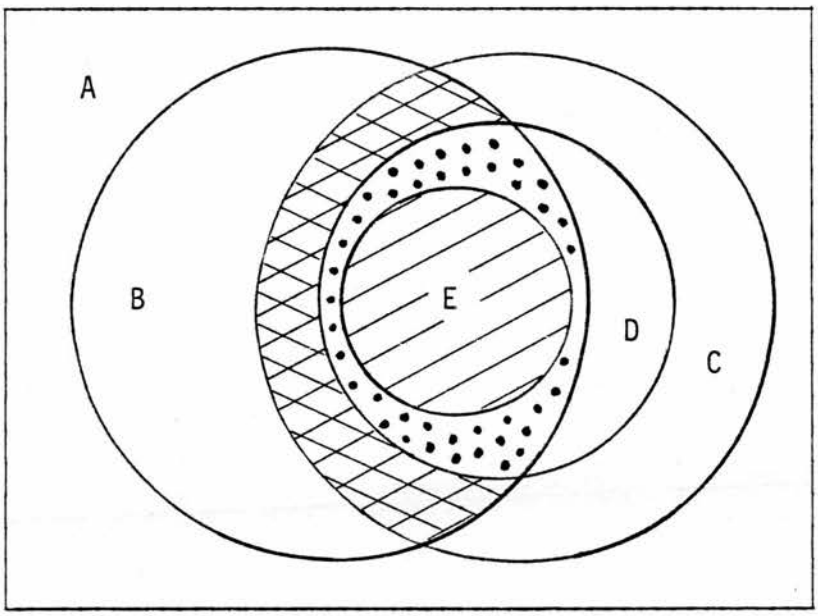
A distinction will now be discussed which has important consequences for the identification part of the model presented in this chapter, in which more than one reference voice pattern is involved. It is perhaps best explained by an everyday example. If the telephone rings in my office, I may expect the caller usually to be an acquaintance of mine. However, it is possible that the caller is someone I have never met before, in which case I will not have an appropriate reference pattern for the caller. Because this possibility exists, the number of reference voices which need to be taken into account is equal to the total population of reference voices in my memory, and I must also allow an extra category corresponding to "any other speaker". In other words, this situation corresponds to the experimental task of open identification, discussed

in sections 4.4 and 4.8. It is ludicrous from a psychological point of view to suppose that the way in which I recognise the stimulus voice is by comparing it individually with each of my population of reference voices. This process would obviously involve a large amount of redundant psychological processing, because there are many speakers in my reference population who are unlikely to be telephoning me in my office. The process would also take a great length of time to be performed - longer than identification in such a situation normally takes. Some form of streamlining is therefore probable, in the form of the temporary discarding from consideration of reference voices with a low probability of corresponding to the stimulus. This might be done in two ways - and most probably, in fact, by a combination of these two.

4.5.1.1 Situational Probabilities

If the telephone rings in my office, it is possible for the caller to be (literally) anyone in the world. However, it is more likely to be someone I know. Similarly, it is more likely to be someone in Edinburgh; more specifically, in the University; even more specifically, in the Linguistics Department; and so on. In other words, the situation in which the speaker recognition process is called into operation allows the range of expected stimulus voices to be narrowed down. The above example may therefore be expressed conveniently in Venn diagram terms in Figure 4.9.

If the telephone rings in my office, the most probable set of reference speakers is represented as the shaded area in Figure 4.9, corresponding to people I know in the Edinburgh University Linguistics Department; the next most probable set is composed of the previous set plus those in the dotted area, corresponding in total to people I know in Edinburgh University; the next most probable is that set plus those in the cross-hatched area, in all representing people I know in Edinburgh; and so on. It may then



rectangle A = the universe (all people in the world)
circle B = people I know
circle C = people in Edinburgh
circle D = people in the University
circle E = people in the Linguistics Department

Figure 4.9 Situational probabilities for reference candidates.

be hypothesised that the initial comparison of the stimulus voice is with the relatively small number of reference voices contained in the most probable set - in this case, the shaded area.

If that analysis fails to produce a satisfactory match between the stimulus voice and one of the references, the set of reference voices under consideration is enlarged to include the

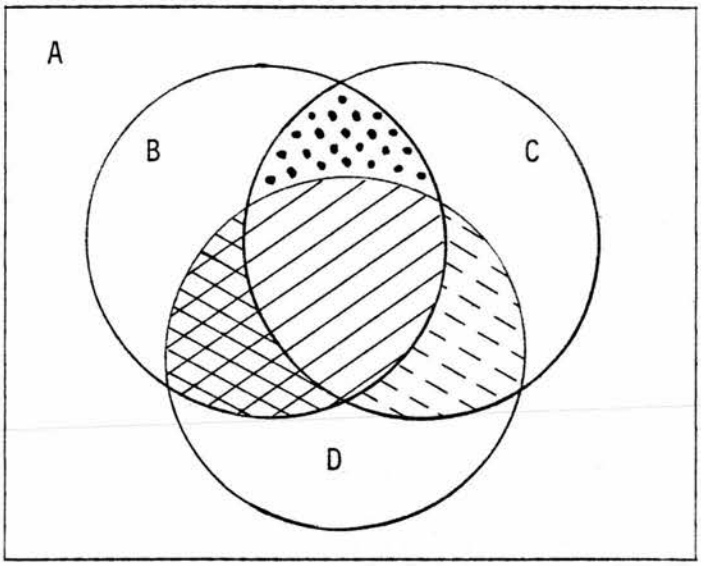
next most probable set (the dotted area in the diagram). The total set is now composed of the shaded area (people I know in Edinburgh University Linguistics Department) and the dotted area (people I know in Edinburgh University, but outside the Linguistics Department), although the former will already have been considered and rejected. If an analysis of the latter set fails to produce a match, the set is expanded again to include the cross-hatched area, and so on.

An attractive feature of this process is that the initial set taken into consideration is relatively small, and each expansion of the set is likely to produce an increase in the number of new reference voices to be taken into account. Economies are therefore made by starting with a small number of very probable reference candidates before proceeding to any larger population.

4.5.1.2 Parametric Probabilities

The second approach to improving the efficiency of reference voice recall in identification tasks involving large populations of possible candidates can be taken from the point of view of features of the stimulus voice. The stimulus voice is analysed parametrically, and only those reference voices which approximate closely in parametric features to the stimulus are called up initially for comparison. Thus, if a stimulus voice is analysed as having a high mean pitch, low mean loudness and extremely breathy phonation, only those reference voices which are characterised by similiar pitch, loudness and phonation values are called up. Again, a Venn diagram captures well this narrowing-down of possibilities (Figure 4.10).

The most probable set of reference voices from the total population corresponds to the shaded area, and these reference voices are taken into consideration first. If the analysis finds that the set is empty (i.e. there are no reference voices with high mean pitch, low mean loudness and extremely breathy phonation), the set is enlarged



rectangle A	=	the universe (total population of reference voices in memory)
circle B	=	voices with high mean pitch value
circle C	=	voices with low mean loudness value
circle D	=	voices with extremely breathy phonation

Figure 4.10 Parametric probabilities for reference candidates.

by the inclusion of further voices which approximate closely to that set and are thus the next most probable - perhaps the dotted area in the diagram (voices with high mean pitch, low mean loudness, but not extremely breathy phonation), or the cross-hatched or broken-shaded areas, etc.

This process of narrowing-down the range of probable reference voices is made all the more efficient if it can be shown that the stimulus utterance contains an abnormal value in relation to a parameter, i.e. one deviating to a large degree from some

physiologically and culturally determined norm. By virtue of the fact that it is an abnormal value, any such occurrence will cause the number of probable candidates to be reduced drastically. In terms of the Venn diagram, an abnormally high mean pitch value will produce a much smaller circle B in the above diagram. Since the probability of any reference voice falling within this (now much smaller) circle B is small, it will be a strong characteristic of any reference voice if it does fall within it. Therefore consideration of any parameter which does not have an abnormal value in the stimulus may be suspended while it is checked whether any reference voice contains the abnormal parameter value. However, it may be that this reduction as a result of an abnormality value leads to an empty set (i.e. there are no reference voice patterns with similar abnormal parameter values). In this case, the process will return to the above procedure for non-abnormal parametric values.

The same principles of economy operate here as with the first suggestion; namely, analysis starts with a small number of very probable reference voices and this set is enlarged only if there is a need to, because recognition has not occurred using the small set.

4.5.2 Multiple Task Situations

So far, the different elements of the categorisation of speaker recognition tasks (identification, differentiation, verification) have been taken as independent categories. In the experimental situation, this view is justified; listeners are aware of what is required of them, how many response alternatives are available, whether they are allowed a "don't know" response or if it is instead a forced-choice format, etc. In the real world, on the other hand, it is not such a clear-cut categorisation to apply. Whether a given situation constitutes one kind of task rather than another, or whether the situation may involve a combination of tasks depends largely on the listener's expectations, as was discussed for real world verification

tasks (section 4.5). The fact that expectations are involved, and that these are often covert and difficult to determine empirically, means that any categorical statement of this sort is impossible. However, we may assume that our intuitive analytical classifications are often correct.

This point is illustrated by a hypothetical situation contained in the following plausible dialogue:

Karen is Maurice's wife. She works in an office with Sara, Pam and Marie, all of whom Maurice knows. She is senior to these people and therefore usually answers the phone, which is on her desk. Maurice is trying to contact her by phone. He dials the office number.

TEXT	ANALYSIS
1. Switchboard girl: "Ryan Associates."	
2. Maurice: "Extension 215, please."	
3. Switchboard girl: "One moment, please." (pause)	<——— (D) (for Maurice)
4. Pam (on extension 215): "Hello." (pause)	<——— D/V/I (for Maurice)
5. Maurice: "Hello ... Pam?"	<——— I (for Pam)
6. Pam: "Yes. Hello, Maurice."	<——— (V) (for Maurice)
7. Maurice: "Hello. I was phoning to see if Karen was in."	<——— (V) (for Pam)
8. Pam: "No. I'm afraid she's out at the moment."	<——— (V) (for Maurice)
etc.	

(I = Identification
D = Differentiation
V = Verification)

Let us first concentrate on the decision processes which Maurice has to perform during the pause between utterances 4 and 5. He hears a female voice. The immediate task for him is to check that it is not the switchboard girl again, saying that she cannot get a reply on that extension. This is a short-term memory differentiation task (unless the switchboard girl's voice is already known to him). Having ascertained that the voice is not that of the switchboard girl, the next candidate, and an extremely probable one, is Karen; firstly, it is normally she who answers on that extension, and secondly, it is her that Maurice wants to talk to. It is therefore a possible source of great economy of mental effort if he first of all verifies whether it is Karen or not. The task is thus one of verification; there is only one reference voice under immediate consideration (and Maurice has strong expectations that the stimulus voice will correspond to that reference) and it is stored in his long-term memory. However, having verified that it is not in fact Karen, Maurice's next task is to decide who it is, and there are three very probable candidates - Sara, Pam and Marie. The task thus becomes one of identification since there are more than one reference voice under consideration, and their patterns are all stored in his long-term memory (i.e. he already knows them all). He decides that the stimulus voice is Pam's, although he is not certain of this, as is shown by the questioning rise in intonation of his next utterance (5). The correctness of his decision is confirmed by Pam's utterance (6). Of course, the stimulus might not in fact have corresponded to one of these three references. However, they are by far the most probable in the particular situation.

He thus arrives at the correct recognition of the stimulus speaker of utterance 4. This decision may only take a second to be reached, but it is probable that it is the result of a sequentially ordered combination of more than one task, as described above - initially differentiation, then verification, and finally identification. The second process comes about as a result of a negative response to the first, and similarly the third is brought about by a failure to verify that the stimulus is Karen.

In the above dialogue, Maurice's utterance 5 will constitute the stimulus for an identification process for Pam (assuming that she does not have strong expectations that the caller will be Maurice). However, she is obviously more confident about her decision than Maurice is, since she responds positively with his name (6).

The differentiation task for Maurice (utterance 3) is symbolised in brackets because it is probably a less conscious process than those described above. Its function is confirmatory (see section 1.5), reassuring Maurice that it is still the switchboard girl to whom he is talking. It may be that he has no good reason to doubt this; however, this confirmatory process must be assumed to take place, for otherwise, one would have no means of explaining why Maurice would react immediately to a different voice on the line. In other words, it is a continual process occurring at each utterance, to confirm the very strong expectation that one is still talking with the same person. It is a differentiation task in the case of the switchboard girl above, since hers is not a familiar voice for Maurice, whereas in utterances 6, 7 and 8 it is verification since Maurice and Pam are acquaintances.

There are certain limitations to the above categorically stated analysis. Two examples relating to the above dialogue will be sufficient to illustrate these restrictions.

(i) The function, as far as speaker recognition is concerned, of Pam's utterances 6 and 8 is confirmatory, and has just been categorised as verification, since Pam's reference pattern is stored in Maurice's long-term memory. However, this categorisation assumes that the stored reference pattern of her voice plays a more important part in the speaker recognition process than any information added to that pattern, extracted from her utterances 6 and 8. That is, whether these confirmatory decisions are categorised as verifications or instead as differentiations depends upon whether the stored pattern is considered to be more important

than the short-term information, or not. This taxonomic problem is more acute in certain circumstances. If the speaker has a temporary intrinsic condition such as a head-cold, the reference pattern for this confirmatory task will need to include this information as a fundamental feature. If the speaker is someone the listener has not spoken with for a long time, the reference pattern will be less distinct than that of someone he meets every day, and the short-term information from contemporary utterances will play an important part in the decision. Even in normal situations, it is intuitive to suppose that information derived from a speaker's most recent utterance does not play a totally insignificant part in the verification/differentiation of that speaker.

(ii) In experimental formats, the reference voices in any one task have always been stored either all in short-term memory or all in long-term. In the real world, however, it is possible for a situation to arise where a mixture of these two are involved. For example, in the above dialogue situation, Maurice may subsequently hear a voice on the line which is not Pam's. He decides this by the above verification/differentiation task. There are then three (or possibly four) very probable reference candidates: Sara and Marie (and perhaps Karen), all of whose reference patterns are stored in Maurice's long-term memory, and the switchboard girl again, whose pattern is in his short-term memory (assuming a not too long time-delay since he last heard her voice). Thus patterns in long-term and short-term memory are involved. The task must be categorised as identification since more than one reference pattern is taken into account, although this classification overlooks the fact that both memory systems are used.

The above two examples serve to illustrate the difficulty in applying the clear-cut distinctions of the experimental situation to real world circumstances.

4.6 A LOGICAL MODEL OF THE SPEAKER RECOGNITION PROCESS

In the previous sections of this chapter, the criterial background categorisations were discussed of the speaker recognition process which takes place in the listener in the experimental and real world situations. In the following sections the formal logical model of that process is presented and discussed.

A model is a representation; a model of a psychological process is a representation which attempts to explain the operation of that process; a logical model is a representation which deals in a reasonably abstract way with the components of a process and the logical connections which exist between those components - it does not necessarily imply that there is a direct relationship between the logical components of the process and the physical, neurological structure which has the ability to perform that process. Since the following model carries no implications about neurological implementation but is rather a model couched in the terminological framework of artificial intelligence, and since knowledge about neurological implementation is still at a rudimentary level, it is a problem that the structure of the model is not readily available for experimental justification. However, the model is not inconsistent with any of the findings or proposals of the categorisations contained in the preceding sections of this chapter.

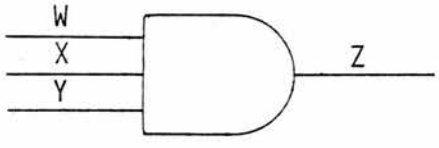
The component parts of the model are referred to as black boxes. A black box is the name given to a device whose output is known for a given input, but whose internal operation is not known. Any complex process can be broken down analytically into component black box processes, and the greater the number of component black boxes the process is analysed into, the smaller the amount of internal machinery assumed to be contained within each black box. Thus, an idealistic way of viewing scientific progress in the sense of the advancement of knowledge is as an analytical decomposition of a complex of few, large black boxes into one of more, smaller, component black boxes. The ultimate stage reached in this process of

discovery would then be a connected network of primitive operations ("primitive" in the sense that they cannot be decomposed further into component operations).

No model of the speaker recognition process has been proposed anywhere in the literature along the same lines as the one presented here. Since this is a first model, it has been kept as a simplified representation of the overall structure of the process. No-one would deny, however, that the speaker recognition process is an extremely complex one, and that a more complex representation would therefore be possible and desirable as a subject of future research.

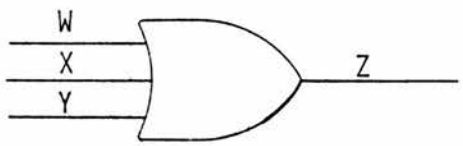
The relations which hold between the various black box processes of the model are symbolised by Boolean logic gates. Two different kinds of these are used - the conjunctive and the disjunctive gates. The meanings of these gates are explained in the following diagrams.

Conjunction



The conjunctive gate is used to express the following relationship: "If, and only if, all the input conditions (W, X and Y) obtain, then Z also obtains". W, X and Y may therefore represent prerequisite input conditions which must be fulfilled before the output condition Z, for example an operation which is dependent upon all three of them, also obtains. This gate is generally referred to as the conjunctive gate, or AND gate.

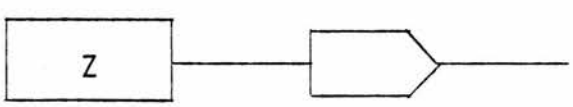
Disjunction



This gate implies the relationship "If one, and only one, of the input conditions obtains, then the output condition also obtains". This is generally referred to as the exclusive disjunctive gate, or EXCLUSIVE OR gate.

The logical nature of these gates is not of paramount importance in the following representations, because the situations are rare in which logically contrary conditions obtain. The function of the gates is rather to determine the sequence in which processes take place. It should be remembered that this model deals with the logical sequence of processes not with the length of time which they require to be performed in the listener.

There is another symbol used in the diagram which needs to be explained. It is represented in the following diagram.



This abbreviatory symbol is used to mean that the process has not come to an end, but that the output of process Z becomes the input for some later process, although that fact is irrelevant for the purposes of the explanation contained within the particular diagram.

The model has a two-part hierarchical structure. The hyponymous processes deal with the tasks of identification, differentiation and verification, as defined previously in this chapter. The superordinate process concerns

(i) the conversion of decisions obtained in the hyponymous processes into responses to be given in experimental tasks, or corresponding psychological reactions occurring in real world situations, and

(ii) the accommodation of factors such as time and fatigue into the determination of the depth of analysis to which the whole speaker recognition process is taken. It is easier, for the description of the model, to explain the lower-level processes before the higher-level.

The relevance of the categorisations discussed in section 4.5.1 can now be seen. The range of probable reference voices may be narrowed down by considering situational expectations and reference-stimulus parametric similarities, producing economies of psychological processing. Both of these suggestions are aimed at accounting for the fact that the speaker recognition process takes place relatively quickly even when the set of possible reference voices is very large, and allows us to avoid the undesirable solution that each reference voice is called up in sequence for comparison. However, both of the above theories have been described in abstract terms as ways in which this streamlining might be effected. It is probable that it is brought about rather by some combination of the two, and that they have a complex interactive effect. I shall not discuss further the first of these suggestions (that the situation narrows down the range of probable references) beyond the simple analysis given above, mainly because it is difficult to specify what the relevant features of any given situation are. Instead I shall concentrate on the more phonetically relevant, second suggestion. This second method is important for the following model because it involves economies being made by using features of the stimulus sample to guide the search for suitable reference voices to be called up for comparison.

An extreme version of this streamlining is produced if it can be shown that the stimulus utterance contains an abnormal value in relation to a particular parameter, and this parameter is considered first. However, if the population of possible reference voices is small, as for example in many experimental closed identification tasks, this guided search procedure may not lead to a streamlining at all in terms of economy of psychological processing. In other words, for small reference populations, recall of voices for comparison in sequence does not seem overly burdensome. In the following model, therefore, identification is handled as two separate processes:

(i) small-population identification, which does not involve a search guided by normal or abnormal features of the stimulus sample, and

(ii) large-population identification, which does.

The former has the simpler structure, and for this reason alone, will be discussed first.

The small-population identification process is represented by Figure 4.12, which will be found at the end of this section. The diagrams for the four parts of the model are assembled together there to facilitate cross-reference, so that the similarities between processes and the relation between the hyponymous processes and the superordinate process may be seen. In this and all subsequent diagrams, the box labelled STIMULUS is an abbreviation for the network displayed in Figure 4.11.

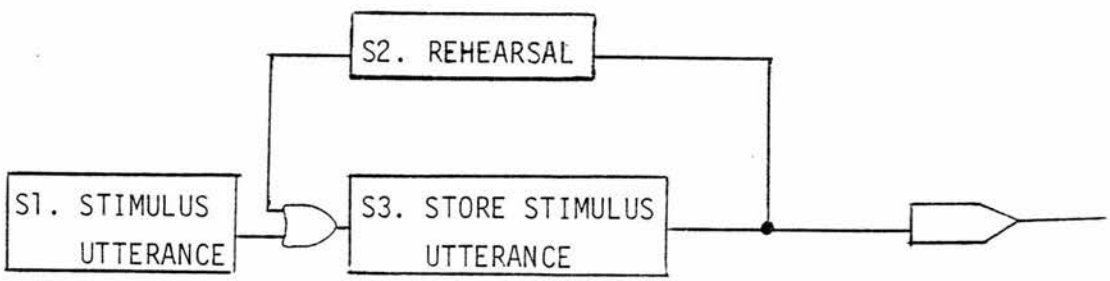


Figure 4.11 The stimulus storing process.

The representation of the stimulus sample which is stored in a very short-term form of memory is derived from the stimulus utterance. This representation may be enriched in two ways: firstly, by further exposure to the stimulus voice, and secondly, by rehearsal - the process, discussed in section 4.3, of "going over" information stored in memory to prevent this information from being lost (forgotten). Owing to the limited capacity of short-term memory, the representation stored will not be one of, in some sense, the whole speech signal, but rather an abstraction of criterial features. This process is thus a form of classification; whether the categories into which the signal is classified correspond to those parameters which are used in the speaker recognition process is a psychological question beyond the scope of the present discussion. This question has been discussed for the processing of speech sounds by many psychologically oriented experimenters (Crowder & Morton, 1969; Darwin & Baddeley, 1974; Fujisaki & Kawashima, 1968).

The small-population identification process takes place in the following manner. From the representation of the stimulus utterance stored in short-term memory is extracted the first parameter in relation to which the utterance is to be analysed. The phrase NEXT PARAMETER in A1 is an abbreviation to be interpreted as meaning the first parameter if the process is just starting. The parameter, after being mechanically extracted from the stored representation, is analysed so as to produce a value for the utterance in relation to that parameter. This value is then related to the corresponding values for the small set of reference voice patterns in sequence. The process RELATE is to be interpreted as a simple pairing of the two values and the assignment of a measure of their similarity; the notion of a threshold is not involved. The output of A2 is thus a parameter similarity measure, which is temporarily stored by process A3.

Process A4 refers to the discussion in Chapter 3 of the different ways in which parameters can be said to have strength as speaker-characterising features. Various such criteria were laid out, and it was suggested that those parameters which ranked highly in terms of many of these criteria should be called first-order parameters (section 3.5); this means that it is highly probable that, by virtue of their overall characterising strength, these parameters are the first in relation to which stimulus utterances are analysed for speaker recognition. Those parameters which do not rank so highly in terms of these strength criteria may be called second-order parameters, and so on. Process A4 therefore checks whether the stimulus utterance has been analysed in relation to all first-order parameters. There are two possible outputs: either

(i) a negative signal, meaning that not all first-order parameters have been considered, in which case recursion takes place and the utterance is re-analysed, this time in terms of the next first-order parameter, or

(ii) a positive signal, indicating that all first-order parameters have been employed.

In this latter case, all the stored parameter similarity measures pass to process A5. The process COMPARE differs from the process RELATE (in the present terminology) in that it involves the notion of a threshold. Thus the sum of the similarity measures for any one reference voice pattern is compared with this threshold, and the output of this process depends on the overall acceptability for all parameters of that reference voice pattern. The process COMPARE, like the process CHECK, may therefore have two possible outputs: either

(i) a negative signal, implying that none of the reference voices exceeds this threshold (the progress of this negative signal is followed in the superordinate process (Figure 4.15)), or

(ii) a positive signal, indicating that one or more reference patterns exceed the threshold.

In the case of more than one reference pattern exceeding the threshold, process A6 selects that one which is highest-ranking (exceeds the threshold by the greatest amount). This positive signal also reappears in Figure 4.15; numbers are given to these signals merely to aid reference between the two diagrams.

The large-population identification process was defined above as differing from the small-population process in two respects:

(i) features of the stimulus utterances are used as guides for the search for probable reference candidates, and

(ii) as an extreme form of this streamlining, the stimulus utterance is initially scanned for abnormal parameter values to guide this search.

Both of these features are included in the model of the large-population identification process (Figure 4.13).

Since the stimulus utterance is first scanned for abnormal parameter values, some information about what constitutes the norm for parameter values must be available. These norms are determined by cultural and physiological factors, and are called up by process B1. Process B2 is then able to check whether the stimulus utterance in fact contains any abnormal values. If there are no such abnormal values, a negative signal is produced (this signal is taken up later in this discussion). If there <u>is</u> an abnormal value, process B3 uses this information in order to call up only those reference voice patterns which are marked as having similar abnormal parameter values. These references are then compared with the stimulus in relation to the particular parameter. If the correspondence does not fall within a certain threshold, the negative signal (discussed below) is taken up. Where certain reference voices do produce a sufficiently close match, process B5 selects the highest-ranking in the case of more than one. The progress of this positive signal is followed in the superordinate process diagram (Figure 4.15).

A negative signal from either process B2 (indicating that the stimulus utterance contains no abnormal parameter values) or process B4 (meaning that no reference pattern approximates closely enough to the abnormal values in the stimulus) causes the process in the bottom half of Figure 4.13 to be initiated. B6 extracts the next parameter (excluding from the reckoning any parameter already analysed because of its abnormal value in the stimulus) and analyses it, to produce a stimulus parameter value. This value is then used to guide the search for reference voice patterns with similar values for that parameter, in exactly the same way as an abnormal value guided this search above (process B3). The rest of this "non-abnormal" branch of the large-population identification process takes the same form as the corresponding part of the small-population identification process (A2-A6). The final output of this branch of the process may therefore be either a positive or a negative signal, both of which figure again in the superordinate process (Figure 4.15).

The difference between the differentiation task and the verification task is, as was explained in section 4.4, that for differentiation the reference voice pattern under consideration is stored in short-term memory, while for verification it must be recalled from long-term. This is an insignificant feature of those tasks as regards the decision processes being modelled in Figure 4.14, and for this reason, the two tasks have been conflated into the one diagram; in all other respects the processing is the same for both tasks.

The upper half of the differentiation and verification process diagram is similar to the upper half of the large-population identification process diagram just described, in that the process involves a search guided by abnormal parameter values. However, there are two major differences between the two processes.

Firstly, and the more importantly, the search in differentiation and verification tasks is guided by abnormal values found in the reference voice pattern - not in the stimulus utterance, as in identification. It follows that the search for corresponding abnormal values in differentiation and verification is made in the stimulus utterance, and not amongst the reference voice patterns, as in identification. In this way, economies are gained in that the stimulus utterance may need to be analysed in relation to only a minimal number of parameters.

Secondly, it is obvious that a box corresponding to process B5 (SELECT HIGHEST-RANKING REFERENCE VOICE PATTERN) is not required since there is only one reference voice pattern under consideration in differentiation and verification tasks. Again, there may be two outputs, one positive, one negative, from the COMPARE process at the end of the "abnormal" branch of the differentiation and verification process.

If the reference voice pattern is found to contain no abnormal parameter value, the lower half of Figure 4.14 comes into effect. It can be seen that this half corresponds exactly to the small-population identification process, except for two differences, both of which are due to the increased number of reference voice patterns in identification. The first is that there is only one reference voice pattern input to the process and therefore the process implied in C6 and C7 is less complex than in A2 and A3. Secondly, there is again no need for a box which selects the highest-ranking reference voice pattern, corresponding to A6 in the identification task. There are similarly two possible outputs from this branch, one positive and one negative.

The positive and negative outputs from the small- and large-population identification, and differentiation and verification tasks all reappear in the model of the superordinate process (Figure 4.15).

This process deals with the conversion of these outputs into experimental responses or real world psychological decisions. Throughout the discussion these will be referred to together as responses, although in real world situations the response is usually not made aloud but may instead be manifested merely by some appropriate psychological reaction.

The progress of these outputs in this conversion may be handled in five groups:

(i) outputs 1, 3 and 4. In all circumstances, a positive identification on the basis of either normal or abnormal parameter values leads to the appropriate reference voice pattern being given as the response (R1) in an experimental closed or open identification task.

(ii) outputs 2 and 5. A failure to recognise any match in an identification task may lead to recursion. Whether this takes place or not depends upon factors external to the reference-stimulus comparison. Such factors will include time and pressure in experimental tasks, and motivation in real world situations (D1). By considering these factors, the desirability of increasing the depth of analysis is evaluated (D2). If it is considered desirable, recursion takes place by changing the analytical parameters from those of the first order to those of the second (D4), and so on until all orders of parameters have been exhausted (this is monitored by processes D5 and D3). If this further analysis is not desirable, for whatever reason, the response given will depend upon the nature of the experimental task. In a closed identification task, the "don't know" alternative will be given (R4), whereas in an open task the response "stimulus does not correspond to any of the references" will be chosen (R5).

(iii) outputs 6 and 8. A match found in a differentiation or verification task leads to a "same speaker" response (R2) (= "stimulus is speaker X"; see Figure 4.6).

(iv) output 7. Failure in a differentiation and verification task to find a value in the stimulus utterance which approximates closely enough to an abnormal parameter value found in the reference voice pattern leads directly to a "different speakers" response (R3).

(v) output 9. If the listener fails to find a match between a reference voice which is not characterised by any abnormal parameter value and the stimulus, a process similar to that for group (ii) above is initiated; namely, it is decided on the basis of external factors whether the analysis should be extended to parameters other than those of the first order. If this deeper analysis is desirable, recursion by switching to second-order parameters occurs, as for outputs 2 and 5. If it is not desirable, a "different speakers" response (R3) will be produced.

4.6.1 Assumptions Underlying the Model

The model has been presented formally since any increase in formality produces an increase in the explicitness of the units of the model and of the relationships which hold between them. However, this explicitness has been achieved by overlooking certain factors and by making certain assumptions about the process being modelled. These assumptions and their justification will now be examined, before a more general discussion of features of the model.

The first such assumption is that the differentiation process takes place on a reference-stimulus basis (as with the identification and verification processes). However, it was shown in section 4.3.1 that the simultaneous presentation task may be performed either on a reference-stimulus basis or by a separability principle ("if the two voices can be separated, then they are different").

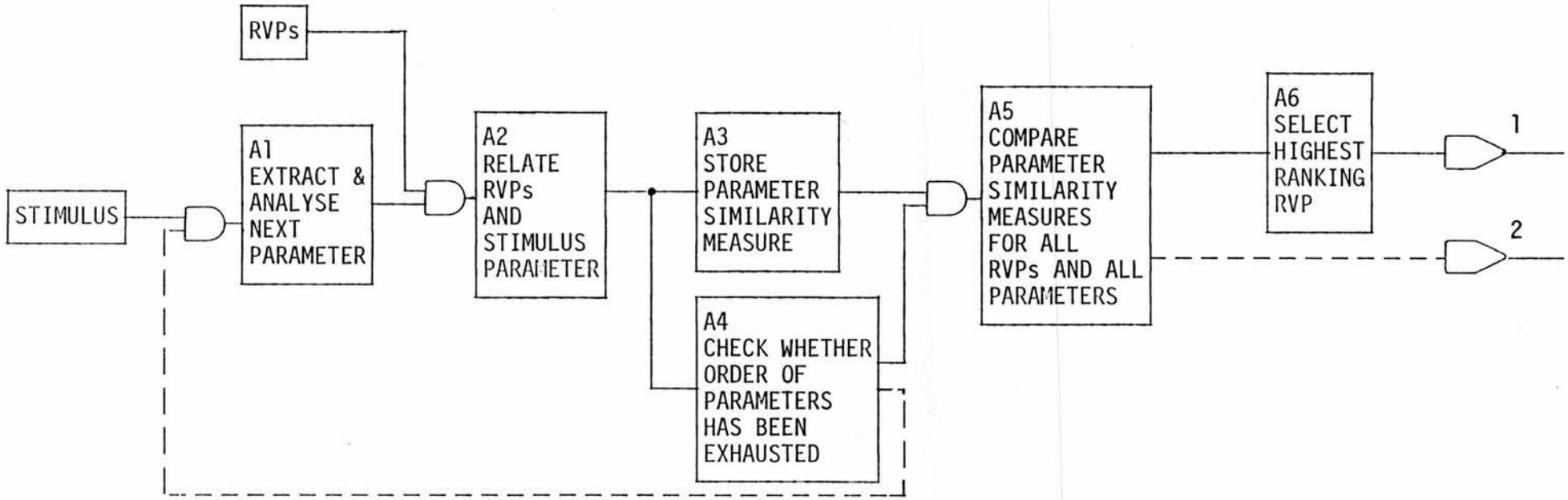


FIGURE 4.12 SMALL-POPULATION IDENTIFICATION

RVP = reference voice pattern

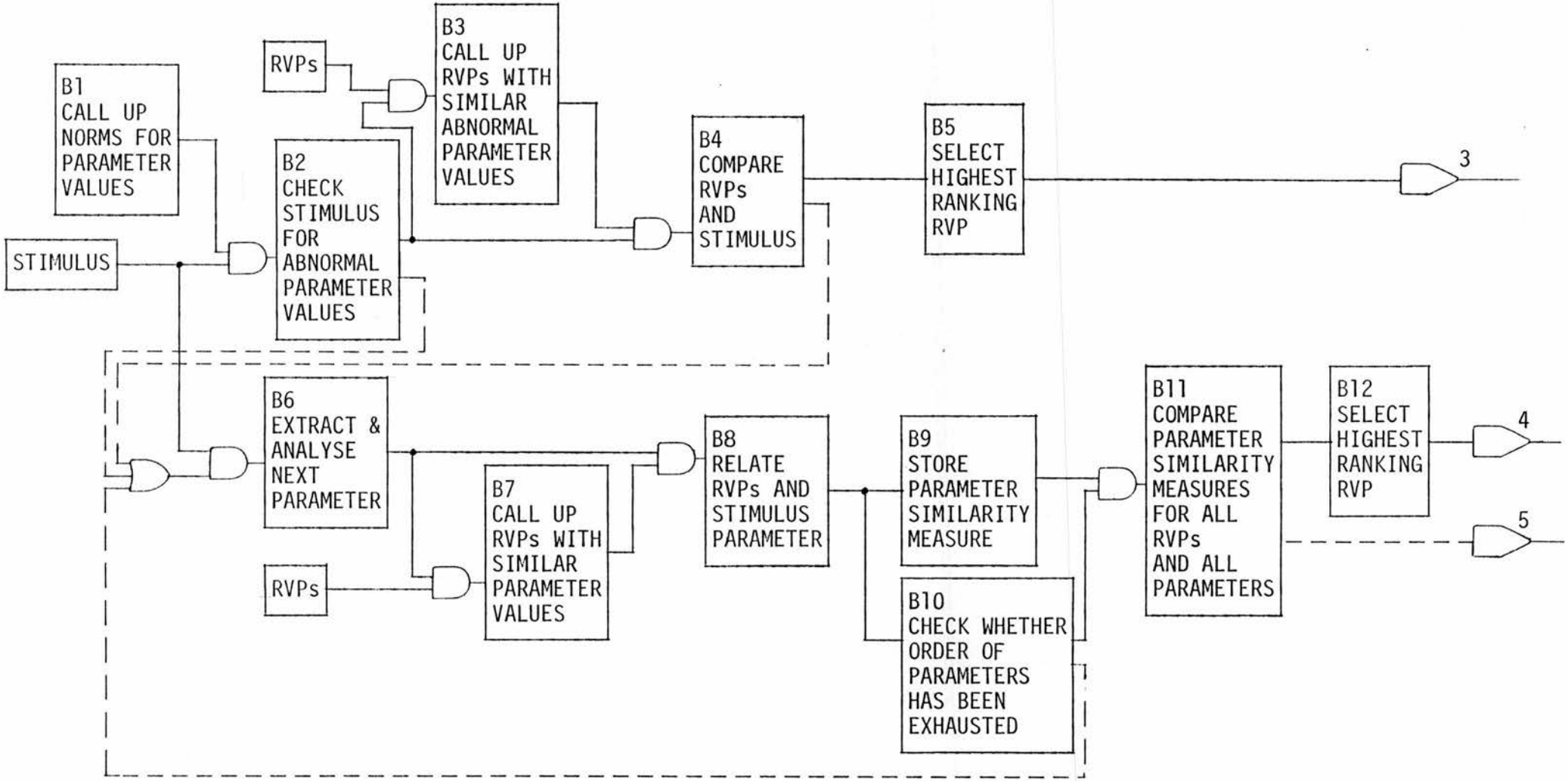


FIGURE 4.13 LARGE-POPULATION IDENTIFICATION RVP = reference voice pattern

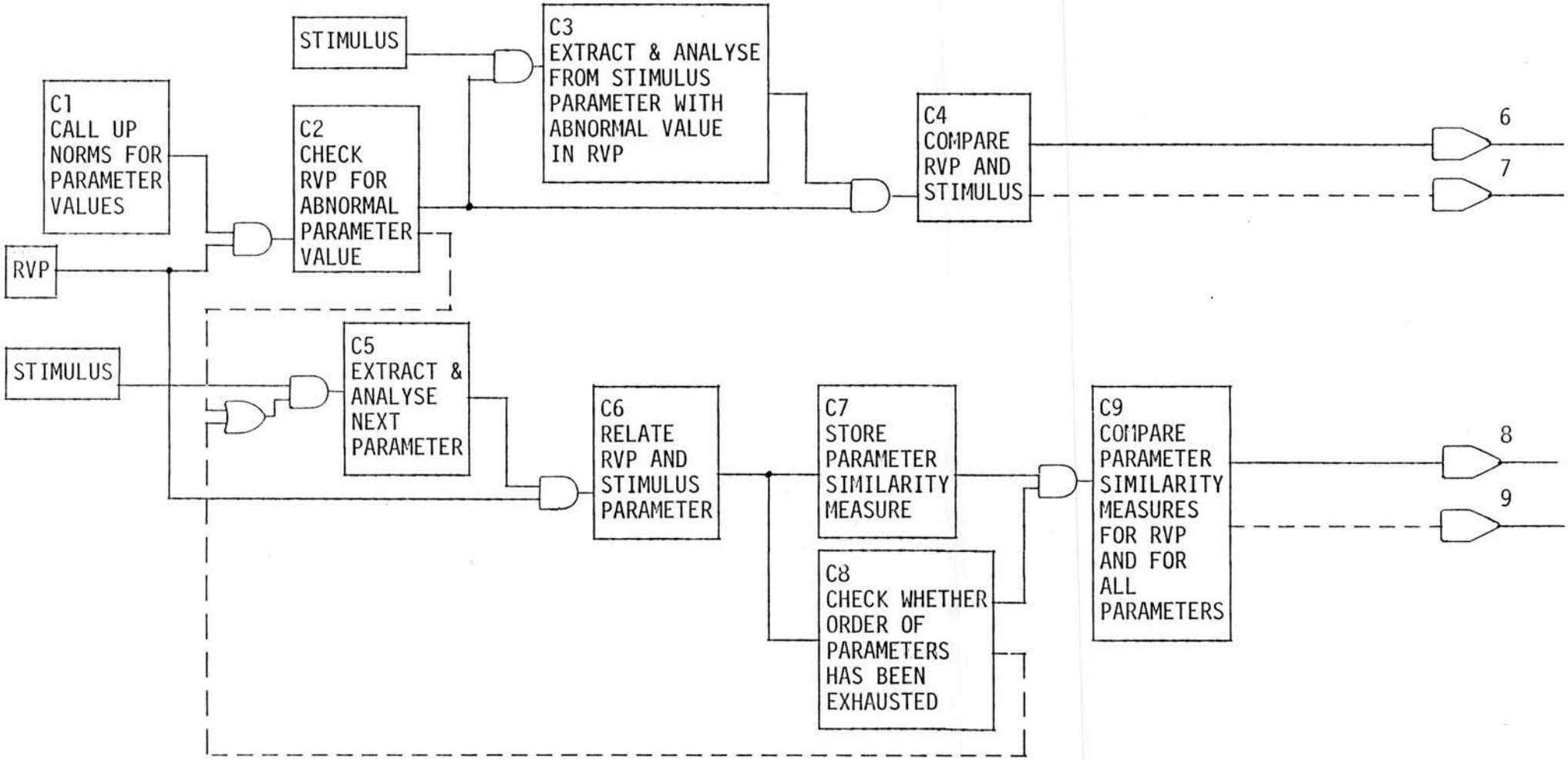


FIGURE 4.14 DIFFERENTIATION AND VERIFICATION

RVP = reference voice pattern

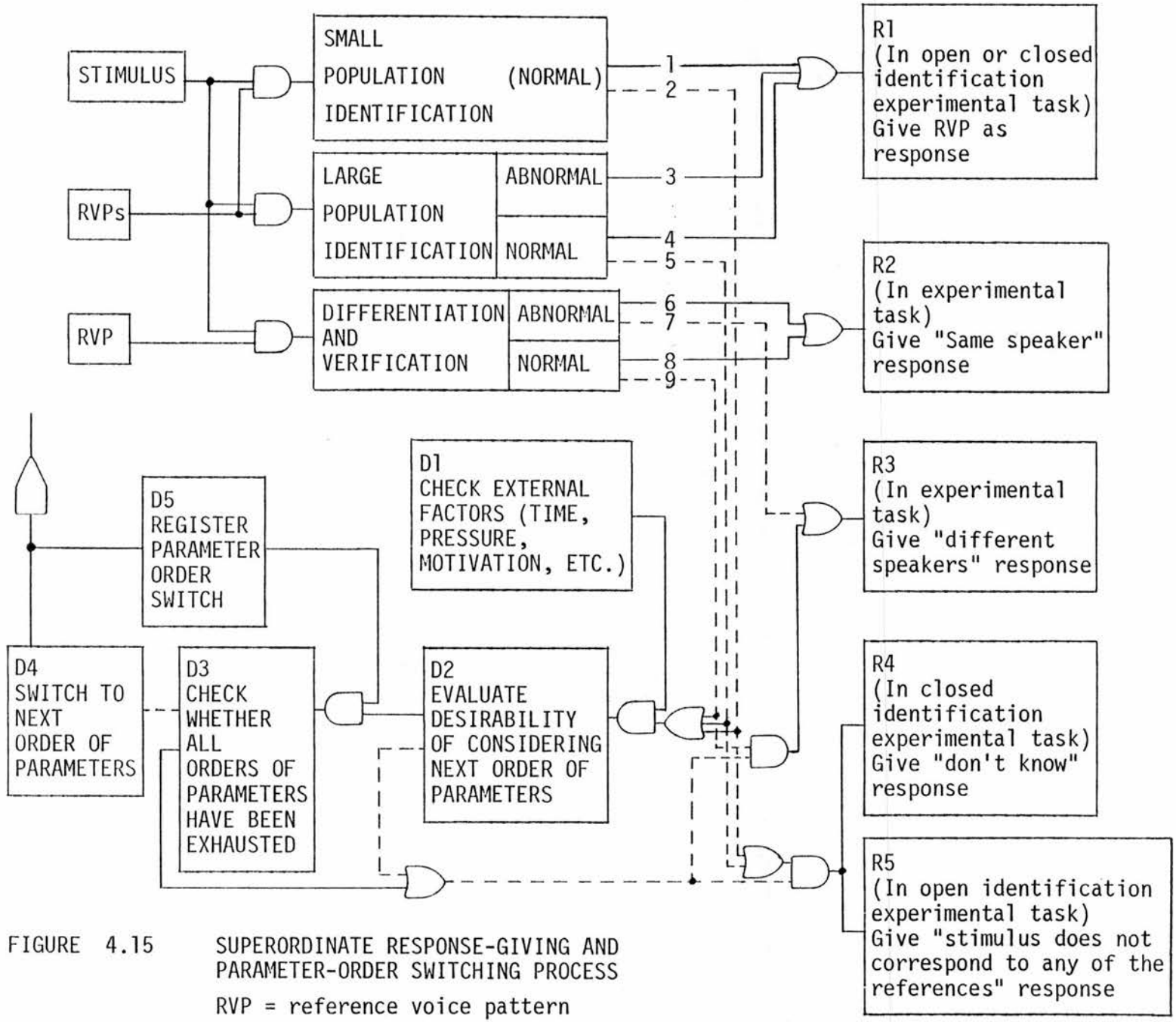


FIGURE 4.15 SUPERORDINATE RESPONSE-GIVING AND PARAMETER-ORDER SWITCHING PROCESS

RVP = reference voice pattern

If this latter strategy is adopted, then the process is not handled by the model, which deals in all cases with kinds of comparison. As was discussed in section 4.3.1, comparison of the two voice patterns may play an insignificant part, if any, in the separability approach.

The next two assumptions refer to the justification allowed for the acceptance of a reference speaker as the stimulus. It is assumed that recognition on the basis of an abnormal parameter value is sufficient for accepting a reference speaker; there is no need for further processing to take place in the form of confirmation from parameters with non-abnormal values. Making this assumption allows us to explain why certain speakers are instantly recognisable, no matter how many other speakers they may be grouped with as possible references, while others are not.

Recognition by first-order parameters is assumed to be sufficient for the acceptance of a reference speaker; second- and lower-order parameters need not be considered in this case. The justification for this view lies in the definition of first-order parameters; they belong to the first order because they are the strongest, the most speaker-characterising (see sections 3.3 and 3.5), and therefore recognition by these parameters is highly valued. On the other hand, failure to recognise by first-order parameters is not enough for the rejection of a reference speaker. Further analysis is desirable for confirmation of this negative result (if this extension is allowed by the external factors of D1). Recognition is thus a positive decision, whereas failure to recognise is considered as the absence of this positive decision. If, in the common real world identification task, the listener should encounter a stimulus parameter value which is abnormal and is not matched at all closely by a corresponding reference value, he is unlikely to dismiss that reference from consideration straightaway since it is always possible, if not exactly probable, that the abnormal value is the result of some nonce, paralinguistic effect.

It is assumed in the model that the only way in which the depth of analysis may be increased is by switching to second- and lower-order parameters. However, it is clear that by the time first-order parameters have been exhausted, the listener may well have had further exposure to the stimulus. This is more probable in the real world than in experimental tasks, where sample duration is usually controlled. The listener may then work on this further exposure (by first-order parameters) rather than analyse the sample which he already has stored (by second- and lower-order parameters). From this point of view, it seems that lower-order parameters may be of limited practical importance.

The processes of identification, differentiation and verification, and their major decision-making component parts, may all be classified as forms of comparison; that is, one stimulus voice is compared with one or more references. In the model, it is assumed that one of the references (in an identification task) will correspond more closely to the stimulus than all the others. A contingency which is not allowed for explicitly in the model is the situation where two or more references contain values which correspond equally well. However, this situation is unlikely to be a significantly common one. It might occur in relation to one parameter but it is improbable that it would occur for all; if the two voices corresponded to the stimulus equally well in relation to all parameters, then the two would be identical for speaker recognition purposes. This seems a highly improbable situation even for such similar instances as identical twins, close family relations, impersonators, etc.

The final assumption to be considered here is that the "don't know" response is available only for closed identification tasks. As was explained in section 4.4.1, the "don't know" alternative is necessary in closed identification to avoid largely unprincipled guesses on the part of the listener.

However, for all the other tasks, the absence of a "don't know" alternative does not place an undue burden on the listener, in that other alternatives exist which carry much the same implication of uncertainty. Incorporation into the model of a "don't know" response for these other tasks would best be handled by introducing a second decision threshold. Thus, in a differentiation task for example, a stimulus voice whose approximation to the reference voice was less than this threshold would trigger a "different speakers" response; whereas an approximation closer than this threshold but not as close as that required for an acceptance decision, would produce a "don't know" response. In other words, a "don't know" response would result from a situation where the two voices corresponded neither closely enough to produce a "same speaker" response, nor so badly that a "different speakers" response resulted. The value of this threshold would be an idiosyncratic feature for each listener, dependent upon the strictness or laxness of his acceptance criteria. It can be seen now that incorporation of a "don't know" alternative causes the decision process to become much more complex. From a two-way process (acceptance/failure to accept) it turns into a three-way (acceptance/rejection/failure to accept or reject). Most importantly, non-acceptance now has to be considered as a positive decision, in contradiction to the view previously expressed. If this new standpoint is accepted (and there seems no alternative explanation in the circumstance), then it follows that the separability approach cannot be adopted for simultaneous presentation tasks where there is a "don't know" response available; or, at least, if the response is available, it will never be given. "Don't know" responses ought to provide us with as much information about the human speaker recognition process as the positive identification, differentiation or verification responses. However, the fact that it is difficult to interpret "don't know" responses in such a way is perhaps another reason why forced-choice formats are favoured by experimenters.

4.6.2 Discussion of the Model

In this section will be discussed some general points concerning the model, and some issues which are associated with it by implication.

In this thesis reference voice patterns have been considered to be composed of specification values for as many parameters as are required for successful speaker recognition. It is these values which are called up from memory for comparison with the corresponding contemporarily calculated values in the stimulus. The reference values are best considered as being composed of two parts - a value proper and a threshold of variability. The value proper will be some absolute measure while the threshold will contain statistically defined information about the stability of that value, i.e. about the variability of the reference person's speech in relation to that parameter. This might be clarified by a hypothetical, simplified example: let us suppose that the pitch mean value for a particular reference speaker took the form (150 ± 10 Hz) (this should more strictly be stated in terms of the psychological unit, the mel). The value proper would then be taken as the absolute value 150 Hz, and the threshold of variability would contain the information about the stability of that value (± 10 Hz). This example is a simplified account, since the threshold information is likely to be of a much more complex probabilistic statistical nature than a mere statement of range extremes. However, the example is sufficient to illustrate the theory. Reference values (values proper and thresholds) will have been calculated on the basis of the listener's previous exposure to the reference speaker's voice. Because the listener's exposure to the stimulus voice in experimental tasks is often limited to a short sample of a couple of seconds' duration, the corresponding stimulus values are likely to contain far less statistical variability information, and may be composed of only an absolute measure. The threshold of variability contained in reference values reflects not

only the variability of the speaker but also that inherent in the parameter. The speaker-oriented factors include all idiosyncratic variation in extrinsically determined factors. Parameter-oriented factors relate to the variability inherent in the nature of a parameter. Such factors may relate to intrinsically determined production characteristics, such as the inertia inherent in the speaker's vocal apparatus, which govern the control possible for the speaker over the manipulation of a particular parameter. They may also reflect cultural restrictions and the effects of perceptual categorisation.

Since reference indices are taken to be specified in a complex statistical way, the COMPARE processes of the model must be similarly complex. There are two main reasons for this complexity:

(i) The process operates in a statistical way, performing a probabilistic decision on the basis of probabilistically defined inputs. However, further discussion of this statistical complexity is beyond the scope of this thesis.

(ii) If, as its input, it receives two high-ranking reference patterns, it must be capable of deciding, for example, whether one which ranks highly in terms of formant structure is stronger than the other, whose formant parameter correspondence is less but whose pitch correspondence is greater. In other words, it must have access to information regarding the relative importance of parameters for speaker recognition (see section 3.3); this information is exactly what is investigated in the experiments reported in Chapter 6.

Reference has been made above to the specification of a reference voice pattern. What is meant by this is the amount of parametric detail which the pattern contains. It is reasonable to assume that the amount of this detail is related to the amount of listener

exposure to the reference voice. In other words, the specification of a reference voice pattern in a verification task for example, which may be a reference familiar to the listener through repeated everyday exposure by social or business contact, is likely to be far greater than the specification of a reference voice pattern in a differentiation task, where the reference speaker is a complete stranger to the listener, and where the listener's exposure to the reference voice may constitute a sample of only a couple of seconds' duration. This is a case of extreme difference. However, two speakers, one of whom is the listener's brother (to whom he talks every day) and the other of whom is a distant relative to whom he has not talked for many years, although both familiar in the technical sense introduced in section 4.3, will have a great difference in the specifications of their reference voice patterns for that listener. This difference is not handled in the model presented, although it may have an effect on the performance of the task. This is not a great criticism however, if one considers the probabilities of reference candidates discussed in section 4.5.1. Under normal conditions, the real world probability of a stimulus speaker being the listener's brother is far greater than of him being his distant relative. Indeed, one might hypothesise a relationship between the specification of a reference voice pattern and the probability of a stimulus corresponding to that reference, under unexceptional circumstances.

A very short-term form of memory is relevant when one considers the speed and manner in which neural processing takes place for speaker recognition as with many other processes. Knowledge of this field is still at an elementary level and so the following discussion takes the form of a short consideration of the alternative possibilities. There are two major alternatives.

(i) Parallel processing implies that different analyses can be performed at the same time. This might be to analyse a stimulus in terms of more than one parameter at a time, or to recall and compare more than one reference voice pattern with the stimulus at the same time. This is an attractive possibility for the model presented because one could then posit that, except in the case of parameters with abnormal values, all first-order parameters are considered at the same time, then all second-order (where applicable), and so on.

(ii) Serial processing means that analyses can only be performed one at a time, although the actual speed of processing may be so rapid that the process might be considered "quasi-simultaneous". This is a less attractive possibility than parallel processing for speaker recognition, because one then has to specify the sequence in which parameters or reference voice patterns are ordered. There are two approaches to this:

(a) an absolute ordering. For the selection of parameters for analysis, the sequence would presumably follow the lines of that given by the strength evaluation technique of Chapter 3.

(b) an ordering dependent upon features of the stimulus. This has already been assumed for the recall of reference voice patterns by the incorporation of the guided search principle into the model (and of the abnormality principle as an extreme form of this).

Whether processing takes place in series or parallel is not fully understood at present. Therefore whether the selection of parameters and the recall of reference voice patterns have been correctly taken to be absolutely or dependently ordered respectively in the model remains to be seen.

For identification tasks, where more than one reference voice is involved, these patterns are assumed to be called up in a dependent sequence. This is especially true for large-population identification, in which it is unreasonable to suppose that all possible references are recalled at the same time (even if this were neurologically possible in all cases. So far, nothing has been proposed along these lines for small-population identification, since dependent sequencing of recall does not lead to great economies of processing over independent sequencing. This is not to say categorically that small-population identification employs independent rather than dependent ordering. If one assumes independent ordering, one still has to face the question "In what order are reference voice patterns called up?", and since the population of reference speakers changes from one occasion to another, it is impossible to answer this question in absolute terms. In addition, since large-population identification is held to entail dependent ordering, it would be surprising if a totally different system were employed in the case of small-population identification. What is being implied here is that the difference in effect is minimal and that there is less justification for assuming a dependent ordering system.

An implicit feature of the model presented is that a greater degree of certainty is required in identification tasks before a response rejecting a reference speaker is given, than it is in the differentiation and verification process. One way in which this difference is manifested is in the progress of a negative signal in the case of abnormal parameter values. For differentiation and verification tasks, this signal is taken as justification for the giving of a "different speakers" response ("stimulus is not speaker X"); if a reference speaker is characterised by an abnormal parameter value, then there is a strong statistical significance to a corresponding value found in the stimulus. In large-population identification tasks, however, failure to find an abnormal value in a reference pattern corresponding to one found in the stimulus is taken to be less

significant owing to the much larger number of references under consideration. In this case, recursion to a consideration of non-abnormal, first-order parameter values occurs. In small-population identification, it has been argued that a consideration of abnormal parameter values does not play a significant part, if any, and therefore this argument does not apply.

From studies of various forms of speech error (such as tongue-slips (Fromkin, 1973) and those induced by delayed auditory feedback (Fairbanks, 1955; Fairbanks & Guttman, 1958)), it has been shown that feedback plays an important role in the speech production and perception processes. In relation to speaker recognition, feedback is probably present at most stages in the process. However, there are two places where the importance of feedback is great:

(i) In forced-choice closed identification tasks, listeners are told how many, and which, reference voice patterns are to be taken into consideration, and that the stimulus voice will correspond to one of these references. However, after analysis, it may be found that no reference voice pattern corresponds closely to the stimulus voice pattern, in which case recursion will occur, probably taking the form of a relaxation of threshold levels. This is probably what takes place when the stimulus samples have been modified acoustically either deliberately or by transmission characteristics, so that the stimuli will not correspond exactly to the reference voice patterns.

(ii) Most experimenters use practice sessions to help the listener to get used to the task, and to allow his performance to reach a reasonably stable and optimal level before the test proper begins. In these practice sessions, listeners are informed of the accuracy of their responses; in identification tasks the identity of the stimulus speaker is revealed, while in differentiation and verification tasks the correct "same"/"different" response is given. Both these kinds of information constitute feedback. Its importance in practice sessions is to allow the listener to modify his reference values and thresholds, where necessary, and this is one of the major purposes of practice sessions.

4.7 TASK DIFFICULTY

Since identification involves the comparison of a stimulus voice with more than one reference voice, one might suppose that this task were more difficult to perform than differentiation or verification, where only one reference voice is considered, i.e. that multiple comparisons are more difficult than single comparisons (section 4.4). Whilst this is probably true in a general sense, identification tasks may not be so much more difficult as one might imagine, for at least the following reasons.

(i) By definition, identification tasks involve reference voice patterns stored in long-term memory. As was mentioned above, the specifications of reference voice patterns stored in long-term memory are typically greater than those stored in short-term. Thus identification tasks have this advantage over differentiation tasks (but not over verification tasks, which also use long-term memory).

(ii) In identification tasks, the listener is told the size and identity of the population of reference speakers with whom stimulus voices are to be compared. In this way, these reference voice patterns to be considered may be primed, i.e. called forward from the total inventory of reference voice patterns stored in long-term memory. Any advantage gained in differentiation tasks from the immediacy of retrieval of reference voice patterns is thus diminished.

(iii) One of the most complex points in the model is the COMPARE process at the end of each task. In closed identification tasks, its function is merely to select the highest-ranking reference voice pattern from amongst the candidates (except in the case where no reference voice pattern approximates at all closely to the stimulus, and where recursion is probable). This decision is therefore categorical, with little relevance being assigned to the degree of approximation between the reference and stimulus voice patterns.

However, it is precisely this degree of approximation which is the crucial factor in the mode of operation of the COMPARE process in differentiation and verification. This decision is not categorical but probabilistic, and therefore more complex. The case of open identification is considered in the next section.

It is easier to rank the relative difficulty of differentiation and verification tasks. In terms of the model, they involve the same processing; however, verification is the less difficult for at least the following reasons.

(i) As was discussed above, reference voice patterns stored in long-term memory are typically more fully specified than those in short-term.

(ii) The short-term memory capacity in both differentiation and verification tasks will contain the stored representation of the stimulus utterance and the parametric values extracted from it. However, in differentiation tasks, the reference voice pattern will also be stored there, so that differentiation involves the retention of a greater amount of information in short-term memory. Advantage can hardly be taken, in differentiation tasks, of long-term memory with its much larger capacity.

(iii) It was mentioned above that any advantage due to the immediacy of retrieval of reference voice patterns in differentiation tasks could be diminished by the priming of reference voice patterns from long-term memory in identification tasks. This factor also applies to verification.

Therefore, as a very general statement, the three kinds of task may be ranked in terms of difficulty of performance as verification (least difficult), differentiation (more difficult) and identification (most difficult).

The strategies adopted in the performance of these tasks can all be viewed as forms of simplification of the process. The reduction of an ABX into a BX format (section 4.4) is a simplification of a multiple comparison into a single comparison. The separability approach adopted by some listeners in simultaneous presentation tasks (section 4.3.1) avoids the whole process of reference-stimulus comparison.

4.8 OPEN IDENTIFICATION

The open identification task was introduced in section 4.4. Experimentally, the listener has to decide whether the stimulus utterance was produced by one of the references, or by a speaker who is not represented in the reference population. Although the task has not figured strongly in the experimental literature, it corresponds to an extremely common situation in the real world. Examples of this occur every time the telephone rings and you are not expecting somebody to be calling you. When you answer the telephone, you may expect the caller often to be a familiar voice. However, it may equally well be someone you have never heard before, in which case you will not already have an appropriate reference voice pattern for the caller in long-term memory. This task is involved in all cases where you are not certain that you are dealing with a finite population of reference speakers.

Very little discussion is contained in the literature of the processes to be undergone when this experimental task is set. Among those who have written about the task, there are two viewpoints as to how it is performed. The first is illustrated by the following three quotations.

> 'The first task of the subjects was to determine whether a presented sample was or was not by one of the eight talkers and to give a rating of the confidence with which he made his decision.

> Secondly, if the subject decided that the test sample was by one of the eight talkers, his task was to identify the talker and rate the confidence with which he made his identification.'
>
> (Stevens et al., 1968:1601)

> 'He has to decide whether or not the unknown sample is one of the known ones and if he decides positively, must determine which known talker is same as the unknown.'
>
> (Tosi, 1979:7)

The third quotation in this first category refers to open identification tasks in the visual comparison of spectrograms.

> 'The tasks of the examiners in the open trials consisted of deciding whether the "matching" spectrograms were or were not produced by one of the "known" speakers - and if they were, which "known" speaker produced them.'
>
> (Tosi et al., 1972:2036)

The viewpoint expressed by these three quotations implies that the process involves two decisions:

(i) whether the stimulus is represented in the reference population, and

(ii) if so, which reference corresponds to the stimulus.

However, these cannot be thought of as two separate decisions. How can one decide that the stimulus is represented in the reference population unless one has already selected the specific corresponding reference voice? In other words, the two are parts of the same process - one cannot answer the first question without also being able to answer the second.

The second viewpoint is expressed by Hecker (1971).

> 'The listener carries out two tasks in succession; first he decides which reference sample is most similar to the test sample, and then he decides whether the two samples are similar enough to have been produced by the same speaker. If the listener should find that none of the reference samples resemble the test sample, or that the selected reference sample and the test sample are not similar enough, he reports that he cannot identify the speaker of the test sample.'
>
> (Hecker, 1971:35)

Because, in the way in which he describes it, this task has some of the features of an identification task and some of a discrimination task, Hecker refers to this experimental format as the identification-discrimination test. Since this name is not consistent with the categorisation of task names given in section 4.4, I shall not use it to refer to this task. Instead, I shall follow Stevens et al. and Tosi et al. in calling it open identification. It is identification because it involves several reference voice patterns.

The performance of this task is therefore again seen as a two-part process:

(i) the reference corresponding most closely to the stimulus is selected, and

(ii) the degree of correspondence between the two is assessed against some threshold value.

If the correspondence falls above this threshold, the reference is given as the response; if it falls below, the "stimulus does not correspond to any of the references" alternative is given. This two-part categorisation is useful in that the difference between closed and open identification tasks can be described quite simply. Closed identification involves only the first part and therefore requires a categorical decision from the listener. Open identification,

on the other hand, makes use of probabilistic information concerning the reference-stimulus correspondence. This difference relates to the way in which the COMPARE processes in the model operate (A5, B4, B11). Although the process involved in open identification has been described above as a two-part decision, it is still one single process.

If the correspondence between the selected reference and the stimulus is sufficient (on first-order parameters), the listener will give the reference as the response. If the correspondence is not great enough, Hecker considers that the listener will give a "stimulus does not correspond to any of the references" response. In terms of the model presented in this chapter, it is preferable to say that the analysis switches to second- and lower-order parameters (if external factors allow) for confirmation of this lack of correspondence before the response is given. Thus the full range of possible parameters is exhausted, and not merely those of the first order, which are habitually used. This explanation accounts for the observation that a "stimulus does not correspond to any of the references" response will take longer to be reached than the "speaker A" type of response.

To look at this same process from a converse angle, it is probable that, knowing that the task is one of open identification and that the stimulus therefore does not necessarily correspond to any of the references, the analysis will whenever possible be taken to a depth of analysis greater than first-order parameters.

Incorporation of a "don't know" alternative into the model may be handled, as for the differentiation and verification process, by the introduction of a second threshold value. Thus a "don't know" response would be given if a reference-stimulus correspondence was lower than the threshold for positive acceptance, but higher than this

second threshold for positive rejection. However, since the stimulus sample is being analysed in relation to the full range of parameters (if external factors allow) if a satisfactory correspondence is not found, the listener will be more certain of his decision and "don't know" responses will be rarer than in other tasks.

CHAPTER 5

REVIEW OF THE LITERATURE

CHAPTER 5

REVIEW OF THE LITERATURE

5.1 INTRODUCTION

A review of the literature may take many forms. For an experimental field, it may take the form of a catalogue of experiments performed - their methodology, results and conclusions. Usually this is a reasonably exhaustive list of the relevant work, and is accompanied by an interpretive commentary which attempts to link the enormous variety of findings into an integrated overview of the field.

The review of the literature contained in this chapter will not attempt to have such a typical exhaustive form, for three reasons.

Firstly, it has hopefully been shown that no writer has yet proposed a theoretical framework for speaker recognition, along the same lines or in such depth as that already discussed in the preceding chapters. For example, no writer has even given a rigorous definition of exactly what is meant by speaker recognition (and what is not included in it), with reference to the sorts of factors discussed in Chapter 1. It is therefore no surprise that, in my opinion, much of the experimentation reveals little about the speaker recognition process since results are difficult to interpret meaningfully; experimenters use differing numbers of speakers and listeners, different tasks, different methods of assessing listener performance, etc. It is therefore difficult to compare the results of one experiment with those of another which has been investigating the same feature but has used a substantially different experimental design. (Indeed, one cannot call it a "substantially"

different design unless one has a fair idea of the influence of the difference in experimental factors). Therefore further investigation of the relative effects of such factors is necessary.

Secondly, it has already been stated in many places that the central interest of this thesis is human speaker recognition in everyday situations, and that the experimental situation differs in several significant respects from the everyday, real world situation. By investigating speaker recognition by the use of experimental formats, one is necessarily introducing an unnaturalness into the situation. Of course, it is impossible to avoid this unnaturalness since one requires the control afforded by experimental formats. However, an appreciation of the nature of this unnaturalness is necessary if one wants to be able to interpret results in such a way as to advance our understanding of the human speaker recognition process. No writer has really approached the question of this introduced unnaturalness and therefore, again, one cannot interpret experimental results meaningfully for human speaker recognition in everyday situations.

Thirdly, exhaustive reviews of the experimental speaker recognition literature already exist. Both Hecker (1971) and Bricker & Pruzansky (1976) provide comprehensive catalogues of experimental work in the field. Both attempt to be interpretive, although too great an emphasis, to my mind, is laid on the experimental side of the field at the expense of any consideration of non-experimental factors.

The review of the literature which follows therefore does not attempt to be an exhaustive listing, but concentrates instead on the major findings concerning the whole range of experimental variables. Bricker & Pruzansky's (1976) categorisation of how the various stages of the speech chain relate to manipulable factors in the experimental format (see section 2.2) may be taken as our starting point. Of the

five manipulable elements in that categorisation (speakers, material, transmission parameters, listeners and tasks), only the first three will be discussed in this chapter. Experimental tasks have already been treated at some length in sections 2.3, 4.3 and 4.4, and listener-oriented factors are handled in Chapter 7, although, as is pointed out in those sections, very little investigation of these factors has been reported in the literature.

Reviews of the speaker recognition literature generally divide the field into three separate aspects (see section 1.1):

(i) human auditory speaker recognition

(ii) human speaker recognition by the visual examination of spectrograms

(iii) automatic speaker recognition by machine

Very little will be said about the last two kinds of speaker recognition. It may be argued that the visual examination of spectrograms, although performed by trained human observers, has more in common with the automatic process than with the listening process, in that spectrographic examination and automatic recognition both bypass the human auditory perceptual process. Discussion of these two processes will therefore be limited to a listing of the pros and cons of each, along with an examination of the relationship of each technique to the human listening process, and any implications of experimental results for the human process.

5.2 AUDITORY SPEAKER RECOGNITION

5.2.1 Speakers

The two main factors to be considered in relation to the selection of speakers to be used as references and/or stimuli in experimental tasks are (i) the size of the group, and (ii) the homogeneity of the group.

As regards the first of these factors, Pollack et al (1954) used groups of up to 16 speakers uttering monosyllabic words. Since the limit of information transmitted was not reached, their data suggests that it is possible to use a greater number of speakers reliably in identification tasks. In contrast, Williams (1964) tested three listener groups with speaker ensembles of 4, 6 and 8 voices respectively. The drop in performance between the 6-speaker and 8-speaker tests was considerably larger than that between the 4-speaker and the 6-speaker, leading Williams to the conclusion that 'considering such factors as training time, test time, amount of information, etc., it may well be that an ensemble of 5 or 6 speakers would be optimal for testing speaker identification' (1964: 14). The opposing nature of these two conclusions suggests that what is important in experimental tasks may not be so much the size of the speaker group, but the homogeneity within the group.

In another part of Williams' (1964) study, he divided a speaker population of 12 randomly into two groups of six speakers ("A" and "B"). The discrepancy in the final test scores (50% and 62% correct respectively) indicates that

> 'speaker identification depends not only on the individual characteristics of each of the speakers but also on the characteristics of the other speakers with whom he is being compared.'
>
> (1964:22)

In short, speaker group "A" was obviously more homogeneous than speaker group "B". Specifying those characteristics which contribute to this homogeneity is a major question still to be answered. In a sense, this argument is circular: it is hoped that a procedure may be devised for selecting homogeneous speaker groups, so that these groups can be used in experimental tasks designed to investigate which are the features by which speakers differ and by which they may be recognised. It therefore comes as no surprise that these features which

contribute to homogeneity have so far largely eluded attempts at specification (Carbonell et al, 1965; Clarke & Becker, 1969). It may well be found that they must again be assigned to idiosyncratic factors (see section 1.3), which is simply a cover-term for features which do not correlate with any element of our categorisation. In experiments, researchers have controlled as well as possible the speaker group homogeneity by controlling the gross factors of sex, age and dialect (usually by selecting male undergraduates, 18 - 25 years old, native speakers of Standard (British or American) English). It has been shown that listeners can reliably identify speakers' sex (Coleman, 1971) and age (Ptacek & Sander, 1966) from their voices, although the effect of this ability on the recognition of speakers' identity has not been investigated. Coleman (1973), in an experimental format where differences in glottal source characteristics were eliminated by the use of an electro-larynx (see below), found that significantly more errors were made in female-female differentiation tasks than for male-male pairings (using ten male and ten female speakers). He concludes that:

> 'male speakers retain a greater degree of identifiability when the discriminations are based on speech characteristics other than phonatory ones. One possible interpretation of this is that the quality of the substituted glottal tone, with its low fundamental frequency of 85 Hz, is more like that of males as a group and this may have contributed a more unnatural quality to the speech of the females. A more likely explanation is that the males in this study differed more among themselves than did the females in speech characteristics such as speaking rate, pause-time, phrasing or syllabification. These would be unaffected by the equalisation of the glottal sound and may have resulted in a perceptually significant variation among the male speakers. The possibility of between-sex differences of this type needs to be investigated further.'
>
> (Coleman, 1973:1743)

5.2.2 Materials

The semantic content of the materials which speakers are required to read is controlled by the use of standard phonetic passages, or articulation- or audiology-test lists.

The factor which has often been used as the experimental variable is the duration of speech samples. This is usually more critical for stimulus samples than for references. Pollack et al (1954) demonstrated that listeners' performance reached a level of over 90% with stimulus sample durations of approximately one second; there was no appreciable improvement in performance with increases of sample duration beyond one second. In contrast, Compton (1963), using isolated steady-state [i] vowels, found that listeners' performance was significantly greater than chance with sample durations of as little as 25 msec. However, Pollack et al add that

> 'on the basis of exploratory tests, we believe that the duration of the speech sample per se is relatively unimportant, except insofar as it admits a larger or smaller statistical sampling of the speaker's speech repertoire.'
>
> (1954:406)

That it is not the duration of the sample but the range of phonemes presented which is important was also demonstrated by Bricker & Pruzansky's (1966) experiments which used sentences, disyllables, monosyllables, consonant + vowel excerpts and vowel excerpts as stimuli. The same finding has been produced in similar formats by Williams (1964) and Stevens et al (1968). Clarke et al (1966) found an insignificant difference when using sentences containing from three to eleven syllables. However, Hecker (1971) points out that since the duration of a 3-syllable sentence may well be almost one second, this finding does not contradict that of Pollack et al (1954).

5.2.3 Transmission Parameters

Since the Second World War, recording techniques have been sufficiently advanced for signal degradation not to be a problem for the experimenter. It is, of course, desirable to use a tape-recorder in experiments since it allows exact repetition of samples, control of sample durations, amplitude levels, etc. In the first systematic experimental study, McGehee (1937) used "live" speakers screened visually from the listeners. In her second series of experiments (McGehee, 1944), and in all experiments since then, electronic recording techniques have been employed.

However, the above is not intended to be implied primarily by the use of the term "transmission parameters". Instead, it refers to any features of a signal which, whether manipulated by the experimenter or not, may affect listeners' performance in a speaker recognition task.

5.2.3.1 Pitch

Development of the laryngograph (Fourcin & Abberton, 1971) has allowed glottal source characteristics to be investigated directly from vocal cord vibration rather than from the full speech signal. Using laryngographic recordings as stimuli, Abberton (1974) carried out an identification task, concluding from her results that listeners can use glottal features, especially mean fundamental frequency, as clues in speaker recognition tasks.

5.2.3.2 Filtering

Experiments in which frequencies are selectively filtered are based on the supposition that speaker-characterising information may be considered to be concentrated in certain frequency ranges. In the earliest such study, the results of Pollack *et al* (1954) indicated that high-pass and low-pass filters with gradual cut-off

characteristics at 500 Hz and 3,000 Hz respectively have an insignificant effect on speaker identifiability. Clarke et al (1966) produced slightly different results, showing that energy outside the 500 - 3,000 Hz range may also affect speaker identifiability. Results of Peters (1954), who used octave band-pass filters, indicate that, of the bands selected, the 1,200 - 2,400 Hz range is the most critical for speaker recognition. Dukiewicz (1970) showed that the importance of this frequency range may depend on the particular vowel being uttered. Using two three-octave band-pass filters (128 - 1,024 Hz and 1,024 - 8,192 Hz), recognition for the vowel [i] was more successful in the higher range, but for [u] was better in the lower. This effect is simply explained since the second formant for the vowel [u] is generally lower than 1,024 Hz, while that of [i] is substantially higher. This explanation is supported by the finding of Compton (1963) that high-pass filtering of the vowel [i] at 1,020 Hz has an insignificant effect on listener performance, while low-pass filtering at that frequency substantially reduces correct identification scores. Gross filtering may therefore overlook predictable segmental differences.

5.2.3.3 Segmental Differences

Various experimenters have investigated the hypothesis that different speech sounds convey differing amounts of speaker-characterising information (and therefore produce differing recognition success rates). Ramashvili (1966), using isolated Russian phonemes, specified a rank order of (i) vowels, (ii) voiced consonants, and (iii) voiceless consonants in descending order of speaker identifiability. Mean scores ranged from 90% for /ɛ/ to 30% for /k/. One exception to the above ranking was the phoneme /u/, which had a score lower than for other vowels. Similarly, Stevens et al (1968) found recognition to be superior for front vowels (including [i]) than for back vowels (including [u]). Since front vowels in general differ from back vowels by having higher second formants, Stevens et al concluded that this was

the major contributor to the differential effect, in keeping with Dukiewicz's (1970) finding reported above, and that of Doehring & Ross (1970) regarding the back vowels [ɑ] and [u].

5.2.3.4 Backward Speech

Use of the tape-recorder makes it simple to play recorded samples backwards. Experiments employing backward stimuli have shown that this produces a significant effect on listener performance as against normal, forward speech, although the size of the degradation caused varies. Bricker & Pruzansky's (1966) results show a decrease in correct identification of approximately 10% on average for various types of stimuli. This compares with the 25% figure produced by Williams (1964). Using various filtering conditions, Clarke et al (1966) found that recognition was better for backward speech when high-pass filtered at 250 Hz than when unfiltered.

Conclusions regarding backward speech are therefore mixed. These conflicting results may derive from various factors. Most importantly, it is not fully understood (or at least has not been fully described) exactly what features of the speech signal are disrupted by backward presentation and how this disruption relates to listeners' perception. It is clear that all temporal sequential features are reversed. The semantic content of the utterance is thus eliminated. Another product of this reversal is that through-time fluctuations such as fundamental frequency contours are disturbed; for example, a statement uttered with a final fall in fundamental frequency and amplitude (which is a normal feature in Standard English) will become a sentence with an initial rise in fundamental frequency and amplitude (which is not). Parameters which do not relate to through-time fluctuations, such as mean fundamental frequency, will remain unaffected. This point seems to be overlooked by Williams (1964) who claims that backward speech retains 'essentially just spectral content' (p.17).

Since backward speech produces inferior performance, the conclusion is drawn that temporal features are important in speaker recognition. However, as Hecker (1971) points out,

> 'it is not clear whether listeners use temporal clues *per se* or whether their judgments depend on a perceptually realistic speech signal. The ability to identify speakers has been learned with natural speech over many years; this ability may not be readily transferable to a novel form of speech distortion that can exist only in the laboratory.'
>
> (Hecker, 1971:42)

Backward speech, like the simultaneous presentation format (see section 4.3.1), cannot exist in the real world and results of experiments using backward speech are therefore difficult to interpret.

5.2.3.5 Whispered Speech

A simple way to eliminate glottal characteristics which result from voicing is to have speakers whisper stimulus utterances. Ignoring the fact that different types of whisper exist (Catford, 1964), the result of this modification is that listeners' responses are based essentially on spectral information. This reduction in information produces a significant degradation in correct identification scores. Williams (1964) found that listeners performed only 57% as well for whispered speech as compared with normal voiced stimuli, a figure which was even lower than for backward speech. The results of Abberton's (1974) experiment show a much smaller degradation for whispered as against normal speech (roughly 10%). Pollack *et al* (1954) showed that samples of whispered speech needed to be over three times as long as normal samples for equivalent recognition scores to be achieved. It is clear therefore that the glottal characteristics of voicing contribute to the identifiability of speakers, although the deterioration produced by absence of voicing may be difficult to

quantify. A converse way of interpreting the results of experiments using whispered speech is to say that a reasonable degree of recognition is still possible when only spectral (vocal tract) features are available. (For some of the listeners in Abberton's experiment, scores were as high for whispered speech as for normal).

5.2.3.6 Vocal Tract Features

Two experiments have used other ways of eliminating glottal features associated with voicing. In Coleman's (1973) study, the glottal pulse was generated by an electro-larynx with a fixed fundamental frequency of 85 Hz (and fixed pulse-shape). In a differentiation task, correct response scores reached 90%, leading Coleman to the conclusion that most of the information on speaker characteristics is preserved when differences in glottal source are eliminated.

Glottal source characteristics were eliminated by Shearme & Holmes (1959) by the use of a channel vocoder with a fixed fundamental frequency of 120 Hz. Formant values were either kept as in the original utterances, or raised by 100 Hz (F1) and 300 Hz (F2 and F3). Raised and unraised samples were used in a differentiation task. Results are taken to indicate that spectral features such as formant positions play a part in speaker recognition tasks.

5.2.3.7 Relative Importance of Vocal Tract and Glottal Features

A more sophisticated study than that of Shearme & Holmes was performed by Miller (1964). Using inverse filtering in synchronism with fundamental frequency, the vocal tract transfer functions and glottal source spectra from recorded utterances were separated by computer. Hybrid samples were then produced by combining the glottal

source spectrum of one speaker with the vocal tract transfer function of another. In two-reference identification tasks, listeners judged the hybrid samples to correspond more closely to the reference owners of the vocal tract component rather than the reference owners of the laryngeal component. Further experiments using substituted artificial glottal waveforms consolidated the finding that vocal tract features are more important to speaker recognition than laryngeal features. (Miller's (1964) abstract contains only a brief account of the procedure used. The series of experiments is reported in greater detail in Hecker (1971)). The experiment of Matsumoto et al (1973) used a similar procedure, and their findings agree with those of Miller.

There is thus general agreement that vocal tract features are most important in speaker recognition. The importance of laryngeal features is generally considered to be secondary, especially in view of the results of Coleman's (1973) study, by which

> 'the maximum reduction in speaker identifiability that might be expected to result from attempts to disguise the voice by modifying the laryngeal tone, would be something less than 10% assuming a similar listening task to that utilised in this study.'
>
> (Coleman, 1973:1743)

An exception to this general agreement is Clarke et al (1966) who

> 'tentatively conclude that pitch is the single most important characteristic of the speech waveform used by the human observer in speaker recognition.'
>
> (Clarke et al, 1966:42)

All of the experiments reported so far in this review uphold one point - that none of the features investigated in the experiments contributes nothing to speaker identifiability. That is, all the features contain a certain degree of speaker-characterising information

and are potentially usable in performing speaker recognition tasks. This leads us back to the distinction stated in section 1.8.3 between those parameters which listeners are capable of using and those which they habitually use in everyday situations, and to Clarke et al's (1966) hypothesis concerning the flexibility and adaptability of the human ability. If we believe in this hypothesis (and no data yet produced contradicts it), then it may be argued that most of the literature tells us little about how listeners recognise speakers in everyday situations. The experiments which are relevant in this respect are those which attempt to rank the relative importance of features of the speech signal (e.g. Miller, 1964; Matsumoto et al, 1973). It is more useful to know that vocal tract features are more important than glottal features than to know that vocal tract features are important (Shearme & Holmes, 1959) or that glottal features are important (Abberton, 1974). The experiments reported in Chapter 6 are an attempt at further specification of the relative importance of parameters.

5.3 SPEAKER RECOGNITION BY THE VISUAL EXAMINATION OF SPECTROGRAMS

The sound spectrograph (Potter, Kopp & Green, 1947) is a machine which converts a speech signal input into a visual display. The three parameters represented on the two-dimensional spectrogram are time (horizontally), frequency (vertically) and amplitude (by the blackness of the pattern). The bandwidth of the analysing filter may be set at wide (300 Hz) for examining formant structure, or at narrow (45 Hz) for examining harmonic structure. Since the spectrograph might be thought of as providing an objective representation, it was soon appreciated after its invention that it could be applied to the question of speaker recognition. It was believed that the spectrogram could reliably capture the speaker-characterising features of an utterance and so it was given the name voiceprint (Gray & Kopp, 1944) on analogy with the uniqueness of human fingerprints. In the early 1960's, interest in this aspect of spectrography increased when

voiceprint data was produced as evidence in legal cases, and claims were made as to its infallibility (Kersta, 1962). During this period, the contour spectrogram unit was invented (Prestigiacomo, 1962), whereby points of equal amplitude are joined by lines, as in altitude contour maps. However, after experimentation, doubts were cast as to the reliability of spectrograms (Ladefoged & Vanderslice, 1967; Bolt et al, 1969, 1970, 1973; Black et al, 1973; Jones, 1973) and they are now not thought reliable enough to be produced as anything greater than supportive evidence.

Hecker (1971) and Tosi (1979) give fuller descriptions of the sound spectrograph and of the results of visual presentation experimentation. There are various possible reasons for the generally inferior performance of observers in visual rather than aural recognition tests.

(i) There is no reason to suppose that those features which are most important in speaker recognition are revealed in either wide-band or narrow-band spectrograms. The relationship between auditory and spectrographic patterns is a complex one, and auditory prominence does not necessarily imply visual prominence.

(ii) The type of spectrogram most widely used experimentally is the 4 kHz wide-band bar or contour. However, this may not be the most appropriate for revealing speaker characteristics.

(a) Wide-band spectrograms show virtually no fundamental frequency information, which is seen well on narrow-band spectrograms. Conversely, formant structure is difficult to deduce from narrow-band spectrograms, but is clear on wide-band.

(b) For electronic reasons, contour spectrograms have better amplitude resolution but worse temporal resolution than bar spectrograms. Kersta (1962) showed that bar spectrograms give higher recognition scores than contour.

(c) It has been claimed that first formant movement is a highly characteristic feature. This will be displayed more clearly on spectrograms where frequency is plotted on a logarithmic rather than the usual linear scale (Anonymous, 1965), although this will obscure characteristics occurring at higher frequencies.

(d) A certain, admittedly small amount of useful information may be present at frequencies of 4kHz and above.

All the above arguments make it clear that only certain features of the speech signal (and thereby only certain speaker-characterising features) can be represented well on any one type of spectrogram. Use of a series of spectrograms of the same utterance, differing in frequency scale, filter setting, etc., may contribute significantly to the accuracy and confidence of observer responses.

(iii) A crucial factor in spectrographic recognition tasks is the length and nature of the training given to the observers. For example, Young & Campbell (1967) gave subjects a minimum of 2½ hours' training. The International Association of Voice Identification (Tosi, 1979) has laid down recommendations for the training of observers in legal cases. Smrkovski (1976), in a small-scale experiment, found I.A.V.I. qualified observers to perform more reliably than untrained subjects. Differences between aural and visual recognition scores may largely be due to differences in experience of the two kinds of stimuli. Humans have spent the whole of their lives listening to voices, whereas spectrographic displays are novel to them.

The major work comparing aural and visual recognition of voices is Stevens et al (1968). Their conclusions summarise the similarities and differences between aural and visual presentation of voices. My comments are included in square brackets.

'(1) Aural identification of talkers based on utterances of single words or phrases is more accurate than identification from the spectrograms, using a matching-from-sample technique. Furthermore, the performance of subjects improves more rapidly with learning for aural identification tests than for visual identification tests, and the subjects are more confident of their identifications for the aural tests.

(2) For visual identification, longer utterances increase the probability of correct identification [whereas for aural identification the improvement was minimal and was in keeping with Pollack et al's (1954) findings].

(3) It is easier to identify a talker when he utters a word containing a front vowel than when he utters a word containing a back vowel [for both aural and visual presentation].

(4) There are large differences in aural identifiability of voices, even when the voices have been selected to be reasonably homogeneous on the basis of judgments of attributes of the voices by a panel of listeners. [Furthermore, the relative identifiability of voices is maintained whether they are presented aurally or visually. For example, the voice recognised least successfully on spectrograms also produced the lowest aural recognition score].

(5) There are large differences in the ability of subjects to identify voices on either a visual basis or an aural basis. [For subjects, the relative ability to identify was not maintained for aural and visual presentation. For example, the least successful subject at visual recognition was not also the worst aural listener].

(6) Indirect evidence suggests that talker identification scores based on responses from a panel of subjects are much better than scores for individual subjects. Further improvement can be obtained if the votes of members of the panel are weighted according to their confidence ratings.

(7) Indirect evidence also suggests that a matching-from-sample technique in which the comparison items consist of several repetitions of the utterance by each talker leads to improved scores relative to the case in which only a single comparison utterance is available from each talker.

(8) Authentication of voices is much poorer on a visual basis than on an aural basis.'

(Stevens et al, 1968:1607)

5.4 AUTOMATIC SPEAKER RECOGNITION BY MACHINE

There are two basic approaches to the machine recognition of voices. The choice between the two is determined largely by the application to which it is hoped the machine may be put.

Firstly, the computer is programmed to compared stimulus and reference matrices of amplitude, frequency and time, using specific utterances. These utterances may be of two kinds:

(i) specific sentences, phrases or words, in which case difficulties of temporal adjustment will be encountered since separate tokens of the same phrase, even if spoken by the same speaker, may not always exhibit identical temporal characteristics, (see section 1.8.2), and

(ii) specific segments (for example, the vowel [i]). The latter case naturally requires that the relevant segment is preliminarily detected and extracted from the stimulus signal. This preliminary analysis is by no means simple and has often had to be performed manually by the experimenter.

Secondly, recognition may be based on the statistical long-term analysis of parameters, such as fundamental frequency. The stimulus text must therefore be substantially longer than that required for the above method.

The recognition process may be considered as being composed of three distinct processes:

(i) preliminary analysis extracts those segments which are to comprise the input to the system,

(ii) the acoustic parameters or features which are thought to contain the speaker-characterising information of the signal are extracted, and

(iii) a decision-making algorithm uses these parameters as input in order to reach a recognition response.

The first automatic recognition approach outlined above, where specific segments are analysed, necessitates all three processes; the second approach which treats the long-term features of the signal, does not require the preliminary analysis, and only the last two subparts are entailed.

It is thought by some writers that a study of the findings of machine recognition experiments will lead, if somewhat indirectly, to a greater understanding of the human speaker recognition process. However, if the ultimate objective of an automatic recognition device is taken to be a 100% correct recognition rate (or, more realistically, a 0% false acceptance rate), then parallels between the machine and human processes may be only superficial. Human listeners do not achieve a 100% correct recognition or 0% false acceptance rate except with the most trivial of tasks. In addition, there is no a priori reason why those parameters which prove to be the most reliable for automatic speaker recognition should correspond to those which are exploited most in the human process. Considerations such as these will have ominous implications for any automatic system modelled on the human process.

Studies which investigate the relationship between automatic recognition scores and human recognition scores for comparable tasks indicate that the major factor distinguishing the two is the ever increasing sophistication of computer techniques.

In Clarke & Becker's (1969) study, the same speech samples were used for recognition by machine and by human listeners. In (i) identification and (ii) differentiation tasks, the average correct recognition scores for the human listeners were (i) 63 - 67% and (ii) 90% respectively. The corresponding figures for machine recognition using a variety of parameters were (i) 28 - 63% and (ii) 58 - 83%.

In a similar experiment, Rosenberg (1973), using mimicking impostors, found that human listeners could recognise speakers as well as, and some better than, an automatic system. More significantly, however, the false acceptance rate for the automatic process was significantly lower than that of the human listeners. Using an identical twin impostor, Lummis (1973) found the same result using only fundamental frequency and amplitude information for the automatic procedure. Human listeners accepted the twin in most cases, whereas he was rejected correctly by the automatic system.

It can be seen therefore that machine recognition procedures have overtaken the human ability, in some respects at least. There are many reviews of the automatic speaker recognition literature (Su & Fu, 1973, ch.2; Atal, 1976; Rosenberg, 1976).

CHAPTER 6

EXPERIMENTATION ON SPEAKER-DEPENDENT FACTORS

CHAPTER 6

EXPERIMENTATION ON SPEAKER-DEPENDENT FACTORS

6.1 INTRODUCTION

The review of the literature contained in the previous chapter has emphasised the fact that a distinction should be drawn between those parameters which listeners *can* use for speaker recognition, and those which they habitually *do* use in everyday situations. Most of the experiments reported relate to the former category, and may therefore be useful for automatic speaker recognition considerations, whereas the experiments of Miller (1964) and Matsumoto *et al* (1973), which investigate the relative importance of various parameters, relate to the latter and are thus more important in human speaker recognition. The experiments reported in this chapter constitute an attempt at a more detailed specification of the relative importance of parameters. Both Miller's and Matsumoto *et al*'s experiments deal with the separation of vocal tract transfer function and glottal source characteristics. The present experiments go further than these, in two respects. Firstly, component features of the formant structure and glottal source are investigated. Secondly, a wider range of features is considered - eight in all, including some suprasegmental parameters, which were not varied in Miller's and Matsumoto *et al*'s work.

6.2 EXPERIMENT 1

It was decided to use a format in which listeners were required to judge the similarity between pairs of synthetic voices. Synthetic speech was used since the stimulus samples needed to be very strictly controllable in terms of their acoustic structure. However,

there are drawbacks to the use of synthetic speech for such experiments.

(i) The mechanical design of the synthesiser imposes restrictions on the choice of variables to be used.

(ii) Mechanical structure differs to varying degrees from one speech synthesiser to another. This may make future replication or adaptation of experiments difficult except on the same machine.

(iii) In Williamson's (1961b) experiment, which used both synthetic and live stimuli in a speaker differentiation task, listeners found the synthetic stimuli to sound so mechanical that they were unwilling to accept them as plausible human voices. The problem became apparent in trial sessions:

> 'The results of these trial-runs were so disappointing that it was not felt worthwhile carrying the test further ... Many subjects reported that quite often the test decision was such a difficult one that they had had to guess, and the results seemed to bear this out.'
>
> (Williamson, 1961b:30-31)

However, Williamson's experiment differed significantly from the present ones, in that hers involved the recognition of speakers, whereas the present ones require the listener to judge the similarity of the voices (see below).

The PAT synthesiser (an acronym for a Parametric Artificial Talker) was developed in the 1950's (Lawrence, 1953; Anthony & Lawrence, 1962), when interest arose among communication engineers as to the minimum amount of information which needed to be conveyed for the successful transmission of intelligible speech. For this reason, the number of acoustic variables (parameters) was kept to a minimum of

eight. A simplified diagram of PAT is given in Figure 6.1. The eight parameters used by PAT are as follows:

(1) Larynx amplitude (AØ) controls the amplitude of the larynx pulse. It may be set at zero (off) or at any setting up to maximum amplitude.

(2) Larynx frequency (fundamental frequency, FØ) controls the frequency of the larynx pulse and therefore only has effect when the larynx amplitude is higher than zero. Under normal circumstances, the FØ range on PAT is 50-250 Hz.

(3,4,5) Centre frequencies of the first three formants (F1, F2, F3). Specifying the centre frequencies of F1, F2 and F3 has been found to be sufficient for the characterisation of vowel qualities. There is a further formant frequency control which compensates for the fourth and higher formants.

(6) Hiss-through-formants amplitude (AH1) controls the amplitude of randomly generated noise, which is then subjected to the formant structure specified by the F1, F2 and F3 parameter settings. In this way [h] sounds (realisations of phonemic /h/, devoiced vowels and aspiration of stops) are simulated.

(7,8) Fricative hiss amplitude and frequency (AH2, FH2) also control the amplitude of randomly generated noise. However, since this noise is used to simulate fricative sounds (fricative realisations of phonemes such as /s/, /f/, and affricated release of stops), it is not subjected to formant structure. The place of articulation of such friction is characterisable by the lower frequency limit of the noise. The FH2 parameter is therefore used to differentiate, for example, between [s] and [f].

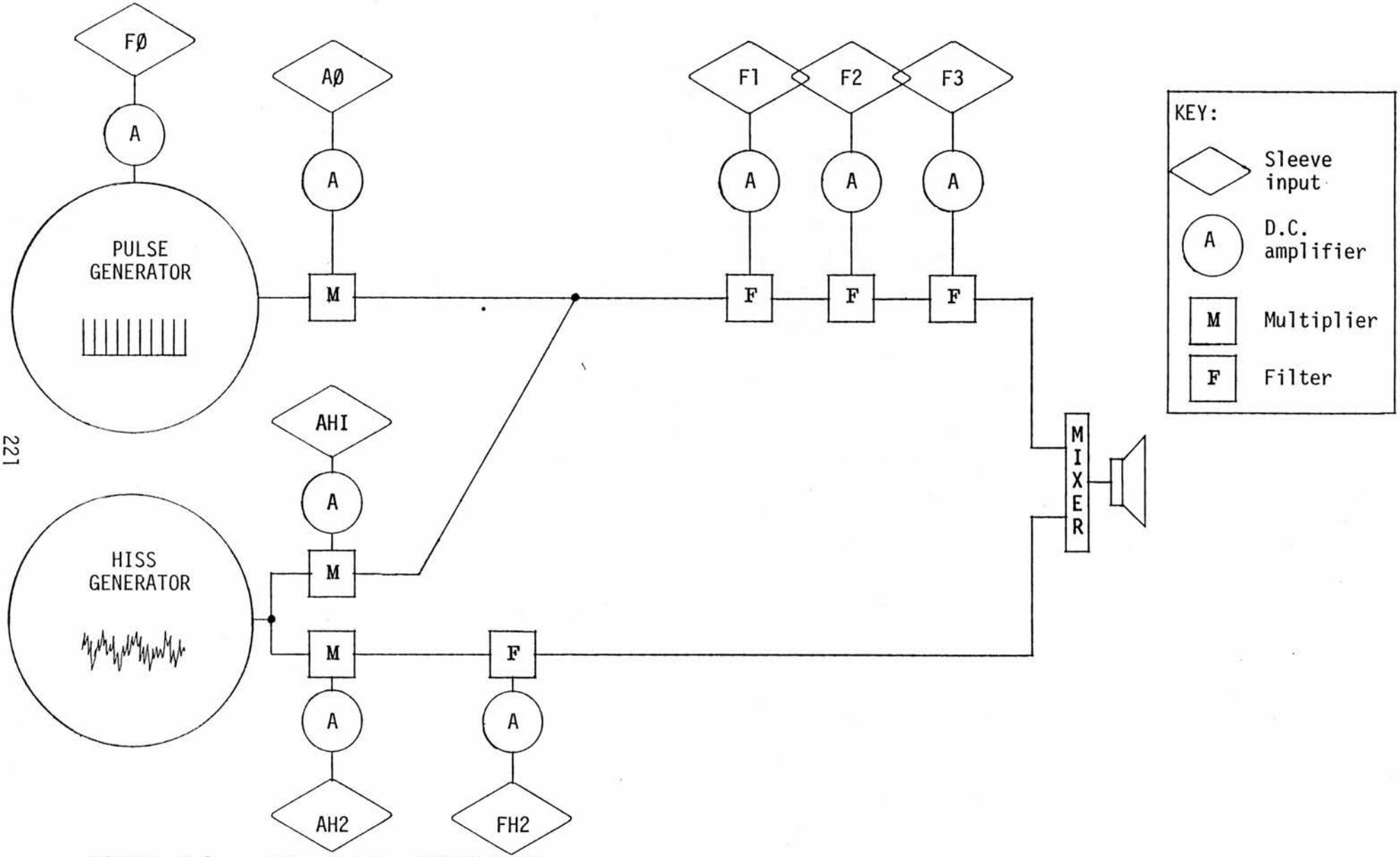


FIGURE 6.1 THE P A T SYNTHESISER

In order to produce the appropriate through-time transitions which are necessary for the perception of certain speech sounds, the eight parameters of PAT may be controlled in either of two ways:

(i) manually, by operating switches coupled to visible dials calibrated in the units relevant to the parameter. Naturally, this method allows the operator to control only two parameters at once, and since in connected speech all eight parameters need to be variable, this method cannot be used for the practical synthesis of anything but steady-state values.

(ii) automatically. The instructions for the eight parameters are represented as lines drawn in silver conductive ink on a transparent sheet of plastic. This sheet is drawn between two rollers, one of which is wire-wound and has a potential difference of 16 volts, falling off linearly. Thus the position of the line on the sheet determines the voltage it receives from the roller. The apparatus used is a later version of, but along the same lines as, that designed by Fourcin (1960). This voltage is converted into a value for one of the synthesiser's parameters. There are eight such lines on the sheet, each controlling one parameter. Since the information for the drawing of the lines is taken from spectrograms of a live utterance, this method is known as speech synthesis by analysis. An electrical motor draws the sheet through the rollers at the same speed as that of the Kay Spectrograph (5 inches per second), enabling features of the spectrogram to be traced directly. The voltage picked up by the silver line passes through a D.C. amplifier, which allows the operator to tune the synthesiser, by making sure that the values of the parameters correctly represent the drawn position of the silver lines. This is achieved by drawing calibration lines at the beginning of the sheet, which correspond to values at $\frac{1}{4}$ and $\frac{3}{4}$ of the range of the parameters. The amplifiers are then adjusted to produce the appropriate reading on the parameter dial.

The task in the present experiments required the listener to judge the similarity of pairs of synthetic voices, on the assumption that the more similar a pair of voices are judged to be, the more difficult they will be to differentiate; conversely, if two voices are judged to be very different, they will be easy to differentiate in a recognition task. This similarity judgment task differed from that used in previous experiments using synthetic stimuli (such as Williamson's, 1961b). In such experiments, the task was either one of identification (selecting one speaker from a specified population of reference speakers) or of differentiation/verification (judging speakers as the same or different). Consequently, the stimulus samples are presented to the listener as synthetic versions of specific live speakers. This situation has consequences, which may take either of two forms:

(i) The listener may be required to judge the voices as if they belonged to live speakers. Indeed, in Williamson's case, listeners were not even informed that any of the samples might not be of live speakers. As she discovered, listeners may be unwilling (consciously or subconsciously) to accept synthetic stimuli as possible utterances from live speakers.

(ii) The listener may be required to respond with the live reference speaker of whom he thought the stimulus was a synthetic version. This situation is influenced by the fact that listeners may adapt at differing rates to the novelty of synthetic speech, and by the differing amounts of their experience of synthetic speech. This format may not be so much a test of the listener's ability as of that of the experimenter (in programming the synthesiser) or of the design of the synthesiser (ever to be able to produce such an imitation).

To sum up therefore, listeners in the present experiments were required to rate the degree of similarity (or conversely of difference) between pairs of synthetic stimuli. They were informed

in the instructions that all the samples were synthetic, so that no element of speaker recognition was necessarily involved. That is, they were asked to judge the samples as voices rather than as live speakers (see section 1.4), and to give a scalar response (similar - different) rather than a categorical one ("same"/"different" or "speaker A"/"speaker B"/etc.).

The differences between the voices were represented by combinations of high, control and low values in relation to eight acoustic parameters. These eight were selected to represent a cross-section of the full range of parameters described in Chapter 3. The magnitude of the parametric changes (between control and high, and control and low values) is of importance if the results are to be indicative of the relative contribution of these changes to listeners' perception. However, there are no means to ensure that changes made in one parameter are of equal perceptual magnitude to those made in another. For example, there are no means for stating that a certain change in FØ and a certain change in tempo are of equal perceptual magnitude. The parametric changes were therefore decided upon such that the modifications produced a perceptual distance which was distinguishable but not large. Special attention was paid to two factors:

(i) For statistical purposes, the differences between the control and the high value for one parameter, and the control and the low value for the same parameter, had to represent the same proportional numerical distance.

(ii) Certain parametric changes interact, such as FØ mean and FØ range. Care was taken that, for example, a high FØ mean and a wide FØ range did not result in an upper extreme which was physiologically implausible.

The eight parameters selected for manipulation were as follows:

(1) Formant mean The possibility of manipulating values independently for each formant was rejected since (i) it would have led to a set of parameters which was unmanageably large and greater than the eight required for the statistical design (see below), and (ii) extrinsic articulatory factors causing changes to one formant are likely to produce concomitant effects in the other formants. Therefore the three parameters relating to supralaryngeal quality (this and 2 and 3 below) took the form of simultaneous changes to all three manipulable formants of PAT.

The formant mean parameter involved a 15% increase or decrease from the control value, this being the closest linear approximation to the auditory result of the raising or lowering of the larynx. The percentage change adopted in this experiment is larger than the 6.3% reported in Lindblom & Sundberg (1971), which was found to produce an effect of lowering of the larynx which was unconvincing and of minimal perceptual difference. The present percentage is closer to the values found in Sundberg & Nordström (1976), although segmental differences in percentage had necessarily to be overlooked. In addition, a difference in the percentages for the different formants would have been preferable but was impracticable. The 15% change was achieved mechanically by altering the D.C. amplifier settings to give a 15% higher or lower reading at the $\frac{1}{4}$ and $\frac{3}{4}$ calibration points. A compensatory adjustment was made to the parameter controlling the fourth and higher formants. This parameter is manually operated, i.e. is not determined by a silver line drawn on the sheet, and is therefore at a fixed setting for any one utterance.

(2) Formant range For this factor, changes were made to all three formants to create a wider or narrower range. For a wide range, the $\frac{1}{4}$ calibration point was lowered by 10% and the $\frac{3}{4}$ raised by 10%; for a narrow range, the $\frac{1}{4}$ was raised by 10% and the $\frac{3}{4}$ lowered by 10%. Articulatorily, this is the result of less extreme,

peripheral gestures, which corresponds to an auditory impression of laxness. Since the test samples involved the simultaneous alteration of both formant mean and range values, the effects of 15% raising or lowering and 10% widening or narrowing had to be combined for each sample. Since the parameter controlling the fourth and higher formants is fixed for any one utterance, it has no range value, and therefore its range cannot be widened or narrowed.

(3) Formant bandwidth is governed by a separate control on PAT, not by a silver line on the sheet. It is therefore not automatically controlled, but fixed for any one utterance. For the control sample, it was set at the standard value of 100 Hz; for the high setting, 150 Hz; for the low, 50 Hz. This represents the damping factor of the vocal organs, which corresponds to the auditory impression of dullness or sharpness. The perception of nasality may also result from wider bandwidths since 'the bandwidths of nasal formants are generally larger than in vowellike sounds' (Fant, 1962).

(4) Whisperiness The only phonatory factor of quality to be manipulated was the amount of whisperiness in the voice. This was produced mechanically by coupling the larynx pulse amplitude and hiss-through-formants amplitude parameters. (A similar technique for the synthesis of whispery voice was employed by Laver (1964, 1976)). The loudness and frequency of the combined signal required for the particular segment was still controlled by the AØ and FØ silver lines on the sheet (i.e. louder for a voiced vowel, zero for voiceless segments, etc.); however, the proportion of whisper to voicing was controlled by the manual D.C. amplifier setting, and could thus be set at a fixed amount for any utterance. The alterations required to change from PAT's normal, non-whispery voice to one with a controllable amount of whisper are summarised in Figure 6.2. Thus all the samples for the experiment were produced using different degrees of whisperiness. The low, control and high values used corresponded to agreed auditory

categorisations of slight, moderate and extreme whisperiness (see section 3.4.2.1; Esling, 1978). This is thus the only parameter in which the control sample differed from the live utterance of which it was a synthetic version.

(5) The FØ mean value of the utterance was raised or lowered by altering the D.C. amplifier setting for the ¼ and ¾ calibration points by 20% from its standard, control position.

(6) FØ range A similar offsetting of the D.C. amplifier control for the ¼ and ¾ calibration points enabled the FØ range to be widened or narrowed. For a wide range, the ¼ point was lowered by 15% and the ¾ raised by 15%; for a narrow range, the ¼ was raised by 15% and the ¾ lowered by 15%. Since there is an interaction between the raising or lowering of the FØ mean value and the widening or narrowing of the FØ range value, the appropriate settings for the ¼ and ¾ calibration points were calculated for the four possible combinations.

(7) The larynx amplitude mean value was altered by increasing or decreasing the D.C. amplifier setting from the control value. The amplitude of fricative noise (AH2) was not altered, so that changes in this factor also involved changes in the relative proportion of larynx amplitude to fricative amplitude. An interaction exists between the amplitude mean and whisperiness factors, and the correct settings of the ¼ and ¾ points for the four conditions were calculated. Since the amplitude ranges on PAT extend from a maximum value to a minimum value of zero, any change in the mean amplitude value will necessarily cause a concomitant change in the amplitude range value. Thus a loud mean amplitude value implies a wider range, and a quiet mean amplitude value implies a narrower range.

(8) Changes in the tempo mean value were achieved mechanically by causing the sheet to pass through the rollers at a faster or slower speed. Two replacement drive wheels were added to the electrical motor,

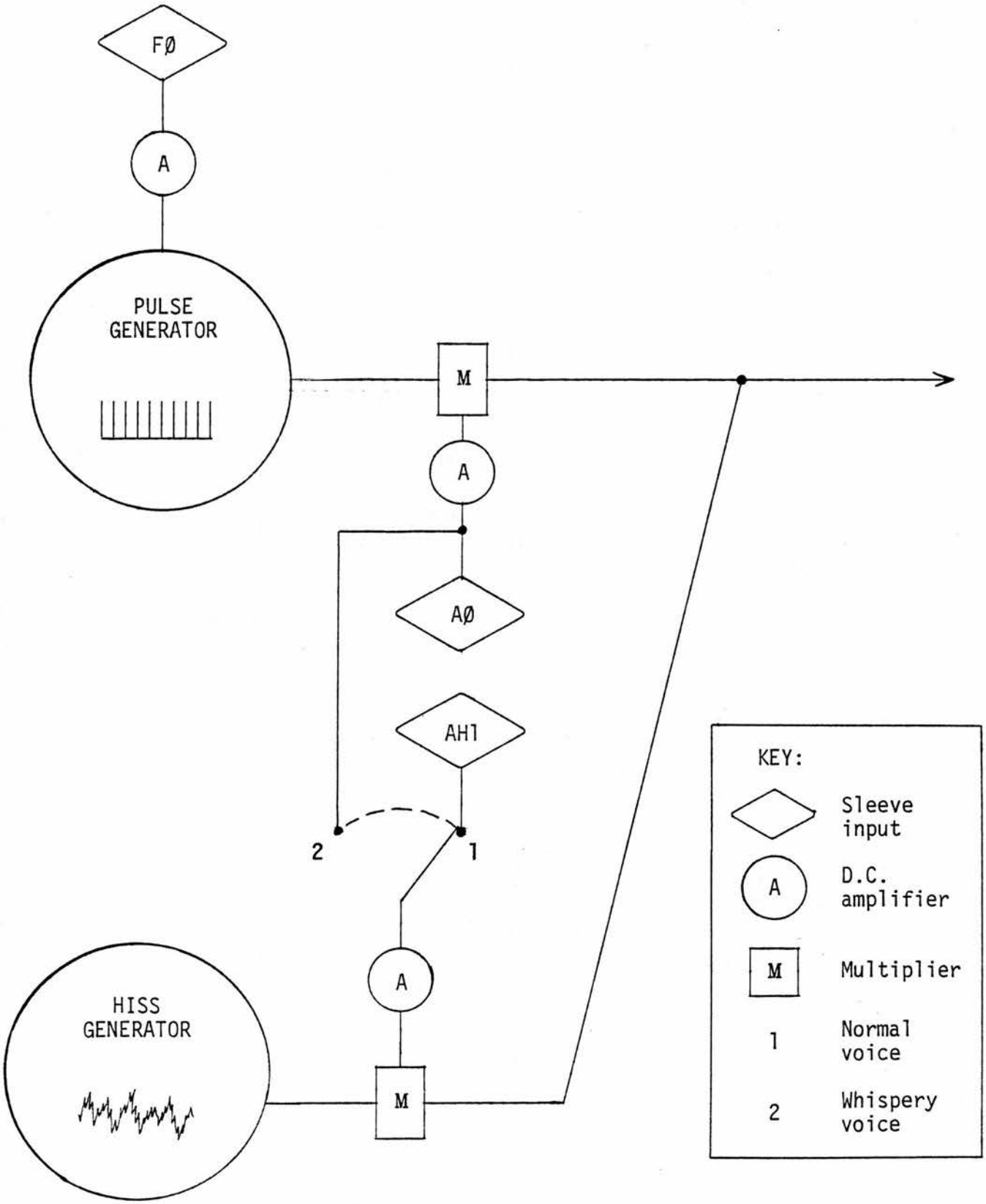


FIGURE 6.2 CIRCUIT MODIFICATIONS FOR WHISPERY VOICE

with circumferences 10% larger and 10% smaller than the control wheel. This modification produced a 10% increase or decrease in the tempo of the utterance. For convenience, the fact was ignored that a 10% increase or decrease in tempo may, in a live utterance, be accompanied by other alterations (segmental simplification, changes in intonation, rhythm, etc.).

Listeners were presented with pairs of synthetic utterances, and were required to judge the similarity or difference between the two. In each trial pair, the first utterance comprised the control sample where all the parametric values were as in the live utterance of which it was a synthetic version (with the exception of the whisperiness parameter; see above). The second utterance comprised the stimulus sample, which consisted of varying combinations of high and low values for the eight factors).

The choice of the live utterance to be analysed for the synthesis of the control sample was made in accordance with the following two criteria.

(i) The utterance needed to be long enough to allow listeners to extract the "speaker-characterising" features of the voice, but not so long that listeners might become bored or fatigued with what would be quite a demanding task. Pollack et al (1954) have shown that a sample duration of approximately one second is sufficient for listener performance to reach an optimal level of 90% for normal speech, and approximately three seconds for whispered speech. Therefore the utterance should be slightly longer than three seconds since an extremely whispery phonation type (one of the factorial values in the experiment) may have similar masking properties to full whisper.

(ii) The utterance had to be able to be synthesised without the use of PAT's hiss-through-formants parameter (AH1), since this was no longer controlled by a silver line on the sheet (see Figure 6.2).

The utterance selected was the first sentence of the Rainbow passage (Appendix 2): "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow". The present author, a standard Southern British English speaker, acted as the live speaker. The utterance was said with neutral (emotionless but not monotonous) intonation and stress contours. With two short pauses, the live utterance ran to five seconds; with the increase and decrease in tempo (the high and low values for one of the experimental factors), this figure became approximately 4.5 and 5.5 seconds respectively.

Listener homogeneity was ensured by selecting subjects from the same sex, age and dialect group. All were male speakers of standard Southern British English, 25 - 35 years of age, with no hearing disabilities. All recording and playback for this and all other experiments was performed with high-quality equipment (Revox A77 tape-recorder, Sennheiser MKH 815T microphone, Decca Deram loudspeaker, Ampex PRT tape, soundproof room).

It is impracticable to employ a full factorial design for an experiment using eight factors, since this entails the presentation to each listener of stimulus samples containing all 256 (2^8) possible combinations of high and low values of these factors. Instead a $\frac{1}{4}$ replicate (2^8 factorial in 64 units) was employed (Cochran & Cox, 1957: 287; plan 6A.16 in blocks of 16 units); four listeners were presented with 16 samples each, a number found to be very satisfactory in terms of avoiding listener fatigue. Because of the restricted nature of the design, some statistical information is lost. However, the selected design ensures that none of the lost information relates to either main effects (average deviation from the overall mean attributable to the influence of one particular factor) or to second-order interactions (interactions between two factors which either reinforce or suppress the effect of each, thereby enhancing or reducing the listener's perception of voice difference). The lost information

therefore refers to higher-order interactions (between three or more factors), whose influence is likely to be of limited importance.

In order to increase the number of listeners and the total number of responses elicited, a second replicate of the design was administered to a new group of four listeners. In this second replicate, the stimulus samples presented to listeners V, VI, VII and VIII were the same as those presented to listeners I, II, III and IV respectively, but in a different randomised trial presentation order per listener. Retrospectively speaking, it would have been preferable to use a single $\frac{1}{2}$ replicate design, thereby avoiding replication and obtaining more statistical information for the same experimental effort. However, at the time, this was not employed owing to doubts as to the immediate availability of eight homogeneous subjects.

Listeners were asked to give their responses as strokes disecting a line representing a scale from SIMILAR to DIFFERENT. For each trial, the more similar the stimulus sample seemed to the control sample, the further the stroke should be placed towards the SIMILAR end of the scale; the more different, the further towards the DIFFERENT end. The positions of the strokes were measured and the distances from the SIMILAR end were expressed as percentages of the total length of the scale, thereby converting the line in effect into a 100-point scale. (This procedure was made very simple by drawing the line 10 cm. in length, and reading the positions of the strokes in mm.).

Four practice trials preceded the test proper in order to allow the listeners to acquaint themselves with the experimental task and with the control voice. To ensure that listeners' responses to stimuli were not affected by those stimuli immediately preceding them, a short-term memory task was set between trial presentations (counting backwards in threes from a specified number, and writing these numbers

down). Each trial presentation therefore consisted of the following sequence on the experimental tape.

(1) Voice "Number A1".
(2) Control sample.
(3) One second's silence.
(4) Stimulus sample (A1).
(5) Three seconds' silence for response.
(6) Voice "Count from (specified number)".
(7) Ten seconds' silence for counting and writing down.
(8) Voice "Stop counting. Number A2".
etc.

The following are the instructions which were given to the listeners.

> Thank you for agreeing to take part in this experiment, in which I would like you to listen to some voices. They will be presented in pairs and you are required to judge on an overall impression how similar the voices are to each other, or how different. The first voice in all the pairs which you will hear is the same - only the second voice changes. All the voices are speaking the same words - "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow".
>
> The voices have all been produced electronically on a speech synthesiser. For electronic reasons, all synthetic voices sound similar to a certain degree. However, some are more similar than others, so don't be afraid to use the full range for your responses. Give your response as illustrated in the examples at the top of the first page. Imagine that the line represents the full range of similarity/difference, and put a vertical stroke to cross this line at the appropriate place. In example one [of the given illustrations], therefore, the response represents a fair degree of judged overall similarity; in example two, a fair degree of difference.

Between each of the pairs, you are required to do a small mathematical exercise. You will be given ten seconds to count backwards in threes from the number given, as also illustrated in the examples.

For each trial, therefore, you will be given

(i) two voices (the first one always being the same)

(ii) three seconds to give your response

(iii) ten seconds to count backwards from the given number, and to write these numbers down.

Give your response deliberately and do not go back and alter a response once you have been given another pair of voices.

To help you to get used to the task, there are four trial runs before the test proper begins.

The eight factors of the experiment are assigned the following letters.

Pitch mean : a	Tempo mean : e
Pitch range : b	Formant bandwidth : f
Loudness mean : c	Formant mean : g
Whisperiness : d	Formant range : h

The calculated main effect A (capitalised to denote the effect of a factor, not the factor itself) represents the average deviation from the overall mean attributable to the use of pitch mean at the high level. The negative of this value represents the average deviation attributable to the use of pitch mean at the low level. The second-order interaction AB represents the addition contribution to the mean arising from the combined effect of using both pitch mean and pitch range at the same level, high or low. The negative of this value represents the contribution from the combined effect of using one of the pair at the low level and the other at the high. For third-order interactions (e.g. ABC), the effect represents the contribution attributable to using either all three factors at the high level, or any two at the low level and the third at the high level.

One of the consequences of the lack of complete balance in the design used is that certain of the values calculated are attributable to more than one kind of effect. Thus, the value of the second-order interaction EG is equally attributable to the third-order interaction ABC. ABC is then called the alias of EG. There are many more aliases than those listed in Table 6.1, although, relating to higher-order interactions, they are likely to be of limited importance. The listed third-order interactions which are aliases are assumed to be zero.

A similar consequence of the restricted design is that the values of the third-order interactions ACD, BEF and CEH may also be attributable jointly to differences in the mean level of performance of the listeners. This is known as confounding.

The assumption underlying this design and method of analysis is that listener performance can be represented by a linear model where a listener's response for a factorial combination is expressed as the sum of the mean response value for that listener, the values of the appropriate main effects and interactions and an error factor; for example:

$$\text{Score}_{I,ab} = \mu_I + A + B + AB + \text{Error}_{I,ab}$$

The factorial effect totals are calculated by Yates' method (Cochran & Cox, 1957:158) for each replicate of the design. The results from the two replicates are pooled in the following way. The sum of, and the difference between, the totals from the two replicates for each effect are calculated. These values are grouped into six: main effects, second-order interactions and higher-order interactions for both combined effects (sums ; Table 6.1) and differences between replicates. From these values are calculated the sum of squares (S.S.) and mean square (M.S.) for each grouping (Table 6.2) by the following formulae.

$$S.S. = \frac{\Sigma\ (\text{ sum or difference })^2}{N}$$

$$M.S. = \frac{S.S.}{d.f.}$$

where N = total number of responses (here 128)
d.f. = degrees of freedom associated with each grouping

Since the mean square values for all three groupings of differences between replicates are negligibly different, the mean square of the total sum of squares for differences between replicates may be taken as a reliable datum value by which the mean square values for combined effects are divided to obtain significance (F) values. This method allows the effects "higher-order interactions (combined effects)" to be examined, although this is only partial owing to the aliasing and confounding inherent in the design. The S.S. value for "between subjects" is the sum of the S.S. entries for each replicate separately (calculated in the above manner, but on third-order interactions confounded with subjects). The critical values for F are given in Table 6.2. Thus, for pooled replicates, the effects "between subjects", "second-order interactions" and "higher-order interactions" are significant at the 5% level, and "main effects" are highly significant at the 1% level. The individual (combined) effects in Table 6.1 may therefore be examined for significance above the critical values given by the formula

$$\pm\ t\ \times\ \sqrt{N\ \times\ \text{M.S. (differences between replicates)}}$$

These critical values are shown in Table 6.1. Positive figures in excess of these critical values indicate that a change from the low to the high value of the factor produces a significant shift in listeners' responses towards the DIFFERENT end of the scale; negative figures denote a shift towards the SIMILAR end. The actual values (point estimates) of the shifts along the 100-point scale produced by the effects and interactions are reached by dividing the (combined)

		summed effect total	point estimate		summed effect total	point estimate
Main effects	A	-451 *	-3.52	E	-819 **	-6.40
	B	-285	-2.23	F	-573 **	-4.48
	C	225	1.76	G	753 **	5.88
	D	69	0.54	H	239	1.87
Second-order interactions	AB	85	0.66	CE	3	0.02
	AC	143	1.12	CF	269	2.10
	AD	-205	-1.60	CG	-377	-2.95
	AE	355	2.77	CH	-119	-0.93
	AF	17	0.13	DE	-341	-2.66
	AG	103	0.80	DF	645 **	5.04
	AH	-187	-1.46	DG	63	0.49
	BC	-311	-2.43	DH	-271	-2.12
	BD	85	0.66	EF	-405 *	-3.16
	BE	-151	-1.18	EG	-145	-1.13
	BF	3	0.02	EH	37	0.29
	BG	133	1.04	FG	-47	-0.37
	BH	167	1.30	FH	-237	-1.85
	CD	31	0.24	GH	-639 **	-4.99
Third-order interactions	BCD	127	0.99	BCH	625 **	4.88
	DEG/CFH	-87	-0.68	ACH	-93	0.73
	ADE	157	1.23	AEF	247	1.93
	BDE	79	0.62	CFG/DEH	67	0.52
	CDG/EFH	-143	-1.12	CEF/DGH	-129	-1.01
	CDE/FGH	117	0.91	BFG	141	1.10
	BDG	43	0.34	AFG	-269	-2.10
	ADG	-263	-2.05	CGH/DEF	-341	-2.66
	ACF	863 **	6.74	BEH	-119	-0.93
	BCF	357	2.79	AEH	319	2.49
	CDH/EFG	-233	-1.82	AGH	-125	-0.98
	CDF/EGH	-109	-0.85	BGH	-331	-2.59

Table 6.1. Experiment 1 : Effect totals for pooled replicates. Critical values for significance : 5% (*): 402 ; 1% (**): 535
a = pitch mean; b = pitch range; c = loudness mean; d = whisperiness; e = tempo mean; f = formant bandwidth g = formant mean; h = formant range

Table 6.2. Experiment 1 : Analysis of variance for pooled replicates

Source	d.f.	S.S.	M.S.	F	Critical values for significance 5% (*)	1% (**)
Between subjects	6	5221	870	2.63*	2.25	3.12
Combined effects						
Main effects	8	15338	1917	6.07**	2.10	2.82
Second-order interactions	28	14894	532	1.69*	1.66	2.08
Higher-order interactions	24	15853	661	2.09*	1.70	2.12
Total of combined effects	66	51305				
Differences between replicates						
Main effects	8	3567	446	1.35	2.10	2.82
Second-order interactions	28	7421	265	0.80	1.66	2.08
Higher-order interactions (error)	24	7950	331			
Total of differences	60	18940	316			
Total	126	70244				

effects totals by the number of responses (128). These are given in Table 6.1. There is thus evidence that the effects of factors e, f and g (tempo mean, formant bandwidth and formant mean) are highly significant, while that of a (pitch mean) is significant at the 5% level. Therefore, for example, stimulus voices with wide formant bandwidths are judged to be less different from the control voice than stimuli with narrow bandwidths are. It is surprising that the effect of factor h (formant range) was not found as significant as either formant bandwidth or formant mean. This may be a product of experimental error; limited justification for this view is that GH (formant mean and formant range at the same level, high or low) was one of the two second-order interactions (along with DF) found to be highly significant at the 1% level. The third-order interactions ACF and BCH are also statistically significant but this may be attributable to the experimental error deriving from the aliasing and confounding inherent in the restricted nature of the design.

6.3 EXPERIMENT 2

One of the findings of Experiment 1 was that the effect of changes in mean tempo values was highly significant. Experiment 1 was a preliminary investigation employing a restricted statistical design, and therefore it is always possible that this result may be affected by experimental error. Certainly, from the phonetic point of view, this was an unexpected finding; listeners showed surprise when informed after the experiment that the stimulus samples had all differed in tempo from the control sample. This finding may be attributable to several reasons. Firstly, a slow sample may sound more different from a normal tempo (control) sample than a fast sample does (because of this fact itself). Such an effect has at least not been found by other phoneticians, and seems an unlikely explanation. Secondly, and more plausibly, a slow tempo implies a longer sample duration (approximately 5.5 seconds at slow tempo, 5 seconds at control,

and 4.5 seconds at fast) and the listener therefore has a greater length of time in which to analyse the sample and make his decision. His responses can therefore be more thorough and discriminating (further towards the DIFFERENT end of the scale). Any tendency towards indecision (e.g. because of lack of time in a fast sample to execute exhaustive analyses) may result in a response towards the SIMILAR end. The possibility of important interactions such as this should not be dismissed too readily, and it was therefore decided to omit tempo mean as a factor for Experiment 2.

Since this experiment was seen as building upon, and seeking further evidence to support, the findings of Experiment 1, it was decided also to omit three factors which were found to be insignificant (pitch range, loudness mean and whisperiness). This left four factors: pitch mean, formant bandwidth and formant mean (which were significant) and formant range (which was surprisingly found not to be significant). This reduction in the number of factors allows a full 2^4 factorial design (Cochran & Cox, 1957:158) to be employed (entailing no confounding). The evidence of Experiment 1 suggested that there is no significant difference between listeners in terms of the values of the effects of the various factors; listeners' scores differed from expectation by a chance factor which had the same distributional properties for all listeners. In short, there is justification for the assumption of a linear model of listener performance, and, for the present full factorial design, different listeners may safely be taken as replicates. Each listener is presented with samples representing all possible combinations of high and low values for the factors (16, or 2^4). Sixteen listeners were employed, who were selected from the same sex, age and dialect group, but were not the same individual listeners, as those in Experiment 1.

The control sample and the 16 different stimulus samples used in this experiment are included on the tape which accompanies this thesis. The order of presentation of the 16 stimuli was

randomised by table for each listener. In all other respects the format of this experiment followed that employed in Experiment 1 (stimulus utterance, control voice, settings for common factors, method of presentation, etc.). Settings for the omitted factors of tempo mean, pitch range and loudness mean reverted to the control settings of Experiment 1. Discarding whisperiness as a factor, all samples (including the control) reverted to PAT's normal non-whispery phonation.

The four factors are identified by the following letters.

Formant mean:	a
Formant range:	b
Formant bandwidth:	c
Pitch mean :	d

As for the previous experiment, presence of a letter in a factorial combination denotes the high value for that factor; absence denotes the low value.

The data obtained was treated in two ways. Firstly, the scores for each sample from the 16 listeners (16 replicates) were pooled, and the same calculations as in Experiment 1 were performed on these pooled scores. These results are shown in the analysis of variance table (Table 6.3) under "combined effects". There is clear evidence in this of a difference between subjects. Therefore, the calculations were performed separately on each listener's scores (each replicate), in order to examine the nature of this difference. These results are shown in Table 6.3 under "between subjects". These two treatments can be summarised as

(i) summing listener scores for each sample, and calculating effect totals from these sums, and

(ii) calculating effect totals separately for each listener, and summing these effect totals.

For the second method, values in the analysis of variance table were obtained from the corrected sums of squares (instead of the sums of squares), derived by the formula

$$C.S.S. = \Sigma Q^2 - \frac{(\Sigma Q)^2}{N}$$

where Q = effect totals obtained by Yates' method
N = number of listeners (here 16)

The F values (between subjects) were calculated by division by the residual mean square. Since these do not exceed the 5% significance value of 2.4, we may conclude that the 16 listeners do not differ in their reactions to the different levels of the four factors and their various combinations, more than can be ascribed to chance. They do, however, differ in the average score recorded. Such findings are to be expected from an experiment where listeners are selected from a homogeneous population. There is thus justification for representing listener performance as a linear model, as above; apart from a constant difference between listeners, they may all be treated as reacting identically on average to the effects and interactions. The main effects and interactions, which may be considered constant factors, may therefore be examined for significance. Main effects A, C and D (changes to the high level in formant mean, formant bandwidth and pitch mean) are highly significant. Factor b (formant range), which was surprisingly found to be insignificant in Experiment 1, is confirmed as being of no importance by itself, although interactions AB (corresponding to GH in Experiment 1, which was found to be significant) and BC are significant here. The significance of interactions AD and BCD is less clear-cut. The point estimates for main effects and interactions are given in Table 6.3.

	Source	d.f.	S.S.	M.S.	F	summed effect total	point estimate
Between subjects	Means	15	15797	1053	3.07*		
	A	15	10334	689	2.01		
	B	15	3189	213	< 1		
	C	15	8874	592	1.72		
	D	15	3900	260	< 1		
	AB	15	4502	300	< 1		
	AC	15	7406	494	1.44		
	AD	15	2899	193	< 1		
	BC	15	1459	97	< 1		
	BD	15	3524	235	< 1		
	CD	15	5909	394	1.15		
	ABC	15	1838	123	< 1		
	ABD	15	4700	313	< 1		
	BCD	15	5274	352	1.02		
	ACD	15	5108	341	< 1		
Total between subjects		225	84712				
Combined effects	A	1	16113		46.9**	-2031	-7.93
	B	1	699		2.0	-423	-1.65
	C	1	25941		75.5**	-2577	-10.07
	D	1	6310		18.4**	-1271	-4.96
	AB	1	8801		25.6**	1501	5.86
	AC	1	433		1.3	-333	-1.30
	AD	1	1575		4.6*	-635	-2.48
	BC	1	7843		22.8**	-1417	-5.54
	BD	1	383		1.1	313	1.22
	CD	1	802		2.3	-453	-1.77
	ABC	1	32		<1	91	0.36
	ABD	1	476		1.4	349	1.36
	BCD	1	1898		5.5*	-697	-2.72
	ACD	1	579		1.7	-385	-1.50
	ABCD	1	32		<1	91	0.36
	Residual (= ABCD between subjects)	15	5152	343			
Total		255	161782				

Table 6.3. Experiment 2: Analysis of variance and effect totals for pooled replicates. Critical values for significance: (between subjects, $F_{15,15}$) 5% (*), 2.4; 1% (**), 3.5 (combined effects, $F_{1,15}$) 5% (*), 4.5; 1% (**), 8.7 a = formant mean; b = formant range; c = formant bandwidth; d = pitch mean.

6.4 DISCUSSION

The format used in the above experiments seems an extremely fruitful one. The task which listeners are required to perform is one of voice similarity judgment. It must be pointed out that there is a basic assumption underlying the applicability of the above results to speaker recognition theory; namely, that if two voices are judged to be very different in the above experiments, then they will be easy to differentiate in a speaker recognition task. Conversely, if they are judged to be very similar, the speaker recognition task will be difficult. However, this is at present no more than an assumption.

The use of synthetic speech allows the acoustic structure of the voices to be governed precisely, and a factorial design to be used.

There are three major findings to the experiments. Firstly, a significant effect was shown in changes in tempo mean. Although this may be partly explained by reference to the length of time available to the listener for analysing samples and arriving at decisions, this result suggests that tempo merits greater attention as a speaker-characterising feature than it has been paid so far.

Secondly, although the average level of performance differs from listener to listener (see Chapter 7), their reactions to factorial changes do not. In other words, there is justification for the adoption of a linear model of listener performance, as above. However, although such a conclusion may be reached on sound statistical grounds, it overlooks various associated issues. The most serious objection is that, although listeners may be instructed to judge the similarity of the samples as overall voices, they may still perform the task in a variety of slightly different ways. For example, one of the methods which one subject found he was using to perform the task

was to imagine plausible real owners of the synthetic voices and to carry out a personal comparison of these hypothesised speakers. Similarly, although the listeners were asked to judge the voices on the basis of overall impressions, several said that it was impossible to avoid a parametric approach (consciously concentrating, for example, on comparing the mean pitches of the two samples, at the probable expense of analysis of other parameters). The reasoning behind asking the listeners to perform the task on as holistic a basis as possible was as follows. All the listeners were informed in the instructions that the samples were not of live speakers, but of synthetic speech. Knowing this, they might be expected to realise that some factorial design employing a limited number of variables would be used and that the task might therefore be facilitated by focussing on those parameters in which the differences between the voices lay. It is therefore hardly avoidable that some listeners adopted this componential approach, consciously or otherwise. However, the process of recognising live speakers in everyday situations seems intuitively much more holistic, in which case the applicability of experimental results based on componential analyses to everyday speaker recognition may be limited. It is not intended to imply that listeners in experiments adopt either one approach or another. It is more likely that a mixture of strategies *is* adopted. What is important is that such differences between listeners are difficult, if not impossible, to control and that global conclusions may overlook unavoidable, but perhaps significant, differences.

Mention might also be made of a similar difference in listeners' abilities to perform the task. One subject's data had to be discarded since he found it impossible to provide scalar responses. Instead, only binary responses were given; voices were either SIMILAR or DIFFERENT, and since all stimuli differed to some extent from the control, his responses were clustered at the DIFFERENT

end of the scale. Naturally, for the rest of the listener group, the distributedness of responses along the SIMILAR-DIFFERENT scale differed from listener to listener. However, such variation may be dealt with and investigated by the statistical analysis, and in the above experiments was not too great, as shown by the justification for the adoption of a linear model of listener performance. Nevertheless, instances of inability to perform scalar judgment must be omitted for such an analysis.

To sum up, although an experimenter may instruct listeners to perform a task in a certain way, they may nevertheless reach their responses by a variety of differing strategies, which must be taken into account in analysing results.

The third finding concerned the importance of factors. Pitch mean, formant mean and formant bandwidth were found to be significant factors in both experiments. In view of the almost unanimous agreement among experimenters as to the importance of these factors (see references in Chapter 5), this does not come as a surprising result. Changes in formant range were consistently found to have an insignificant effect. This was unexpected since previous experimentation has demonstrated the overall importance of formant structure, as confirmed here by the significant results for formant mean and bandwidth. Interpretation of these parametric significances is confused, however, by two facts. Firstly, the inter-speaker differences represented by the experimental factors may derive from different kinds of source. They may result from:

(i) organic differences in the various intrinsic dimensions of the vocal apparatus,

(ii) differences in the extrinsic settings habitually adopted for normal speech,

(iii) the effects of paralinguistic adjustments,

(iv) the requirements for linguistic segmental and suprasegmental articulation.

The above are in order of decreasing time-domain and probability of influence. As is argued in section 3.4.1.1, the first two, extralinguistic, sources, being omnipresent or at least long-term features, are of the greatest realistic importance in speaker recognition, and they alone will be discussed below. However, as noted in section 1.8.1, a basic problem arises in the relating of acoustic/perceptual phenomena to articulatory correlates. It may be impossible to designate any observed feature as the result of extrinsic manipulation or as deriving instead from intrinsic, anatomical sources. Thus, for example, intrinsic and extrinsic aspects of nasality may produce perceptually identical effects on the speech signal.

Secondly, the acoustic parameters used as factors in the experiments do not stand in a strict one-to-one relationship with (intrinsic or extrinsic) articulatory features of speakers. Differences in formant mean may be associated with two long-term articulatory causes. As an organic feature, they may reflect differences in overall vocal tract length, while they may also be connected with habitual settings which affect overall vocal tract length (raised or lowered larynx, and lip protrusion).

Variation in mean pitch may result partly from organic differences in the size and elasticity of the laryngeal mechanism; extreme differences of this kind derive from differences in speaker sex, age and physique (see section 7.1). Variation in mean pitch may correlate with variation in formant mean: speakers of large physique tend to have both large vocal tract lengths and large larynges, being conducive to low formant and pitch means respectively. A low mean pitch may also represent a habitual setting, although the intrinsic organic foundation is a more influential factor in the determination of this absolute value for most speakers.

Long-term formant bandwidth reflects the damping coefficients of all the organs of the vocal apparatus. As extrinsic settings, the variations in overall tension associated with tenseness and laxness have an influence in determining formant bandwidths, as do the individual voice quality settings of nasalisation, faucalisation and pharyngalisation (see section 3.4.2.1). However, the intrinsic damping factors of all the organs of the vocal apparatus contribute to the overall coefficient and, being by definition outside the speaker's control and relatively stable, may be considered at least equally speaker-characterising as those associated with manipulable settings.

There is thus partial justification for claiming that the three factors consistently found to be significant in the voice similarity judgment experiments constitute reliable indicators of the intrinsic organic foundation of a speaker's vocal apparatus. It may also be claimed that the fourth factor, formant range, which was shown to be unimportant, differs in this respect from the other three: it is associated with less peripheral articulatory gestures and an auditory impression of laxness (this use of the term laxness refers specifically to tongue movement, not to the constellation of settings in local features subsumed by the term in Laver's (1980:155) descriptive system (see section 3.4.2.1)). Although this may be influenced to some extent by the intrinsic dimensions and musculature of the tongue and oral cavity, less peripheral articulatory gestures more probably represent a habitual tendency. The traditional notion of a tongue centring tendency may be adopted for this, implying an overall decrease in the extent of radial departures from the tongue's centre of mass.

On the basis of the above reasoning, the experimental results suggest that parameters which derive from intrinsic features of the speaker are of the greatest importance in speaker recognition. However, the relating of acoustic/perceptual features with consistent articulatory correlates is a field where further experimentation is needed.

CHAPTER 7

LISTENER-DEPENDENT FACTORS

CHAPTER 7

L I S T E N E R - D E P E N D E N T F A C T O R S

7.1 INTRODUCTION

Bricker & Pruzansky's (1976) categorisation of the speaker recognition process chain (see section 2.2) deals with those elements which may be used as experimental variables. There are five of these:

(i) the speaker ensemble,
(ii) speech material,
(iii) transmission system parameters,
(iv) listeners, and
(v) tasks and performance measure.

However, they note that very little of the work which they summarise in their review of the literature has investigated the listener-dependent factors affecting performance in speaker recognition tasks. This section argues that there is no a priori reason for this, and that the fourth category of operational element (differences in characteristics of the listeners taking part in the experiment) may be exploited as experimental variables as easily as the first category (speaker differences).

It is easier to discuss speaker-dependent factors before listener-dependent ones, since the richness of research into speaker differences allows the discussion to be illustrated with reported correlations (Scherer & Giles, 1979). I shall take as the parameter for illustration habitual pitch range which, as was shown in section 1.8.1, has been discussed widely by phoneticians and others. The speaker-dependent factors which are the main determinants of the absolute width and height of a person's habitual pitch range are the following.

(1) Biological differences in sex, age and physique. The average length of male vocal cords is about one-third greater than that of females' (roughly 23 mm. as against 17 mm.; Kaplan, 1960:128). This partly accounts for the lower average pitch of male voices (Smith, 1979). Jones (1964:276) and Schubiger (1958:4, FN) both claim that the female pitch range is somewhat narrower than the male.

Calcification and ossification of the laryngeal cartilages starts at about 25 years of age, so that, by the age of 65 years, the cartilages are entirely bone and the surrounding muscles are less elastic (Kaplan, 1960:144; Ferreri, 1959; Fyfe & Naylor, 1958). These senescent changes give rise to a decrease in both the width and height of the speaker's habitual (and extreme) pitch range (summarised in Helfrich, 1979). Physiological changes occurring at puberty also affect these parameters.

(2) Permanent or temporary biological differences in health (Laver & Trudgill, 1979). O'Connor (1973:297) reports that it has been claimed that epileptics and stutterers characteristically use an abnormally restricted pitch range.

(3) Psychological differences in the speaker's attitude (Scherer, 1979). Christophersen (1956:181) notes that very animated speech will use a wider pitch range than when the speaker is tired or uninterested. Similarly, Jones (1964:275) mentions that pitch range is wider when the speaker is excited than when he is in a serious mood.

(4) Differences in native language or dialect. Abercrombie (1967:99) states that native speakers of Tlingit use a markedly lower pitch range than speakers of English. However, Delattre (1965: 25) claims that language differences are not very influential and that pitch range is dependent 'on the subject rather than on the

language'. (By 'subject' Delattre means 'subject-matter', not 'speaker').

(5) Idiosyncratic differences (see section 1.3).

The above five categories are not homogeneous, in that they overlook differences such as the dichotomies of inter- and intra-speaker variation (see section 1.8.2), and intrinsic and extrinsic factors (see section 1.8.1). However, they are sufficiently sophisticated for present purposes.

In their capacity as features which allow the listener to infer characteristics of the speaker from his voice, features of a person's speech determined by the above categories of factors may be termed indexical, and indeed the above divisions correspond closely to those of Laver's (1968) classification of indexical information (see section 1.3), with the exception of idiosyncratic factors.

The results of the experiments reported in Chapter 6 show that the variation in listener performance is quite large. It was found that, although listeners reacted identically on average to changes in the experimental factors used, there was a significant difference in the average score recorded. However, no writer has investigated the variation in listener performance, or even discussed the factors which contribute to it, apart from Williams (1964) and Doehring & Ross (1972), who demonstrated the importance of the amount of training given to listeners for tasks involving familiarised reference speakers (section 4.3), and of whether listeners are informed of the correctness of their responses during this training.

The following may be proposed as the main determinants of listener differences.

(1) Biological differences in sex and age. It has been found that females achieve significantly better performance than males in verbal and auditory tasks (Maccoby & Jacklin, 1974; Mazanec & McCall, 1975; McGuinness, 1976), and findings consistent with this superiority might therefore be expected in speaker recognition tasks. Similarly, one might expect older listeners' performance on average to be inferior owing to presbycusis (the physiological and mental deterioration in hearing accompanying old age; Melrose et al, 1963) and the associated loss of discrimination ability (Gaeth, 1948). However, there are large inter-personal differences in the presbycusic process.

(2) Permanent or temporary biological differences in health. Loss of hearing due to ill-health will obviously play an important part in listeners' abilities. One would also expect blind listeners, whose major contact with the world is through sound, to perform better on average than sighted listeners.

(3) Psychological differences in the listener's attitude. If a listener is tired or bored, he is likely to perform worse than normal.

(4) Differences in native language or dialect. Two opposing viewpoints might be adopted:

(a) Listeners are accustomed to recognising speakers of their own native language from everyday situations, and their perception has therefore become well-tuned to differences in familiar stress and intonation patterns, vowel qualities, etc. They should therefore perform more successfully with speakers of their native language.

(b) Since listeners are unable to recognise the content of utterances in a foreign language (which they do not speak), their attention is not distracted and they can devote all their concentration to the speaker's voice. They should therefore perform more successfully with speakers of a foreign language.

(5) Idiosyncratic differences (see section 1.3).

It is an immediate observation that the above five categories of listener differences are very similar in nature to those for speaker differences. In formats where the variation between speakers or listeners is not being employed as an experimental variable, the effect of these factors may be easily avoided, with the exception of idiosyncratic factors. Thus speakers and listeners are selected from the same sex and age groups (usually male undergraduate students, 18-25 years old), of normal health and with no hearing disabilities, and speakers of the same native language (almost always American or British English). Recording and test sessions are limited to a reasonable duration to avoid changes in the speaker's or listener's attitude due to fatigue or boredom (see section 4.3). Whilst such a procedure avoids the influence of the group factors of sex, age, etc., the interpreted results of such experiments are relevant specifically only to the speaker/listener group selected. Whether the results may be projected onto different speaker/listener groups depends upon the actual influence of each of the factors of sex, age, etc. - a question which has been deliberately bypassed by the above procedure.

Since idiosyncratic factors, by the definition given in section 1.3, do not correlate with group factors such as sex, age, etc., then it is impossible experimentally to eliminate idiosyncratic inter- and intra-personal variation. While this may be achieved in relation to group factors by selecting speakers or listeners of the same sex, age group, etc., the influence of idiosyncratic factors may only be minimised, not totally avoided. Hecker (1971) summarises the procedure generally adopted by experimenters in relation to idiosyncratic listener differences.

> 'The ability to recognise speakers varies from listener to listener. If only a few listeners participate in a given test, the average score

> will depend greatly on their particular abilities. As the number of listeners is increased, this dependence is reduced, and the test results become more useful for making comparisons between experiments using different listeners. Thus, the size of the listener group can have an appreciable effect on performance. Although this effect has not been studied specifically, most investigators have used at least ten listeners. An even larger listener group may be necessary in order to obtain generally meaningful results.'
>
> (Hecker, 1971:27)

The influence of idiosyncratic factors may therefore be minimised by employing large speaker/listener groups, but not totally avoided unless, to take the argument to its logical extreme, the whole population in question acts as speakers/listeners.

If our objective in studying speaker recognition is to gain a greater understanding of how listeners recognise speakers or, more broadly, of the human productive and perceptual processes in general, then listener differences carry equally great importance as the speaker-characterising features which have been the object of investigation both in the experiments reported in Chapter 6 and in most of the speaker recognition experimentation literature. In the past, phonetic research has concentrated on articulatory rather than on acoustic or auditory description, and study of the process of speaker recognition has therefore likewise focussed on those characteristics in relation to which speakers differ, at the expense of study of the corresponding characteristics for listeners. The study of these listener differences is in principle no more difficult a field of research than that of speaker differences.

The experiments reported in this chapter investigate the importance of listener differences. They test whether there is any difference in listeners' performance if they are required to recognise speakers of a language foreign to them instead of their native language

(which is what experimental formats so far have investigated). This factor (language differences) was chosen in preference to the other four because of the following considerations:

(i) few people would argue with the view that health factors such as hearing loss have an obvious and great effect,

(ii) idiosyncratic factors are, by definition, impossible to correlate, and

(iii) of the remaining three kinds of differences, language differences might be expected to be the most important, and to have the greatest implications for wider fields of linguistic theory.

The experiments must be described, however, as only a preliminary and relatively informal attempt to investigate the listener-dependent effects in speaker recognition.

7.2 EXPERIMENT 1

Eight native speakers of English were recorded reading a standard passage (the Rainbow passage; Appendix 2). There were accent differences between the speakers: four had Edinburgh accents, two had other Scottish accents and two had standard Southern English accents. From these recordings the test tape was produced, which consisted of 40 samples. Each sample consisted of two recordings in a simultaneous presentation format, i.e. one recording was superimposed on the other, by double-tracking on a tape-recorder (see section 4.3.1). In 22 of the samples, the recordings were of different speakers (e.g. speaker A + speaker B); in the remaining 18 samples, of the same speaker (e.g. speaker A + speaker A).

Some of the samples were then modified acoustically. In all there were six conditions:

(1) normal, unmodified.

(2) the relative amplitude of the two recordings was altered (R.A.).

(3) fricative noise, generated by the P.A.T. synthesiser and high-pass filtered at 4 KHz, was added (NOISE).

(4) a high-pass filter set at 440 Hz (HP).

(5) a low-pass filter set at 4560 Hz (LP4).

(6) a low-pass filter set at 3040 Hz (LP3).

In addition, some of the filtering conditions were combined. The samples were then electronically gated, so that the duration of each was 2.5 seconds. This was found to be a satisfactory duration since it allowed listeners to distinguish recordings, but not to pick out and consistently concentrate on individual words and phrases.

In order to allow the effect of each modification to be measured, the two-voice samples were prepared in pairs. The combination of (same or different) speakers in both samples was held constant, while one of the samples underwent an acoustic modification and the other did not. Thus a pair of two-voice samples might be

(1) speaker A + speaker B (normal).

(2) speaker A + speaker B (high-pass filtered)

The samples were then ordered within the following restrictions:

(i) the two samples of each pair were kept as far apart as possible within the following restrictions.

(ii) no consecutive samples involved the same speaker, as a component of either a same or different combination.

(iii) a randomised sequence of "same"/"different" correct responses was maintained.

Twenty listeners, ten native English speakers and ten non-native, were asked to listen to the samples and decide whether the simultaneously presented recordings were of the "same" or "different" speakers. The word "recordings" was used in the instructions in preference to "speakers" or "voices" as this was found to be the least ambiguous and the most helpful for the explanation of the simultaneous

presentation format. Four practice samples, which were not marked and to which listeners were told the correct alternatives after they had given their responses, preceded the test proper. This ensured that listeners understood the task; the non-native listeners' lack of competence in English made the task especially difficult for them to understand. Listeners were pressed not to respond "don't know" unless it would have been a complete guess to answer "same" or "different".

Listeners' success rates were high, with a range of scores from 19 to 32 samples correct out of 40, and a mean of 27 samples (68%). There was no difference between mean scores for native and non-native listeners.

The effect of each acoustic modification is shown in Figure 7.1. The graphs on the right-hand side of that diagram show the scores of correct responses for modified samples plotted as percentages of the scores for the corresponding unmodified samples. None of these effects approach statistical significance at the 5% level.

Different-accent pairings (e.g. an Edinburgh accent + a S. English accent) proved to be easier (mean score 82% correct) than both same-accent pairings (49%) and same-speaker samples (68%). This is highly significant at the 1% level. This is an intuitive expectation, since it is a prerequisite of accent classification that all speakers of one accent share features which speakers of other accents do not. However, an insignificant difference was found between native and non-native listener scores for different-accent pairings.

7.3 EXPERIMENT 2

Six native speakers of Arabic were recorded reading firstly a passage in Classical Arabic and secondly a passage in English. All six were teachers of English and their English reading style was quite

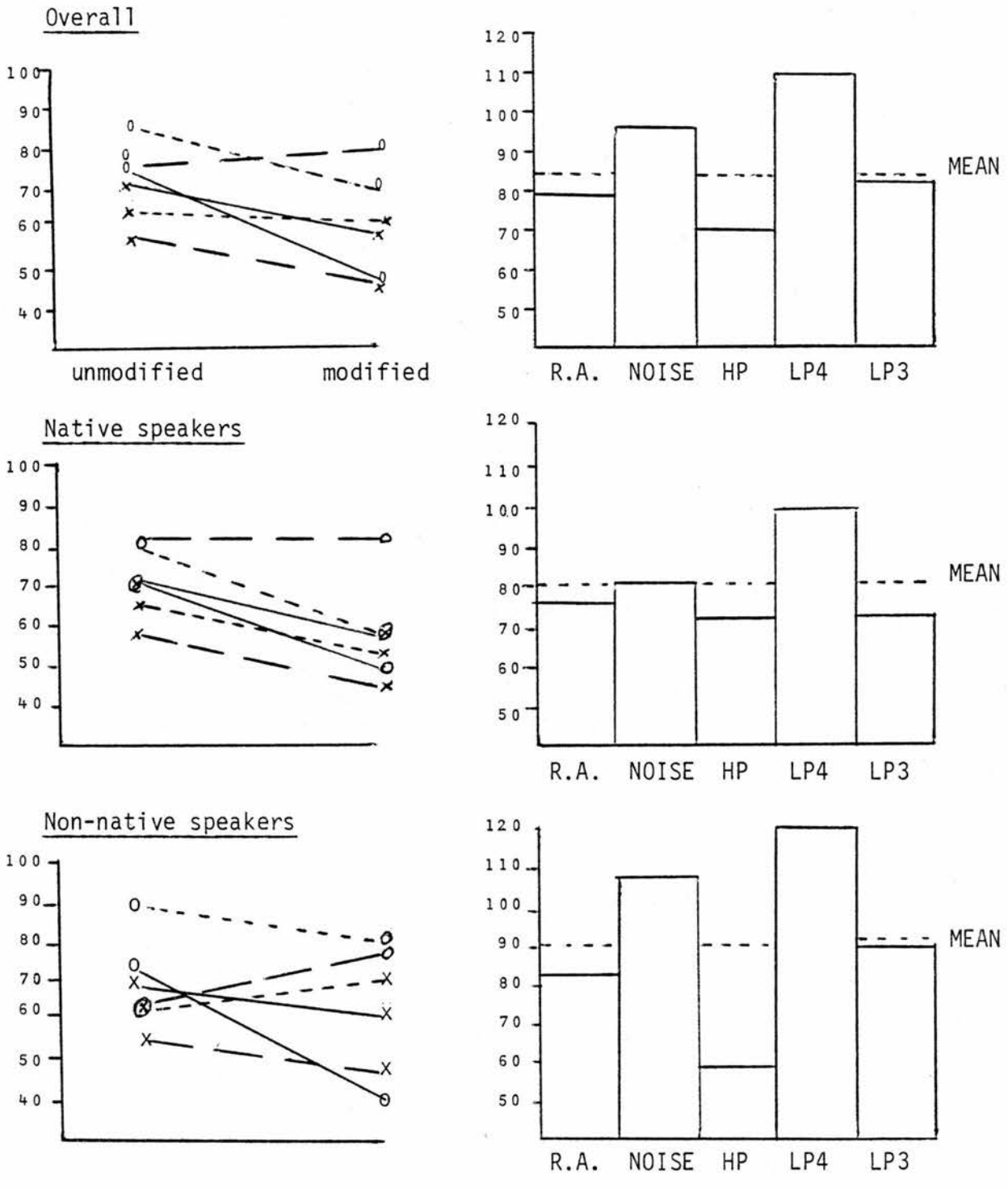


Figure 7.1 Results of Experiment 1. Left hand graphs show correct response percentages for unmodified and modified samples. Right-hand graphs show correct responses for modified samples plotted as percentages of correct responses for unmodified samples. Key to left-hand graph

x——x mean
x— —x R.A.
x— — —x noise
o——o HP
o— —o LP4
o— — —o LP3

fluent. From these recordings the test tape was produced, which consisted of 36 samples, each consisting of two recordings in a simultaneous presentation format. In 18 of the samples, the pairs of recordings were in Arabic; in the other 18, in English. No Arabic-English pairings were involved. Samples were gated to 2.5 seconds duration, as in the previous experiment, and ordered within the following restrictions:

(i) no consecutive samples involved the same speaker as a component in either a same or different combination.

(ii) a randomised sequence of (a) "same"/ "different" correct responses, and (b) Arabic/English stimulus language was maintained.

Sixty listeners, 20 native English speakers, 20 native Arabic speakers and 20 speakers of other native languages, were required to listen to the samples and decide whether the simultaneously presented recordings were of the "same" or "different" speakers. As in the previous experiment, listeners were pressed not to respond "don't know" unless it would have been a complete guess to respond "same" or "different". Four practice samples preceded the test proper.

A 3 x 2 matrix of correct responses in terms of the native language of the listeners and the language of the samples is given in Figure 7.2.

		Listeners			
		English	Arabic	Other	Mean
Samples	English	72%	65%	65%	67%
	Arabic	78%	75%	72%	75%
	Mean	75%	70%	68%	71%

Figure 7.2 Results of Experiment 2. Percentages represent correct responses.

Three main observations may be made.

(i) Listener mean success rates were high (26 samples correct out of 36, or 71%), and comparable with those achieved in Experiment 1.

(ii) The difference found in the mean success rates for English, Arabic and "other" listeners (75%, 70% and 68% respectively) is not significant.

(iii) Arabic samples were performed more successfully than English by all groups of listeners (75% against 67% respectively). This is significant at the 5% level.

Because of the limitations imposed by the availability of listeners, some bias was introduced into the experiment. Namely, both the Arabic and the "other" listeners had a limited knowledge of English, since they were students attending courses in Edinburgh. Thus the Arabic listeners knew Arabic and limited English; the "other" listeners knew no Arabic but limited English; the English listeners knew English but no Arabic.

The results of the experiment support neither of the two hypotheses proposed in section 7.1 (either (a) that listeners would perform more successfully with samples in their native language, or (b) that they would perform better with samples in a foreign language). Whilst success rates for English and "other" listeners were better with Arabic samples (in support of the second hypothesis), the Arabic listeners' performance was also better for Arabic samples (in support of the first hypothesis). Indeed, as far as the nature of the samples is concerned, one can only note that Arabic samples were performed significantly more successfully than English, although it is difficult to account for this with any phonetic justification.

7.4 CONCLUSION

No firm statements may be made on the basis of the data collected in the two experiments. This failure of the experiments to highlight any consistent listener-dependent effects in speaker recognition may be attributed to three factors:

(i) The simultaneous presentation format used in both the experiments may serve only to exaggerate any idiosyncratic listener differences. As was discussed in section 4.3.1, this format is unnatural in the sense that it cannot occur in the real world while preserving the possibility of "same" or "different" alternatives; in the real world, two simultaneously heard voices must belong to different speakers. Certain listeners may therefore have reacted to the unnaturalness of the task while others may have adapted more successfully to its novelty. It was noted in section 4.3.1 that large inter-subject differences in performance may result from a difference in the strategy adopted to perform the simultaneous presentation task.

(ii) It might be argued that the English listeners in both experiments were more used to taking part in such experiments. However, in neither experiment was there a significant difference between the performance of the different groups of listeners. Similarly, all the listeners were university students, all of whom were expected to have had experience in performing similar tasks.

(iii) It may be that, although intuitively both speaker-dependent and listener-dependent factors are important in speaker recognition, the latter show much greater inter-person variability and inconsistency. That is, listeners may vary more widely in perceiving sounds than speakers do in producing them. It may also be that this inter-listener variation has a greater proportion of unaccountable, idiosyncratic variation than the corresponding inter-speaker variation.

Whether any of the above arguments holds or not, it is clear that any future experiments on the listener-dependent effects in speaker

recognition will require larger corpora of listeners and greater statistical formality than those reported above. However, there is no principled reason why the investigation of listener differences should be more difficult than of speaker differences, nor why the results of such an investigation should be less enlightening for the field of speaker recognition.

CHAPTER 8

C O N C L U S I O N

CHAPTER 8

C O N C L U S I O N

This thesis has dealt with speaker recognition, broadly defined as the human ability to decide on a speaker's identity from hearing a sample of his speech. The coverage given to the field here has been by no means complete. Not only would this be impossible in a thesis of this size, but also it would be superfluous for certain aspects of the subject which have been handled adequately elsewhere and are therefore given only a cursory treatment here. Attention has thus been focussed selectively on those aspects which seem to merit greater attention than they have been afforded in the literature, in addition to those aspects for which the attention has, to my mind, been somewhat misdirected.

Study of speaker recognition has been carried out for some years and is a topic of current interest for researchers in several different fields. In general phonetics, it follows the modern interest in the description of the function of non-linguistic elements of speech in the wide field of communication. The practical applications of speaker recognition have an obvious and great importance. With the ever-increasing relevance and sophistication of speech recognition and speech synthesis systems due to the ever greater demands imposed on telecommunications, one can expect automatic speaker recognition to be a subject of ongoing research and interest for many years to come. Whether as an implementable device or as a theoretical explanatory representation, any advanced speech recognition model must include an equally advanced speaker recognition component as a fundamental constituent.

In any field dealing with the perception of speech, the necessary involvement of a voice as the initiator of that speech requires that some form of speaker normalisation take place by the discounting of the solely speaker-characterising features of the speech signal. That is, a process on the same principles as those underlying speaker recognition occurs. The example of this given in section 1.7.1 was child language acquisition, although the argument applies to other processes entailing speech perception, such as foreign language learning. If the results of the experiments reported in chapter 6 can be interpreted as evidence that the most important speaker-characterising features are those which derive from intrinsic factors of the speaker, then it is no hindrance to the speaker normalisation component of these processes that there exist such extrinsic registers as so-called baby-talk and foreigner-talk. Naturally, these registers may not help these learners in acquiring the normal adult form of the language; but, for infants, speaker recognition is an equally important faculty to be acquired.

One of the most serious criticisms to be made of the speaker recognition literature is that few writers have addressed themselves at anything greater than a superficial level to defining speaker recognition and to making explicit a theoretical framework within which experiments may be performed and results interpreted. Chapter 1 of this thesis attempts to define the scope of speaker recognition. The most important amendment to the view generally expressed in the literature is the emphasis laid here on the study of auditory speaker recognition in everyday situations. Criticism is levelled at researchers' concentration on the experimental investigation of speaker recognition, at the expense of consideration of (i) how humans habitually recognise speakers in normal everyday life, (ii) the relationship between the experimental and the everyday situations, and (iii) the consequent relevance of the wealth of experimental findings to real life speaker recognition, communication and general phonetic theory.

The paradigm of speaker recognition experimentation is that the process involves a pattern-matching technique; that is, a representation of the sample uttered by the voice to be recognised (the stimulus voice pattern) is compared with some form of internalised representation of voices previously heard by the listener (reference voice patterns), and a decision is reached on the basis of the degree of similarity between the two. The majority of research has been aimed at specifying those speaker-characterising features which compose stimulus and reference voice patterns. This emphasis is justified since the acoustic/perceptual attributes of voices are likely to have by far the greatest influence on listener performance in both the experimental and the everyday situations. The presupposition underlying most experimental procedure is that a correlation can or cannot be found between the manipulation of a controllable factor (the independent variable) and the consequent variation observed in another measurable factor (the dependent variable). However, in section 1.3, a special set of factors is discussed which cannot be investigated by such a paradigm. Idiosyncratic factors are claimed to be definable as being impossible to correlate with any group factors (sex, age, regional origin, social status, etc.), and are therefore not totally controllable experimentally, although their effect may be diminished by the use of large speaker/listener groups. It is argued that the case of idiosyncrasies is especially relevant to speaker recognition, since they are highly speaker-characteristic (even if their occurrence may be statistically rare). This is naturally an ominous conclusion for speaker recognition experimenters.

The case of idiosyncratic features may be considered an extreme manifestation of the basic distinction which has been stressed throughout the thesis and underlies most of the work reported here; namely, that speaker recognition can be performed by reference to a large number of potentially usable parameters, but that a subset of these parameters is habitually used in everyday life. In chapter 3, a brief examination was presented of the

more obvious parameters of the speech signal by which speakers may be recognised. The main conclusion which may be drawn from this is that the number of potentially usable parameters in the speech signal must be immense. This immensity is appreciated when it is remembered that chapter 3 examines only phonetic (voice quality and voice dynamic) features of speech, and does not deal with other levels (segmental phonetic, syntactic, semantic, lexical), which are mentioned briefly in section 1.6. The number of potentially usable features may be found to correspond, more or less, to all the inter-speaker variation present in speech. In short, any form of inter-speaker variation may be used in speaker recognition. Supporting this view is the fact that the speaker recognition parameters discussed in chapter 3 may be used for the practical description of voices (section 3.6). What is important therefore is not the absolute potentiality of features as speaker recognition parameters, but the relative strengths of speaker-characterising features (section 3.3), which determine their relative weighting in auditory speaker recognition. The review of the literature in chapter 5 was not comprehensive, since it is argued that almost all of the speaker recognition literature deals not with how listeners are recognised, but with how they can be recognised. It was towards a specification of the relative importance of habitually used parameters that the experiments of chapter 6 were directed. The most important habitually used parameters are referred to as first-order parameters, implying that they are the first in relation to which the reference-stimulus comparison is made. Although it must be pointed out that these experiments constitute only a preliminary investigation of a still very rich and fruitful area of study, the consistency and statistical reliability of the results suggest that further research based on the findings of those experiments would prove profitable.

A biased view would be obtained from an examination of an exhaustive review of the overall speaker recognition literature, since a very large amount of research has been devoted to automatic

speaker recognition by machine. This is hardly surprising when one considers the motives behind the sponsorship of speaker recognition research (summarised by Bricker & Pruzansky, 1976). Institutions interested in the development of speech communication systems or limited access security systems have financed automatic speaker recognition work. The interest in the use of spectrograms as legal evidence motivated American legal and police departments to support study of the reliability of "voiceprints". In contrast, research into auditory speaker recognition has not received comparable support. This is an unfortunate situation because parallels and implications cannot be reliably drawn from the relative wealth of study into speaker recognition by machine and by spectrogram, for auditory speaker recognition; as discussed in sections 4.2 and 5.3, there are large fundamental differences between the processes.

The field of auditory speaker recognition is not an easy one to study.

> 'A possible limitation of this method is that it is entirely subjective. No matter how accurate and reliable listeners may be, they are unable to explain the criteria underlying their decisions.'
>
> (Hecker, 1971:2)

Just as there are different tasks to be performed, so there are differences in the methods which listeners can employ to perform those tasks. An obvious example of this (see section 4.4) is the ABX format, in which the listener is first presented with two reference voices (A and B), followed by the stimulus voice (X), which corresponds to either A or B. Listeners may perform this as a two-reference task (comparing X with both A and B) or reduce it to a one-reference differentiation task (ignoring A and comparing X only with B - if they are sufficiently similar, B is given as the response; if not, then A). Certain similar limitations were discussed in section 6.4 on the interpretation of the experimental results of chapter 6,

dependent upon the range of possible methods available to the listener to perform the task. The necessary involvement of the subjective perceptual processes of the human listener causes research into auditory speaker recognition to be a difficult field. It may be difficult for the experimenter to know whether alternative strategies exist for the performance of a task, or, more importantly, to discover whether a certain subject has performed the task in one way or another, and thus to be able to interpret results accordingly.

Of the five manipulable elements (speakers, speech material, transmission parameters, listeners, tasks) of Bricker & Pruzansky's categorisation of the speaker recognition process (see section 2.2), the last two have been largely ignored by writers. For this very reason, coverage of these two elements in this thesis has been extensive. Both are influential and important for a full understanding of the auditory speaker recognition process, even if their influence is smaller than that of the other three elements. The attention given by writers to the category of listeners has been cursory, and the consideration afforded it here constitutes no more than a survey of the more obviously relevant listener-dependent factors. However, most writers are ready to acknowledge the degree of the effect on speaker recognition of factors such as the size and homogeneity of the listener group, the amount of training given to listeners, and their experience of speaker recognition tasks. For example, Tosi (1979) realistically predicts that, despite the ever-increasing sophistication of computer techniques, automatic speaker recognition will not obviate the need for human operators in legal applications, and control over their training and experience will be required.

> 'After these preparatory operations are completed [noise elimination, temporal segmentation, filtering, etc. of the speech signal], the practitioners of the future would decide by perceptual means whether or not the samples can be submitted to computer analysis and, if so, what kind of algorithm

> or combination of algorithms would be the optimal ones in each particular case. Still, the examiner would have the last word for interpretation of computer results. ... It is quite obvious then that the practitioner of the future would be closer to the present type of examiner certified by the I.A.V.I. [International Association of Voice Identification] than to a computer operator. ... These examiners have to be trained, tested, and certified in order that the quality of voice examinations is properly maintained.'
>
> (Tosi, 1979:148-9)

The culmination of the discussion of experimental and real world tasks is the model of the speaker recognition process, represented in Figures 4.12 - 4.15. The model is formal and thereby relatively explicit, and summarises and justifies the difference in classification of the different kinds of task. The differences between tasks are defined by reference to two characteristics: (i) whether one or more than one reference voices are to be compared with the stimulus, and (ii) whether the task involves short-term memory alone or long-term memory. Differences in these characteristics have significant consequences on the manner in which the task is performed. For example, in multiple comparison tasks, where more than one reference voice is involved, the selection of probable candidates from the reference population is seen as a search among the references guided by features of the stimulus sample. For single comparison tasks, involving only one reference, the converse process is entailed; a confirmatory search is made in the stimulus sample, guided by features of the stored reference voice pattern.

Such dichotomies are rigorously applicable to experimental formats. However, certain flexibility of interpretation is required when applying such a formal taxonomy to real life possibilities. Most importantly, the distinction between single and multiple comparison tasks rests largely on the strength of the listener's expectations as to whether the situational probability

of one particular reference speaker corresponding to the stimulus is much greater than that of any other reference(s), or whether a number of reference speakers are equally possible. Any such categorisation is difficult to apply to real world situations since expectations are inherently difficult to quantify.

There are further features of everyday situations which limit the relevance of previous work.

> 'Difficult though it may be to work with, spontaneous speech deserves more attention than it has received from both listener and machine recognition researchers.'
>
> (Bricker & Pruzansky, 1976:322)

Speaker recognition in real life deals almost exclusively with spontaneous conversation, and therefore the findings of experiments using strictly controlled speech material may have limited application to everyday life.

Two experimental formats have been singled out for special examination in this thesis, again for the very reason that they have not been widely discussed in the literature. The first is the simultaneous presentation (differentiation) task, in which the two voices are presented one superimposed upon the other (section 4.3.1). The second is the open identification task, where the stimulus sample may or may not have been uttered by one of the speakers in the reference population (section 4.8). However, the latter task is of much greater general importance to speaker recognition than the former. Simultaneous presentation is impossible in the real world - two simultaneously heard voices must, in real life, belong to two different speakers. In addition, widely differing strategies may be adopted to perform this task. In contrast, open identification is probably the most common everyday task (see section 4.5), and it is not obvious that alternative strategies exist for the listener. However, as an experimental task, it is less easy to score than other formats.

Experimental considerations may be interpreted as being a prime influence on the limited view of speaker recognition presented in the literature. This concentration has resulted in a lack of consideration of the general framework of speaker recognition, and divergent views on the applicability of experimental results to everyday speaker recognition. Without such a theoretical framework, it is often difficult to interpret experimental results meaningfully. It is hoped that the theoretical discussions contained in this thesis might provide such a framework, or at least provoke consideration of such issues by workers in the field.

APPENDICES

APPENDIX 1

'COGNITIVE IMPLICATIONS OF LABELS FOR VOICES' (Brown, forthcoming)

ABSTRACT

An experiment is reported which investigates the acceptability for English speakers of various syntactic constructions in conjunction with labels for voices (Laver, 1974). A regularity is observed in the co-occurrence possibilities; subjects differentiate between labels for temporal aspects of speech (tempo, continuity and rhythmicality) and for non-temporal (pitch, loudness and quality) in terms of the constructions which may accompany them. No purely syntactic motivation is found for this regularity and it is suggested instead that it is indicative of a basic phonetic difference between cognitively governed temporal parameters in speech and physiologically determined non-temporal parameters. This view is supported by examples from other languages.

INTRODUCTION

Laver (1974) deals with the wide and subtle vocabulary which we use, as laymen, for talking about speech. In particular, the kind of lay expression which he examines is 'the formula "a ... voice",

where the adjective preceding "voice" might be one of literally hundreds of labels on the pattern of "dulcet, venomous, hollow, educated, ..." and so forth' (p.62). He does not attempt any exhaustive specification of the phonetic correlates of such labels, not least because any one label may imply different things to different people. Instead he directs his attention towards categorising labels in terms of the aspects of the semiotic process of speech to which the labels refer.

> 'The first and most fundamental semiotic distinction between different types of labels for voices concerns whether the label refers to the sounds produced by the speaker or to the characteristics of the speaker producing the sounds. It will be convenient to call the first sort of label a "descriptive" label and the second an "indexical" label.
>
> Descriptive labels themselves fall into two broad categories "impressionistic" labels and "phonetic" labels. By impressionistic labels I mean labels which need an audible demonstration of the type of voice referred to before the listener or reader can construct an accurate interpretation of the label. ... Phonetic labels, on the other hand, are part of established phonetic vocabulary, and as such have exact and agreed definitions subscribed to by all trained phoneticians (in the ideal situation at least).' (Laver, 1974:63)

Another drawback to any attempt at a specification of the phonetic correlates of impressionistic and indexical labels is that, by virtue of their very nature as phrases in lay usage, they may be ambiguous or vague. It is convenient in considering the phonetic referents of labels to adopt a categorisation which derives from Sapir (1927). The medium of speech is seen as consisting of three

strands - segmental features, features of voice quality, and features of voice dynamics. Labels may refer differently to these three strands, or ambiguously to the parameters which compose the strands. Thus

> 'it is ... segmental features that are in question when we say that someone has "a clear voice" (this may not be immediately apparent, but it is the way segmental features in English are treated that makes for clarity of utterance); but "a hoarse voice", more obviously, is connected with features of voice quality, and so usually (though again this is perhaps less obvious) is "a pleasant voice"; and finally "a loud voice" and "a low voice" clearly refer to features of voice dynamics (though the latter expression is itself ambiguous, since it may mean either low in pitch or low in volume).' (Abercrombie, 1967:90)

All labels may refer either to permanent features of a speaker's voice or to modifications to his normal voice on one particular occasion. The former may reflect both the uncontrollable (though stable, habitual) features of the speaker's vocal apparatus. The latter, however, may only result from the manipulation of controllable features. In communicative terms, the former are extralinguistic (conveying no systematically encoded information), while the latter are paralinguistic (conveying nuances of mood, etc.). This difference is illustrated by the following examples.

1. My mother-in-law has a loud voice.
2. "What the hell do you think you're doing?" he asked in a loud voice.

Both Laver and Abercrombie consider only the frame 'a ___ voice'. However, this frame is inappropriate with certain labels when they are used in the permanent, extralinguistic sense (as in 1 above). Those labels which do not fit in this context all refer to temporal aspects of speech, such as tempo (the speed of utterance) and continuity (the interplay of speech and pauses or hesitations). Thus the following (a) sentences are at least less acceptable than the corresponding (b) examples.

3. (a) My hairdresser has a slow voice.
   (b) My hairdresser is a slow speaker.

4. (a) That salesman has a fluent voice.
   (b) That salesman is a fluent speaker.

This observation does not hold for labels used in the paralinguistic context. Thus the following examples are acceptable.

5. (i) "I'm afraid your father has died," said the doctor in a slow voice.
   (ii) The doctor spoke in a slow voice as he informed him of his father's death.

6. The student spoke in a fluent voice as he talked about his research.

Similarly, the observation does not hold for labels referring to non-temporal aspects of speech, such as pitch and loudness, in either the extralinguistic or paralinguistic context. 7(a) and 8(a) are thus acceptable, while the (b) equivalents are not.

7. (a) That dwarf has a high voice.
   (b) *That dwarf is a high speaker.

8. (a) (i) "Careful with that axe!" he warned them in a high voice.
   (ii) He spoke in a high voice as he warned them of the danger.
   (b) *He was a high speaker as he warned them of the danger.

The acceptable possibilities described above (which reflect the intuitions of the writer as a native speaker of English) can be summarised in the following diagram.

	Temporal	Non-Temporal
Extralinguistic	a ___ speaker	a ___ voice
Paralinguistic	a ___ voice	a ___ voice

Another way in which the adjective labels may be used is in adverbial constructions, generally in English by the addition of the -ly suffix. This applies to labels referring to both temporal and non-temporal aspects of speech, in both the extralinguistic and paralinguistic contexts. Thus all the following are acceptable sentences of English.

9. My hairdresser speaks slowly.

10. The doctor spoke slowly as he informed him of his father's death.
11. My psychoanalyst speaks quietly.
12. The assassin spoke quietly as he outlined his plan.

EXPERIMENTATION

In order to test whether the writer's intuitions corresponded to those of other native English speakers, an experiment was carried out using different permutations of frames and adjective labels. The labels were chosen in accordance with the following criteria.

(i) Labels referring to features of voice quality and voice dynamics were used, because 'there seem to be very few impressionistic labels for voices which refer specifically to segmental pronunciation, or which include segmental pronunciation in the reference together with other features' (Laver, 1974:66).

(ii) The referents of the labels used represented a cross-section of voice quality and voice dynamic parameters. Following Laver (1980), voice quality parameters were subcategorised into laryngeal features (phonation types) and supralaryngeal (articulatory settings). Division of voice dynamic parameters followed the traditional categorisation into the perceptual phenomena of pitch, loudness and length (temporal features). Temporal parameters were further subdivided into tempo and continuity.

(iii) All the labels used were of the 'descriptive' variety defined above. This avoided the ambiguity as to whether a label refers primarily to features of the speaker or to features of his voice. This ambiguity accounts for the acceptability of both frames - 'a ___ speaker' and 'a ___ voice' - with indexical labels. And, of course, there is a strong implicational relationship between the two. Thus a Liverpool speaker usually has what one would call a Liverpool voice; to call someone an upper-class speaker generally means that he has an upper-class voice; a patronising speaker normally uses a patronising voice; and so on.

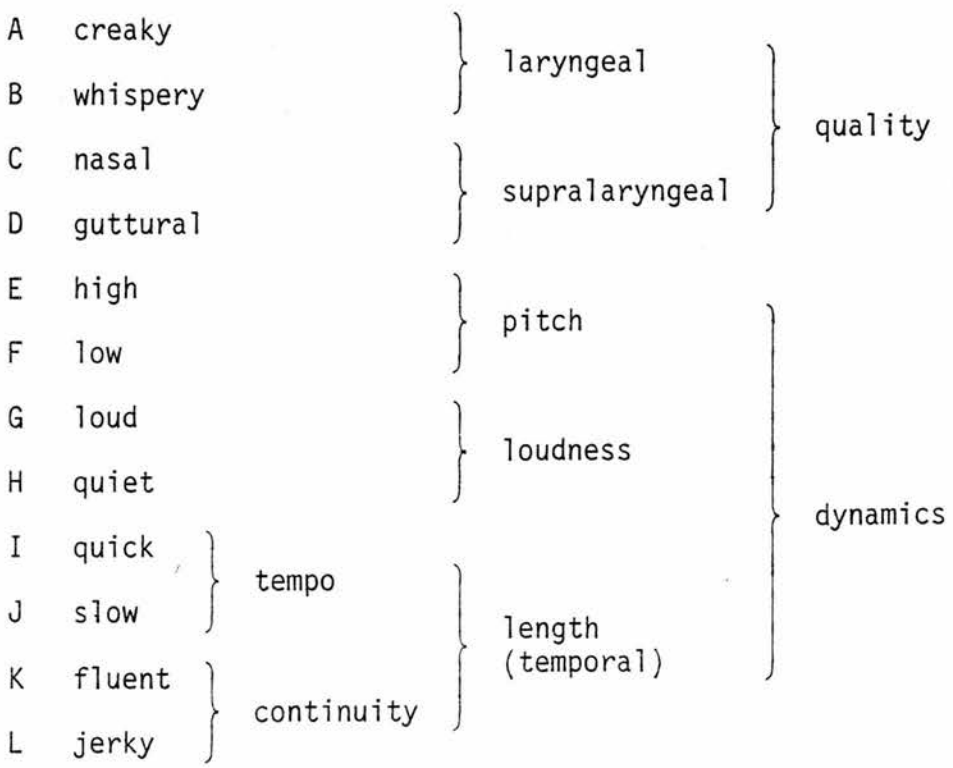
(iv) Labels were chosen which referred least ambiguously to one parameter, and one alone. For some parameters it was easy to find relatively unambiguous labels. Thus a high voice usually refers exclusively to a high mean pitch value. Similarly, the adjective quiet was preferred to soft since the latter often has implications of breathy phonation. For other parameters the task of finding unambiguous labels was not so easy. There is no unambiguous descriptive label in English for a voice with a low mean pitch value. Deep has strong qualitative implications, and bass is never used in paralinguistic contexts. Low was eventually chosen, although as Abercrombie points out above, this may equally well refer to a low mean loudness value.

(v) Adjective labels were chosen which had adverbial counterparts. Furthermore, only those adjectives were chosen from which adverbs are

formed by the addition of the -ly suffix. Thus quick was preferred to fast since the adverb corresponding to quick (quickly) involves the suffixation of -ly, whereas that corresponding to fast (fast) does not.

(vi) Other things being equal, labels were chosen which belonged to everyday vocabulary. There seem to be certain areas, especially tongue settings in the oral cavity, where unambiguous labels (or, indeed, any labels) are extremely scarce in everyday English. Another such area is the temporal parameter of rhythmicality, where, apart from the self-explanatory adjectives rhythmical and unrhythmical, English has no unambiguous expressions.

The following labels were thus chosen for the purpose of the experiment.



The sentence frames used corresponded to those illustrated above. They therefore took into account the following distinctions.

(i) adjectival labels/adverbial labels

(ii) extralinguistic context/paralinguistic context

(iii) (for extralinguistic adjectival labels) 'a ___ voice'/ 'a ___ speaker'.

The six frames took the following forms.

1. X has a ___ voice.
2. X is a ___ speaker.
3. X speaks ___-ly.
4. X spoke ___-ly when ...
5. X spoke in a ___ voice when ...
6. "...," said X in a ___ voice.

Frames 5 and 6 both fill the paralinguistic adjectival category and for this reason were considered to be equivalent. It was expected that the acceptability responses for these two frames would be identical.

The frames used were thus kept low in number and syntactic complexity. Frames 4, 5 and 6 are the most complex in that they are all composed of two clauses. However, it would have been impossible to investigate the effect on subject performance of interchanging the two clauses without increasing the frames to an unmanageable number.

The six sentence frames were combined with the twelve chosen adjectives to give a total of 72 permutations. The stimulus sentences were written using appropriate nouns, etc. to give plausible real world situations. The following are a sample of the stimuli.

4. I The porter spoke quickly when he told them their train was coming.

2. J Our vicar is a slow speaker.

6. G "Get off my land!" said the warden in a loud voice.

3. H Our bank manager speaks quietly.

5. F The president spoke in a low voice when he talked about the serious energy shortage.

1. A Our MP has a creaky voice.

Forty native speakers of English acted as subjects for the experiment. The factors of age, sex and dialect were not controlled. Of these, dialect is the most likely to have an effect on responses, although it was thought that this would be very slight. Subjects were chosen whose phonetic sophistication was not such that the labels would be interpreted in any technical phonetic sense (for example, creaky voice and whispery voice are technical terms in Laver's (1980) classification of voice quality). Each subject was presented with 36 sentences, constituting the permutation of all 6 frames with 6 out of the 12 adjectives. Both subsets of 6 adjectives were made up by taking one from each pair in a parametric category (for example, high was in one subset, low in the other). Where

parametric polarity exists (i.e. in the dynamic parameters), it was ensured that the subsets did not consist of all the positive adjectives (high, loud, quick, fluent) or all the negative. The first subset therefore consisted of adjectives A, D, E, H, J and K, and the second of B, C, F, G, I and L. In this way, responses from 20 subjects were obtained for each combination.

Subjects were asked to judge the acceptability of the sentences. They were allowed three kinds of response: (i) acceptable, (ii) unacceptable, and (iii) an intermediate response for dubious examples. It was emphasised that they should judge the acceptability of the sentence, not its meaningfulness (i.e. whether they could infer what a person meant if he said the sentence to them). It was also stressed that they should judge the sentence in the exact form given, not any version involving a slight paraphrase or word-change.

The results of the experiment are given in Table 1. For each intersection of the adjective labels with the sentence frames, a bar-graph is given of the 20 listeners' responses. From left to right, these represent the number of acceptable, dubious and unacceptable responses respectively.

TABLE 1 : Subject responses of the acceptability of labels in sentence frames.

Bar-graphs indicate the number of acceptable, dubious and unacceptable responses respectively from left to right.

KEY to frames :

1. X has a ___ voice
2. X is a ___ speaker
3. X speaks ___-ly.
4. X spoke ___-ly when ...
5. X spoke in a ___ voice when ...
6. "...," said X in a ___ voice.

FRAME

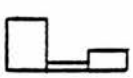
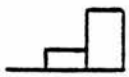
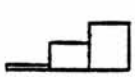
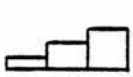
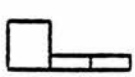
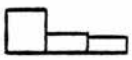
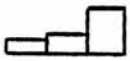
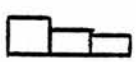
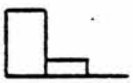
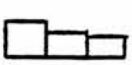
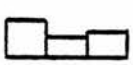
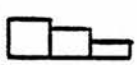
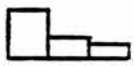
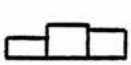
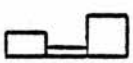
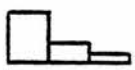
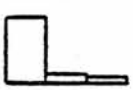
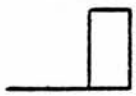
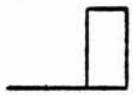
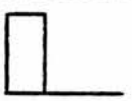
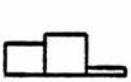
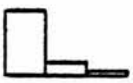
	1	2	3	4	5	6
A creaky	 13 2 5	 0 5 15	 1 7 12	 3 7 10	 12 4 4	 16 2 2
B whispery	 11 5 4	 4 5 11	 1 2 17	 1 2 17	 9 6 5	 11 2 7
C nasal	 16 4 0	 9 6 5	 14 5 1	 9 5 6	 9 7 4	 12 5 3
D guttural	 18 2 0	 5 8 7	 7 10 3	 7 3 10	 17 3 0	 13 5 2
E high	 17 2 1	 0 1 19	 0 1 19	 0 1 19	 16 3 1	 17 2 1
F low	 19 1 0	 0 2 18	 0 0 20	 0 0 20	 19 1 0	 20 0 0
G loud	 20 0 0	 8 10 2	 19 1 0	 16 3 1	19 1 0	20 0 0
H quiet	20 0 0	13 6 1	20 0 0	20 0 0	19 1 0	17 3 0
I quick	1 7 12	12 5 3	20 0 0	18 2 0	3 4 13	4 2 14
J slow	1 8 11	19 0 1	20 0 0	20 0 0	8 10 2	5 11 4
K fluent	1 7 12	15 5 0	16 4 0	20 0 0	2 7 11	6 7 7
L jerky	10 8 2	9 6 5	16 4 0	14 6 0	14 4 2	16 4 0

TABLE 1

An immediate observation is that there are few stimuli which produced a unanimous response from the subjects. This may be attributable to three reasons: (i) that subjects' ability to perform the task was impaired by the number of stimuli presented to each of them, although no time limit was set upon the task and no subjects spontaneously complained of this fact, (ii) that there is a large amount of inter-subject variation, although no consistent dialectal trends were observed, and (iii) that such expressions are not a common feature of everyday speech but belong rather to the language of literature, journalism, etc. In other words, it is more fruitful to deal with this data not in categorical terms but as reflecting general trends towards acceptability or unacceptability, with a relatively large number of examples reflecting neither. The generally more acceptable stimuli are characterised by a clustering of responses towards the left of the graph, the generally less acceptable towards the right, and examples about which there was no definite trend produced a spread of responses across the three categories.

The frame 'a ___ voice' used extralinguistically (column 1) is seen to be acceptable only with non-temporal labels (A-H). The only exception to this observation is the high acceptability response for jerky, an adjective which may be interpreted as containing reference to non-temporal aspects (a rapidly fluctuating pitch and/or loudness value, as in a tremolo voice).

The frame 'a ___ speaker' used extralinguistically (column 2) is acceptable with temporal labels (I-L), although the response for jerky

is again unconvincing. A possible ambiguity with this frame is that the adjective may be interpreted as referring not to the manner of speaking but to the person who is speaking. Thus a quiet speaker is acceptable if the label is taken to refer to the person's shy, introverted nature (with possible implications thereby for his voice) as opposed to the low loudness value of his voice. This may explain the high acceptability responses for quiet and loud, although the response for the latter may also be affected by the confusion caused by the existence of the noun loudspeaker. The uncertainty of the nasal and guttural responses is difficult to account for.

Sentences using the corresponding adverbs (columns 3 and 4) were judged acceptable with all temporal labels (I-L). High and low both produced high unacceptability scores since the corresponding adverbs do not exist; this is due presumably to the existence of the adjective lowly and the verb speak highly of. Subjects reported their dislike of the qualitative adverbs creakily, whisperily, nasally and gutturally. This may be due in part to the relative lack, and thus rarity, of descriptive labels for features of voice quality in English.

The frame 'a ___ voice' produced acceptable sentences with all non-temporal labels (A-H) in paralinguistic contexts (columns 5 and 6). The predominance of 'dubious' and 'unacceptable' responses for the temporal labels (I-L) is perhaps a perseverative effect from the general dislike of constructions of the form a quick voice (see column 1).

Ignoring counterexamples, which may be explained as individual cases, the above co-occurrence regularities may be summarised in the following rules.

(i) In extralinguistic contexts, non-temporal labels produce acceptable sentences in the frame 'a ___ voice', while the appropriate frame for temporal labels is 'a ___ speaker'.

(ii) In paralinguistic contexts, all labels except those referring to features of voice quality may occur as adverbs.

(iii) In paralinguistic contexts, only non-temporal labels produce fully acceptable sentences with the frame 'a ___ voice'.

There is thus evidence to conclude that the acceptable possibilities above form a co-occurrence regularity in English. That is, there is a strong tendency for native speakers to differentiate between labels for temporal and non-temporal aspects of speech in terms of the grammatical constructions which can accompany them. The explanation for this regularity might come from one of two sources: firstly, a purely syntactic motivation might be found for it, or secondly, it might reflect a phonetic difference in the referents of the labels. These two possibilities are now examined.

SYNTACTIC DESCRIPTION

It was shown above that, as a general rule, excluding instances where unacceptability can be accounted for as an individual peculiarity of the particular adjective, labels are acceptable in their adverbial form (loudly, slowly, etc.) in all circumstances. It is clear that these forms bear a close syntactic relationship to the nominal constructions (a loud voice, a slow speaker, etc.). Study of the grammar of nominalisations dates back at least to Jespersen (1933), who pointed out the tendency among English speakers to use deverbal noun constructions in preference to the full verbs from which they derive.

> 'Such expressions instead of the simple verb are in accordance with the general tendency of modern English to place an insignificant (auxiliary) verb ... before the really important idea.'
> (Jespersen, 1933:71, quoted in Liefrink, 1973)

Examples of this are the following (b) sentences, which are preferred to the (a) equivalents.

13. (a) The referee looked at his watch.
    (b) The referee had a look at his watch.

14. (a) The mother pushed the swing.
    (b) The mother gave the swing a push.

It follows that any adverb used in the first kind of sentence will occur as an adjective in the second kind.

15. (a) The referee looked quickly at his watch.
    (b) The referee had a quick look at his watch.

16. (a) The mother gently pushed the swing.
    (b) The mother gave the swing a gentle push.

Therefore there is a tendency among English speakers to avoid adverbs and use deverbal noun constructions instead. It is only a tendency, since both are acceptable English constructions.

Quirk *et al.* (1972) describe the relationships which hold for one kind of deverbal noun - where the noun refers to the performer of an activity.

> 'Predication of the adjective is ... blocked when the noun head is agential and the adjective refers to the activity:
>
> A kind writer - The writer is kind
>
> A hard worker - { He works hard / *The worker is hard'
>
> (Quirk *et al.*, 1972:906)

The corresponding examples in the present discussion are the following.

17. (a) A loud voice.
    (b) The voice is loud.

18. (a) A slow speaker.
    (b) He speaks slowly.
    (c) *The speaker is slow.

There is, however, a further peculiarity to the present situation, which is not accounted for by Quirk *et al.*, and which differs from

their example. It is that while 19(b) is unacceptable as an equivalent of 19(a), 20(b) may be used as a semantically equivalent version of 20(a), to which it bears an obvious close syntactic relationship (in the same way as the examples in 18 above).

19. (a) A kind writer.
    (b) *He writes kindly.

20. (a) A loud voice.
    (b) He speaks loudly.

Some transformationalists would say that the above phenomena are to be treated not merely as close relationships, but that phrases such as 18(a) are transformationally derived from underlying structures corresponding to 18(b).[1] The problem, then, for these syntacticians lies in describing why 21(a) should develop into 21(b) and not 21(c), while 22(a)'s development is exactly the opposite.

21. (a) He speaks loudly.
    (b) A loud voice.
    (c) *A loud speaker.

22. (a) He speaks slowly.
    (b) *A slow voice.
    (c) A slow speaker.

The writer has found no reference to this particular problem in the literature, and, since this phenomenon is not directly analogous

[1] This derivation seems to be working in a different direction from the relationship implied in the quotation from Quirk et al. This may be due to the fact that Quirk et al.'s grammar concentrates more on the speech production process from the language learner's point of view (i.e. it is not possible to say the worker is hard on analogy with the writer is kind).

to the similar kind of construction which writers have analysed (a hard worker), it is assumed that there is no syntactic motivation for it, and that any attempt at a purely syntactic explanation will encounter the same problems as those discussed above. There are thus two alternative solutions: either (i) the phenomenon is simply dismissed as a peculiarity of English syntax, an exception to the rules (which is an undesirable last resort), or, preferably, (ii) an extra-syntactic determining factor must be sought. Therefore, in the next section, the phonetic category of temporal parameters is considered, in the hope that a phonetic distinction may be what is reflected in the syntactic differentiation.

PHONETIC DESCRIPTION

A suitable way to begin this phonetic examination is to define what is meant by the temporal parameters of tempo, continuity and rhythmicality. Throughout, it will be the perceptual characteristics of these parameters which are under consideration, not their physical manifestation.

Tempo may be defined as the speed at which a person speaks, measurable in physical terms as the amount of speech output[1] per unit time.

---

[1] For our purposes, we may disregard the question of what should be taken as the units of speech output - segments, syllables, morphemes, etc.

> 'Closely connected with tempo is continuity, which refers to the incidence of pauses in the stream of speech - where they come, how frequent they are, and how long they are. The incidence of pauses, whether they are hesitations or whether they are deliberate cessations of talking for the purpose of taking breath, seems to be a highly idiosyncratic matter, and there is a lot of variation from speaker to speaker. Under the conditions of ordinary conversation nobody's speech is fluent, and it is probably true to say that the more thought there is behind what one is saying, the less fluent will be the speech.' (Abercrombie, 1967:96)

Rhythmicality may be defined as the variation in the temporal regularity of stresses in English, or their relative isochrony (Abercrombie, 1964a).

When one comes to consider what determines the value of these parameters, as speaker-characterising features of an individual, it is clear that they depend upon very different factors from the other, non-temporal parameters of pitch, loudness, etc. Among the main determinants of pitch and loudness are anatomical and physiological factors such as the strength of the intercostal muscles, the size of the lungs, the size of the vocal cords, the efficiency of vocal cord vibration, etc. Thus, for example, shorter-than-average vocal cords are directly conducive to a speaker having a high mean pitch. Anatomical factors of this nature are unlikely to play any great part in the determination of temporal parameter values for the normal, healthy speaker. Instead, as Abercrombie suggests in the previous quotation, temporal parameter values depend rather on the ability of the speaker to pre-plan stretches of speech and to continue planning

while articulation is taking place. If a speaker's ability to do this is relatively inefficient, then this will be a main factor not only for the fluency of his speech, but also for the other temporal parameters of tempo and rhythmicality. Similarly, if his ability is temporarily impaired by the need to put a lot of thought into the speech, to be continually searching for le mot juste, this will disrupt these temporal characteristics.

It is not an insignificant factor that one can talk about there being something efficient about talking fluently, quickly and rhythmically, while this does not apply to the other dynamic parameters. Thus one would not consider a rise in mean pitch to be either an increase or decrease in efficiency.

Speakers have little difficulty in raising or lowering pitch and loudness values on a short-term basis. Thus it is simple to speak on demand with a higher or lower mean pitch, or more loudly or quietly; this is probably why these parameters figure strongly among those which are manipulated for linguistic and paralinguistic purposes. Similarly, speakers have little difficulty in producing a momentary decrease in, for example, tempo, and this may be why speaking more slowly is a common strategy for adding emphasis or importance to what one is saying. On the other hand, it is more difficult to increase tempo at will for anything other than a very short stretch of speech.

There is a schoolboy game, to succeed in which one has to speak for half a minute without breaks and at a fast tempo. The effect is

similar to that achieved by horse-racing commentators, whose speech is very fast and fluent, since they have to convey a lot of information in a short space of time. There is also a BBC radio game "Just a Minute", in which contestants have to speak continuously for a minute on a given topic. Strategies are often adopted by the participants in an effort to maintain fluency and avoid hesitations. One of the major of these is to speak at a reduced tempo (but high fluency) rate. The speech then approaches what one would call a drawling style. It would be nonsense to think of similar games where one had to speak at a higher or lower pitch, or greater or quieter loudness, or indeed slower tempo or less fluent continuity rates. There is thus justification for stating that it is more difficult to alter the values of temporal parameters on a short-term basis except for very short stretches of speech and that this is attributable to the fact that such changes require an increase in efficiency of the speaker's inherent speech-planning process. Increases in tempo may be achieved over longer stretches, but only in situations where the content of the speech is already given, such as reading aloud, repeating word-for-word.

It is owing to the above factors that increases in temporal aspects are seldom used effectively for paralinguistic purposes. This may help to explain the inconsistent acceptability scores for temporal labels in paralinguistic contexts (see Table 1), in contradiction to the writer's intuitions represented in sentences 5 and 6 above. However, when such increases are employed to effect by an individual, this is probably the

result of practice since the effect is useful in his profession, e.g. public speaker, stand-up comedian, disc-jockey.

Returning to the extralinguistic function of temporal parameters as speaker-characterising features, it is clear that, since these parameters depend not upon anatomical factors but upon the ability to organise speech, the stretch of speech required for the determination of the value of these parameters for any individual will be relatively long. Thus one can usually derive sufficient information about quality, pitch and loudness from a stretch of a few segments, or even one; but, to gain enough information about a speaker's tempo, rhythmicality and continuity, one requires a stretch of at least sentence-length and perhaps what one would call a whole discourse.

We can therefore conclude that there is a basic phonetic difference between temporal and non-temporal aspects of speech. Variation in temporal parameters is inhibited by the rhythm with which cognitive activity takes place. This cognitive rhythm imposes restrictions on the temporal patterning of the events of articulation, as of other physical activity. As Lenneberg (1967) puts it, 'the rhythm is the grid, so to speak, into whose slots events may be intercalated' (p.119). This distinction between cognitively governed temporal aspects of speech and physiologically determined non-temporal aspects is by no means a new one, as the following quotations from reports of experimental study show.

'Henderson et al. (1965) argued that if planning goes on in the hesitation pauses then this alternating pattern might represent a kind of cognitive rhythm, the way cognitive activity advances during speech.'

(Goldman-Eisler, 1967:122-3)

'[Speech-error data] suggests that syllable structure and rhythm are ... more than just linguistic constructs, and can be plausibly considered to be central aspects of the neural control programme in speech.'

(Boomer & Laver, 1968:9)

'The tendency towards the equalisation of phrases cannot be derived from the physiology of breathing. The explanation must be found on a higher level.'

(Fonagy & Magdics, 1960:189)[1]

If this is true, one might expect this to be what is reflected in the syntactic alternation between 'a ___ voice' (which tends to focus on the articulatory aspects of speech) and 'a ___ speaker' (which refers rather to the cognitive aspects). One might also expect that this difference is similarly manifested in other expressions referring to the organisation rather than the articulation of speech. As a small test of this hypothesis, the construction

X has a ___ style of speech

was also incorporated as one of the frames in the experiment reported above. The acceptability responses for this frame with all 12 adjectives used are given in Table 2.

---

[1] 'The equalisation of phrases' refers to the phenomenon that longer phrases are spoken more quickly than shorter ones, thereby reducing the difference in temporal duration.

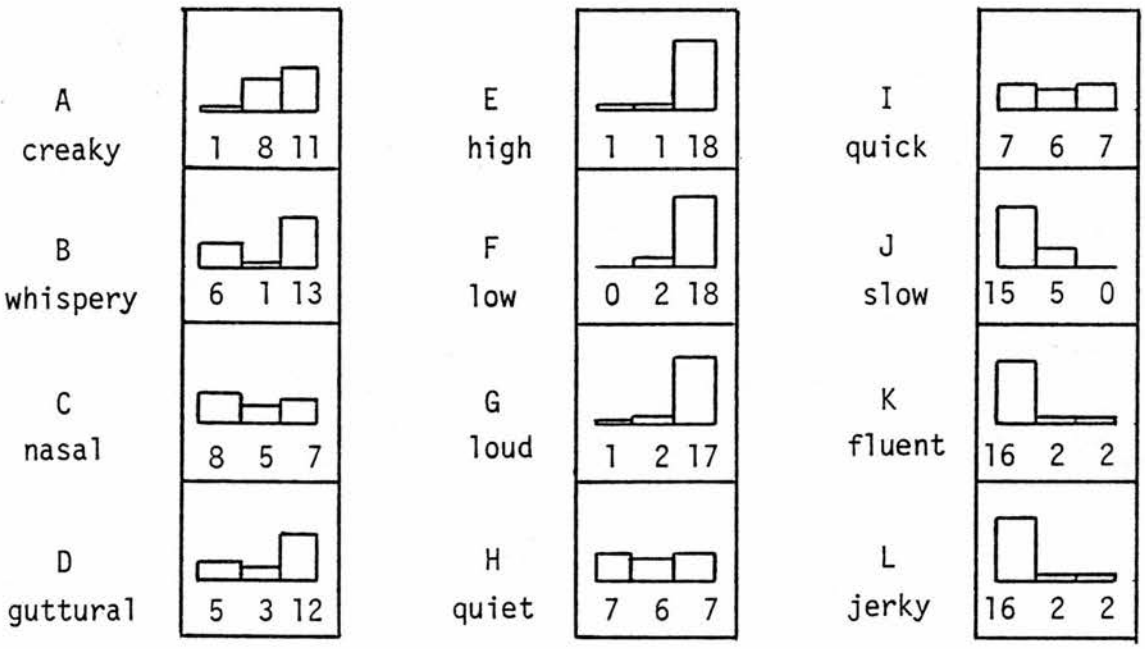


TABLE 2 : Subject responses of the acceptability of labels in the frame X has a ___ style of speech .

Bar-graphs indicate the number of acceptable, dubious and unacceptable responses respectively from left to right.

Again, the results are not conclusive, but indicate a general trend in keeping with the above hypothesis, with three exceptions. Responses for the adjective quick varied widely, although this may be explained by the general dislike of that adjective for the description of speech; some subjects commented that they would have found the adjective fast more acceptable. The varied responses produced by the adjectives quiet and nasal are similar to those of quick, and are difficult to account for, although it is presumably not coincidence that these two also produced high acceptability responses for the frame 'a ___ speaker' (see column 2 of Table 1). That the above hypothesis is generally correct is borne out by the fact that adjectives which include temporal aspects of speech in their reference produce acceptable sentences when inserted into the above frame.

X has a drawling/monotonous/sombre style of speech

SUPPORT FROM OTHER LANGUAGES

We have seen that the co-occurrence regularity may be explained not by any syntactic motivation, but as reflecting a phonetic distinction. In this case, one would expect the distinction to be manifested syntactically in languages other than English. In an informal survey, no language was found which did not have a co-occurrence regularity comparable to that seen in English. Of course, different languages express concepts in different ways and one cannot always

expect direct equivalents of the English constructions to be found. Most languages prefer adverbial constructions, unlike Jespersen's claim for English. However, in all the languages investigated where unambiguous translation equivalents exist, a distinction was found between the acceptability of the expression translating the English 'a ___ voice' with non-temporal labels in extralinguistic contexts, and its unacceptability with temporal labels (as in the English X has a loud voice but *X has a fast voice).

Language	Non-temporal	Temporal
Thai	sieng sung 'voice - high'	* sieng reu 'voice - fast'
Japanese	hikui koe 'low - voice'	* osoi koe 'slow - voice'
Danish	en sagte stemme 'a - quiet - voice'	*en hurtig stemme 'a - fast - voice'
French	une voix forte 'a - voice - loud'	*une voix lente 'a - voice - slow'
Swahili	sauti nene 'voice - low'	* sauti mbio 'voice - fast'
Arabic	sot misarsaʔ 'voice - high'	* sot batiʔ 'voice - slow'
Maltese	vuci tgħajjat 'voice - she shouts' (= 'a loud voice')	* vuci tgħaġġel 'voice - she hurries' (= 'a fast voice')

APPENDIX 2

T H E R A I N B O W P A S S A G E (Fairbanks, 1960:127)

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbow. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, and the width of the coloured band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of superposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green lights when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

APPENDIX 3 Factorial combinations for stimuli in the full 2^4 factorial design used in Experiment 2 (section 6.3). Factorial values for the control voice were midway between the high and low levels. Recordings of the stimulus voices are included on the tape which accompanies this thesis.

Stimulus number	a formant mean	b formant range	c formant bandwidth	d pitch mean
1	low	low	low	low
2	high	low	low	low
3	low	high	low	low
4	low	low	high	low
5	low	low	low	high
6	high	high	low	low
7	high	low	high	low
8	high	low	low	high
9	low	high	high	low
10	low	high	low	high
11	low	low	high	high
12	high	high	high	low
13	high	high	low	high
14	high	low	high	high
15	low	high	high	high
16	high	high	high	high

REFERENCES

R E F E R E N C E S

(JASA = Journal of the Acoustical Society of America)

ABBERTON, E, (1974) 'Listener identification of speakers from larynx frequency', Work in Progress, Department of Phonetics and Linguistics, University College, London.

ABERCROMBIE, D, (1963) 'Pseudo-procedures in linguistics', Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 16:9-12.
Also in D. Abercrombie (1965), pp.114-119.

_____(1964a)'A phonetician's view of verse structure', Linguistics 6: 5-13.
Also in D. Abercrombie (1965), pp.16-25.

_____(1964b)'Syllable quantity and enclitics in English' in D. Abercrombie et al (eds., 1964), pp.216-222.
Also in D. Abercrombie (1965), pp.26-34.

_____(1965) Studies in Phonetics and Linguistics, Oxford University Press, London.

_____(1967) Elements of General Phonetics, Edinburgh University Press.

_____(1968a) 'Paralanguage', British Journal of Disorders of Communication 3: 55-59.
Also in J. Laver & S. Hutcheson (eds., 1972), pp.64-70.

_____(1968b) 'Some functions of silent stress', Work in Progress, Department of Phonetics and Linguistics, Edinburgh University, 2: 1-10.
Also in A.J. Aitken, A. McIntosh & H. Palsson (eds., 1971) Edinburgh Studies in English and Scots, Longman, London, pp.147-156.

ABERCROMBIE D., FRY, D.B., MacCARTHY, P.A.D., SCOTT, N.C., & TRIM, J.L.M. (eds., 1964) In Honour of Daniel Jones, Longman, London.

ANONYMOUS (1965) 'Voice print identification', in W.W. Turner (ed.) Criminalistics, Aquaduct Books, Rochester.

ANTHONY, J. & LAWRENCE, W. (1962) 'A resonance analogue speech synthesiser', Proceedings of the 4th International Congress on Acoustics, Copenhagen, Paper G43.

ARMSTRONG, L.E. & WARD, I.C., (1931) A Handbook of English Intonation, Heffer, Cambridge (2nd edition).

ATAL, B.S. (1976), 'Automatic recognition of speakers from their voices', Proceedings of the IEEE 64: 460-475.

BEYN, E.S. & KNYAZEVA, G.R. (1962), The problem of prosopagnosia', Journal of Neurology, Neurosurgery and Psychiatry, 25: 154-158.

BINNIE, C.A., MONTGOMERY, A.A. & JACKMAN, P.L. (1974) 'Auditory and visual contributions to the perception of consonants', Journal of Speech and Hearing Research 17: 619-630.

BLACK, J.W., LASHBROOK, W., NASH, E., OYER, H.J., PEDREY, C., TOSI, O.I. & TRUBY, H. (1973), 'Reply to "Speaker identification by speech spectrograms: some further observations", (Bolt et al, 1973)', JASA, 54: 535-537.

BOLT, R.H., COOPER, F.S., DAVID, E.E. Jnr., DENES, P.B., PICKETT, J.M. & STEVENS, K.N. (1969) 'Identification of a speaker by speech spectrograms', Science, 166: 338-343.

_____ (1970) 'Speaker identification by speech spectrograms: a scientist's view of its reliability for legal purposes', JASA, 47: 597-612.
Also in E.E. David, Jnr., & P.B. Denes (eds., 1972) Human Communication: A Unified View, McGraw-Hill, New York, pp.369-398.

_____ (1973) 'Speaker identification by speech spectrograms: some further observations'. JASA, 54: 531-534.
Also in M.E. Hawley (ed., 1977), pp.430-433.

BOOMER, D.S. & LAVER, J., (1968) 'Slips of the tongue', British Journal of Disorders of Communication, 3: 2-12.
Also in V. Fromkin (ed., 1973), pp.120-131.

BORNSTEIN, B. & KIDRON, D.P. (1959) 'Prosopagnosia', Journal of Neurology, Neurosurgery and Psychiatry 22: 124-131.

BRAZIL, D., (1975) Discourse Intonation, English Language Research, Birmingham University.

BRICKER, P.D. & PRUZANSKY, S. (1966) 'Effects of stimulus content and duration on talker identification'. JASA, 40: 1441-1449.

_____ (1976) 'Speaker recognition', in N.J. Lass (ed., 1976), pp.295-326.

BROWN, G. (1977) Listening to Spoken English, Longman, London.

BROWN, R. (1978a) 'Two categorisations of speaker recognition tasks', Work in Progress, Department of Linguistics, Edinburgh University, 11: 52-62.

_____ (1978b) 'Paratones: their reality and realisation'. Social Science Research Council Project 'Intonation of Scottish English', Report for 1978, pp.57-84.

_____ (1979a) 'Memory and decision in speaker recognition'. International Journal of Man-Machine Studies, 11: 729-742.

_____ (1979b) 'An observation on labels for voices: data for the cognitive nature of temporal aspects of speech'. Work in Progress, Department of Linguistics, Edinburgh University, 12: 82-94.

_____ (forthcoming) 'Cognitive implications of labels for voices'. Submitted to the Journal of the International Phonetic Association. Reproduced here as Appendix 1.

CARBONELL, J.R., GRIGNETTI, M.C., STEVENS, K.N., WILLIAMS, C.E. & WOODS, B. (1965) 'Speaker authentication techniques'. Report 1296, Contract DA-28-043-AMC-00116(E), Bolt, Beranek & Newman Inc., Cambridge, Massachusetts.

CATFORD, J.C. (1964) 'Phonation types: the classification of some laryngeal components of speech production', in D. Abercrombie et al (eds.), pp.26-37.

CHERRY, E.C. (1953) 'Some experiments on the recognition of speech, with one and with two ears'. JASA, 25: 975-979.

_____ (1957) On Human Communication, Wiley, New York.

CHERRY, E.C. & TAYLOR, W.K. (1954), 'Some further experiments upon the recognition of speech, with one and with two ears'. JASA, 26: 554-559.

CHRISTOPHERSEN, P. (1956) An English Phonetics Course, Longman, Green, London.

CLARK, H.H. & CLARK, E.V. (1977) Psychology and Language: An Introduction to Psycholinguistics, Harcourt Brace Jovanovich, New York.

CLARKE, F.R. & BECKER, R.W. (1969) 'Comparison of techniques for discriminating among talkers'. Journal of Speech and Hearing Research, 12: 747-761.

CLARKE, F.R., BECKER, R.W. & NIXON, J.C. (1966) 'Characteristics that determine speaker recognition'. Report ESD-TR-66-636, Electronics Systems Division, Air Force Systems Command, Hanscom Field.

COCHRAN, W.G. & COX, G.M. (1957) Experimental Designs, Wiley, New York (2nd edition).

COLEMAN, R.O. (1971) 'Male and female voice quality and its relationship to vowel formant frequencies'. Journal of Speech and Hearing Research, 14: 565-577.

_____ (1973) 'Speaker identification in the absence of inter-subject differences in glottal source characteristics'. JASA, 53: 1741-1743.
Also in N.J. Lass (ed., 1974), pp.315-320.

COMPTON, A.J. (1963) 'Effects of filtering and vocal duration upon the identification of speakers, aurally'. JASA, 35: 1748-1752.
Also in M.E. Hawley (ed., 1977), pp.415-419.

CROWDER, R.G. & MORTON, J. (1969) 'Pre-categorical acoustic storage (PAS)'. Perception and Psychophysics 5: 365-373.

CRYSTAL, D. (1963) 'A perspective for paralanguage', Maitre Phonetique 120: 25-29.

_____ (1964) 'An approach to a reply', Maitre Phonetique, 122: 23-24.

CRYSTAL, D. & QUIRK, R. (1964) Systems of Prosodic and Paralinguistic Features in English. Mouton, The Hague.

CURRIE, K.L. (1979) 'Contour systems of one variety of Scottish English' Language and Speech, 22: 1-20.

DARWIN, C.J. & BADDELEY, A.D. (1974) 'Acoustic memory and the perception of speech'. Cognitive Psychology, 6: 41-60.

DELATTRE, P. (1965) Comparing the Phonetic Features of English, French, German and Spanish. Harrap, London.

DOEHRING, D.G. & ROSS, R.W. (1972) 'Voice recognition by matching to sample'. Journal of Psycholinguistic Research 1: 233-242.

DUKIEWICZ, L. (1970) 'Frequency-band dependence of speaker identification'. In W. Jassem (ed., 1970) Speech Analysis and Synthesis (volume 2). Institute of Fundamental Technical Research, Polish Academy of Sciences, Warsaw.

EGAN, J.P., SCHULMAN, A.I. & GREENBERG, G.Z. (1959) 'Operating characteristics determined by binary decisions and by ratings'. JASA, 31: 768-773.

ESLING, J.H. (1978) Voice quality in Edinburgh: a sociolinguistic and phonetic study. Ph.D thesis, University of Edinburgh.

FAIRBANKS, G. (1955) 'Selective vocal effects of delayed auditory feedback', Journal of Speech and Hearing Disorders 20: 333-346. Also in G. Fairbanks (1966), pp.10-23.

_____ (1960) Voice and Articulation Drillbook. Harper, New York. (2nd. edition).

_____ (1966) Experimental Phonetics: Selected Articles. University of Illinois Press, Urbana.

FAIRBANKS, G. & GUTTMAN, N. (1958) 'Effects of delayed auditory feedback upon articulation'. Journal of Speech and Hearing Research, 1: 12-22. Also in G. Fairbanks (1966), pp.24-34.

FANT, G. (1962) 'Descriptive analysis of the acoustic aspects of speech', Logos, 5: 3-17. Also in I. Lehiste (ed., 1967), pp.93-107.

FEIBLEMAN, J.K. (1946) An Introduction to Peirce's Philosophy, Allen and Unwin, London.

FERRERI, G. (1959) 'Senescence of the larynx', Italian General Review of Oto-rhinolaryngology 1: 640-709.

FLOYD, W. (1964) 'Voice identification techniques', Report AD-606-634, Intelligence Applications Branch, Rome Air Development Centre, Research and Technology Division, Air Force Systems Command, Griffiss Air Force Base, New York.

FONAGY, I. & MAGDICS, K. (1960) 'Speed of utterance in phrases of different lengths'. Language & Speech 3:179-192.

FOURCIN, A.J. (1960) 'A potential dividing function generator for the control of speech synthesis'. JASA, 32:1501 (A).

FOURCIN, A.J. & ABBERTON, E. (1971) 'First applications of a new laryngograph'. Medical and Biological Illustrated 21: 68-78.

FRIEDLANDER, B.Z. (1968) "The effect of speaker identity, voice inflection, vocabulary, and message redundancy on infants' selection of vocal reinforcement". Journal of Experimental Child Psychology, 6:443-459.

FROMKIN, V.A. (ed., 1973) Speech Errors as Linguistic Evidence Mouton, The Hague.

FRY, D.B. (1955) 'Duration and intensity as physical correlates of linguistic stress'. JASA 27: 765-768. Also in I. Lehiste (ed., 1967), pp.155-158.

_____ (1958) 'Experiments in the perception of stress'. Language and Speech, 1: 126-152. Also in D.B. Fry (ed., 1976), pp.401-424.

_____ (1965) 'The dependence of stress judgments on vowel formant structure'. Proceedings of the 6th International Congress of Phonetic Sciences, Karger, pp.306-311. Also in D.B. Fry (ed., 1976), pp.425-430.

_____ (ed., 1976) Acoustic Phonetics: A Course of Basic Readings. Cambridge University Press.

FUJISAKI, H. & KAWASHIMA, T. (1968) 'The influence of various factors on the identification and discrimination of synthetic speech sounds'. Reports of the 6th International Congress on Acoustics, Tokyo.

FYFE, F.W. & NAYLOR, E. (1958) 'Calcification and ossification in the cricoid cartilage of the larynx with annotation on the mechanism of change of pitch'. Proceedings of the Canadian Otolaryngological Society, pp.67-69.

GAETH, S. (1948) A study of phonemic regression associated with hearing loss. Unpublished doctoral dissertation, Department of Speech, Northwestern University.

GARVIN, P.L. & LADEFOGED, P. (1963) 'Speaker identification and message identification in speech recognition'. Phonetica 9: 193-199.

GIMSON, A.C. (1962) An Introduction to the Pronunciation of English. Edward Arnold, London.

GOLDMAN-EISLER, F. (1967) 'Sequential temporal patterns and cognitive processes in speech'. Language & Speech 10: 122-132.

GRAY, C.H.G. & KOPP, G.A. (1944) 'Voice print identification'. Unpublished report, Bell Laboratories, New Jersey.

GREENLEE, D. (1973) Peirce's Concept of Sign. Mouton, The Hague.

HALL, R.A. Jnr. (1964) Introductory Linguistics. Chilton, Philadelphia.

HARTSHORNE, C. & WEISS, P. (1931-5) C.S. Peirce: Collected Papers. Harvard University Press.

HAWLEY, M.E. (ed., 1977) Speech Intelligibility and Speaker Recognition. Wiley, New York.

HECKER, M.H.L. (1971) 'Speaker recognition: an interpretive survey of the literature'. American Speech and Hearing Association Monograph 16, Washington, D.C.

HELFRICH, H. (1979) 'Age markers in speech', in K.R. Scherer & H. Giles (eds.) pp.63-107.

HENDERSON, A., GOLDMAN-EISLER, F. & SKARBEK, A. (1965) 'Temporal patterns of cognitive activity and breath control in speech'. Language & Speech 8:236-242.

HOLMGREN, G.L. (1967) 'Physical and psychological correlates of speaker recognition'. Journal of Speech and Hearing Research, 10: 57-66.

HONIKMAN, B. (1964) 'Articulatory settings' in D. Abercrombie et al (eds.), pp.73-84.

JASSEM, W. (1952) Intonation of Conversational English (Educated Southern British). Nakladem Wroclawskiego Towarzystwa Naukowego, Wroclaw.

JESPERSEN, O. (1933) Essentials of English Grammar. Allen & Unwin, London.

JONES, D. (1950) The Pronunciation of English. Cambridge University Press (3rd edition).

_____ (1964) An Outline of English Phonetics. Heffer, Cambridge (9th edition).

JONES, W.R. (1973) 'Danger - voiceprints ahead'. The American Criminal Law Review, 11: 549-573.

KAPLAN, H.M. (1960) Anatomy and Physiology of Speech. McGraw-Hill, New York.

KERSTA, L.G. (1962) 'Voiceprint identification', Nature 196: 1253-1257.
Also in M.E. Hawley (ed., 1977), pp.425-429.

KINGDON, R. (1958a) The Groundwork of English Intonation. Longman, London.

KINGDON, R. (1958b) English Intonation Practice. Longman, Green, London.

LADEFOGED, P. (1962) Elements of Acoustic Phonetics. Oliver and Boyd, London.

LADEFOGED, P. & VANDERSLICE, R. (1967) 'The voiceprint mystique'. Working Papers in Phonetics, University of California at Los Angeles, 7.

LASS, N.J. (ed., 1974) Experimental Phonetics MSS Information Corporation, New York.

_____ (ed., 1976) Contemporary Issues in Experimental Phonetics. Academic Press, London.

LAVER, J. (1964) 'The synthesis of voice-quality'. PAT Report for 1964, Phonetics Department, University of Edinburgh, 33-38.

_____ (1967) 'The synthesis of components in voice quality'. Proceedings of the 6th International Congress of Phonetic Sciences, Prague. Academia, Prague, pp.523-525.

_____ (1968) 'Voice quality and indexical information'. British Journal of Disorders of Communication, 3: 43-54. Also in J. Laver & S. Hutcheson (eds., 1972), pp.189-203.

_____ (1974) 'Labels for voices'. Journal of the International Phonetic Association, 4: 62-75.

_____ (1975) Individual features in voice quality. Ph.D. thesis, University of Edinburgh.

_____ (1976) 'Language and nonverbal communication' in E.C. Carterette & M.P. Friedman (eds.) Handbook of Perception (volume VII). Academic Press, London, pp.345-361.

_____ (1979) 'The description of voice quality in general phonetic theory'. Work in Progress, Department of Linguistics, Edinburgh University, 12: 30-52.

_____ (1980, in press) The Phonetic Description of Voice Quality. Cambridge University Press.

LAVER, J. & HUTCHESON, S. (eds., 1972) Communication in Face to Face Interaction. Penguin, Harmondsworth.

LAVER, J. & TRUDGILL, P. (1979) 'Phonetic and linguistic markers in speech' in K.R. Scherer & H. Giles (eds.), pp.1-32.

LAWRENCE, W. (1953) 'The synthesis of signals which have a low information rate', in W. Jackson (ed.) Communication Theory, Butterworths Scientific Publications, London, pp.460-469.
Also in D.B. Fry (ed., 1976), pp.208-218.

LEE, W.R. (1960) An English Reader, MacMillan, London.

LEHISTE, I. (ed., 1967) Readings in Acoustic Phonetics, MIT Press, Cambridge, Massachusetts.

_____ (1973) 'Rhythmic units and syntactic units in production and perception'. JASA, 54: 1228-1234.

LENNEBERG, E.H. (1967) Biological Foundations of Language, Wiley, New York.

LIEFRINK, F. (1973) Semantico-Syntax, Longman, London.

LINDBLOM, B. & SUNDBERG, J. (1971) 'Acoustical consequences of lip, tongue, jaw and larynx movement'. JASA, 50: 1166-1179.

LUMMIS, R.C. (1973) 'Speaker verification by computer using speech intensity for temporal registration'. IEEE Transactions on Audio and Electroacoustics AU-21: 80-89.

LYONS, J. (1972) 'Human Language', in R.A. Hinde (ed.) Non-verbal Communication, Cambridge University Press, pp.49-85.

_____ (1977) Semantics (Volume 1), Cambridge University Press.

MacCARTHY, P.A.D. (1944) English Pronunciation, Heffer, Cambridge.

MACCOBY, E.E. & JACKLIN, C.N. (1974) The Psychology of Sex Differences, Stanford University Press.

MATSUMOTO, H., HIKI, S., SONE, T. & NIMURA, T. (1973) 'Multidimensional representation of personal quality and its acoustical correlates'. IEEE Transactions on Audio and Electroacoustics AU-21: 428-436.

MAZANEC, N. & McCALL, G.J. (1975) 'Sex, cognitive categories and observational accuracy'. Psychological Reports 37: 987-990.

McDERMOTT, B.J. (1969) 'Multidimensional analyses of circuit quality judgments'. JASA 45: 774-781.

McGEHEE, F. (1937) 'The reliability of the identification of the human voice'. Journal of General Psychology 17: 249-271.

_____ (1944) 'An experimental study in voice recognition'. Journal of General Psychology 31: 53-65.

McGUINNESS, D. (1976) 'Sex differences in the organisation of perception and cognition', in B. Lloyd & J. Archer (eds.) Exploring Sex Differences, Academic Press, London.

McGURK, H. & MacDONALD, J. (1976) 'Hearing lips and seeing voices'. Nature 264: 746-748.

MELROSE, J., WELSH, O.L. & LUTERMAN, D.M. (1963) 'Auditory responses in selected elderly men'. Journal of Gerontology 18: 267-270.

MILLER, J.E. (1964) 'Decapitation and recapitation: a study in voice quality'. JASA 36: 2002 (A).

NORMAN, D.A. (1969) Memory and Attention: An Introduction to Human Information Processing. Wiley, New York.

O'CONNOR, J.D. (1973) Phonetics, Penguin, Harmondsworth.

O'CONNOR, J.D. & ARNOLD, G.F. (1961) Intonation of Colloquial English: a Practical Handbook. Longman, London.

OSGOOD, C.E., SUCI, G.J. & TANNENBAUM, P.H. (1957) The Measurement of Meaning. University of Illinois Press, Urbana.

PALMER, H.E. & BLANDFORD, F.G. (1924) A Grammar of Spoken English on a Strictly Phonetic Basis. Heffer, Cambridge.

PEIRCE, C.S. (1940) The Philosophy of Peirce: Selected Writings (ed. J. Buchler) Kegan Paul, Trench & Trubner, London.

PETERS, R.W. (1954) 'Studies in extra messages: listener identification of speakers' voices under conditions of certain restrictions imposed upon the voice signal'. U.S. Naval School of Aviation Medicine, Joint Project NM001-064-01, Report 30, Pensacola, Florida.

PIKE, K.L. (1945) The Intonation of American English, University of Michigan Press, Ann Arbor.

POLLACK, I., PICKETT, J.M. & SUMBY, W.H. (1954) 'On the identification of speakers by voice'. JASA 26:403-406. Also in N.J. Lass (ed., 1974), pp.251-258.

POTTER, R.K., KOPP, G.A. & GREEN, H.C. (1947) Visible Speech, Van Nostrand, New York.

PRESTIGIACOMO, A.J. (1962) 'Amplitude contour display of sound spectrograms'. JASA 34: 1684-1688.

PTACEK, P.H. & SANDER, E.K. (1966) 'Age recognition from voice', Journal of Speech and Hearing Research, 9: 273-277.

QUIRK, R., GREENBAUM, S., LEECH, G. & SVARTVIK, J. (1972) A Grammar of Contemporary English, Longman, London.

RAMASHVILI, G.S. (1966) 'Automatic voice recognition', Engineering Cybernetics, 5: 84-90.

REES, M. & URQUHART, A.H. (1976) 'Intonation as a guide to readers' structuring of prose texts'. Work in Progress, Department of Linguistics, Edinburgh University, 9: 19-26.

ROSENBERG, A.E. (1973) 'Listener performance in speaker verification tasks'. IEEE Transactions on Audio and Electro acoustics, AU-21: 221-225.

_____ (1976) 'Automatic speaker verification: a review' Proceedings of the IEEE, 64: 475-487.

SAPIR, E. (1921) Language: an Introduction to the Study of Speech. Harcourt Brace, New York.

_____ (1927) 'Speech as a personality trait'. American Journal of Sociology, 32: 892-905. Also in D.G. Mandelbaum (ed., 1949) Selected Writings of Edward Sapir in Language, Culture and Personality, University of California Press, Berkeley and Los Angeles, pp.533-543.
Also in J. Laver & S. Hutcheson (eds., 1972), pp.71-81.

SCHERER, K.R. (1979) 'Personality markers in speech' in K.R. Scherer & H. Giles (eds.) pp.147-209.

SCHERER, K.R. & GILES, H. (eds., 1979) Social Markers in Speech, Cambridge University Press.

SCHUBIGER, M. (1958) English Intonation, Max Niemeyer, Tübingen.

SEARLE, J.R. (1969) Speech Acts: an Essay in the Philosophy of Language, Cambridge University Press.

SHEARME, J.N. & HOLMES, J.N. (1959) 'An experiment concerning the recognition of voices'. Language and Speech, 2: 123-131.

SMITH, P.M. (1979) 'Sex markers in speech' in K.R. Scherer & H. Giles (eds.), pp.109-146.

SMRKOVSKI, L. (1976) 'Study of speaker identification by aural and visual examination of non-contemporary speech samples'. Journal of Official Analytical Chemists 59: 927-931.

STEVENS, K.N. (1972) 'Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds'. Proceedings of the 7th International Congress of Phonetic Sciences, Montreal, Mouton, The Hague, pp.206-232.

STEVENS, K.N., WILLIAMS, C.E., CARBONELL, J.R. & WOODS, B. (1968) 'Speaker authentication and identification: a comparison of spectrographic and auditory presentation of speech material'. JASA, 44: 1596-1607. Also in N.J. Lass (ed., 1974), pp.259-285.

STEVENS, S.S. & DAVIS, H. (1938) Hearing: its Psychology and Physiology Wiley, New York.

STUART, D.G. & GODFREY, J.J. (1970) 'The specification of individual speech-voice characteristics'. Working Papers, Georgetown University School of Languages and Linguistics, 1: 103-114.

SU, L.-S. & FU, K.S. (1973) 'Automatic speaker identification using nasal spectra and nasal co-articulation as acoustic clues'. Report AEOSR-TR-74-0114. Air Force Office of Scientific Research.

SUNDBERG, J. & NORDSTRÖM, P.-E. (1976) 'Raised and lowered larynx - the effect on vowel formant frequencies'. Quarterly Progress and Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 2-3: 35-39.

SWEET, H. (1906) A Primer of Phonetics, Clarendon Press, Oxford (3rd. edition).

TOSI, O.I. (1975) 'The problem of speaker identification and elimination', in S. Singh (ed.) Measurement Procedures in Speech, Hearing and Language, University Park Press, Baltimore.

_____ (1979) Voice Identification: Theory and Legal Applications, University Park Press, Baltimore.

TOSI, O.I., OYER, H., LASHBROOK, W., PEDREY, C., NICOL, J. & NASH, E. (1972) 'Experiment on voice identification'. JASA, 51: 2030-2043.
Also in N.J. Lass (ed., 1974), pp.286-314.

TRAGER, G.L. (1958) 'Paralanguage: a first approximation'. Studies in Linguistics 13: 1-12.

_____ (1964) 'Paralanguage and other things'. Maitre Phonetique, 122: 21-33.

ULDALL, E.T. (1971) 'Isochronous stresses in R.P.' in L.L. Hammerich, R. Jakobson & E. Zwirner (eds.) Form and Substance; phonetic and linguistic papers presented to Eli Fischer-Jørgensen, 11th February, 1971. Akademisk Forlag, Copenhagen pp.205-210.

VAN RIPER, C. & IRWIN, J.V. (1958) Voice and Articulation, Prentice-Hall, Englewood Cliffs, New Jersey.

VOIERS, W.D. (1964) 'Perceptual bases of speaker identity'. JASA 36: 1065-1073.

WARD, I.C. (1929) The Phonetics of English, Heffer, Cambridge.

WELMERS, W.E. (1959) 'Tonemics, morphotonemics and tonal morphemes'. General Linguistics 4: 1-9.

WILLIAMS, C.E. (1964) 'The effects of selected factors on the aural identification of speakers'. Section III of Report ESD-TDR-65-153, Electronics Systems Division, Air Force Systems Command, Hanscom Field.

WILLIAMSON, J.A. (1961a) An investigation of several factors which affect the ability to identify voices as same or different. Diploma dissertation, University of Edinburgh.

_____ (1961b) 'Further speaker-recognition work', PAT Report for 1961, Phonetics Department, University of Edinburgh, 29-31.

WOLF, J.J. (1972) 'Efficient parameters for speaker recognition'. JASA, 51: 2044-2056.

YOUNG, M.A. & CAMPBELL, R.A. (1967) 'Effects of context on talker identification'. JASA 42: 1250-1254.