

Reliability of Voice Prints

PAMELA S.

Reg.No. M2K16

A Dissertation Submitted in a part fulfillment of
Final Year M.Sc. (Speech and Hearing),
Univeristy of Mysore, Mysore

ALL INDIA INSTITUTE OF SPEECH AND HEARING
MANASAGANGOTTHRI
MYSORE - 570006

May - 2002

DEDICATION

To my

beloved

Savithri Maam

and

Papu

CERTIFICATE

This is to certify that this Dissertation entitled "**RELIABILITY OF VOICE PRINTS**" is a bonafide work in part fulfillment for the degree of Master of Science (Speech and Hearing) of the student (Register No. M2K16).

Mysore,
May, 2002



Director

All India Institute of
Speech and Hearing,
Mysore - 570 006

CERTIFICATE

This is to certify that this Dissertation entitled "**RELIABILITY OF VOICE PRINTS**" is a bonafide work in part fulfillment for the degree of Master of Science (Speech and Hearing) of the student (Register No. M2K16).

Guide



Dr. S.R Savithri

Reader & Head in-charge,
Department of Speech - Language Science,
All India Institute of Speech & Hearing,
Mysore - 570 006

Mysore,

May, 2002

DECLARATION

This Dissertation entitled "**RELIABILITY OF VOICE PRINTS**" is the result of my own study under the guidance of Dr. S.R.Savithri, Reader & Head in-charge, Department of Speech - Language Science, All India Institute of Speech and Hearing, Mysore and not been submitted earlier in any other University for the award of any diploma or degree.

Mysore,

May, 2002

Reg.No. M2K16

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to my guide Dr. S.R.Savithri, Reader and Incharge of Speech-Language Sciences, for all her help and guidance.

I would like to thank Dr. M. Jayaram, Director, All India Institute of Speech and Hearing, for allowing me to conduct this Dissertation.

I would like to thank my parents for everything.

I would like to thank Papu, without you life would have been really boring. Thanks for all your support.

I will like to thank Sridevi maam, for all her help and guidance.

Thanks to Yeshoda maam.Rohini maam and Santhosh for all their help.

Thank you Ananthi, you had been a great inspiration for me for all these two years.

Thanks a lot Beula, for all your useful advice.

Hey, Scratchy girls, thanks a lot for all the nice moments that we spent together.

Special thanks to Mukesh, Siddhartha, Sharad and Mukunthan.

Thanks to Karthik, Yatin and Uday for being my subjects.

Many thanks to my classmates for all their help and support.

Thanks to the Library staff for their co-operation.

I will like to thank Manjula maam and Madhusudhan Sir, for their wonderful typing work.

Lastly I will like to thank Shivappa Sir for Xeroxing and Binding work.

TABLE OF CONTENTS

			Page No
1	Chapter I	Introduction	01-04
2	Chapter II	Review	05-102
3	Chapter III	Method	103-105
4	Chapter IV	Results & Discussion	106-142
5	Chapter V	Summary & Conclusion	143-145
6		BIBLIOGRAPHY	146-159

LIST OF TABLES

Table No.	Description	Page No.
1	Sources of between speaker variations.	102
2	Material for study.	103-104
3	F2, onset of burst and frication Noise in all the Subjects (Hz).	106
4	CD, TD, TDF2 in all the subjects in (msec).	107
5	Percent times the values were same.	107
6	F2 values in 29 words.	108-112
7	Shows the significant difference between F2 values.	113
8	Mean, S.D, Minimum and Maximum value of onset of burst (Hz).	114-116
9	Significant difference between onset of burst.	117
10	Mean, SD, Maximum and minimum frequency of frication onset (Hz).	118
11	Significant difference between onset of frication Noise.	119
12	Mean, S.D, Minimum and Maximum total duration (msec).	120-122
13	Significant difference between total duration.	123
14	Mean, SD, Minimum and Maximum closure duration (msec).	125-127
15	Significant difference between closure duration.	128
16	Mean, S.D, Minimum and Maximum TD,F2 for 29 words (msec).	129-133
17	Significant difference between TDF2.	134
18	Shows a summary of inter-intra subject differences across all the words.	135

LIST OF FIGURES

Figure No.	Description	Page No.
1	Overview of model of sources of between speaker differences.	34
2	The segmental strand.	48
3	The suprasegmental strand.	60
4	Integration rules, phonetic representation, implementation rules and physical constraints.	67
5	Schematic impression of part of a phonetic representation.	76
6	Schematic representation of the use of a prestige value of a sociolinguistic variable according to class and style.	
7	Shows the spectrograms of the word/bacna /for all the six subjects.	136
8	Variations in closure duration.	137
9	Range of F2 values of different vowels.	137
10	Variations in total durations of phonemes across words.	138
11	Variations in burst frequency range across plosives.	138
12	Variations in F2 transition duration across words.	139
13	Variations in frication noise across words.	139

CHAPTER I

INTRODUCTION

Way back in 1944 Gray & Kopp had coined the term "voice print" in a report discussing the identification of speaker by visual inspection of spectrograms and concluded that this method seemed to offer good possibilities. They had aimed at helping the military. After the World War II got over there was no need of voice print as such. Again in 1962 Kersta had reexamined Voice Print and he had claimed that spectrograms of several utterances of the same words by a given speaker always contain more similar spectral features than those produced by different speakers. According to Kersta speaker recognition by visual inspection of spectrogram consists of subjectively matching similarities found in pairs of spectrograms from the same person that are not found in pairs of spectrograms from different persons.

Subjective matching of eye-brain similarities found in pairs of spectrograms obtained the term as "Voice Printing". The spectrograph was commercialized by Kersta under the trade name of "Voice Print". The "Voice print" had a play back facility allowing continuous listening of samples prior to be processed. Speakers read the selected words or phrases, then spectrograms of different words are prepared from them and are presented to a trained observer who attempts to determine whether some utterances were produced by a given speaker. The observers try to perform a recognition task; they match

spectrograms that represent **the** same speaker and are instructed to examine features such that : similar mean frequencies of vowel formants, formant bandwidth, gaps and type of vertical striation, slopes of formants, durations, and characteristic patterns of different speech sounds.

There have been various methods of speaker identification. The classification of these methods according to (Hecker 1971) is as follows:

- a) Speaker identification by listening
- b) Speaker identification by machine
- c) Speaker identification by visual examination of spectrograms.

All of these three procedures are based on the assumption that inter-speaker variability is always greater or different than intra-speaker variability, regardless of parameters involved in these variabilities. To prove scientifically that inter-speaker variability is greater or different than intra-speaker variability is by inference. An inference thus derived might be affected by effects both from speakers and from the method of identification used. Speaker identification by listening only is far from being 100% accurate. It is a subjective method, an expert using only that method would be unable to justify his conclusions. The task of comparing voices purely by listening becomes a difficult one when several speakers are involved. In this case, the method necessitates that the examiner relies a great deal on auditory memory.

The second method is speaker identification by machine, is less accurate or developed than any other method, involving human examiners. In the future hard research might bring knowledge to overcome the present limitations of speaker identification by machine. It is quite difficult to predict just when or even if totally reliable machines of voice identification will become available. Even if such a machine were available the human expert, trained in phonetics, spectrography, and related areas, would be required to select the proper samples from the unknown and the known voices to feed the machine output and possibly to check the results by using an alternative method.

The third method of speaker recognition is based on the visual examination and the comparisons of spectrograms. Here the observer has to compare the different spectrograms and has to report that whether the spectrogram are same or different.

The spectrogram portrays three main parameters of speech; time (on the horizontal axis), frequencies (on the vertical axis), and relative amplitude (degree of darkness of different spectrographic regions) Each of the isolated phoneme, word or phrase is correlated with a characteristic spectrographic pattern. The general aspects of patterns corresponding to different utterances of the same word are similar in such a way that a person specially trained in "reading" spectrogram, who also knows the "Statistics" of the language, would determine with more or less accuracy which words or phrases were portrayed by a particular spectrographic pattern. However the inter-speaker and intra-speaker

variabilities are also portrayed by the spectrographic patterns. Spectrograms of different utterances of the same word or phrase by the same or by different speakers are never exactly alike. The examiner selecting samples of the voices to be spectrally compared must first listen to the samples in order to properly label the spectrograms.

At the All India Institute of Speech and Hearing, forensic evaluations have been conducted in the past. However, it is not known as to what percent of matching would indicate similarity/dissimilarity of speakers. In this context, the present study was planned. The aim of the study was to find out the reliability of spectrograms for forensic evaluation. Specifically intra and inter subject reliability would be measured for acoustic parameters in six subjects.

CHAPTER II

REVIEW OF LITERATURE

A persons voice is a complex acoustic signal which encodes various kinds of information, among them a reflect of some of the anatomy and physiology of the speaker [Corsi, 1982].

The idea that someone could be identified by the sound of his voice had its origins in the work of Alexander Melville Bell. Alexander Melville Bell, developed a visual representation of the spoken word. This visual display of the spoken word conveyed much more information about the pronunciation of that word than the dictionary spelling could ever suggest. His depiction of speech sounds demonstrated the subtle differences with which different people pronounced the same words. This sort of speech sound analysis developed by Bell is the phonetic alphabet, which he called "visible speech". His method of encoding the great variety of the speech sounds was by handwritten symbols and was language independent. This code produced a visual representation of speech which could convey to the eye the subtle difference in which words were spoken. This system was used by both Bell and his son, Alexander Graham Bell, in helping the hearing impaired population to learn to speak.

It was in the early 1940's that a new method of speech and sound analysis was developed. Potter, Kopp, and Green working for Bell Laboratories in Murray Hill, New Jersey, began work on a project to develop a visual representation of

speech using a sound spectrograph. This research was intensified during World war II when acoustic scientists suggested that enemy radio voices could be identified by the spectrograms produced by the sound spectrograph. The war ended before the technique could be perfected.

A voice print models physiological characteristics of a particular persons voice and can be used to authenticate that person's identity. Authentication against a voice- print is based on inherent properties of the speakers voice, so it provides a higher level of security than prompting for a password or personal identification number. The recognition of individuals from their speech is an area of speech science which reliably arouses public attention. Interest in scientific controversies is always greatest when the issue concerned has direct practical consequences in every day life. Thus, interest in speaker recognition peaks when voice recordings seem to offer the hope of identifying the perpetrator of some well publicised crime. But, however pressing the practical needs, understanding of the bases of speaker recognition has remained primitive; and because of this; attempts to satisfy those needs are fraught with danger (Nolan, 1983).

The notion that an individual has 'a voice' by which he can be recognised is a natural one, given on "day-to-day" experiences of successfully recognising people by their speech alone- typically over the telephone. It is so natural that it was adopted by many speech scientists without fundamental scrutiny, with the result that the usual questions posed was not whether individual could be

uniquely recognised from their voices, but how this recognition could be most effectively and reliably carried out in an objective way (Nolan, 1983).

The kind of activity covered by the term speaker recognition is conceptually straight forward, and definition abundant. Hecker (1971) suggests that speaker recognition is any decision making process that uses the speaker dependent features of the speech signal and Atal (1976) offers the formulation any decision making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance.

As pointed out by Brown (1982) different aspects of the identity of an individual may be successfully accessed as a result of the matching process between an input voice stimulus and a stored reference voice-aspects such as the individual's name, physical appearance, or description (e.g. role in society). The everyday process of recognising a speaker from a voice sample involves these aspects to greater or smaller degree- it is possible simply to recognise a voice as familiar, but not recall details of its producer (if these were indeed ever known); to associate the voice with a description (e.g. the telephone receptionist); and so on. Clearly any speaker recognition task (apart from, in human terms, the simple question have you heard this voice before?, or in machine terms, 'does this input voice sample match one of a number of stored, but unlabelled, reference pattern?') involves accessing some sort of identity characteristics. These various processes will be treated as logically subsequent to, and therefore secondary to,

the initial decisions process which confirms or denies that two voice samples were the product of the same vocal apparatus (Nolan, 1983).

The definitions of speaker recognition above leave unstated the linguistic levels at which speaker recognition may exploit speaker specific information. Syntactic and lexical clues to identify are undoubtedly frequently present in utterances, and are clearly worthy of exploration for speaker recognition (Nolan, 1983).

Types of speaker recognition:

Under the overall heading of speaker recognition, it is necessary to distinguish a number of distinct fields of study. Bricker and Pruzan sky (1976) recognise three major divisions: speaker recognition by listening, by machine, and by visual inspection of spectrograms (SRL, SRM and SRS). In this categorisation, SRL involves the study of how human listeners achieve the task of associating a particular voice with a particular individual or group, and indeed to what extent such a task can be performed. SRM encompasses the attempts to develop automatic and semi-automatic strategies, standardly computer based, for associating voices with speakers; SRM is therefore often thought of an objective in comparison with SRL because of its relative freedom from human decision making. The third category, SRS; comprises efforts to make decisions on the identity or non-identity of voices on the basis of visual examination of speech spectrograms by trained observers. The importance of this type of work stems

from its practical application; since the mid 1960's there has been a continuing and heating debate as to whether visual spectrographic evidence should be admitted as legal evidence, and if so, what its status should be (Nolan, 1983).

There are, however, reasons for preferring a two fold division of a slightly different nature. The characteristic of SRL, as investigated by most studies, which sets it apart from all other types of speaker recognition is not so much the fact that the recognition is performed by listening, but rather that it is performed by untrained observer in real-life (or experimentally simulated real-life) conditions. On the other hand SRM and SRS both involve the application of analytic techniques to the problem, whether humanly acquired or automatically programmed (Nolan, 1983).

There are further consideration favouring a two fold categorisation-technical speaker recognition and naive speaker recognition. Firstly, the division between automatic methods and SRS is contingent resulting from the history of the methods concerned, rather than essential in the way that the distinction between technical and naive speaker recognition is given an accurate (probably computer-based) spectrograph, it should be possible for an observer to make reliable measurements on the given spectrogram which he could then use as input to objective decision strategies. This is similar to the kind of semi-automatic recognition strategy developed, for example, by Broderick, Paul and Rennick (1975) where a human operator selects specific speech events, by visual observation, as input to statistical decision procedures. A continuum of potential

methods exists, therefore, with technical speaker recognition, whereas the division between technical and naive speaker recognition is a fundamental one based on the two recognition situations (Nolan, 1983).

Secondly the traditional three fold categorisation does not readily provide a place for technical speaker recognition by listening -that is, the application of auditory techniques acquired through phonetic training to making decisions about the identity of speech samples. This approach to recognition is quite different from the recognition processes which are normally studied under the SRL heading. The latter involve decisions made on the basis of largely subconscious generate impressions about the similarity or dissimilarity of given speech samples; on the other hand the phonetician engaged in a speaker recognition task (Baldwin, 1977) is not concerned with general impressions unless they are supported by phonetic description, and is all the time applying a detailed system of analysis. The tendency in discussions of speaker recognition techniques is not to draw any distinctions within speaker recognition relying on aural capabilities. The result of this, even with writers who are aware of the limitations of visual inspection of spectrograms, is an underestimation of the relative value of careful auditory analysis compared with spectrogram observation for example Tosi (1975): "Typically all types of aural examination of voices and visual examination of speech spectrogram are considered subjective methods, although the latter is closer to the objective part of the spectrum of methods than the former". A generalization of this kind is not possible without specifying exactly

the degree and kind of analysis implied in the aural and the visual examination. In short, a categorisation of speaker recognition tasks is proposed which is based on whether only normal everyday human abilities are exploited or whether specialised techniques - aural, visual or electronic - are brought to bear (Nolan, 1983).

Identification and Verification :

Within technical speaker recognition a distinction is generally drawn on the basis of assumptions under which decisions about speakers identity have to be made. In the real world task of speaker verification (or authentication), and its experimental stimulations, an identity claim by an individual is accepted or rejected by comparing a sample of his speech against a stored reference sample spoken by the individual whose identity he is claiming, and making a decision on the basis of a predetermined similarity threshold. Speaker verification have applications in security checking e.g. where it may be desired to establish the identity of a person seeking admittance, or in banking, where an automated money dispenser might test the voice of the customer wanting to withdraw money against a sample of the voice of the owner of the account in question. Speaker verification involves the comparison of a test sample of speech with a reference sample from just one speaker, requires a preset similarity threshold, and usually yields one of four kinds of decisions correct acceptance, correct rejection, false acceptance, false rejection (although a 'no decision' response may also be permitted). The relative acceptability of one or other kind of errors

determines the tolerance at which the similarity threshold will be set-a system which cannot be permitted to accept impostors will almost certainly reject true identity, claims from time to time. The assumptions underlying speaker verification tasks are that both test and reference samples will be from cooperative speaker, so that vocal mimicry on the part of an impostor, but not vocal disguise on the part of the 'true' speaker, may be encountered; and that the utterance type(s) on which verification is to be performed may be specified (Nolan, 1983).

In speaker identification (and elimination) an utterance from an unknown speaker has to be attributed, or not, to one of a population of known speaker for whom reference samples are available. Speaker identification is usually considered to include the kind of recognition which forensic work entails - a sample of speech recorded during the commission of, or constituting a crime must often be compared with samples of speech from a number of suspects. Here the number of decisions increases with the size of the reference population; and the cost, in practical applications, of errors of identification or elimination is so high as to necessitate a 'no decision' option. It is necessary to assume the possibility of attempted disguise in the test or reference samples; and the same utterance type may not be available in both test and reference samples (Nolan, 1983).

Under speaker identification three types of recognition test can be carried out: Closed tests, Open tests, and discrimination tests (Tosi, 1979). In a closed

test it is known that the speaker to be identified is among the population of reference speakers, whilst in an open test, the speaker to be identified may or may not be included in that population. Thus in the closed test, only an error of false identification may occur, whilst in open tests there is the additional possibility of incorrectly eliminating all the reference population when in reality it included the test speaker. In a discrimination test, the decision procedure has to ascertain whether or not two samples of speech are similar enough to have been spoken by the same speaker; errors of false identification and false elimination are possible (Nolan, 1983).

It is apparent that an open test is simply an interactive discrimination test, in which the test sample undergoes a discrimination test with each of the reference samples in turn; and that in both open and discrimination tests some form of acceptance threshold is required. In the closed test such a threshold is not needed as the 'nearest' reference speaker is automatically selected.

It is also apparent that speaker discrimination most closely resembles speaker verification in the nature of its decision problem - a point, which seems to have escaped comment. In both tasks a test sample and a reference speaker is automatically selected.

It is also apparent that speaker discrimination most closely resembles speaker verification in the nature of its decision problem - a point which seems to have escaped comment. In both tasks a test sample and a reference have to be

evaluated, and designated as produced by the same or different speakers, according to an acceptance threshold. As far as the nature of the decision problem is concerned, the usual forensic situation should be classed as a type of speaker verification-typically an incriminating sample has to be attributed, or not, to a suspect. The fact that it is universally dealt with under the heading of identification (Bolt et al 1979; Tosi 1979) has to do with the circumstantial characteristics associated with the two categories of recognition - the fact that, as mentioned above, lack of co-operation, and disguise attempts, may be expected in the two categories of recognition- the fact that, as mentioned above, lack of co-operation, and disguise attempts, may be expected in the forensic case; in contrast to, for instance, access control, where genuine claimants can be expected to be co-operative, but impostors attempting mimicry must be guarded against (Nolan, 1983).

Experiments assessing the value of the particular parameters for speaker recognition have most frequently adopted the closed test design. The reason for this is not that this design best approximates real life applications - it is in fact the one least likely to occur in forensic cases- but rather that it gives the most straight forward comparison of parameters without the complication of choosing a threshold by Atal (1976) "Both specific recognition identifications and verification, have been investigated in past experimental studies. Of the two, the identification task is the more suited for comparing the performance of different parameters. In [closed test] speaker identification a single error rate can provide

a measure of the performance, while in speaker verification, two kinds of errors namely, the probabilities of false verification and false rejection as functions of a threshold parameter, determine the performance. Also, the identification accuracy is a more sensitive indicator of the ability of a parameter for discriminating speaker" (Nolan, 1983).

Auditory identification by phoneticians:

In the United Kingdom evidence produced in courts of law to establish speaker identity has been almost exclusively auditory. Widespread press coverage was given to a case in Winchester Magistrates Court (November 1967) where a man was convicted of making five hoax calls. The coverage implied that spectrographic evidence, voice pictures, had constituted crucial evidence. However, it appears that in fact the phonetician called as an expert witness by the prosecution based his opinion on auditory judgements, and produced spectrograms in court only in response to a request from the prosecution to present relevant speech samples in visual form; and so the case was not fully comparable to those in the USA where spectrograms had been used as the primary means of identification.

Considerable alarm was felt among phoneticians in the UK test, despite a lack of theoretical justification and empirical validation of the techniques, a precedent to be set for the use of evidence based on spectrograms, and this alarm was voiced in, for example, a letter drafted by Trim and signed by the majority of

phoneticians in the UK. A copy was sent to the Home Secretary, and in Scotland contact was made by Anthony with the hard Advocate, explaining with supporting evidence phoneticians disquiet at speaker identification based on spectrograms (Nolan, 1983).

For a number of years, however, there has been a practice of calling on phoneticians and others considered to be competent in auditory analysis of speech to assist the police in investigations, and to appear in court in the role of expert witness to give opinions on speech samples.

Little explicit discussion, with the exception of Baldwin (1977), and certainly nothing detailed or comprehensive, has been published on the methods employed by those phoneticians who have undertaken such work. They have worked largely as individuals, without co-ordination, further more presenting their evidence with varying assessments of the general reliability of the technique. It seems, however, that the methods used are essentially those of the traditional dialectologist; noting detailed realisational differences of elements (both segmental and supra-segmental) of the phonological system, and differences in the system itself; by repeated listening, and analysis according to the established auditory/articulatory phonetic framework of classification (Nolan, 1983).

Whilst it seems reasonable to assume that trained listeners with an analytic framework for speech at their disposal should be able to offer more

reliable auditory judgements in speaker recognition than untrained listener, a number of factors have caused such applications to be gravely questioned in the phonetic community, and the issue is currently highly controversial. Among these factors are the following :

In the absence of an integrated theory of the origins and nature of speaker dependent characteristics used by phoneticians, and the extent to which they may vary in the speech of an individual, opinions on the reliability of the technique are prone to be based on incomplete information. Secondly, phonetic training does not train the listener to set aside the default human ability to normalise across speakers - the ability which enables him to hear as the same sounds from different speakers which are adjectively acoustically distinct: it might be therefore, that a decision made by the phoneticians principally on the basis of phonological factors would be altered, or at least given different weight, if supplemented by objective acoustic information. Associated with this is the problem that whilst the ideal of phonetic training is to free the phonetician's perception totally from the habits and biases ingrained by experience of his native language(s) and accent(s), it is unlikely that this ideal state is ever achieved; consequently a phonetician's sensitivity to fine distinctions between speakers is in practice likely to be highly correlated with his familiarity with the accent of the speakers. Thirdly there are no commonly agreed methods of listening and analysis, allowing potentially great inconsistency across cases; further, no specified professional qualification or standard of proficiency is

required before a person may offer an expert opinion. Above all, there has been a lack of empirical research directed to demonstrating the reliability or otherwise of this method of speaker identification (Nolan, 1983).

In 1978, the Colloquium of British Academic Phoneticians, prompted by concern at instances of its members and others being called upon to give opinions in court on speaker identity, set up a committee to report on forensic application of phonetics. A survey of phoneticians conducted by the committee elicited a variety of views on auditory speaker identification by phoneticians in legal cases, most replies stressing that at the very least the limitations of the technique need to be made clear before opinions are given. At the 1980 colloquium a special session on the topic revealed considerable disagreement over the weight that should be attached to evidence given by phoneticians - disagreement understandable, but less than fruitful, in the absence of empirical research. The motion that phoneticians should not consider themselves expert in speaker identification until they have demonstrated themselves to be so was carried by 30 votes to 12 with eight abstentions. This motion clearly expressed the need for scientific evaluation of phoneticians' auditory judgement in speaker identification, and prompted the setting up of a project in the area.

This project is being carried out by Marion Shirt at Leeds University, and is directed specifically to the question of whether phoneticians do in fact perform better than untrained listeners in a number of speaker recognition tasks.

The first experiment took the following form. Studio quality recordings were made of pairs of male speakers discussing similar pictures out of sight of each other, their task being to decide whether the pictures were identical. Voice samples of approximately five seconds duration were excerpted and grouped, the different groups containing voices of various degrees of accent homogeneity. The task comprised of six closed identification tests, in which a test sample, had to be matched to one of the six references; a closed test where 10 samples from a total of five speakers had to be matched; a closed test, where a match known to exist among 10 samples had to be found; and two open tests, in which the listeners had to decide if any matches existed among 10 samples. Three discrimination tests were included using samples of around 20 second duration.

Phoneticians and phonetically naive subjects took part in the experiment. Preliminary indications are that whilst the phonetician did on average achieve better accuracy than the non phoneticians (53% compared with 46%) even the best performance of phonetician (76%) fell well short of 100% accuracy; and the group of phoneticians as a whole exhibited a wide range of performance (down to 38%), as did the non-phoneticians (Nolan, 1983).

Two kinds of limitations in the experiment should be noted. Firstly relating to the condition of the experiment, although both groups of listeners were allowed unlimited time to make their decisions, the naive subjects had in practice to complete the task in an afternoon, whilst the phoneticians could spread their listening over a longer period in some cases totaling many hours of

listening, and also the naive subjects were provided with twin cassette players, whereas the phoneticians were allowed to use listening facilities of their choice (e.g. tape loop repeaters). Secondly relating to the task, the five second samples were too short to permit systematic phonetic and phonological comparisons to be made between samples, and thus precluded the phoneticians bringing to bear many of the strategies they would standardly employ when assessing the similarity of speech samples. These limitations notwithstanding the results of the study will be of use in evaluating auditory identification by phoneticians, and will serve as a starting point for further much needed research into the reliability of the technique (Nolan, 1983).

Voice Print identification :

The term voice-print was promoted by Kersta (1962) who argued the parallelism of spectrograms and fingerprints. Kersta (1962) cited, in support of his claim that spectrograms could be used for speaker identification, an experiment in which high school girls were trained in spectrogram reading and then presented with spectrograms of 10 frequently occurring monosyllables. Tests in which these examiners were given a matrix of few voiceprints for each speaker and then had to sort test utterances into piles for each speaker (closed identification), were carried out for populations of five, nine and twelve males, yielding promising 99.6%, 99.2% and 99% identification rates respectively. When words exempted from the context of a cue sentence instead of spoken in isolation were used, the deterioration in the lumped error rate was merely from

0.8% to 1%. It might be inferred that the very high identification rates indicate optimum conditions for speaker recognition, Kersta's account lacks details of the procedure, and so it is not clear that the margin by which his results exceed those of other experiments did not result from, for example, a less rigorous choice of speakers from the point of view of dialect variation (Nolan, 1983).

Young and Campbell (1967) set out to examine the effect of taking the words on which visual spectrographic identification might be based from the context of a sentence. They used some of the same words as Kersta and had five speakers record them, both in isolation and embedded in sentences. They trained 10 observers, all familiar with spectrograms, pointing out possible 'unique clues' to speaker identity such as the frequency, intensity and bandwidth of the formants, and the regularity of the vertical striations as an indication of the melodiousness of the voice. It was Young and Campbell's thesis that if 'unique clues' to speaker identity did exist, the level of identification performance for words in differing context should be similar to the level for words spoken in isolation. The results showed that observers had much greater difficulty identifying speakers by means of words taken from a sentence context than from words spoken in isolation, the respective rates being 37.3% and 78.4%. This is in considerable contrast to Kersta's (1962) difference of 0.2% for the two contexts. There is also an appreciable discrepancy between error rates in the comparable task with a five speaker population and words spoken in isolation, where Kersta obtained 99.6% to Young and Campbell's 78.4%. This discrepancy may well be

accountable for in terms of the speakers used in the two studies, as Young and Campbell choose speakers who were quite homogenous with respect to sex, dialect, age and education.

In an attempt to assess the artificiality of using data from read sentences, Hazen (1973) used as his data words extracted from spontaneous speech obtained in interviews with 60 males, and then used spectrograms of these words in open and closed identification trials. The observer were given a 'file-card' for each speaker in the population which consisted in two examples of the word in question, chosen as the visually least similar of the examples available. Identification was carried out in two strategies: reduction of the population to 'suspects' and positive identification and elimination. The test word came from the same context, as one of the file card examples, or from different context; these two conditions providing correct identification rates of 57.4% and 16.8% respectively. Hazen concludes that 'given the condition of this study, accurate identification of speaker by visual comparison of spectrograms is not possible' - a conclusion that has serious implication for the forensic application of the technique, where spontaneous speech is usually involved (Nolan, 1983). The most extensive of the investigations carried out with the intent of checking Kersta's claims and estimating the validity of such procedures in forensics was that of Tosi et al (1972). The experiments extended over a two year period, used recordings from 250 speaker randomly selected from a population of approximately 2500 male students at Michigan University, and involved 34996

trials of identification performed by 29 examiners with a months training. They were asked to grade their degree of confidence in each decision on a few point scale.

Although the large number involved in the investigation appear to lend it an impressive scale, and lead to it being frequently cited as if it gave definitive evidence on 'voiceprinting' caution is needed in its interpretation. Hollien (1977) gives a reminder that identification trials were carried out on subsets of between only 10 and 40 speakers drawn from the 250 for whom recordings were available, and Thomas (1975) points out that if the 250 speakers were chosen by a successfully random selection procedures, they would constitute a 'heterogenous group representatives of all elements comprising the population', whereas it would be more relevant to establishing the reliability of speaker identification if the speaker were as homogeneous as possible with respect to accent. He also draws attention (1975) to the fact that the 'continuous speech' in the experiment consisted of readings of 'nonsensical' sentences containing the nine key words *it is on, me, and the, I, to, you*, it is far from obvious that read non sense bears a close relationship to meaningful spontaneous speech (Nolan, 1983).

The experiments included investigation of the effect of using non-contemporaneous reference and test samples, as well as the open/closed nature of the test and the context from which the compels were taken. Overall, the tests, which best replicated the forensic situation (open tests with non-

contemporaneous samples taken from 'continuous speech') yielded 6-4% false identification and 12.7% false elimination. It was argued that as 60% of wrong answers (though also 20% of correct answers) were graded 'uncertain', had the examiner had the option of expressing no opinion when in doubt, false identification errors would have been cut to 2.4% and false elimination to 4.8%. These results together with those of a field study conducted for Michigan State Police to discover the relation between laboratory experiments and the actual situation a professional examiner encounter when handling forensic situations (Tosi, 1975). This led Tosi to the opinion that, if certain conditions are fulfilled, identification by visual examination of spectrograms can offer reasonable reliability (Tosi, 1975). These conditions specify that visual examination should be combined with listening; examiners should be qualified, including a training in phonetics and a two year apprenticeship in field work; they should avoid positive conclusions if the slightest doubt exists; and they should be entitled to ask for as many samples of speech, and as much time, as is needed.

Tosi's at least qualified approval of 'voice-printing' as a means of establishing a speaker's identity contrasts with the unqualified championing of the technique by Kersta : "Voice print identification is a method by which people can be identified from the spectrographic examination of their voice. Closely analogous to fingerprint identification, which uses the unique features found in people's finger prints, voice print identification uses the unique features found in

their utterance (1962), experiments showed that professional ventriloquists and mimics cannot create voices or imitate others without revealing their own identities (1962)".

The fragility of the specific evidence associated with such claims is well illustrated by Ladefoged and Vanderslice (1967), who include a critical representation of the voiceprints on which Kersta based a positive identification in a case in California (People vs King). Not only are the claimed points of similarity between pairs of spectrogram often highly dubious, but, as Ladefoged and Vanderslice point out the evidence even includes a blatant and basic errors of miss labelling in the case of one of the spectrograms used (Nolan, 1983).

Nevertheless, faced with an increasing need to identify speaker from recordings, a number of states in the USA, including Michigan and California, began to accept evidence based on voice prints, a move which brought forceful protest from phoneticians and speech scientists (Vanderslice, 1969, Bolt et al, 1970, Hollien, 1974a). The objections to the use of voice print techniques may be classified into three kinds concerning the interpretation of laboratory assessment, the procedures of decision making, and (most fundamentally) the nature of the information on which those decisions have to be based (Nolan, 1983).

The interpretation of results from laboratory trials is confounded by the conflicting identification rates found by different experimenters. It is clear,

however, that none of the experimenters who have sought to replicate Kersta's original experiments have achieved such high rates. Secondly few of the investigations have concluded the kind of trial, which most closely approximates the common forensic situation, namely the discrimination test. A significant way in which forensic conditions differ from those of the laboratory investigations is the quality of the recordings, which may be available. In practical applications it is likely that low quality equipment will have been used to record a speech signal transmitted through the telephone network from an unknown and perhaps noisy place. The characteristics of the total transmission system are most unlikely to be recoverable in detail, and so its distorting effects and the effects of the various noise sources are irretrievably confused with the speech signal itself. Attempts have been made to use voiceprint methods on a recording which was of such poor quality as to be virtually useless' and in which 'the speech during several parts of the conversation was unintelligible (Hollien, 1974).

The first question to be asked about the procedures entailed in voiceprint identification is whether the visual examination of speech samples gives more accurate result than aural examination. A priori, it might be expected that the human ear, inherently suited to the communication mode which its capacities have helped to shape, and which has been practised in speech skills throughout the observers life, should be more acute than the eye, trained at most for a few years at an unnatural task. On the other hand perhaps the ear is most adopt at achieving the converse of speaker identification - ignoring speaker dependent

information which can be regarded as noise with respect to the linguistic message, and allowing conscious appreciation primarily of that message. The experimental evidence however points strongly to the conclusion that aural identification is more successful than visual. Young and Campbell (1967) point out that the results they obtained for visual identification were worse than comparable results in Bricker and Pruzansky (1966), who investigated the ability of untrained listeners to identify the speakers of utterances having various content and duration. They concluded that humans can extract more relevant information from the unprocessed acoustic signal than they do from a visual representation (Nolan, 1983).

This indirect conclusion is supported by the work of Stevens et al (1968) who compared aural and visual strategies directly. Their judges had to perform a series of open and closed tasks, identifying speakers from samples of their speech presented either aurally through headphones, or visually as spectrograms. The error rates were found to be about 6% for aural presentation, and about 21% for visual. Only within the verification task, as opposed to the identification task, has the ears capacity for speaker recognition been surpassed, as demonstrated by Rosenbeg (1973); and there significantly, by an automatic verification scheme not human inspection of voiceprints. So it seems clear then, that the voiceprint procedure can at best complement aural identification, perhaps by highlighting acoustic features to which the ear is insensitive, and at worst it is an artifice to

give a spurious of 'scientific' authority to judgements which the layman is better able to make (Nolan, 1983).

The other major cause for concern relating to the procedures of voice-printing stems from their subjective nature. Tosi (1975) concedes that the crucial problem with subjective methods of testing the honesty and reliability of the examiner, and it is easy to suspect that a voiceprint examiner who is employed for his ability to identify and eliminate speakers would be tempted to make a positive decision on inadequate evidence if faced with a whole series of cases where a 'no decision' response was appropriate. The concern is the more acute for attempts, in the face of strong opposition from phoneticians, by a self appointed set of voice-printers to gain monopoly (in USA) over court testimony; Hollien (1974) opines that "it would appear that, if the proponents of voiceprints are successful, a subculture would develop expressly for the judicial system, where only certified professional examiners would certify in the courts of law. Further, since presumably they would be the only individuals empowered to certify new examiners, as uncertified scientist, no matter how distinguished and well regarded by his peers, simply could not qualify to testify without their approval".

Successful contesting of voiceprints evidence in the 1970s by prominent phoneticians and speech scientists such as Ladefoged and Hollien to some extent checked such a development and led to reversals in a few states, including Michigan and Pennsylvania, of earlier rulings which had admitted voiceprint

give a spurious of 'scientific' authority to judgements which the layman is better able to make (Nolan, 1983).

The other major cause for concern relating to the procedures of voice-printing stems from their subjective nature. Tosi (1975) concedes that the crucial problem with subjective methods of testing the honesty and reliability of the examiner, and it is easy to suspect that a voiceprint examiner who is employed for his ability to identify and eliminate speakers would be tempted to make a positive decision on inadequate evidence if faced with a whole series of cases where a 'no decision' response was appropriate. The concern is the more acute for attempts, in the face of strong opposition from phoneticians, by a self appointed set of voice-printers to gain monopoly (in USA) over court testimony; Hollien (1974) opines that "it would appear that, if the proponents of voiceprints are successful, a subculture would develop expressly for the judicial system, where only certified professional examiners would certify in the courts of law. Further, since presumably they would be the only individuals empowered to certify new examiners, as uncertified scientist, no matter how distinguished and well regarded by his peers, simply could not qualify to testify without their approval".

Successful contesting of voiceprints evidence in the 1970s by prominent phoneticians and speech scientists such as Ladefoged and Hollien to some extent checked such a development and led to reversals in a few states, including Michigan and Pennsylvania, of earlier rulings which had admitted voiceprint

objectivity than aural or spectrographic identification. But the most fundamental of the objections to voice print identification, based on the nature of the information is speech signal, will still be damaging as long as the observation of Bolt et al (1973) holds true. It is appropriate to turn to a detailed examination of how the speech signal may be differentially determined by individuals.

The bases of between speaker differences:

Familiarity with what has been written on speaker recognition would, by itself, give the impression that no problems exist in understanding the origin of between speaker differences in the speech signal - the task being merely how to extract information from such differences so as to be able reliably to identify speakers. It will however, become clear that a lack of concern for the complexity or through over reliance on even greater technological and statistical sophistication, leave those who advocate the practical application of speaker recognitions schemes open to serious theoretical criticism (Nolan, 1983).

The widely accepted model of between speaker difference divides them into categories according to whether the aspect of speech production underlying them is a structural one or a functional one, that is, whether the difference derives from the shape, size, and inherent dynamic limitations of the speaker's vocal apparatus or rather the manner in which he manipulates it. Glen and Kleiner (1968) commence by stating "Acoustic parameters of speech reflecting speaker identity must be derived either from the physiological characteristics of the

speaker vocal apparatus or from idiosyncrasies in his manner of speaking and this is echoed by, among other, Wolf (1972)". Differences in voices stem from two broad bases: organic and learned differences, Atal (1976), "Speaker related variations in speech are caused in part by anatomical differences in the vocal tract and in part by the differences in the speaking habits of different individuals", Bricker and Pruzansky (1976) "specifically, speaker information is latent in the speaker in the form of anatomical features and neurally stored habit patterns".

The model of between speaker differences outlined above has been rejected by (Nolan, 1983). Firstly, the plasticity of the vocal tract means that in few, if any, cases, does a given organic feature leave an invariant imprint on the acoustic signal. Whilst it is true to say for example, that there is considerable between speaker variation in the size and mass of the vocal folds, and that this has a determining influence on the fundamental frequencies used by a speaker, the determination is by no means absolute. There may be a physiologically determined maximum and minimum to a given speaker fundamental frequency range; and his preferred range may in some sense be the optimal one given his particular larynx; but he nevertheless has at his disposal a variety of other fundamental frequency ranges within the absolute physiological limits. The case nearest to an invariant organic characteristic may be that of the nasal cavities, which would appear to be invariant and perhaps bestow invariant cues on nasal sounds; but even here, although nasal sounds have been used successfully in

speaker recognition experiments the spectral properties of nasals are affected by coupling through the velic orifice to the variable oral and pharyngeal vocal tract. And in the extreme, of course, a speaker can choose not to reveal any information about his cavities by speaking with fully denasalised voice that is, with the velum raised all the time.

The second complication is that whilst organic characteristics of a speaker set the limits to variation in a particular dimension such as fundamental frequency, or height of the second formant, information about these limits is conflated with linguistic information, which exploits exactly the same dimensions. Much more needs to be known about a sample of speech than just its fundamental frequency statistics before reliable inferences can be drawn about the laryngeal properties of the speaker and hence his identity.

On the other side of the dichotomy, it will become clear below that what is "learned" by a speaker of a language is of far greater complexity than is apparent, from the discussion of 'habit patterns' found in work on speaker recognition. According to Wolf (1972), features of 'learned' origin are the result of differences in the patterns of co-ordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variation in the dynamics of the vocal tract such as the rate of formant transition and co-articulation effects.

Whilst variation of this kind are of considerable theoretical interest and will form the focal point of the research, to limit the domain of what the speaker has learned to such low-level phonetic performance strategies is to ignore the vast core of knowledge the speaker has about the phonetics and the phonology of his language, and about how these may be modulated according to the situation in which he is speaking. The variety of this knowledge will become apparent.

To be fair, in existing work on speaker recognition there are occasional insights into the complexity of the problem of the sources of speaker dependent information. Atal (1976) writes that "speech is produced as a result of a complex sequence of transformations occurring at several different levels, semantic, linguistic, articulatory and acoustic. In general, differences in these transformations are likely to show up as differences in the acoustic properties of the speech signal".

Regardless of the speaker, some aspects of the sound are non essential in that they are not always used to identify words, so speakers are free to produce them in various ways. Different speakers will develop characteristically different habits in using these non essential aspects, or a single speaker will show considerable variation in their use from one utterance to another. This freedom allows a speaker substantiate latitude in fitting speech to a situation, to a mood, to the interpersonal relationship of the speaker and the listener and even to a contemporary emotional state and to health.

Nolan (1983) suggests a model depicting sources of difference between speakers figure 1 : shows the schematic diagram of the model.

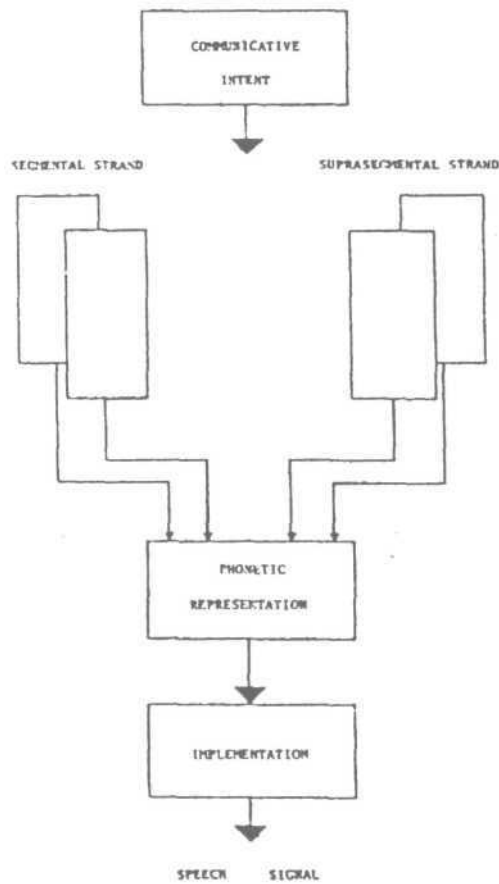


Figure 1 : Overview of model of sources of between - speaker differences

At first approximation, it may be thought of as a model of a speaker producing speech, that is as a performance model in the sense of Chomsky which attempts to describe the processes involved in the actual use of language in concrete situations.

However, the higher one removes in the speech production process from the physically observable acoustic and product and the vocal tract features immediately preceding it, the more is accessible become the actual processes involved, and the greater must be the reliance placed on the type of abstract linguistic descriptions normally thought of in Generative Linguistics as a 'competence' model - that is, a description of the knowledge of a native speaker/hearer concerning his language, but which does not necessarily map isomorphically onto structures and processes in the speaker/hearer using language. The psychological reality and its role in speech production, of a higher level element such as the phoneme in the present model is open to question.

Overview of the Model:

At the top of the model is the communicative intent of the speaker. It is important to recognise that this is complex and many sided, it is partly for this reason that the term meaning has been avoided, since this is often associated with simplistic views of speech communication as a process whereby a speaker conveys a single meaning or message to a listener - for example Tosi (1979): "This speech wave, modulated according to a phonetic code or language, conveys a message to the listener".

Such a concept of 'meaning' or 'message' ignores the fact that at the same time as communicating the bald cognitive content of an utterance, the speaker is communicating information about many aspects of his attitude towards

himself (self-image), towards the status relationship between himself and others present, towards the situation in which the utterance is spoken, and so on. Of course not all that a listener can infer from speech is intended by the speaker; here a distinction drawn by Lyons (1977) between 'communicative' and 'informative' signals is apt. This distinction depends on the intention of the speaker: "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, whereas 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 the signal is communicative or not rests, then, upon the possibility of choice, or selection, on the part of the sender". (Lyons initial restriction of 'communicative' to 'factual, or propositional information' will not be adopted here). It is, of course, not possible to determine merely by inspection of some aspects of the signal whether it is functionally communicative or merely informative.

The phonetic resources onto which the speaker maps his communicative intent have traditionally been regarded as structured into two strands -to use the term of Abercrombie (1967); both the dividing line between them and their labels are not unproblematical, but here they are termed segmental and suprasegmental. Other terms for the latter include 'prosodic' (eg., Crystal 1969) and the rather confusing 'voice dynamics' - the term in fact adopted in Abercrombie (1967).

There are intuitively un-problematical clear cases where the distinction between the two strands would give rise to general agreement; 'phonemes' of a language such as /i./, /p/,/n/ belong to the segmental strand, and intonation contours to the suprasegmental strand. The distinction may be approached from a number of viewpoints. From that of function, Crystal (1969) suggests : "We may define systems as sets of mutually defining phonological features which have an essentially variable relationship to the words selected, as opposed to those features [for examples the (segmental) phonemes, the lexical meanings] which have a direct and identifying relationship to such words".

This has the undesirable consequence of excluding lexical tone (in those languages such as Chinese, Igbo, etc., where a change in tonal pattern over identical segments may change the lexical identity) from the suprasegmental strand, with which they have most in common realisationally, including an independence from segmental occurrence (see the mention below of 'autosegmental phonology). In terms of realisation, either auditory or articulatory. Crystal (1969) claims that prosodic features may be defined as vocal effects constituted by variations along the parameters of pitch, loudness, duration and silence. This then excludes vocal effects, which are primarily the result of physiological mechanisms other than the vocal cords.

Lehiste (1970) remarks that suprasegmental features are established by a comparison of items in a sequence (ie syntagmatic comparison), whereas segmental features can be defined without reference to the sequence of segments

in which the segment appears, and their presence can be established either by inspection or by paradigmatic comparison.

This ignores on the one hand the paradigmatic identifiability of certain tone patterns (e.g the fall-rise English intonation nucleus), and on the other the role of, for example, formant transitions in adjacent vowels in the identification of consonantal place of articulation.

The definition favoured here (which follows closely that tentatively offered by Lehiste (1970) is that the supra-segmental strand comprises phonetic systems whose contrastive patterns occupy a linear domain greater than the extent of a segment; the norm is for supra-segmental contrasts to be realised over units of the extent of a syllable up to the tone unit (or greater-of. the work of Lehiste (1975, 1979) on 'paragraph intonation').

Despite these problems of definition, the two strands represent a fundamental division of the spoken medium. The traditionally recognised independence of representations in segmental and supra-segmental strands has had its most recent normal recognition in 'autosegmental' phonology. Here, supra-segmentals receive a representation quite autonomous of the segmental phonemes string; this accounts, for example, for the perseverance of lexical tone patterns despite deletion of segments by phonological rule or the underlying unity of an intonation a pattern which may be realised over a variety of segmental strings.

The two strands are not kept separate in the speech signal, nor can they form independent production targets for the speaker, since the temporal overlaying of the two strands is not arbitrary (there are an unlimited number of incorrect time alignments of a string of segments and an intonation contour, for example). The integration of the two strands yields the phonetic representation. This contains all details of an utterance which are of potential linguistic relevance. It may be thought of as specifying all the aspects of an utterance about which there is public agreement by virtue of a culturally shared language. Such aspects will include those determined in any way by the communicative intent, and also those which are purely informative in that they characterise a particular language or language subdivision.

Finally, the specifications of the phonetic representation are acted upon by the implementation rules, of which the output is neuromuscular commands, yielding movements of the vocal organs and their acoustic transform.

Omitted from mention so far have been the other two inputs to the phonetic representation in Fig 1. Abercrombie (1967) writes of three, rather than two strands composing the phonemic medium; the third comprises 'features of voice quality': "The term voice quality refers to those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his mouth".

For Abercrombie, such features can be divided into those which are outside the speaker's control (by virtue of being determined by some aspect of his vocal structure, or his transient condition) versus those which are within the speaker's voluntary control-corresponding to Laver's distinction of intrinsic versus extrinsic features; for example Laver (1976) : "Intrinsic features... derive solely from the invariant absolutely uncontrollable physical foundation of the speaker's vocal apparatus. They contribute only to voice quality... extrinsic features are made up of all aspects of vocal activity which are under the volitional control of the speaker, were 'consciously' or not".

Causing much more of a problem is the second kind of physiologically controlled communication, namely the phenomenon of voice quality of voice set.

It is also frequently used, without any implication of 'quasi-permanency', to refer specifically to the mode of vibration of the larynx, for which the term 'phonation type' will be employed here.) It follows from Laver's definition that extrinsic voice quality is susceptible to exploitation by the speaker in conveying (part of) his communicative intent-as, for example, when the speaker indicates the secrecy of what he is saying by using a whispery voice-and so a third strand' parallel to the segmental and suprasegmental might be expected in the upper part of the model.

But on closer examination it appears that the 'voice quality' strand, or as it will be called here the long term strand, is of a rather different status from the

other two. Long term characteristics are derivative of the two main strands; so a long term component such as 'palatalised voice' exist not in isolation, but by virtue of a tendency towards palatalisation recurring through a substantial proportion of elements in the segmental strand. Similarly an impression of a speaker as having a 'high pitched' voice, or a 'monotonous' voice, stems not from an isolated suprasegmental element, but from a 'cumulative abstraction' (Laver 1980) from an appreciable proportion of the suprasegmental strand. Thus a component of long term quality can be thought of as resulting from a configurational trend (or possibly a dynamic trend - a particular quality might derive for instance from characteristic rates of pitch change, or transition between segments) in the action of the vocal apparatus; this trend is referred to as a 'setting' of the vocal apparatus.

As Laver (1980) puts it "It is not proposed that the settings and segments are complementary divisions of phonetic quality... The analysis of phonetic quality into settings is a second-order analysis, abstracting data from a prior segmental analysis".

Since it is clear that such abstraction is equally possible from the supra-segmental strand, the present model incorporates two second-order long term strands, corresponding to the two primary strands, and each forwarding long term target specifications to the phonetic representation.

As a descriptive device the notion of long term settings is justified firstly behaviourally because it corresponds to the capabilities of listeners to make judgements of this kind; and secondly because it offers the potential of descriptive economy in cases where, for example, in comparing two varieties of a language or two languages, a parallel difference is noted to recur throughout a number of segments which is compatible with the effects of a single long term quality specification (cf. Trudgill 1974; Labov 1972) (It may well be that phoneticians have always covertly acted in this way, though without explicit recognition of their practice, given the surprisingly large number of languages which from their descriptions appear to have vowels approximating to the extreme qualities of the Cardinal Vowels). But the hypothesis implicit in the present model is that long term properties have reality for the speaker. As in all questions of the reality of linguistic constructs, the impenetrability of mental activity to direct investigation leaves justification of the constructs dependent on their ability to predict observed behaviour, and on the overall economy of the model of which they form a part. Taking as an example the use of long term effects for paralinguistic communication, such as the reported use of strong nasalisation in (especially American) English to signal irony, or of whispery voice to convey conspiratorially, a model without a long term mechanism would equally well predict the manifestation of these segmental modifications on every segment, on every second segment, on every second pair of segments, on completely arbitrary segments, and so on. In fact (Setting aside the question of

blocking by conflicting segmental specifications, such as oral stops (nasalisation) voiceless segments (Creaky voice) there is no evidence that any but the first of these actually occurs, this is precisely the result predicted by the present model, where the speaker has the facility to set a target in a particular phonetic dimension which remains in force until cancelled. In the alternative model the value alteration for each successive segment would be a separate operation to be performed, unrelated furthermore to identical operations on preceding and following segments.

The affective part of communicative intent is taken to refer to the attitude of the speaker, as, for instance, when a person speaks using a wide pitch range on specific contours to indicate friendliness, or speaks loudly to convey anger.

In fact, consideration of the next subpart isolated here, social intent, at once demonstrates the impossibility of drawing clear-cut dividing lines. Sociolinguistics has recently made considerable progress in describing those features of speech which are informative of a person's group membership (socioeconomic, ethnic, regional etc) but it has also revealed and quantified the degree to which a person's speech changes with the context in which he is speaking (or more accurately with his interpretation of that context); specifically, as the context becomes more formal, so a speaker will tend (in many urban communities at least) to change values for sociolinguistic variables in the direction of those of people of higher status. One important aspect of context is the addressee (or addressees); depending on his interpretation of the relative

status of speaker and addressee, and on their roles defined in a particular interaction, a speaker may 'converge' or 'diverge' to make his speech more or less like that of the other participant(s). increasingly it is being observed that sociolinguistic markers are not invariant, but depend on (the participants' interpretations of) social aspects of the interactional context (Brown and Levinson 1979). Returning to the question of borderlines between categories of cognitive intent, since context clearly plays an important role in determining features of speech, it is no longer certain that, for example, the loudness of angry speech is isolatably the result of affective communicative intent; it might also be communicating the speaker's understanding of the interactional context- for instance a dispute or the exercising of authority.

A fourth subdivision of communicative intent is the self-presentational. A wide variety of information may be encoded by a speaker in order to project a personality corresponding to his self-image (in a particular context); Argyle (1967) : "Certain aspects of behaviour during social encounters can be looked at as consequences of the participants having self images. They present themselves in a certain way, adopt a particular 'face', and try to get others to accept this picture of themselves".

Personality dimensions such as extroversion-introversion, dominance dependence, masculinity-femininity, or their perceptual correlates are associated with particular ways of speaking, and these can to some extent be intentionally adopted.

The last subdivision of communicative intent to be considered here concerns the control and structuring of any verbal interaction. In conversations, it is generally found unsatisfactory if both participants speak, and are silent, simultaneously. The participants therefore manage the interaction so that 'speaking turns' are allocated to each; and it is likely that a speaker communicates, perhaps through his overall pitch level, loudness, and rate of utterance whether he is (in his interpretation) approaching the end of a turn, or conversely is 'in full flow' and unwilling to be interrupted; compare: signals for yielding the role of speaker to the other participant are given by eye-contact behaviour, particular intonation patterns and body movements, for instance.

Quite possibly the work of Lehiste (1975) which showed that listeners were to some extent able to tell whether excerpted read sentences were paragraph-level or not, was exploring cues similar to those used in interaction management.

The above categorisation cannot claim to be exhaustive; but it begins to indicate the complexity of communicational functions which are encoded in the speech signal, and which must be considered when an apparently speaker-specific variable is selected for speaker recognition.

I The segmental strand:

1.1 Segmentation and the phoneme : The widely accepted view that spoken language consists in a linear succession of discrete segments, reflected in

alphabetic writing systems, cannot be induced from the speech signal. There are sharp discontinuities of the speech signal in time, but these are by no means in one-to-one correspondence, in position or number, with segment boundaries. The hypothesis stems rather from speaker's intuitions about where in a word a change in a sound will change the identity of the word. Thus, in the word *bid*, there are three and only three such points. At the beginning, a, b (or k, l etc) could be substituted; in the middle, an e (or a, u etc); and at the end a, b (or t, n etc). In each of these positions of choice, by such a process of commutation, a system of distinct elements is discovered; each of these elements exists merely by being significantly different from, or in opposition with, the others. Identity can be established between particular elements occurring in different positions in the word when at each position the particular element enters into a similar relationship (in phonetic terms) with the other elements that can occur in that position, for example, initial p can be identified with final p by virtue of entering into a similar relationship in terms of phonetic properties with b, and the other elements in both positions - in spite of the fact that in absolute phonetic terms the initial p and the final p are not identical ([p^h] versus [p^l], for example, in certain dialects of English). The abstract oppositional element /p/ realised in different environments as [p^h], [p^l] and other positional variants or allophones, is known as a phoneme.

There are many theories of the phoneme; and many of the basic tenets of phoneme theory (though not, in practice, segmentation) have been rejected by

'Generative' phonologists. There is not the space here to debate such issues, and it is proposed to adopt a traditional 'classical phonemic' model, as this is likely to be familiar to workers in speaker recognition, and to be used by phoneticians when analysing voice samples from the point of view of speaker identity. The sentiment expressed by Wells (1970) dealing with accent differences still seems appropriate:

The material presented here is formulated in phonemic terms. This would seem to make for easier understanding than a possible alternative presentation in terms of generative rules, particularly when the proper formulation of phonological rules is still a matter of some dispute.

12 Structure of the segmental strand : Figure 2 shows the primary and secondary (long term) segment strands in more detail, an utterance in phonemic representation is input to realisation rules, which specify the (segmental) phonetic properties the speaker has to achieve when producing the utterance. As an example from a variety of English, the word teal, phonemically /ti:l/, would be subject to rules specifying a realisation including, but not exhausted by, the following degree of detail:

[tshhIjt]

that is, an aspirated alveolar stop with slightly affricated release; a diphthong gliding from half front half close to just short of close front; and a strongly pharyngalised lateral. In fact such a representation is quite inadequate, as it is

still largely bounded by the constraint of a segmental transcription, whereas clearly if the speaker is to produce the utterance correctly a time base is needed.

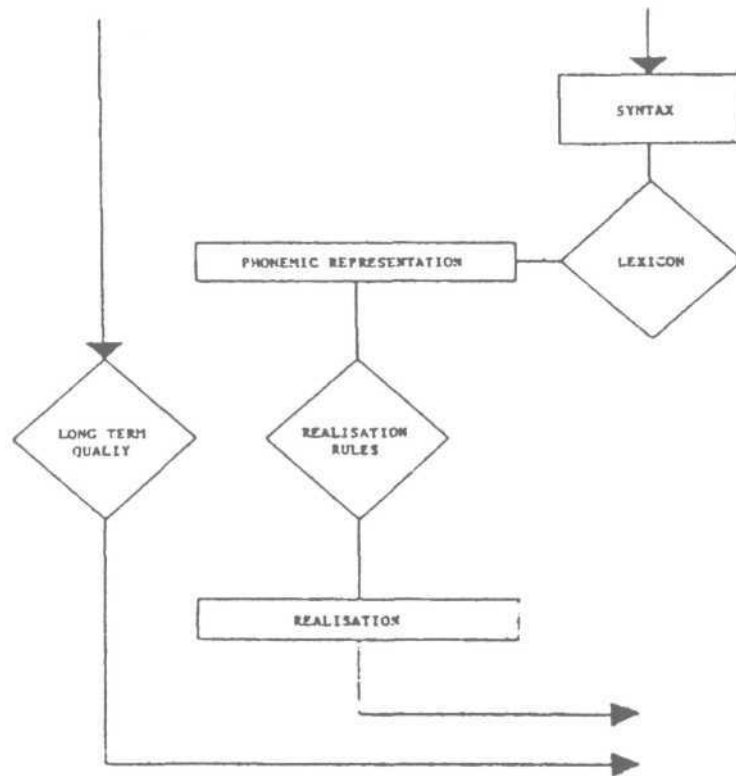


Figure 2 : The segmental strand

The phonemic representation is composed of items from the lexicon chosen, and appropriately concatenated by the syntactic component, in accord primarily, though not exclusively, with the cognitive intent of the utterance. The lexicon is like an ordinary dictionary in that it stores the words known by **the** speaker, associating their meanings, connotations, syntactic properties, and (all non-predictable) information about their phonological form; though undoubtedly

the retrieval mechanism will be far more complex than the alphabetical ordering used in dictionaries. Associated with the lexicon will be an inventory of the segments - phonemes -which can function to differentiate lexical items, and also rules specifying restrictions on their linear combination (/stnai/ and /taetf/ are not possible words of English, though/splai/ and /taeft/ are, even though they happen not to occur).

The other component in Fig 2 is labelled long term quality. It is assumed that the complex structural mechanism of the primary segmental strand does not have a parallel in the long term strand, and that communicative intent is mapped directly onto components of long term quality (nasalised voice, etc). The phonetic representation will incorporate the effect of the long term quality specifications; but the diagram does not show the input from the long term strand before or during operation of the realisation rules, which entails the claim that no realisation rule depends for its operation on a particular value of a long term dimension; no cases are known, but in the absence hitherto of even this degree of formalisation of a 'voice quality' strand the problem may not have been recognised.

13 Segmental strand categorisation of between-speaker differences: Wells (1982), on which the following draws heavily, employs a four-way categorisation: systemic, phonotactic, incidental and realisational differences; similar categories are used by others, for example O'Connor (1973). Of these, the

realisational category relates, appropriately, to the realisation rules; the other three have more to do with the lexicon.

14 Systemic differences: Two speakers may have differing phonemic systems - in terms of the total number of phonemes, or phonemes of a particular kind (e.g. front, short, vowels). A Scottish English speaker may have only /u/, where received pronunciation has the opposition /u:-/ʊ/, so that for the former good and food rhyme; a South Yorkshire speaker may have three different phonemes /i:/, /ɪ/ and /eɪ/ in the words three, tea and teach which all have a vowel in RP. Within RP a few speakers still have a phonemic opposition between /ɔ:/ and /ɒ:/, whereas for most there rhymes with law: less likely is a system lacking the opposition /ɔ:/-/ɒ:/ because although your and yore, sure and shore Shaw undoubtedly rhyme for many RP speakers, /ɔ:/ would be found in a few (mainly rarer?) words (e.g. sewer /s(j)ɔ:/, Ruhr /rʊɔ:/). This means that it will not be possible for the phonetician analysing limited samples for speaker recognition to arrive at firm systemic statements; rather, observations of potentially systemic import will have to be treated as incidental. Among the consonants, a number of RP speakers lack /r/, /w/ being substituted; /h/ on the other hand is a popular social shibboleth, its systemic absence certainly a non-RP feature. Concessions to foreign pronunciation may add phonemes of marginal status to a speaker's system: Jean and salon as /ʒɛn/ and /sælɔ̃/ rather than /ʒɛ:n/ and /sælən/ add marginal /ʒ:/ / (if not the two French phonemes involved, /ɑ̃/ and /ɔ̃/).

1.5 Phonotactic differences: When a speaker has a phoneme which is equivalent to a phoneme of another speaker (in terms of its systemic relationship to the other phonemes), but the range of phonological environments it can occur in differs, the difference is considered to be phonotactic. For example /r/ does not occur pre-consonantly or pre-pausally in RP (hence car /kɑ/ ~~fierce~~, etc), but is permitted in these environments in many other varieties of English. For some speakers /ʒ/ does not occur word-finally being replaced by /dʒ/ in words such as garage /gɑrɑ:dʒ/ ~~ʒ~~ ~~ʒ~~ ~~ʒ~~, beige /beɪdʒ/. Word-initial stop plus fricative clusters are not normally permitted, but some speakers apparently have /ps-/ in words like pseudonym as well as psi (Jones 1975). However in many cases a phonotactic classification seems no more appropriate than an incidental one; and for practical purposes, since an absolute phonotactic difference cannot be established without examining all a speaker's lexical items, phonotactic differences are more likely to have to be treated as incidental.

1.6 Incidental differences: Incidental (or lexical-incidental) refers to the incidence of a phoneme with respect to individual lexical items or groups of lexical items. Some incidental differences in particular words are popularly commented on- the variation of /ɪ:/ with /aɪ/ in either, neither, or /ɪ:/, /i/ and /e/ in economics, and presence or absence of /h/ in hotel (for a speaker who elsewhere pronounced /h/) being cases in point. In other cases alternatives operate over a morphemic class of words: RP /æ/, or /a:/ in stressed trans- (transport, transfer..), /æ/, /ɑ/ or /ə/ in unstressed-graph (telegraph, spectrograph); /ɪ/ or /ə/

in -less, -ness (hopeless, goodness...), in be-, de- (besides, decide...) and in -age, -es, -ed (manage, dances, battled...); and /ʌ/ or /ə/ in unstressed subservience). (It is by no means certain, of course, that a speaker will treat all items of a class in the same way). The following contrived example demonstrates some of the incidental features which might be used by a phonetician asked to assess the similarity of voice samples (cf: Baldwin 1977):

asəʊsɪrɪtɪd ɡærɑ:zɪz telɪgræpt ɒn tʃu:zdr ɔ:ɪd sɒld əz kɒst
 ʃ ə ɪdʒ ə a: tʃ eɪ əv ɔ:

prɒblɪm
 ə

As well as their possible speaker-diagnostic value at the phonemic level, incidental variations must also be taken into account in any scheme which exploits the phonetic quality of a particular phoneme. It would be unfortunate to weight the decision against identification by mistakenly comparing the /u:/ of Ruth and proof in a recording A, with the ʊu:/ of tooth and roof in a recording B, where the speaker in fact had used /u/, which is occasionally heard in these words.

1.7 Stress: In its realisation -in dimensions such as pitch, amplitude, duration and phonation type (the last under-researched as yet in this context) - it resembles suprasegmental features; but it seems that the presence or absence of stress regularly conditions a number of segmental realisation rules, such as

aspiration and vowel quality changes, so it must be represented in the segmental strand.

Stress placement in some languages is highly predictable- in Finnish it always comes on the first syllable of a word, and in Welsh (with certain exceptions) on the penultimate. Even in English, where stress placement is apparently free, Chomsky and Halle (1968) have demonstrated that it is possible to predict it in a large number of cases using appropriate phonosyntactic rules; for discussion here, however, the more traditional position, that free stress is marked on words in the lexicon, is adopted.

In quite a number of words in English speakers may choose from alternative common stress patterns (')ex(')quisite, (')for(')midable, (')dis(')pute (noun), (')con(')troversy, (')con(')tribute, and in compound words such as (')ice-(')cream, (')shop-(')steward; However it is important to recognise that one speaker may change his stressing of a given word according to its rhythmical context, usually so as to avoid two adjacent stressed syllables; thus she's 'just fifteen', but 'fifteen' years.

1.8 Realisation rules, allophones, and coarticulation: The realisation rules convert an abstract string of phonemes into a representation which contains specifications for all the culturally shared segmental phonetic properties which are controlled by the speaker and have a potentially informative capacity, including those which inform that the speaker is exploiting a particular variety of a

language. Thus (wo) speakers may have phoneme systems which are identical in every respect, but in realising the phoneme string /tu:/ one produces [tʰʊʊ] and the other [tʰɪu]. Gimson (1980) discusses, for each of the phonemes of RP, the variant realisations which a learner of English may expect to encounter.

Realisation rules, then, are triggered by the phonemic representation to supply phonetic detail, a phoneme normally being thought of as comprising the minimum specification of phonetic detail which will distinguish it from any other phoneme. But the phonetic detail supplied is crucially dependent, on the position in which the phoneme occurs—its position in relation to higher order structures such as the linear sequence for segments of which it is a part, particularly adjacent segments—its environment. Phonetic detail of this kind is traditionally termed allophonic, and since Wang and Fillmore (1961) two kinds of allophone have been distinguished termed extrinsic and intrinsic. For Wang and Fillmore extrinsic allophones reflect speech habits of a particular community, whilst intrinsic allophones reflect (universal) constraints of the vocal apparatus. Thus use of these terms for allophones seems to be equivalent to that of Laver for voice qualities.

Clear cases of extrinsic allophones are those where the phenomenon only occurs in a limited number of languages (therefore is not universally constrained), and which are contextually determined (in the sense above) and so lack an explanation in terms of smoothing between segments. Classic examples are the post syllable-nucleus 'dark' lateral [ɫ] which occurs in some varieties of

English regardless of the quality of the preceding vowel (as opposed to the relatively 'clear' [I] allophone before the syllable nucleus); and again for certain varieties of English the (roughly speaking) word-final glottalised allophones of the voiceless stops.

The definition of an intrinsic allophone, and hence the cut-off point between the two, is problematical, however. At first sight the tongue-body accommodation of the initial 'clear' [I] allophone in English to the quality of a following vowel (Bladon and Al-Bamerni 1976) would appear to be ascribable to purely mechanical, 'automatic' smoothing between segments. But the finding that neither is such lateral-vowel accommodation constant across extrinsic allophones in one language (Bladon and Al-Bamerni 1976) nor does it necessarily occur in another language (Ni Chasaide, 1977), nor is it constant in degree across speakers of the same language indicates that the situation is considerably more complex. The direction in which solutions will have to be sought is suggested in Tatham (1969), who argues convincingly that between the two categories of 'extrinsic events', which 'do not occur except under direct voluntary control', and 'uncontrollable intrinsic events' which are bound to occur when an intrinsic event takes place', there is an intermediate category of events which will take place, due to mechanical smoothing and the like, unless they are specifically inhibited by 'extrinsic resistance'.

In summary, then, the realisation rules will have to specify extrinsic allophones, and (where necessary in a particular language) the limits on the

freedom of the speaker to indulge in mechanically natural but resistable assimilation processes. The notions of 'allophone' and 'coarticulation' have been given a perhaps disproportionate amount of attention in this treatment of the segmental strand since they are basic to an understanding between-speaker differences in coarticulation.

1.9 Realisational differences : Realisational between-speaker differences are of three types. In the first case, all realisations of a phoneme for one speaker are different from all those of another speaker in a regular way. Clear examples of this are the realisation of RP /r/ by instead a labio-dental frictionless approximant [v] (which is nevertheless still distinct from [w] < /w/) or by a velar or uvular sound; of the realisation of RP /s/ as a voiceless lateral fricative [ɬ] (which would be counted a speech defect); less clear cases are the realisations of certain vowels; the realisations of RP /ɪ/ vary from nearly back to nearly front vowels, and though the environmentally determined range of one speaker's realisations may partially overlap with that of another, it is possible in such a case to draw a generalisation that in the same environment one speaker uses, for example, a fronter vowel than the other; similar are the cases of /u:/ (varying across speakers from almost back to front of central), and the diphthong /aʊ/ (varying considerably in frontness from speaker to speaker, but also subject to environmental determination- especially before [t]).

In the second case, the realisation rules determine a speaker-specific allophone in one or more contexts. RP /p/, /t/, /k/ may be aspirated to greater or

lesser degree before stressed vowels, or glottalised or not word-finally; /I/ may have an extrinsic allophone in post syllable-nuclear position which is 'darker' than other allophones in greater or lesser degree. Of course, if the number of contexts in which a different realisation occurs is large, and the difference between speakers in each context is in the same direction, there may be ambiguity between this case and the preceding one. Both categories underlie information which may be exploited by an auditory phonetician using a detailed phonetic descriptive framework.

Thirdly, the realisation rules govern the (culturally shared) adjustment of segments according to their environment—their coarticulation with each other. Su et al (1974) were the first to explore the speaker-specificity of coarticulation, in a study of the effect of vowel quality on a preceding nasal. In general the between-speaker differences resulting from this category of realisation are rather fine and amenable only to instrumental investigation. The realisation rules are also responsible for determining the durational relationships of the elements of the segmental sequence, which are sensitive to both context (e.g. Umeda, 1977) and environment (English vowels before voiceless obstruents are as little as half as long as when they occur in other environments (e.g. Gimson, 1980) - though their final duration will be subject to further modification by suprasegmental factors. According to Lehiste (1970), "Under otherwise identical conditions, a speaker produces durations that are normally distributed within a range characteristic of the speaker. Differences between speakers are

often quite large, and Umeda (1977) in an appended brief comparison between speakers of durations in /st(r)/ clusters found that "the closure time is fairly well regulated while the aspiration period seems to have a good deal of room for options"; but there does not seem to have been a systematic study of such durations from the point of view of speaker recognition.

1.10 The long term segmental strand : The contribution of the long term segmental strand to between-speaker differences is in the form of a set of default values for the various segmental phonetic dimensions. These default values normally fall within the ranges defined by what is acceptable in the speech community (if they do not, judgements such as 'his horrible nasal/adenoidal voice; may be made, for values which would pass without comment in another speech community); and they apply unless the requirements of the communicative intent of some utterance map onto the long term strand and select some specific values instead. They determine (the segmental aspect of) what Laver (1976) refers to as 'concurrent features'. The concurrent features make up the extrinsic contribution to voice quality. They provide the background, quasi-permanent auditory colouring to a person's voice which together with the intrinsic give a person his characteristic overall voice quality.

Among the dimensions involved will be nasality and other resonance characteristics such as palatalisation and pharyngalisation; and also phonation types such as breathy, creaky, or falsetto phonation. these latter are regarded here as part of the long term segmental strand as they represent long term

modulations of what is essentially an inherently segmental property, voicing, rather than of a suprasegmental pattern.

In the case of long term properties phonetics has only recently addressed the problem of their objective description, and so phoneticians have not had quite the same advantage over the lay listener (who may have a wide range of labels for 'voice qualities' which he can often apply confidently, even if their import is not generally agreed on) which they have enjoyed in the short term strand, where they could bring to bear a well established technical framework in which they had been trained; though now such a framework is available for long term quality. It is mainly concerned with the segmental long term strand, and discusses speaker recognition schemes which have exploited it; it then investigates acoustic correlates of long term qualities. A distant objective of such work might be an automatic classification of voices according to phonetic categories of long term quality.

II The suprasegmental strand :

II.1 Structure of the suprasegmental strand : Figure 3 shows the suprasegmental strand, long and short term aspects, and its relation to syntax and the lexicon. It can be questioned to what extent discreteness of form parallel to that of phonemes, exists in the suprasegmental strand. Crystal 1969 however it is assumed here that at least some of the suprasegmental systems do involve discrete contrastive primes, whilst in other cases communicative intent may map

into continuous variables (such as the precise height from which the voice might fall in 1 was SHOCKED, according to the degree of shock to be conveyed) by direct input to the realisation rules.

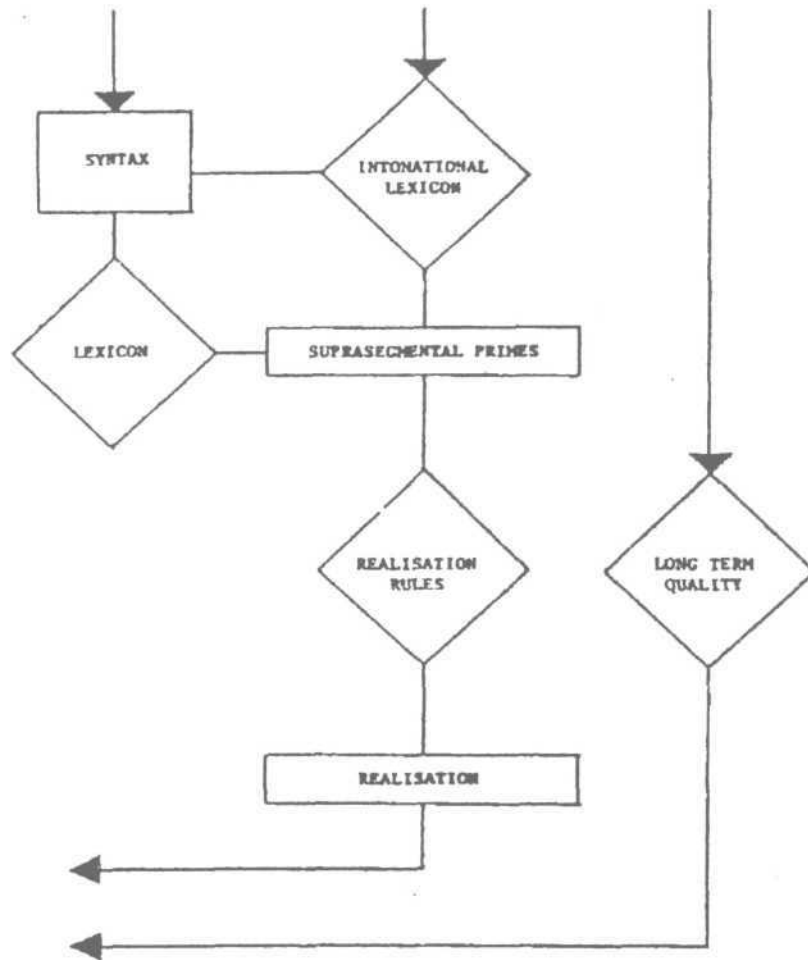
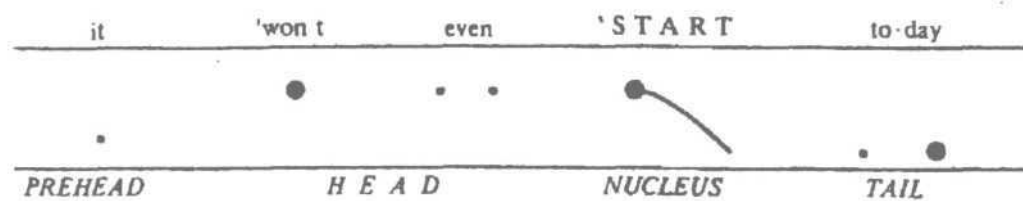


Figure 3 : The suprasegmental strand

The analysis assumed here for intonation, which is at the most systematic end of the continuum of discreteness in suprasegmental effects, is one which takes as its basic unit of analysis the tone unit. One of a number of contrastive tunes is chosen in each tone unit, the tune being comprised of subcomponents

which correspond to four divisions of a tone unit: prehead, head, nucleus and tail. The nucleus occurs on the stressed syllable of a particular word, and makes it the most prominent, usually by virtue of a pitch glide; the other subcomponents are optional and depend on the presence of appropriate syllables for their realisation. As an example



Although the interlinear graph gives an impression of the pitch movement and by large dots implies greater loudness on the stressed syllables of lexical items, the pattern will also be realised in durational effects and possibly phonation type.

At least three different aspects of intonation must be distinguished Crystal (1975) placement of tone unit boundaries, possible positions being derivable from syntactic structure and actual occurrence being dependent on factors such as rate of utterance; placement of nucleus, which (Crystal, 1975) is primarily determined by lexical or semantic factors, sometimes by specific structures, and sometimes by affective information; and thirdly selection of tune type, or at least of nucleus type, which may be determined by syntactic properties and directly by communicative (eg affective) intent -a particular tune may be 'in statements;

grudgingly admitting..., in questions... interested and concerned as well as surprised (O'Connor and Arnold 1973).

It may be possible, as suggested by Liberman (1978) and discussed by Ladd (1980), to draw analogies between the segmental lexicon and an intonational lexicon, where tunes are associated with meanings. Complete tunes would be comparable to words, which may be made up of morphemes; thus a tune with a rising head and a high fall nucleus shares 'the definiteness and completeness of all the falling tone groups', but 'adds an attitude of protest' (O'Connor and Arnold 1973), in the way that words containing same morphemes may thereby share some component of meaning, even though the meanings of the complete words are not totally predictable from the constituent morphemes.

The systemic suprasegmental representation derives from the selection made from the intonational lexicon of tunes on the basis of syntactic properties and direct affective information. It will consist in a number of layers, one being the tunes, another being a representation of the structure of the utterance in terms of stressed and unstressed syllables, in which tone unit boundaries, and nucleus placement, might be indicated, and another, in the case of a tone language, being the succession of tonemes. Schematically for English.

He didn't "want to buy it/ but I per'suaded him ||

the representation at the level of suprasegmental primes might be

H-M H-LM H-L
 sSs *S s S s | s s s *S s s ||

Realisation rules will then perform a number of operations. One of these will be the specification of integer pitch values for the tonal and intonational primes, which underlying may be specified with only three or four levels- whatever the minimum needed to describe the suprasegmental contrasts. Thus a high fall nucleus may be underlyingly H-L and a low fall M-L, but the realisation rules will specify (according to communicational requirements) how wide the fall will be. Some such specifications will be context-sensitive, determining for example the step down in pitch which occurs between heads in successive minor tone units (Trim., 1959; Crystal, 1969). Similarly in tone languages they will make language specific adjustments to tones according to their tonal environment. They will also be responsible for associating suprasegmental representations correctly with the syllable string- given the stress pattern indicated above, He did n't want to buy it/, and many others, are not possible associations.

II.2 Between-speaker differences : As far as locating the points at which speaker-specific information may be based is concerned, at least some parallels may be drawn with the segmental strand. Again, the differences between speakers are the same in kind as those between accents.



In the intonational lexicon, the system of primes from which the 'words', the tunes, are composed, may be different. It is theoretically possible that a speaker would lack a particular nucleus type used by another speaker, and so his system of elementary contrasting primes would be non-isomorphic. Evidence on

this within a speech community seems scant, and from the point of view of speaker recognition, as with phoneme-systemic difference.

Parallel to segmental phonotactic differences are differences between speakers in the co-occurrence restrictions between types of prime (head and nucleus, etc). Similar reservations about data limitations are in order here, but in the writer's experience teaching intonation analysis such differences do seem to exist. O'Connor and Arnold (1973) suggest that in RP the combination high head plus low rise nucleus is normal yet some (near) RP-speaking students find such a combination very hard to produce, the natural tendency being to replace it by high head plus fall-rise nucleus.

Incidental differences are not directly paralleled in the suprasegmental lexicon. To establish an incidental difference (recall e.g. /I:/ versus /ai/ in either), a notion of 'same word', in which the different phonemes occur, is implicit. The knowledge of 'sameness' derives from factors such as semantic equivalence, phonological similarity in other than the crucial respect, spelling, and etymology. In the case of an intonational 'word' the latter two factors are absent, and the first three are considerably more obscure, given the less discrete nature of intonational function, and form. A more feasible approach would be to examine the selectional frequency of particular intonational words, for example, with respect to syntactic types such as questions, non final statement clauses, etc. Clearly work of such statistical nature would only be possible given rather large data samples.

The discussion which follows realisation sources of speaker differences is perforce largely hypothetical, as no work in speaker recognition has attempted to isolate suprasegmental primes (as opposed to the frequency used segmental phonemes) and compare their realisations across speaker it is an approach worth exploring.

The suprasegmental realisation rules provide a phonetic interpretation of the sequences of contrastive primes. A fall-rise nucleus, for example, may be specified underlyingly as H-L-M, but the operation of the realisation rules will determine the detailed pitch movement; Australian English, for example appears often to have a realisation where the pitch rises back almost to the starting point  which is unlike a frequent RP realisation with very little rise ; less dramatic differences will exist between speakers of the same accent. There is also scope for idiosyncrasy in syntagmatic pitch relationships - for example the degree of step down between successive high heads mentioned above, or the relationship between the last syllable of a head and the start of a falling nucleus. The realisation rules will specify the co-occurrence of phonation type correlates of suprasegmental primes; different speakers will vary in their predilection for creaky voice in the lower part of a fall-rise, or their preparedness to adopt falsetto as an adjunct of an 'H' specification. Likewise, syllables will be assigned an amplitude factor and duration factor which will interact with segmental durations and amplitudes. It is possible that individuals exploit freedoms in the associating of tunes with syllables; a rise-fall nucleus (perhaps M-H-L) on a phrase such as

'two in-deed may, according to O'Connor and Arnold (1973:12) be 'spread over two or three syllables. Both patterns being commonly heard' thus -

It is not known whether tone-language realisational operations such as Sandhi and downstep leave scope for idiosyncrasy.

Finally, the contribution of the long term suprasegmental strand to between-speaker differences consists in a set of default values for suprasegmental dimensions, including perhaps mean pitch, pitch range, mean loudness, and information about normal speaking rate, which apply throughout a given speaker's vocal production unless they are manipulated by the requirements of communicative intent. They comprise the suprasegmental aspect of Laver's concurrent features'.

III Integration Rules :

When a speaker produces speech he must achieve a correct integration of the segmental and suprasegmental specifications; a particular language imposes limits on variation in this integration, as discussed by Lehiste (1970)

Languages seem to differ with respect to the distribution of the fundamental frequency contour over the voiced portion of the syllable. [A deaf subject] produced the word fell with a fundamental frequency movement that contained into the final /i/; the result sounded nonnatural and nonnative. On the other hand, in languages such as Lithuanian the fundamental frequency contour clearly includes both a vowel and a postvocalic resonant.

It is an empirical question how much tolerance a language permits in this integration for speaker idiosyncrasy.

The integration rules (see Fig 4) will have to perform at least the following tasks; adjust segmental durations; and align segmental and suprasegmental specifications in time.

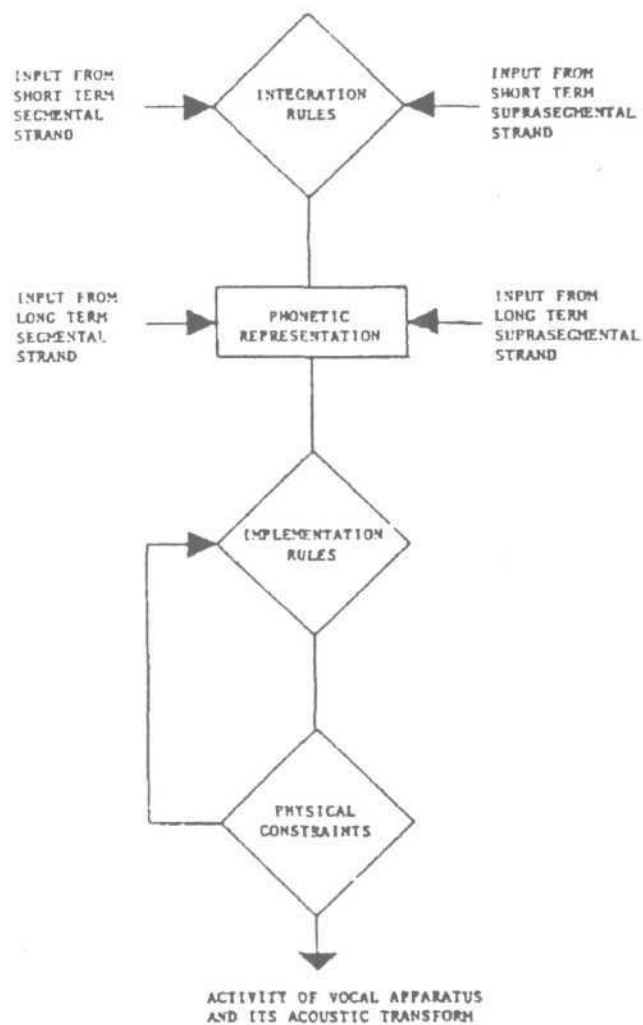


Figure 4 : Integration rules, phonetic representation, implementation rules and physical constraints.

In considering segment duration three aspects have to be taken into account. A phoneme may be thought of as having firstly an intrinsic duration associated with it, which is then adjusted by the realisation rules according to segmental environment (e.g. the shortening of vowels before voiceless obstruents in English); and thirdly each syllable in the suprasegmental strand will have a duration factor associated with it, affected by nucleus placement and type, and other intonational features, which will define the rhythm of the utterance. These factors will interact with the segment durations (probably in a complex way - vowels and consonants, syllable initials and finals, may not all be lengthened proportionally in, for example, nuclear position) to produce the durational properties of the segments.

The necessity and nature of temporal integration has already been exemplified in the quotation above from Lehiste, by a deaf speaker's failure in this respect. Loudness as well as pitch will have to be correctly aligned:

Kratochvil (1973) shows how Modern Standard Chinese tone 3 before another tone 3 receives the same fundamental frequency shape, under Sandhi, as tone 2, but retains a difference in amplitude shape.

IV Phonetic representation :

So far questions about the nature of the phonetic representation have not been addressed. The whole concept of a phonetic representation is problematical; it is dealt with more extensively in Nolan (1982). At first it might

seem that the nature of the speaker/hearer's phonetic representation is a matter of purely theoretical interest, but in fact it has considerable bearing on the central issues of speaker recognition- the character of between-speaker differences, and within speakervariation.

A first question concerns what information the phonetic representation contains. Andrew Crompton (personal communication) has suggested a formulation which corrects the view implicit in a lot of work, and reinforced by segmental 'narrow phonetic' transcriptions, that a phonetic representation is merely a string of phonemes with some allophonic detail added; rather, according to him, the phonetic representation specified.

All the linguistically relevant features of an utterance, where by linguistically relevant I refer to anything language users can make use of or react to: this therefore includes, for instance, voice quality differences of a dialectal or idiolectal kind.

However there is tension in this formulation between 'linguistically relevant' and 'anything language users can make use of or react to', since listeners can react to aspects of the speech signal which neither derive from any facet of the speaker's communicative intent, not constitute part of the particular linguistic system used by the speaker, but are purely intrinsic; for instance, it was shown by Lass et al (1978) that listeners can judge a speaker's height and weight from speech samples to within, on average., 1.5 inches and 4 pounds respectively

presumably on the basis of absolute formant nequencies and bandwidths, fundamental frequency, and other such features. These intrinsic properties are not linguistically relevant, given the premise that any individual can potentially control the complete resources of the linguistic system as a vehicle for the transmission of the types of communicative intent.

Crompton (1981) recognises this tension: "Such things are by their nature outside the control of speakers, and the properties of the speech signal to which they give rise are neither universal nor part of any individual language. This suggests that they should not be represented in the phonetic representation. On the other hand, it is a fact that listeners are able to identify talkers on the basis of their personal quality, of which these biologically determined characteristics are a part. It would therefore appear that our linguistic knowledge includes details of the personal characteristics of individual talkers, and this must presumably be accounted for in a comprehensive theory of linguistic abilities. How these two conflicting arguments are to be reconciled is not clear to me".

However, it is argued in Nolan (1982) that one of the two ways in which a phonetic representation must be remote from the physical acoustic signal is that it is the product of a process of normalisation (across speakers) - it is the level at which all linguistically relevant information in completely equivalent utterances is identically represented for speaker and listeners. That language users have available the kind of information of intrinsic origin exemplified above is a bi-

product of this normalisation process, not part of the phonetic representation itself.

Is such purely intrinsic information were to be included in the phonetic representation it would mean firstly that the speaker would be redundantly programming himself to do what he can't avoid doing; and, more importantly, there would be no possibility of regarding the phonetic representation as a level of information neutral to different language users - the level at which they can judge two utterances by two physically different speakers as being in all respects linguistically equivalent.

A complicating corollary of this viewpoint, which must be noted, is that for the same physical signal-token, the phonetic representations of speaker and hearer (or those of two hearers need not be identical - the absence of nasal resonances in the signal might be due to intrinsic adenoidal denasality, and therefore not specified in the phonetic representation of the speaker; but be incorrectly inferred by a listener as specified in it (i.e. intended) leading at a higher level to inferences about its possible informative import (regional sociolinguistic , for example), or communicative intent (where the speaker intended to transmit regional information). And since the sources of intrinsic feature (e.g. long vocal tract) are usually imitable within limits (eg. by larynx-lowering), listeners will never know that two utterances are equivalent in phonetic representation - they can only hypothesis that they are.

The second question concerns the domain in which the phonetic representation exists. It may be that as part of a model of the knowledge that a speaker has about his language the phonetic representation's dimensions are purely abstract, and of the same theoretical status as syntactic constructs such as sentence, noun, and so on, and would have equally specifiable mappings into perceptual, acoustic, and articulatory domains (Crompton 1981). However, this merely sidesteps the issue of whether speakers' behaviour indicates that the knowledge they have of phonetic properties is, perhaps, in one domain rather than another. If mapping between the three domains were absolutely isomorphic, the issue would not be resolvable; but the mapping is not one-to-one, in so far that each successive transformation in the direction articulation-acoustics-audition involves information loss-compare Jakobson et al (1952).

Each of the consecutive stages, from articulation to perception, may be predicted from the preceding stage. Since with each subsequent stage the selectivity increases, this predictability is irreversible and some variables of any antecedent stage are irrelevant for the subsequent stage. The exact measurement of the vocal tract permits the calculation of the sound wave, but the same acoustical phenomenon may be obtained by altogether different means-and therefore it seems legitimate to ask, when speakers believe themselves to be producing utterances with the same phonetic properties, in which domain(s) the sameness exists. If evidence could be found that speakers producing phonetically same effects exploit different articulatory strategies, this would

argue in favour of the phonetic representation being an auditory 'goal' which the speaker is free to implement as best he can.

In fact such evidence does exist. Care must be taken in interpreting likely evidence, as some of it ambiguously indicates either sameness of auditory target or of vocal tract configuration, as indicated by MacNeilage (1979). This is probably the case with several experiments; with Lindblom et al's (1979) demonstration that speakers achieve accurate vowel formant frequencies even immediately at voice onset despite bite blocks anchoring the jaw at abnormal degrees of opening; also with Bell-Berti's (1975) evidence that some American English speakers use muscular action to expand the pharynx to sustain glottal airflow in voiced stops, while others allow it to expand passively; and with the finding of Bell-Berti et al (1978) that American English speakers could be sorted into two categories according to genioglossus activity in front vowels - decreasing activity corresponding either to decreasing vowel height ($\dot{I} > I > e > \epsilon$) or to the tense vowel/lax vowel distinction ($\dot{I} > e > I > \epsilon$); and with the presence in one but absence in another speaker of interarytenoid muscle activity in controlling glottal opening, measured by Sawashima et al (1978).

More clearly indicative of auditory goals is the finding of Harshman et al (1977), using factor analysis of vocal tract cross-sectional area over a set of vowels, that different speakers used different proportions of the two principal 'movement' factors (anatomical differences, they suggest). This study is complemented by that of Perkell (1979), who shows (through direct

palatography) considerable variation across subjects in tongue-palate contact for particular vowels; he contends that his results, along with those of other authors, suggest (1979) that each individual does what is necessary to produce an appropriate acoustic output.

Delattre (1967) uses X-ray and spectrographic evidence to show that some American English speakers achieve a retroflex quality without in fact raising the tongue tip, but by bunching the tongue-the so-called 'molar' r. Lieberman (1967) produces evidence that different speakers producing similar fundamental frequency curves can have different patterns of subglottal pressure variation, and suggest they may therefore be compensating with different patterns of laryngeal tensioning (e.g. 1967) Most telling is the finding of Riordan (1977) that if lip rounding is artificially prevented from occurring on rounded vowels, speakers nevertheless achieve a lowering of formants by compensatory larynx lowering. An auditory goal is being achieved as best possible by an alternative implementary strategy.

Input in the phonetic representation will be from the integration rules, already discussed, and from the long term strands. The input from the latter will be in the form of a value in each of the phonetic dimensions which will be stated at the beginning of an utterance. A long term value may, of course, be reset during the course of an utterance for communicative effect. A distinction may exist between two kinds of long term value: null versus non-null. A null value would be the equivalent, in Tatha. m's (1969) categories of articulatory event, of

distribution along the basilar membrane; subject further to a linguistic -specific transformation to a speaker-neutral domain).

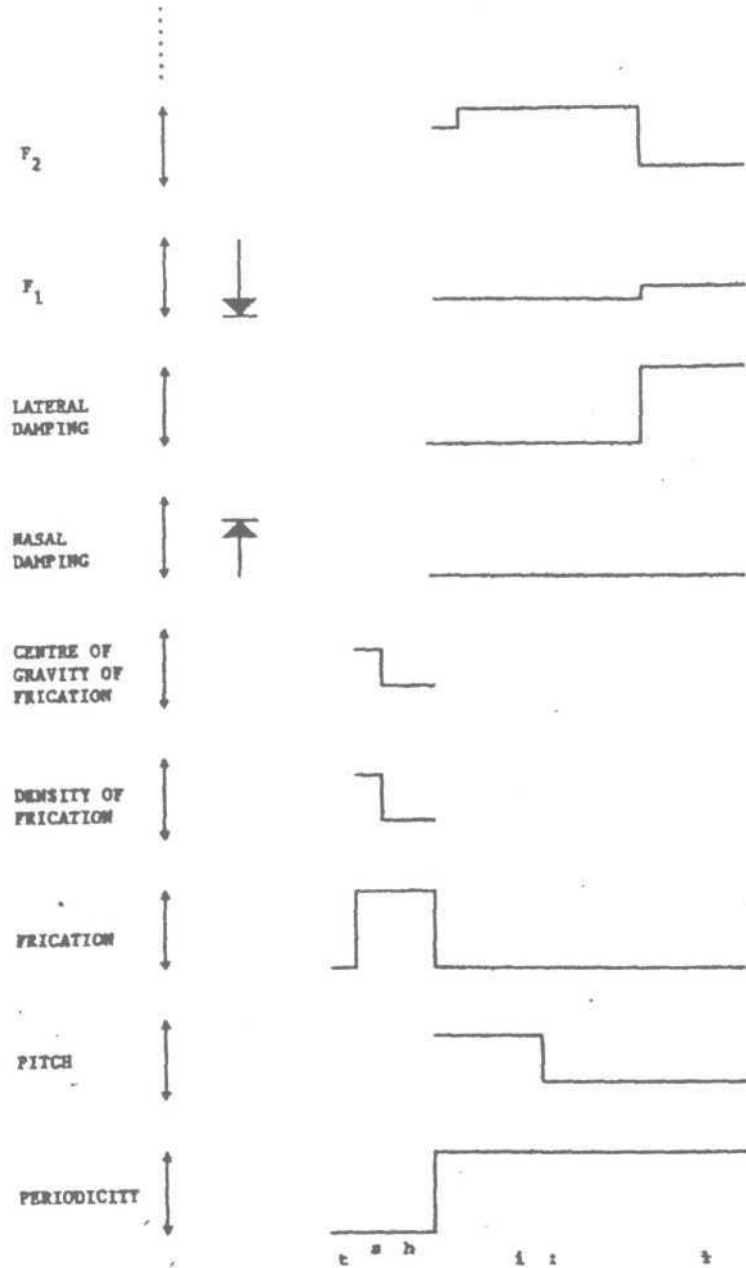


Figure 5 : Schematic impression of part of a phonetic representation.

palatography) considerable variation across subjects in tongue-palate contact for particular vowels; he contends that his results, along with those of other authors, suggest (1979) that each individual does what is necessary to produce an appropriate acoustic output.

Delattre (1967) uses X-ray and spectrographic evidence to show that some American English speakers achieve a retroflex quality without in fact raising the tongue tip, but by bunching the tongue-the so-called 'molar' r. Lieberman (1967) produces evidence that different speakers producing similar fundamental frequency curves can have different patterns of subglottal pressure variation, and suggest they may therefore be compensating with different patterns of laryngeal tensioning (e.g. 1967) Most telling is the finding of Riordan (1977) that if lip rounding is artificially prevented from occurring on rounded vowels, speakers nevertheless achieve a lowering of formants by compensatory larynx lowering. An auditory goal is being achieved as best possible by an alternative implementary strategy.

Input in the phonetic representation will be from the integration rules, already discussed, and from the long term strands. The input from the latter will be in the form of a value in each of the phonetic dimensions which will be stated at the beginning of an utterance. A long term value may, of course, be reset during the course of an utterance for communicative effect. A distinction may exist between two kinds of long term value: null versus non-null. A null value would be the equivalent, in Tatha m's (1969) categories of articulatory event, of

distribution along the basilar membrane; subject further to a linguistic -specific transformation to a speaker-neutral domain).

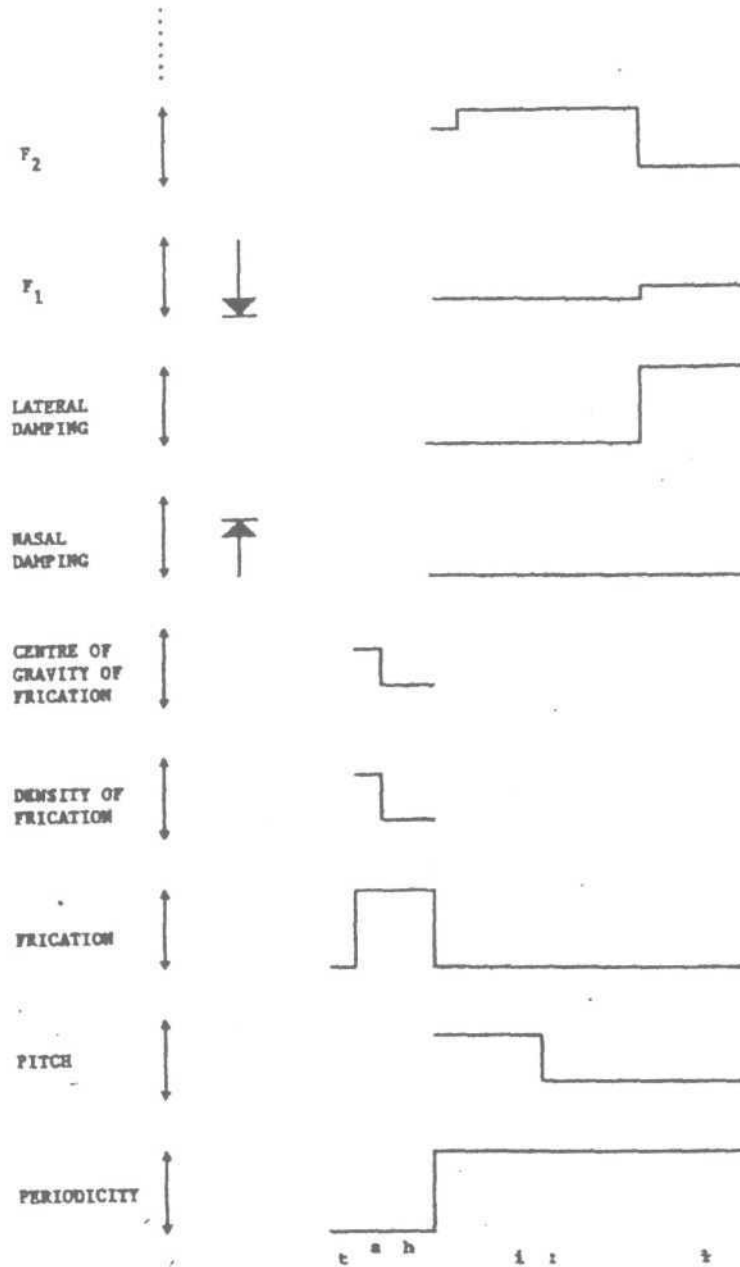


Figure 5 : Schematic impression of part of a phonetic representation.

In each dimension specifications are in the form of a duplex (v,t) where v is the integer value in the dimension and t its temporal domain. Note that although the representation still consists in successive discrete values, phonemic segmentation has been broken down. The model therefore represents a departure from the traditional kind of 'segment' and 'coarticulation' based models of which the assumptions, according to Fowler (1980) exclude the dimension of time from having an essential role either in defining the phonological units themselves or their relations in a planned utterance.

At the left of Fig 5 any non-null long term values are represented on the diagram by arrows, indicating the effect to be aimed at when not in conflict with the short term specifications. The utterance represented is of the English word teal, spoken with affrication and aspiration of the initial stop and strong pharyngalisation of the lateral; all long term values are null except for an instruction to attain lowest possible F1 at all times (likely to be implemented by larynx lowering), and moderately high nasal damping [likely to be implemented by velic opening, though as Laver (1980) points out there are a number of ways of implementing auditory nasality].

V Implementation rules :

The phonetic representation serves as one input to a set of implementational strategies. These will have access to the knowledge the speaker has acquired about the relations between muscular behaviour and activity

of the vocal apparatus, and in turn between such activity and the predictable acoustic result. They have a store of information about the physical limitations of the apparatus, and part of this store is constantly being updated so that, where possible, short term perturbations of the vocal apparatus can be compensated for in advance of production by different articulatory strategies and, apart from abnormal perturbations, to achieve future goals correctly the implementational rules will require feedback from present and past activity of the apparatus (for example very different muscular activity will be required to produce the [i] in the two sequences [ci] and [qi]).

If the whole point of the implementational rules is to achieve identical auditory effect for identical phonetic representations in the face of individuals diverse vocal apparatus, it seems at first paradoxical to claim that the implementation rules contribute speaker-specific acoustic features. The reason, however, is that strategies will strive to implement as accurately as possible certain 'primary' auditory specifications in the phonetic representation; different implementation strategies may achieve these equally well, but have different 'secondary' effects. A possible example of this is the tongue-tip versus molar strategies for achieving auditory retroflex quality (Delattre); a curling up and back of the tongue will tend in an intrinsically 'natural' gesture to cause following alveolar sounds to become retroflex, which is not the case after molar r. In this case different implementary strategies for an effect of high priority lead to different secondary consequences where the phonetic representation has a high

tolerance. Another situation to consider might be the juxtaposition of two targets of equal priority but conflicting value, where a number of transitional strategies will be equally tolerated.

VI Physical constraints :

The vocal apparatus itself is perhaps the most obvious source of phonetic differences between speakers. What it is essential to recognise, however, is that it does not determine particular acoustic characteristics of a person's speech, but merely the range within which variation in a particular parameter is constrained to take place. Thus it is certainly true to say that the dimensions of a person's vocal tract, or the length and mass of his vocal folds, will in some sense 'determine' his formant frequencies and fundamental frequency, respectively, and may even define 'optimum' values for his in these parameters; but the plasticity of the vocal tract is such that his scope for variation in these parameters is considerable (eg by raising/lowering the larynx, and imposing a greater/lesser degree of tension on the vocal folds).

There is, in fact, no acoustic feature which escapes the plasticity of the vocal tract. Nasals such as [m], [n] are often spoken of as if they did, since nasal resonances are thought of as depending on the nasal cavity of which the dimensions cannot be wittfully altered. But the spectrum of a nasal depends not just on the nasal cavities, but on the complete pharynx-nasal tube and are

sidebranch; changes in tongue body position will therefore alter (especially the frequency of the antiresonances in) the complete output spectrum.

In this taramount, then to claiming that there are in fact no 'intrinsic features' which derive solely from the invariant absolutely uncontrollable physical formation of the speaker's vocal apparatus (Laver 1976) It means rather that intrinsic features in general take the form of ranges within which variation may take place, and of complex interactions, for example, though a person may lower his formant frequencies by lowering his larynx, muscular interdependencies may occasion an alteration in phonation type; or the top end of an individual's frequency range may only be attainable both with a change of phonation type to falsetto, and a raising of the larynx consequent upon having to tension the vocal folds in an extreme way.

Intrinsic features may be classified along a continuum of permanence. Some factors underlying them, such as the size, mass, composition and innervation of the organs of speech, change only slowly through time (e.g. as the result of aging). Others, such as states of health last days or weeks, whilst effects of fatigue and diurnal rhythms, emotional states, and experimental phonetic intervention, are even more transient.

An independent cross-cutting classification of intrinsic constraints divides them into configurational and dynamic constraints. Configurational constraints comprise the physical limits on the size and shape of the vocal tract, and the

relative position of the articulators. Under configurational constraints would also be included differences of composition of the vocal tract-for example, the acoustic boundary effects at the walls of the vocal tract might vary according to the thickness of fatty tissue in the boundary walls. The main acoustic dimensions, which have limits imposed on them by intrinsic constraints are formant frequencies and bandwidths, fundamental frequency, presence and frequency of antiformants, and intensity and frequency distribution of fricative energy.

Dynamic constraints impose upper limits on the rate and acceleration of articulators, and on the rate of change of vocal tract configuration, including upper limits on the transmission of neural impulses. Little is known about such constraints, but it is reasonable to assume that different speakers may have differential agility in speech production, in the same way that speed of movement and coordination differ in other physical skills such as gymnastics or playing a musical instruments.

VII Mapping of communicative intent:

It is time now to consider the ways in which communicative intent is mapped onto the resources of the linguistic mechanism, which has been outlined above. The crucial importance of this mapping from the point of view of speaker recognition lies in the need to be aware of potential changes in the phonetic signal of a particular speaker according to his communicative intent and in

interaction with the situation in which he is speaking. There has been a very inadequate amount of attention paid to this mapping in speaker recognition work.

VII.1 Cognitive intent : The cognitive part of communicative intent provides perhaps the main exception to this neglect, since it is readily apparent that a change in the cognitive meaning of an utterance, causing a change either in selection of lexical items and/or their syntagmatic sequencing, will cause utterances to be phonetically non-equivalent. Speaker verification schemes, therefore require the speaker to produce some pre-agreed sequence of words, which is then compared with a stored reference token or tokens of the same words; and in legal applications for speaker identification, attempts are made to elicit the same sequence of words from the suspect as occur at some point in the recording of the criminal (e.g Bolt et al 1979; Tosi 1979)

Cognitive intent is also mapped indirectly onto the suprasegmental strand, since syntactic structure is one of the factors which determines the choice of intonation patterns. However, it cannot be assumed that a particular syntactic structure will be associated with a given intonation pattern, since choice of intonation pattern is also determined partly by, for example, affective factors - as Crystal points out (1969) the two patterns 'What are you doing?' and 'What are you doing?' are equally possible, but the second is "generally more serious and abrupt in its implications - at least for British English than the more friendly and interested first pattern".

interaction with the situation in which he is speaking. There has been a very inadequate amount of attention paid to this mapping in speaker recognition work.

VII.1 Cognitive intent : The cognitive part of communicative intent provides perhaps the main exception to this neglect, since it is readily apparent that a change in the cognitive meaning of an utterance, causing a change either in selection of lexical items and/or their syntagmatic sequencing, will cause utterances to be phonetically non-equivalent. Speaker verification schemes, therefore require the speaker to produce some pre-agreed sequence of words, which is then compared with a stored reference token or tokens of the same words; and in legal applications for speaker identification, attempts are made to elicit the same sequence of words from the suspect as occur at some point in the recording of the criminal (e.g Bolt et al 1979; Tosi 1979)

Cognitive intent is also mapped indirectly onto the suprasegmental strand, since syntactic structure is one of the factors which determines the choice of intonation patterns. However, it cannot be assumed that a particular syntactic structure will be associated with a given intonation pattern, since choice of intonation pattern is also determined partly by, for example, affective factors - as Crystal points out (1969) the two patterns 'What are you doing?' and 'What are you doing?' are equally possible, but the second is "generally more serious and abrupt in its implications - at least for British English than the more friendly and interested first pattern".

In general, cognitive intent seems not to be mapped either into the realisation rules of the two primary strands - increasing the frontness of the realisation of a vowel phoneme, reducing the pitch movement of a fall-rise nucleus, using persistent lowered larynx voice or a high pitch range seem unlikely to change the 'factual, prepositional' content of an utterance. However there may be exceptions; there may be processes which in fact or override the cognitive meaning of a lexical item; for example, the lexical entry for wonderful would be unlikely to contain a meaning such as 'bad', but it is quite possible to utter Oh yes he's a wonderful cook in such a way (perhaps with reduced pitch range throughout, heavy nasalisation, and creaky voice) that it makes little sense to consider the cognitive intent of the speaker to be derivable from the literal meaning of the words - in fact a meaning may be reversed.

V11.2 Effective intent : The affective part of communicative intent, the attitudes and feelings that a speaker wishes to convey, is mapped in complex ways. Some are not of direct concern for the phonetic aspects of speaker recognition- choice of lexical item may be influenced, and of syntactic structure, but these will be obvious in the everyday sense of resulting in a 'different utterance'.

From the phonetic point of view, affective information is first and foremost thought of in terms of mapping onto the suprasegmental strand. The mapping involves both direct influence on the choice of discrete contrasting intonation patterns ('intonational words') at the level of the intonational lexicon (which is reflected, for example, in the guidance to learners about attitude

conveyed, given in respect of the various contrasting tone groups of their analysis by O'Connor and Arnold (1973) and also mapping into the realisation rules where less discrete suprasegmental effects are specified, for example, a high fall nucleus according to O'Connor and Arnold (1973) conveys in statements 'a sense of involvement', but it is not counter-intuitive to suggest that the degree of involvement may be reflected in the size of the pitch movement of the high fall.

Affective information may also be mapped into the long term strands. A speaker may replace his default value for phonation type, for example, by one determined by the attitude he wishes to convey- a speaker with a normally creaky phonation type might adopt a more breathy phonation type in order to convey sympathy; and in the suprasegmental strand a speaker with a default pitch range which is narrow might broaden it in order to communicate enthusiasm.

The significance of these mappings from the point of view of speaker recognition is that they underlie phonetic properties which may be selected as speaker-specific parameters; yet unless the affective communicative intent of two otherwise similar utterances is the same there is no guarantee that affective mapping will not confound identification or elimination based on those properties.

VII.3 Social intent : A simplistic, and perhaps popular, view of social information in speech would suggest that social (including geographical) group membership within a language community is reflected in the use of certain

phonetic features, and that by extracting these features it is possible to assign an individual to a particular group. Whilst this would not count as identification of an individual, it would at least be part of a process of subclassification of speakers, and therefore constitute a useful component of the process of speaker recognition. Presumably this is what Bolt et al. (1979) have in mind when they write. "Training in the performance of voice identification should include more extensive instruction in related scientific disciplines than is usually included at present... For example, a knowledge of dialectology would show how shifts in vowel color could produce important differences between voices being examined by listening".

It is, however, necessary to reinterpret what they actually say, which seems only to make sense if 'produce important differences' is replaced by 'constitute differences of social significance'.

It is indeed true that dialectology and more recently sociolinguistics (concerned with social stratification of languages, so far particularly in urban areas) have recorded many phonetic differences between the speech of different groups. However, if such differences were purely in the linguistic system which a speaker of a dialect or sociolect has at his disposal compared with a speaker of another, then the differences would be merely informative rather than communicative, and have no place in the present discussion.

But one of the most striking discoveries of the type of work in sociolinguistics which Labov pioneered is that each speaker has control over a range of styles of speech - Labov (1972): 'As far as we can see, there are no single -style speakers' - and, most significantly for speaker recognition, that the sociolinguistic variables along which speakers of different social strata will employ different values are the same variables along which stylistic variation takes place: Labov (1972): "The same sociolinguistic variable is used to signal social and stylistic stratification".

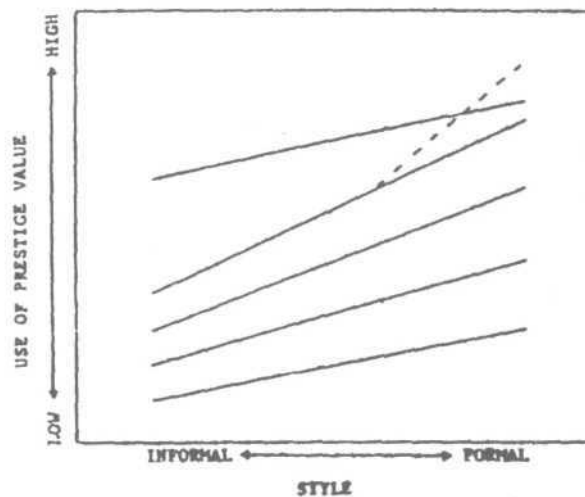


Figure 6 : Schematic representation of the use of a prestige value of a sociolinguistic variable to class and style.

Figure 6 adapted from Labov (1972), shows this situation schematically. The lines each relate to a particular independently defined socioeconomic group; the vertical axis shows that each higher social grouping uses a higher percentage of a prestige form; and moving from left to right shows how in each class use of

prestige variants increases with the formality of the context. (Labov elicited speech under the following situations of increasing formality, or likely attention to speech interviews, reading a text, reading word lists, and reading minimal pairs of words; and he also attempted where possible to obtain as the least formal style recordings of casual conversation when the informants no longer had their attention focussed on the fact that they were being recorded). A detail of Fig 5 which is not of direct concern here is that at the most formal end of the style-range a lower group may surpass a higher group's usage of a prestige value - in Labov's data the 'Lower Middle Class' and 'Upper Middle Class' respectively. He terms this phenomenon 'hypercorrection', and it is indicated by the dotted line. Apparently this behaviour only occurs when the variable in question is involved in a linguistic change in progress (1972). In fact two points are of significance for speaker recognition. Firstly, for many variables, such as the use of /n/ versus /h/ in the -ing suffix in New York (Labov, 1966) and in Norwich (Trudgill, 1974), a speaker in a given context does not produce either 100% or 0% of the variable; as Trudgill (1974) puts it in relation to another of the socially significant New York variables, presence versus absence of postvocalic /r/ '.

"The researcher could not predict on any one occasion whether an individual would say cah or car, but he could show that, if he was of a certain social class, age and sex, he would use one or other variant approximately percent of the time, on an average in a given situation".

Labov (1972) exemplifies this kind of variation by giving the order of occurrence of values for one informant in each of two styles; casual speech: From the view point of speaker recognition, this means that the occurrence of a value of a variable in a limited sample of speech cannot be taken as predicting its reliable occurrence throughout all the speech of that speaker in the given context.

The second point is that, since sociolinguistic stratification and stylistic variation are taking place along the same dimensions (Labov, 1972) it may therefore be difficult to interpret any signal by itself-to distinguish, for example, a casual salesman from a careful pipe fitter.

As a concrete example (from Labov, 1972) the indices for post-vocalic /r/ for two informants were: Miriam L (lawyer) - casual speech 32, careful (interview) speech 47; Doris H. ('lower middle class') casual speech 31, careful speech 31. Thus both speakers, from different strata, attain virtually the same value (31/32) but in different styles. So it is not the case that an individual has one all-purpose manner of speaking which immediately pins him down to a particular group, but rather he has control over a stylistic range which overlaps the stylistic range of at least some other groups. From the perspective of speaker identification, it is not acceptable to assume a priori that, for instance, suspects interpret the provision of samples for voice comparison as a context of equivalent formality to the circumstance in which the incriminating recording was made- at the very least, it is implicit in work within Labov's paradigm that awareness of being recorded itself induces a different style of speaking. It is not clear on what

basis the problem of style-shifting has been ignored in discussions of the feasibility of speaker recognition.

So far in this discussion the focus has been on those variables which a speaker may manipulate, and may therefore be considered as communicative. Not all variables, however, are of this type; Labov (e.g 1972) draws a distinction between indicators and markers.

Indicators are linguistic features which are embedded in a social matrix, showing social differentiation by age or social group, but which show no pattern of style shifting and appear to have little evaluative force [for the speakers themselves]... Markers... do show stylistic stratification as well as social stratification. Though they lie below the level of conscious awareness, they will produce regular responses on subjective reaction tests.

The variables mentioned above are examples of markers; Labov cites the merger in American English of the vowels of hock and hawk in the speech of some, according to region, class and age, but which nonetheless does not undergo style-shifting as an indicator.

Further, certain markers may attain social recognition and be the subject of comment among speakers themselves - these Labov calls stereotypes. In terms of the present model, a marker would result from the mapping of social communicative intent at some point in the linguistic mechanism resulting in a determined value; an indicator would be the product simply of a particular

speaker's default value at some point in the mechanism a default value, of course, within the limits set by the particular variety of which he feels himself to be a speaker.

A specific aspect of the social context in which a person is speaking requires further comment, namely the other participant(s) in the interaction. The context must in fact take its definition in part from the participants interpretation of their interpersonal relationship and relative status; it is not enough for a speaker to know that he is in a situation, such as a chance meeting on the street, where casual conversation might be the norm, in order for him to know what style to adopt- he also needs to assess his relationship to the other person, as for example his friend, his boss, or his subordinate and junior. As Brown and Levinson (1979) point out, a specially important distinction in this respect concerns whether the two participants feel themselves to be members of the same group or not.

It seems a reasonable hypothesis that if both parties to an interaction are drawn from one group then it is likely that the social relationship obtaining between them will be organized around non-group (or subgroup) identities - sex, kinship, role, personality, or whatever the relevant criteria may be. On the other hand, if the parties belong to different groups, then their group identities are likely to be the ones that determine their relationship. So the distinction between in-group and out-group relationships is fundamental to the organization of interaction for any two parties.

An instance of in-group versus out-group interaction relevant to the present topic might be the (tapped telephone) conversation of a criminal with his colleague, versus the verbal interaction of a suspect with a policeman or 'voiceprint' expert. In response to the in-group/out-group nature of this relation to the other participants the speaker may choose to make his speech more similar to that of the other (convergence), or in the case of desired dissociation, more dissimilar (divergence) (Giles et al 1979); further, the degree of convergence may even change during the time course of an interaction with one individual, and according to the topic of conversation (Douglas-Cowie, 1979). It may be that convergence is a very basic part of the human communicative ability, since Lieberman (1967) reports evidence that when talking with a parent, a ten month old boy lowered the fundamental frequency of his babble compared with when he was alone, and that he lowered it more with the father than with the mother. Crystal (1975) suggests that 'This "vocal empathy" seems a normal adult phenomenon also'. Although there seem to have been few if any studies of this kind of convergence, which may be only minimally sociolinguistic in that it involves extrinsic accommodation to a partially intrinsically determined feature of the other's speech, it cannot be ignored in a theory of speaker recognition, and is ripe for research.

The phonetic mapping of a speaker's vocal intent, and likewise the existence of variety-specific default values ('indicators'), is not confined to any one part of the language mechanism; although the majority of sociolinguistic

studies have concentrated on the realisation of segments, the suprasegmental strand is equally implicated (Pellowe and Jones 1978; Knowles 1978) and increasing attention is being paid to the long term 'voice quality' reflexes of these strands (e.g. Labov 1972; Trudgill 1974; Esling 1978; Knowles 1978).

The percentage of phonetic features which are subject to manipulation by the speaker as sociolinguistic markers, and subject to variation according to his interpretation of the context, is hard to estimate; however it is clear from the work published over the last 15 years in sociolinguistics that such markers are far from being isolated phenomena - at least in urban communities. The case presented above that they constitute a problem for speaker recognition may turn out to be overstated, but the onus is properly on those who claim parameters to be successful in identifying speakers in the laboratory to demonstrate that these parameters are resistant to the kinds of variation that occur in different social contexts - particularly since this variation, as is inevitable on the one hand from the ambiguity with respect to class and style which characterise markers, and on the other from the definition of 'convergence' phenomena, will have precisely the effect of making one speaker sound more like another.

If speaker recognition is to be reliable and efficient, then in theory the problems posed by stylistic variation might be circumvented in two ways. Firstly, knowledge of the ways a community operates with its markers might be used to normalise the value obtained in some dimension for a particular speaker in a given context. This solution seems improbable, given the difficulty in practice of

knowing how a speaker interprets any given social context, and the complexity of discovering the way in which markers operate in a community. The second solution would be to ensure that speaker recognition parameters derive only from features of speech which are inert with respect to social context-that is, from indicators (in Labov's sense), or from features which are uncorrelated with social stratification. However, it may be that further exploration of the phenomenon of convergence will be required to ascertain that features do in fact exist which are inert to all aspects of social context in its broadest sense.

VII.4 Accentual versus personal information in speech : Here it is appropriate to consider the division sometimes made of information in the speech signal over and above cognitive information into accentual versus personal (e.g. Ladefoged, 1967). The notion of accent, as a subdivision of a language associated with a particular speech community, and defined by the co-occurrence of a static set of phonetic/phonological properties, is no longer tenable, given the variability of some at least of those properties according to context. Two other possibilities exist.

The term accent could be restricted to the set of properties which occur at some point in the stylistic continuum for a particular speech community - possibly at the least formal style since it is here that the greatest diversity would be manifest. This then produces the problem of labelling what it is that the speakers of that community are speaking in more formal contexts, which is nevertheless distinct from the pronunciation of other speech communities.

A more promising possibility is to consider an accent to consist in the complete stylistic range of pronunciations controlled by members of a speech community. Under this interpretation it might be feasible to maintain a distinction between accentual information, determined by the community-agreed sociolinguistic system, and personal information, where the speaker has chosen idiosyncratic values within the tolerances allowed by the 'accent'. However this presumes that all the speakers have equal control over the stylistic facilities provided by the 'accent' in this broad sense; that is, that their stylistic repertoires are equivalent. This is clearly not the case; as pointed out by Brown and Levinson (1979) membership of particular groups in the community will restrict a speaker to subparts of the total range of variation : "A.. way in which group or category affiliation can be signalled by the code that a speaker utilizes derives from the fact that different groups within the speech community may command different subsets of the total linguistic resources available in the community".

Moreover, the restriction of repertoires of pronunciation does not end at group level, but carries on down to individual level; occasioned by differences in education, breadth of linguistic contact, and so on.

Three points may be made in relation to individual repertoires. Firstly, even given a complete and correct description of the system of social and stylistic variables available to members of a speech community it would still not be possible to extrapolate from a recording of a speaker in one social context to his

performance in a different one, since it will not be known how far his flexibility is hampered by repertoire restrictions. Secondly, it is apparent that a contributory factor to speaker idiosyncrasy, and indeed one which may not be insignificant when listeners categorise speakers, is the stylistic flexibility of the speaker - the extent of the subpart of the 'accent', in the broad sense, that he controls. And thirdly it follows from the fact that one facet of personal quality finds its definition only within the systematic stylistic relationships of the 'accent' that a sharp division between the two kinds of information is not feasible.

VII. 5 Self - presentation : Turning now to the exploitation of phonetic parameters in order to communicate a self-image to others, Scherer (1979) explains "Actors [= participants in interactions] often use behavioural cues for the presentation of self and, given the importance of speech in social interaction, it is not surprising that speech cues are prime candidates for self-presentation purposes".

Self-presentation as used below may turn out to be too umbrella-like a term, intended it is to include aspects of the view of self from bio-physical characteristics through to personality dimensions such as extroversion-introversion, yet it seems to provide a useful category of communicative intent distinct from the communication of purely short term emotions and attitudes, and from the communication of information about interpreted positions within a purely social matrix.

At one end of the continuum speakers may manipulate, within the intrinsic constraints of their own vocal apparatus, characteristics of speech which over the population as a whole will be correlated with bio-physical characteristics. A person wishing to communicate a self-image of a large physique might therefore adopt low formant and fundamental frequencies as would normally be expected from a person of that physique; similarly the voice correlates in terms of formant frequencies and fundamental frequency of maleness and femaleness (see e.g. Coleman 1971, 1976) ,may to some extent adopted by a speaker within his intrinsic limits. It seems probable that these kinds of information, closely related as they are to intrinsic limitations on vocal capability, will be mapped onto the long term strands Scherer's (1979) discussion of personality markers in 'vocal aspects' of speech is confined to long term properties.

At the other end of the continuum, more indirect culturally mediated relationships exist between personality dimensions and phonetic dimensions. Among those in the work cited by Scherer (1979) are positive correlations between (mean) fundamental frequency and self-attributions (on inventories and rating scales) of achievement, task ability, sociability, dominance, and aggressiveness between mean fundamental frequency and self ratings of adjustment, orderliness and lack of autonomy between extroversion and intensity and between breathy voice and introversion, neurotic tendency, and anxiety. It appears that more complicated interrelations may exist, abstract aspects of

personality being mediated in their mapping by the sociolinguistic mechanisms; Douglas-Cowie (1978) reports that the degree of convergence of rural dialect speakers to an outsider's standard pronunciation in interviews 'is often clearly related to their social ambition' being assessed from ratings made by the other informants, all informants knowing each other well. Pellowe and Jones (1978) note how one female informant's self-image in respect of age is reflected in her choice of the suprasegmental primes they were studying amongst women there is an age trend which indicates that younger women are realising rises in more and more tone units in which their elders would have realised falls... This is a trend which seems to be socially significant for members of the speech community. It seems, for example, to be a behaviour being emulated by Ar who in terms of her age should have had a value of +15% or so but who in fact has a value of -26%.

The sociolinguistic mechanism may also provide for expression of sexual identity, though rarely as clearly as in Darkhat Mongolian (Trudgill, 1974) which has a different vowel system for men versus woman; but in many instances Trudgill (1974) notes that 'women consistently use forms which more closely approach those of the standard variety or the prestige accent than those used by men'.

VII. 6 Interaction management : According to Duncan (1973), speakers may produce three kinds of signal in their attempts to manage the progress of 'speaking turns' in a dialogue: (a) a turn signal : (b) a turn-claiming suppression signal : and (c) a within turn signal. The cues for the latter two involve body

movements and syntax in the data analysed by Duncan, but intuitively it seems possible that a speaker will increase his pitch or loudness to 'fightoff' an attempt at interruption by another speaker. 'Turn signals', whereby the speaker indicates that he feels he has completed his 'turn' and is prepared to allow the other to speak, were found to involve cues including particular intonation patterns, a 'drawling' of certain syllables, and a drop in pitch or loudness.

Lehiste (1975) found similar cues correlating with judgements of whether excerpted sentences had been read paragraph-initially, or paragraph-finally - high fundamental frequency peaks cued isolation and paragraph-initial judgements, and low fundamental frequency, perhaps with laryngealisation (creaky voice), paragraph-final judgements. It may be that the organisation of a read text, or a monologue, into 'paragraphs' has an affinity with the management of verbal interaction between two or more participants.

Further research should increase understanding of the cues speakers rely on to direct the progress of an interaction; and it may then be more possible to assess whether the increase in within-speaker variation they occasion, along, for example, fundamental frequency parameters, constitutes a problem for speaker recognition.

VIII Summary:

The model which reveals the bases of speaker-specific information in the speech wave, and the sources of its variability - the two being in a symbiotic

relation, the model is undoubtedly inadequate in many respects, and in some controversial; but if it appears complex, this is not in itself a shortcoming, for it correctly reflects the immense complexity of the linguistic mechanism and the sophistication of the human communicative ability.

As a starting point the frequently quoted dichotomy between 'organic' and 'learned' sources of between speaker differences was taken. The inadequacies of this dichotomy the 'intrinsic' component of speaker idiosyncrasy is in the form not of absolute values, but of limitations on the variation which a speaker can induce in his vocal apparatus. Within speaker differences can also be caused by changes in intrinsic constraints, due to changes in state of health, etc, but these are not considered in detail here.

If all other sources of idiosyncrasy are lumped together under the heading of 'learned', then it is apparent that, at the very least, different kinds of learning are involved. At the lower end of the model, the speaker acquires by trial and error, rather than by learning through direct imitation of what cannot by its nature be accessible to him, a set of implementational strategies for achieving appropriate auditory phonetic effects. Although the notion is not tested here it is conceivable that it is these strategies below the phonetic representation which are least susceptible to volitional alteration by the speaker.

At higher levels the speaker learns, on the basis of the language use he is exposed to, and arguably also on the basis of innate preconceptions as to the

nature of language, a complex mechanism of expression. This mechanism serves for the mapping of different aspects of the communicative intent of the speaker, and this mapping is such that many parts of the mechanism - segmental and suprasegmental, short and long term, primes and realisational rules - can be affected by one aspect (e.g social) of the communicative intent.

At each point where communicative intent is mapped there may be thought of as existing default values, which are peculiar to the speaker, though they (normally) fall within the range permitted by the particular variety of the language he speaks. The point in a hyperspace defined by all a speaker's default values might be thought of as constituting his extrinsic personal quality; but this point is a purely fictional abstraction, because in any utterance a speaker will be mapping communicative intent in such a way as to replace some default values by determined values - for example a speaker may have a long term default value of non-nasalisation, and a default value of [ä] for /æ/, but may change these to nasalisation and [ɛ̃] when communicating an attitude of irony in a social context where he is converging to a speaker with a different pronunciation.

Within-speaker variability is clearly of concern in speaker recognition, but experiments based on the assumption that this variability results purely from random intrinsic changes in time, for example by getting subjects to read a passage several times over a few months, will not permit theoretically sound extrapolation to the real world. The way a speaker speaks on a given occasion is the result of a complex interaction between his communicative intent, the

language mechanism he controls, and the context in which he is speaking. It may be that the within-speaker variation that results is trivial compared with the gross acoustic similarity of utterances from the same vocal apparatus; it may be that the parameters used in 'voiceprint' and automatic speaker identification schemes are just those which are inert to social context, attitude of the speaker, interaction management etc (however great a coincidence this would be); but these hypothetical states of affairs need to be demonstrated, not assumed a priori as at present, if techniques of speaker recognition are to be acceptable outside the laboratory. In the real world, speakers communicate rather than merely exercise their vocal apparatus.

language mechanism he controls, and the context in which he is speaking. It may be that the within-speaker variation that results is trivial compared with the gross acoustic similarity of utterances from the same vocal apparatus; it may be that the parameters used in 'voiceprint' and automatic speaker identification schemes are just those which are inert to social context, attitude of the speaker, interaction management etc (however great a coincidence this would be); but these hypothetical states of affairs need to be demonstrated, not assumed a priori as at present, if techniques of speaker recognition are to be acceptable outside the laboratory. In the real world, speakers communicate rather than merely exercise their vocal apparatus.

<u>Segmental</u>	<u>Suprasegmental</u>	<u>Interpration rules</u>	<u>Mapping of communicative intent</u>	<u>Accentual vs Personal Information in Speech</u>
<ul style="list-style-type: none"> • Systemic differences • phonotactic differences • Incidental differences • Stress • Realization rules, allophones and coarticulation • Realisational differences • Long term segmental strand. 	<ul style="list-style-type: none"> • Types of primes • Phonotactic interpretation of the sequences and contrastive primes • Differences in mean pitch, pitch range, mean loudness, speaking rate 	<ul style="list-style-type: none"> • Phonetic representation • Implementati on rules • Physical constraints 	<ul style="list-style-type: none"> • Cognitive intent • Effective intent • Social intent 	<ul style="list-style-type: none"> • Self presentation • Interaction management

Table 1 : Summary of between - speaker differences

Table 1 summarises the sources of between speaker differences. Clearly, it is impossible to cover all variations of all sorts that is apparently lacking in the literature. However, the literature review warrants study in all the variables. The objective of the present study is limited to find out the systemic differences as applicable to speaker verification. Specifically, acoustic parameters measured from spectrography reflecting the systemic differences are studied here.

CHAPTER III

METHOD

Materials : Twenty-nine bisyllabic (CVC, CVCV,CVCVC,CVCCV) meaningful Hindi words with 16 plosives, five nasal continuant, four affricates and four fricatives in the medial position were selected. These as written one on each card formed the material. Table 2 shows the details of the material.

Sl.No	Key Phoneme	Description	Word
1	Plosives k	Velar unaspirated voiceless stop	bakra
2	k ^h	velar aspirated voiceless stop	dak ^h ra
3	g	velar unaspirated voiced stop	pagla
4	g ^h	velar murmured voiced stop	me:gha
5	t	Retroflex unaspirated voiceless stop	matar
6	th	Retroflex aspirated voiceless stop	baethna
7	d	Retroflex unaspirated voiced stop	padosi
8	dh	Retroflex murmured voiced stop	padhai
9	t	Alveolar unaspirated voiceless stop	patta
10	th	Alveolar aspirated voiceless stop	pathik
11	d	Alveolar unaspirated voiced stop	badam
12	dh	Alveolar murmured voiced stop	ra:dha
13	p	Bilabial unaspirated voiceless stop	paplu
14	ph	Bilabial aspirated aspirated stop	saphal
15	b	Bilabial unaspirated voiced stop	k ^h abar
16	bh	Bilabial murmured voiced stop	ab ^h a:v
17	Nasal continuants n	Velar voiced nasal continuant	pakna
18	n	Palatal voiced nasal continuant	winan
19	n	Retroflex voiced nasal continuant	pranam
20	n	Alveolar voiced nasal continuant	pa:ni
21	m	Bilabial nasal voiced nasal continuant	Kaman

22	Affricates c	Palatal unaspirated voiceless affricate	bacna
23	ch	Palatal aspirated voiceless affricate	kaechi
24	j	Palatal unaspirated voiced affricate	sajna
25	jh	Palatal murmured voiced affricate	ma:jhi
26	Fricatives s	Palatal voiceless fricative	ka:si
27	s	Retroflex voiceless fricative	usa
28	s	Dental voiceless fricative	kasam
29	h	Glottal voiced fricative	pan

Table 2 : Material for study

Subjects : Six normal Hindi speaking male subjects in the age range of 20 to 25 years participated in the study.

Procedure : The subjects were instructed to read the words visually presented into a microphone (H-Legend) kept at a distance of 10 cm from the mouth. They were to read each list (randomized) five times. All these were audio-recorded using the Sony Tape Deck (TC-FX 170).

Acoustic Analysis : The words were digitized and stored into the computer memory using a 12 bit A/D converter at 8000 Hz resolution. Wide band bar type of spectrogram were obtained from which frequency of the second formant, F2 transition, frication noise and noise distribution in the stop consonants were measured. Using the wave display the closure duration and duration of speech sounds were measured. The measurements were done as follows:

- (i) Frequency of the second formant (F2) : Frequency of the second formant was measured by placing the cursor on the second dark band visible on the spectrogram

- (ii) F2 transition: F2 transition was measured as the time difference between the offset of the steady state to the end of the F2 for preceding vowel and as the time difference between the onset of the F2 and the steady state of F2 for the following vowel.
- (iii) Frication Noise: This was measured by placing the cursor at the onset of fricatives as visible on the spectrogram.
- (iv) Noise distribution in stop consonants: The frequency distribution of the burst was measured as the frequency difference between the lowest and highest frequency of the burst.
- (v) Closure duration: CD was measured as the time difference between the offset of the preceding vowel and the onset of the burst for the stop consonant.
- (vi) Total duration of speech sound:
 - For vowels and nasals it was the time difference between the onset of regular waveform till the offset of the same.
 - For stops, fricatives and affricates, it was the time difference between the offset of the preceding vowel to the onset of the following vowel/speech sound.

All the measurements were done by using the SSL software of the voice and speech systems, Bangalore.

Statistical Analysis : ANOVA and non parametric statistics were used to find out the inter-subject and intra-subject variability. Also the percent time a parameter was same across and between subjects was calculated by the following formula.

$$\frac{\text{Number of times a parameter was same}}{\text{Number of times a parameter was measured}} \times 100$$

CHAPTER IV

RESULTS AND DISCUSSION

The results are described under three headings

- I. Intra-subject variability
- II. Inter-subject variability
- III. Inter and intra-subject variability for individual words.

I Intra-subject variability

- (i) Spectral Parameters : Table 3 shows the average values of spectral parameters. No significant differences between F_2 , onset of bursts and frication noise was observed for subjects S_1 to S_6 . There was significance difference in onset of bursts, and onset of frication noise for S_6 .

Average	F_2	Onset of Burst	Onset of frication Noise
S_1	1130	939	1283
S_2	1388	951	1473
S_3	1388	647	119
S_4	1476	291	113
S_5	1317	328	137
S_6	1320	161	1736

Table 3: F_2 onset of bursts and onset of frication noise in all the subjects (Hz)

- (ii) Temporal parameters : Table 4 shows the average values of the closure duration, total duration, TDF₂. No significant differences were observed between any of the temporal measures.

Average	Total duration	Closure duration	TDF ₂
S ₁	85	82	64
S ₂	90	77	66
S ₃	83	76	63
S ₄	83	77	75
S ₅	83	76	73
S ₆	88	85	87

- (iii) Percentage of times the values (Spectral and Temporal) was the same :

Table 5 shows the percent of times the values were the same. The total duration of the phoneme and the closure duration of plosives were same maximally and onset of burst was same in minimum percent.

Subjects	F ₂	Onset of friction Noise	Onset of bursts	Closure Duration	Total Duration	TDF ₂
S ₁	54.4	65	25	85	89	79
S ₂	50	30	28	30	30	41
S ₃	51	40	33	71	71	31
S ₄	40	45	44	26	31	38
S ₅	40	60	28	28	35	45
S ₆	55	50	78	50	45	54
Average	48.4	48	38.6	48.4	52	47

Table 5 : Percent times the values were same.

II Inter-subject variability :

Table 6 shows the F2 mean, standard deviation, F2 minimum and F2 maximum for all the 29 words. It was observed that the F2 of the vowel (except words 6, 7, 15, 18 and 27) was significantly different across subjects.

F2 was considered to be the same when the difference between F2 of the word was within 10Hz. A total of 36 same and 138 different values were obtained. F2 was same in 20.6% of measurements and different in 79.4% of measurements.

Words	Sub	Mean	S.D	Max	Min
W₁	1	1274	32	1215	1294
	2	1175	8	1168	1184
	3	1250	16	1234	1278
	4	1162	27	1125	1195
	5	1231	24	1200	1250
	6	1137	0	1137	1137
W₂	1	1420	90	1354	1568
	2	1500	49	1435	1498
	3	1487	9	1478	1592
	4	1560	24	1529	1560
	5	1536	15	1520	1466
	6	1443	14	1430	1592

Words.	Sub	Mean	S.D	Max	Min
W₃	1	1707	34	1659	1754
	2	1730	21	1700	1756
	3	1353	20	1325	1372
	4	1570	15	1550	1592
	5	1356	38	1309	1403
	6	1307	15	1231	1372
W₄₁	1	1410	85.9	1325	1529
	2	1362	13.9	1341	1372
	3	1368	29.29	1344	1419
	4	1461	44.5	1388	1498
	5	1307	28.01	1290	1356
	6	1332	30.7	1300	1372

W₅	1	1080.2	20.8	1043	1090
	2	1111.4	23.9	1090	1152
	3	1173.4	64.5	1120	1275
	4	1085	10.4	1078	1097
	5	1191	189.6	1096	1529
	6	1512.6	0.52	1512	1513
W₆	1	1249.2	25	1211	1278
	2	1255.8	32	1198	1278
	3	1263	80	1184	1378
	4	1281	27	1250	1325
	5	1254	8	1243	1262
	6	1214	10	1200	1230
W₇	1	1597	194	1249	1692
	2	1431	78.9	1309	1498
	3	1541	87.2	1456	1670
	4	1667	14.9	1650	1686
	5	1606	36.8	1376	1670
	6	1532	11.5	1520	1545
W₈	1	1270	31.4	1215	1284
	2	1295	14.8	1184	1219
	3	1306	26.4	1262	1322
	4	1237	28.4	1200	1278
	5	1374	49.8	1309	1440

	6	1333	15.4	1320	1356
W₉	1	1235	21.8	1215	1262
	2	1321	6.92	1309	1325
	3	1278	31.37	1247	1325
	4	1245	7.96	1231	1250
	5	1266	63.36	1200	1372
	6	1376	23.31	1341	1388
W₁₀	1	1156	16	1129	1168
	2	1169	22	1137	1201
	3	1182	4	1175	1189
	4	1143	7	1137	1152
	5	1219	17	1200	1297
	6	1217	23	1250	1297
W₁₁	1	1516	16	1496	1529
	2	1463	23	1435	1498
	3	1558	66	1421	1577
	4	1433	23	1403	1451
	5	1270	56	1247	1380
	6	1690	67	1545	1700
W₁₂	1	2153	60	2062	2204
	2	2069	22	2034	2099
	3	2081	8	2078	2095
	4	1804	52	1749	2872

	5	1780	63	1733	1890
	6	2134	58	2002	2210
W ₁₃	1	1423	20	1382	1436
	2	1516	78	1430	1639
	3	1746	51	1702	1327
	4	1508	52	1466	1598
	5	1605	42	1529	1629
	6	1931	49	1577	1670
W ₁₄	1	1228	25	1210	1272
	2	1262	22	1229	1278
	3	1441	101	1360	1560
	4	1301	25	1278	1341
	5	1438	17	1419	1466
	6	1442	7	1435	1451
W ₁₅	1	1234	40	1168	1270
	2	1250	11	1235	1262
	3	1264	39	1200	1309
	4	1251	15	1247	1284
	5	1301	66	1215	1372
	6	1425	30	1372	1450
W ₁₆	1	2162	39	2094	2188
	2	2120	53	2031	2156
	3	2051	24	2014	2077

	4	2129	41	2105	2203
	5	1189	274	698	7320
	6	1825	274	1513	2235
W ₁₇	1	1261	14	1247	1278
	2	1301	43	1235	1341
	3	1331	16	1312	1356
	4	1199	15	1184	1216
	5	1261	35	1215	1309
	6	1847	257	1388	1980
W ₁₈	1	1116	442	325	1341
	2	1362	34	1325	1420
	3	1346	21	1322	1331
	4	1303	23	1262	1341
	5	1362	85	1231	1435
	6	1316	18	1372	1419
W ₁₉	1	1325	0	1325	1325
	2	1355	41	1320	1420
	3	1302	21	1290	1341
	4	1300	35	1247	1347
	5	1303	1	1305	1309
	6	1409	48	1327	1450
W ₂₀	1	1241	9	1231	1251
	2	1229	19	1215	1255

	3	1183	31	1152	1216
	4	1145	21	1135	1134
	5	1244	26	1200	1270
	6	1215	17	1200	1236
	1	2627	17	2000	2047
	2	2198	87	2125	2298
	3	1999	74	1274	2062
	4	1349	4	1344	1356
	5	1530	44	1493	1594
	6	1846	243	1567	2031
W ₂₂	1	1228	13	1215	1247
	2	1350	13	1325	1357
	3	1309	39	1347	1341
	4	1193	24	1152	1251
	5	1203	26	1168	1231
	6	1351	35	1120	1391
W ₂₃	1	1387	60	1341	1466
	2	1412	52	1341	1472
	3	1495	43	1451	1576
	4	1285	46	1252	1356
	5	1399	77	1309	1498
	6	1464	23	1420	1498
W ₂₄	1	1134	3.1	1130	1138

	2	1190	31.4	1135	1215
	3	1150	38.6	1121	1200
	4	1121	9.5	1105	1130
	5	1156	5.5	1150	1164
	6	1240	23.1	1200	1260
W ₂₅	1	1315	6.8	1309	1325
	2	1385	44.7	1325	1428
	3	1326	12.2	1309	1342
	4	1179	6.9	1168	1134
	5	1269	39.7	1231	1324
	6	1446	48.6	1388	1482
W ₂₆	1	969	13	849	806
	2	937	57	837	980
	3	817	48	715	862
	4	744	29	713	776
	5	817	94	713	964
	6	1637	31	1600	1672
W ₂₇	1	1122	44.96	325	1388
	2	1347	26.29	1309	1372
	3	1342	8.9	1334	1356
	4	1435	11.86	1425	1456
	5	1422	33.9	1372	1466
	6	1454	41.3	1403	1499

W₂₈	1	1220	13.1	1200	1237
	2	1240	12.1	1219	1247
	3	1188	40.6	1152	1254
	4	1203	6.7	1200	1215
	5	1143	14.63	1721	1155
	6	1344	89.9	1210	1419

W₂₉	1	1153	32	1098	1178
	2	1236	12	1231	1253
	3	1124	63.1	1106	1250
	4	1351	31.8	1300	1388
	5	1525	23.4	1498	1558
	6	1231	22.5	1200	1250

Table 6 :F2 values in 29 words

Table 7 : Shows the significant difference between F2 values. It was observed that except words 6, 7, 18, 27 significant difference were observed for all the other words across subjects.

Word	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1 Vs 2	+	+	+	-	-	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	+
1 Vs 3	-	+	+	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
1 Vs 4	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	+	-	+	-	+	+	+	-	+
1 Vs 5	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1 Vs 6	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 3	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 4	-	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 5	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 6	+	+	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 4	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 5	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 6	+	-	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4 Vs 5	+	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4 Vs 6	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5 Vs 6	+	+	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 7 : significant difference between F2 values of subjects

Table 8 and 9 show the mean frequency onset of stop release and their significant difference respectively. A significant difference between subjects was observed except for words 1, 3, 5 and 10.

A difference of 10 Hz with in a word was considered 'same'. A total of 12 same measurements 84 different measurements were obtained. The frequency of onset of stop release was same 12.5% of times and different 87.5% of times.

Words	Sub	Mean	S.D	Max	Min
W₁	1	286.2	13.0	270	239
	2	331	32.4	824	902
	3	766.8	994.4	283	2543
	4	324.2	240.3	0	636
	5	287.4	18.0	256	300
	6	0	0	0	0
W₂	1	211	80	196	217
	2	344	107.0	270	447
	3	528	39.0	478	559
	4	662	175.0	500	886
	5	613	591.0	0	1309
	6	174	389.0	0	870
W₃	1	332	17	305	350
	2	111	145	572	940
	3	501	132	261	591

Words	Sub	Mean	S.D	Max	Min
	4	349	49	295	390
	5	539	95	400	609
	6	121	272	0	609
W₅	1	708	102	621	886
	2	842	73	709	965
	3	1338	159	1200	1592
	4	930	962	0	215
	5	46	64	0	133
	6	0	0	0	0
	W₆	1	1139	35	1090
2		1248	23	1215	1278
3		1147	36	1105	1184
4		480	409	0	912
5		0	0	0	0
6		120	269	0	603

W ₈	1	847	17	823	870
	2	844	73	713	895
	3	623	83	500	713
	4	0	0	0	0
	5	304	430	0	917
	6	0	0	0	0
W ₉	1	1728	34	1690	1775
	2	844	104	745	979
	3	1210	81	1121	1274
	4	480	688	0	1697
	5	828	21	800	851
	6	0	0	0	0
W ₁₀	1	559.6	21.7	521	572
	2	579	10.1	565	588
	3	548.4	46.2	510	619
	4	349.4	49.2	295	390
	5	514.8	291.9	0	729
	6	488.6	273.8	70	839
W ₁₃	1	1443	14.6	1420	959
	2	1303.8	143.0	1152	1443
	3	1238.8	6.3	1232	1245
	4	202.4	277.1	0	512
	5	0	0	0	0

	6	0	0	0	0
W ₁₄	1	669.8	50.5	64	760
	2	668.4	20.3	651	698
	3	507	158.3	290	698
	4	532.4	7.3	521	540
	5	162.4	222.4	0	412
	6	528.3	176.0	368	807
W ₁₅	1	730.2	58.2	635	780
	2	926.4	42.0	880	989
	3	507	158.3	290	698
	4	184.4	255.0	0	512
	5	161.8	147.7	0	274
	6	567.2	41.0	500	596
W ₁₆	1	912	12	902	953
	2	904	32	854	933
	3	382	7	870	886
	4	328	48	293	360
	5	0	0	0	0
	6	136	189	0	384
W ₁₈	1	2179.2	67.3	2134	2298
	2	2241.4	38.2	2210	2298
	3	2151.6	1417	2134	2172
	4	429	392.3	0	721

	5	0	0	0	0
	6	0	0	0	0
W_{19}	1	2157	20.0	2129	2188
	2	1754	373.0	1451	2164
	3	1304	165.0	1121	1435
	4	2188	72.0	2100	2265
	5	216	297.0	0	572

	6	332	89.0	227	431
W_{20}	1	559	13.0	541	575
	2	579	74.0	520	682
	3	442	247.0	0	572
	4	182	251.0	0	500
	5	0	0	0	0
	6	397	224.0	0	541

Table 8 : Mean, SD, minimum and maximum values of onset of bursts (Hz)

Onset burst	1	3	4	5	6	8	9	10	11	13	14	15	16	18	19	20
1 Vs 2	+	-	+	-	-	-	+	-	-	-	-	+	-	-	+	-
1 Vs 3	+	-	-	-	-	-	+	-	-	+	-	+	-	-	+	-
1 Vs 4	-	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+
1 Vs 5	-	-	+	-	+	+	+	-	+	+	+	+	+	+	+	+
1 Vs 6	-	-	+	-	+	+	+	-	+	+	-	+	+	+	+	-
2 Vs 3	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+	-
2 Vs 4	-	-	+	-	+	+	-	-	+	+	-	+	+	+	+	+
2 Vs 5	-	-	+	-	+	+	-	-	+	+	+	+	+	+	+	+
2 Vs 6	+	-	+	-	+	+	-	-	+	+	-	+	+	+	+	-
3 Vs 4	-	-	-	-	+	+	-	-	+	+	-	+	+	+	+	+
3 Vs 5	-	-	-	-	+	+	-	-	+	+	+	+	+	+	+	+
3 Vs 6	+	-	+	-	+	+	-	-	+	+	-	-	+	+	+	-
4 Vs 5	-	-	-	-	+	+	-	-	+	+	+	-	+	+	+	-
4 Vs 6	-	+	+	-	+	+	-	-	+	+	-	+	+	+	+	-
5 Vs 6	-	+	+	-	-	+	-	-	+	-	+	+	+	-	-	+

Table 9 : Significant difference between onset of burst

Table 10 and 11 show the mean frequency of onset of frication and their significant differences respectively. Significant differences were observed between subjects for all words.

A difference of 10 Hz within a word was considered 'same'. A total of 6 'same' measurements and 18 'different' measurements were observed. The frequency of most frication was same 25% of times and different 75% of times.

Words	Subject	Mean	S.D	Max	Min
W ₁	1	2162	57	2078	2235
	2	2203	90	2105	2302
	3	2057	144	1862	2230
	4	1889	158	1690	2108
	5	1931	306	1720	2470
	6	1883	47	1827	1950
W ₂	1	1321	54	1231	1356
	2	1302	69	1233	1386
	3	1535	7	1526	1344
	4	1745	29	1721	1796
	5	1529	54	1449	1510
	6	1326	91	1168	1400

Words	Subject	Mean	S.D	Max	Min
W ₃	1	850	472	21	1121
	2	1775	218	1482	1984
	3	1203	189	1021	1513
	4	2463	21	2436	2190
	5	1911	89	1780	1984
	6	692	509	337	1592
W ₄	1	1320	53	1230	1354
	2	1301	70	1231	1376
	3	1530	15	1524	1340
	4	1750	90	1720	1790
	5	1520	96	1445	1510
	6	1335	180	1160	1425

Table 10 : Mean, S.D., maximum and minimum frequency of frication onset (Hz)

Frication Noise	W1	W2	W3	W4
1 Vs 2		-	+	
1 Vs 3	-	+	-	+
1 Vs 4	+	+	-	+
1 Vs 5	+	+	+	+
1 Vs 6	+	-	-	-
2 Vs 3	-	+	+	+
2 Vs 4	+	+	+	+
2 Vs 5	+	+	-	+
2 Vs 6	+	-	+	-
3 Vs 4	-	+	+	+
3 Vs 5	-	-	+	-
3 Vs 6	-	+	+	+
4 Vs 5	-	+	+	+
4 Vs 6	-	+	+	+
5 Vs 6	-	+	+	+

Table 11 : Significant difference between onset of frication noise.

Table 12 shows the total duration of speech sounds and Table 13 shows significant difference between the total duration of six subjects. Significant difference between total duration of subjects was observed except for the word 9 and 17.

Total duration was considered to be the same in 38 measurements and different in 136 measurements. It was considered as same when the difference between the measurements was within 1 msec. Total duration was same 22% of the time and it was different 78% of times.

Words	Subjects	Mean	S.D.	Min	Max
W1	1	42.8	2.9	40	47
	2	56.4	15.1	40	77
	3	58	5.0	52	65
	4	91.6	5.1	86	98
	5	87	13.4	70	99
	6	94.8	3.6	89	99
W2	1	20.2	2.5	18	23
	2	47	19.0	19	71
	3	69.8	11.3	56	87
	4	86.4	6.4	78	95
	5	65.2	9.8	56	76
	6	81.4	8.0	75	90
W3	1	42	3	37	45
	2	55.2	12.6	42	70

Words	Subjects	Mean	S.D.	Min	Max
	3	83.2	4.4	79	90
	4	95	5.7	86	100
	5	42.4	14.9	20	60
	6	50.6	11.1	40	65
W4	1	171.8	5.2	164	177
	2	60.2	2.1	58	63
	3	58	3.5	52	61
	4	57.4	16.9	37	80
	5	52.2	13.6	30	62
	6	61.6	3.1	59	65
W5	1	82.4	13.4	63	95
	2	73.6	10.5	64	85
	3	91.4	9.8	80	106
	4	106.6	4.8	100	113

		84.6	8.2	75	95
	6	94.2	2.5	90	97
W6	I	55.8	3.6	52	60
	2	66.2	2.3	64	70
	3	51.4	7.5	45	61
	4	61.2	8.8	47	70
	5	83.2	6.4	75	90
	6	72	2.3	70	75
W7	1	76	9.3	66	85
	2	67	7.2	60	75
	3	67.6	2.0	65	70
	4	73	4.8	67	79
	5	72.6	13.1	54	85
	6	74.4	4.5	70	82
W8	1	117.4	5.0	110	123
	2	102.4	8.0	95	115
	3	102.4	30.2	69	129
	4	50	7.9	41	60
	5	95.2	0.8	94	96
	6	108	7.4	100	120
W9	I	70.4	2.0	67	72
	2	70	5.8	65	80
	3	63.4	7.7	50	69

	4	57.2	4.3	50	62
	5	63.6	4.0	59	69
	6	57.6	3.5	54	63
W10	1	77.8	1.7	76	80
	2	77.4	4.6	73	85
	3	88	11.5	75	99
	4	72.2	8.8	65	85
	5	49.2	9.3	37	60
	6	81.2	7.1	70	89
W11	1	87	4.5	80	91
	2	97.2	7.3	89	109
	3	97.4	11.2	90	117
	4	99.6	2.0	98	103
	5	90.8	6.9	81	97
	6	101.4	8.7	90	112
W12	1	141.2	7.1	132	147
	2	137.8	6.1	130	147
	3	102	5.3	95	109
	4	98.2	2.4	94	100
	5	51.6	4.4	45	55
	6	54.6	4.8	49	60
W13	1	122.8	6.9	112	129
	2	128.6	11.0	114	142
	3	119.6	6.1	114	129
	4	100.2	10.3	90	117

	5	90.20	7.1	82	99
	6	88	6.5	78	96
W14	1	116.2	3.2	112	120
	2	103.8	10.3	93	117
	3	83	20.7	60	102
	4	97.6	13.9	77	116
	5	97	2.9	93	100
	6	103.4	2.3	100	105
W15	1	82.4	12.1	61	90
	2	83.6	2.1	81	87
	3	64.4	5.2	56	70
	4	91.8	8.1	79	99
	5	91.4	9.1	75	96
	6	103.4	2.3	100	105
W16	1	59.8	2.2	56	62
	2	64.6	7.1	56	74
	3	65.4	3.2	60	68
	4	77.8	5.8	69	85
	5	76.6	4.7	72	83
	6	81.6	7.1	71	90
W17	1	57.4	5.8	54	61
	2	53	6.8	6	85
	3	59.2	5.9	52	71
	4	57.4	11.4	38	50
	5	73.2	7.7	57	76

	6	109.4	0.5	62	75
W18	1	37.6	5.8	28	43
	2	38	6.8	30	45
	3	55.2	5.9	45	60
	4	57.4	11.4	40	70
	5	73.2	7.7	60	80
	6	109.4	0.5	109	110
W19	1	53	5.7	44	59
	2	61	12.5	44	75
	3	65.4	3.2	60	68
	4	66	12	52	84
	5	84	7.0	75	92
	6	75	7.0	70	85
W20	1	98.6	5.8	92	105
	2	88.2	8.1	74	94
	3	59.2	7.0	52	71
	4	57.2	7.3	48	65
	5	75.4	6.9	70	87
	6	85	5	80	90
W21	1	42.8	2.9	40	47
	2	56.4	15.1	40	77
	3	58	5.0	52	65
	4	91.6	5.1	86	98
	5	87	13.4	70	99
	6	94.8	3.6	89	99

W22	1	42	3	37	45
	2	55.2	12.6	42	70
	3	83.2	4.4	79	90
	4	95	5.7	86	100
	5	42.4	14.9	20	60
	6	50.6	11.1	40	65
W23	1	20.2	2.5	18	23
	2	47	19.0	19	71
	3	69.8	11.3	56	87
	4	86.4	6.4	78	95
	5	65.2	9.8	56	76
	6	81.4	8.0	75	90
W24	1	82.4	13.4	63	95
	2	73.6	10.5	64	85
	3	91.4	9.8	80	106
	4	106.6	4.8	100	113
	5	84.6	8.2	75	95
	6	94.2	2.5	90	97
W25	1	76	9.3	66	85
	2	67	7.2	60	75
	3	67.6	2.0	65	70
	4	73	4.8	67	79
	5	72.6	13.1	54	85
	6	74.4	4.5	70	82
W26	1	77.8	1.7	76	80
	2	77.4	4.6	73	85
	3	88	11.5	75	99
	4	72.2	8.8	65	85
	5	49.2	9.3	37	60
	6	81.2	7.1	70	89
W27	1	55.8	3.6	52	60
	2	66.2	2.3	64	70
	3	51.4	7.5	45	61
	4	61.2	8.8	47	70
	5	83.2	6.4	75	90
	6	72	2.3	70	75
W28	1	122.8	6.9	112	129
	2	128.6	11.0	114	142
	3	119.6	6.1	114	129
	4	100.2	10.3	90	117
	5	90.20	7.1	82	99
	6	88	6.5	78	96
W29	1	116.2	3.2	112	120
	2	103.8	10.3	93	117
	3	83	20.7	60	102
	4	97.6	13.9	77	116
	5	97	2.9	93	100
	6	103.4	2.3	100	105

Table 12 : Mean, S.D., Minimum and Maximum total duration (msec)

Word	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1 Vs 2	+	+	+	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	+	+	+	+	+	
1 Vs 3	+	+	+	+	-	-	-	-	-	-	-	+	-	+	+	-	-	+	+	+	-	-	+	+	+	+	-	-	+
1 Vs 4	+	+	+	+	+	-	-	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+	+	+	+	+	+	+	+
1 Vs 5	+	+	+	+	-	+	-	+	-	+	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	-	+
1 Vs 6	+	+	+	+	-	+	-	-	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 3	+	+	+	-	+	+	-	-	-	+	-	+	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 4	+	+	+	-	+	-	-	+	+	-	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 5	+	+	+	-	-	+	-	-	-	+	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2 Vs 6	+	+	+	+	+	-	+	-	+	+	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 4	+	+	+	-	-	+	+	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 5	+	+	+	-	+	+	-	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3 Vs 6	+	+	+	+	-	+	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4 Vs 5	+	+	+	-	+	+	-	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4 Vs 6	-	-	+	-	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5 Vs 6	+	+	-	+	-	+	-	-	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 13 : Significant difference between total duration.

Table 14 and 15 shows the mean closure duration and significant differences between closure durations of subjects respectively. Significant difference between closure durations of subjects was observed except for those in words 4, 7, and 14.

Closure duration was considered 'same' when the differences between closure duration of the same word were within 1 msec. A total of 32 same measurements and 64 'different measurements was obtained 33% of times the closure duration was same and 67% of times it was different.

Wards	Subjects	Mean	S.D.	Max	Min
WI	1	42	2	40	47
	2	56	15	40	77
	3	56	5	52	66
	4	91	3	86	98
	5	87	13	70	99
	6	94	3	89	99
W3	1	42	3	37	45
	2	55.2	12.6	42	70
	3	83.2	4.4	79	90
	4	95	5.7	86	100
	5	42.4	14.9	20	60
	6	50.6	11.1	40	65
W4	1	171.8	5.2	164	177

Wards	Subjects	Mean	S.D.	Max	Min
	2	60.2	2.1	58	63
	3	58	3.5	52	61
	4	57.4	16.9	37	80
	5	52.2	13.6	30	62
	6	61.6	3.1	59	65
W5	1	82.4	13.1	63	95
	2	73.6	10.5	64	85
	3	91.4	9.8	80	106
	4	106.6	4.8	100	113
	5	84.6	8.2	75	95
	6	94.6	2.5	90	97
W6	1	55.8	3.6	52	113
	2	66.2	2.3	64	60

	3	51.4	7.5	43	70
	4	61.2	8.8	47	61
	5	83.2	6.4	75	70
	6	72	2.3	70	90
W8	1	117.4	5	110	123
	2	102.4	8	95	115
	3	102.4	3.2	69	129
	4	50	7.9	41	60
	5	95.2	0.83	94	%
	6	108	7.4	100	120
W9	1	70.2	2.4	66	72
	2	70	5.8	65	80
	3	63.4	7.7	50	69
	4	57.2	4.3	50	620
	5	63.6	4.9	59	69
	6	57.6	3.5	54	63
W10	1	77	1	76	80
	2	77	4.6	73	85
	3	88	11.5	75	99
	4	72	8.8	65	85
	5	49	9.3	37	60
	6	81	7.1	70	89
W11	1	87	4.5	80	91

	2	97	7.3	89	109
	3	97	11.2	90	117
	4	99	2.7	98	103
	5	90	6.9	81	97
	6	101	8.7	90	112
W13	1	122	6.9	112	129
	2	123	11	114	142
	3	119	6.1	114	129
	4	100	10.3	98	117
	5	90	7.1	82	99
	6	88	6.5	78	96
W14	1	116	3.2	112	120
	2	103	10.3	93	117
	3	83	20.7	60	102
	4	97	13.9	77	116
	5	97	2.9	93	100
	6	103	2.3	100	105
W15	1	82.4	12,1	61	90
	2	83.6	2.1	81	87
	3	64.9	5.2	56	70
	4	91.8	8.1	79	99
	5	91.4	9.1	75	%
	6	103.4	2.3	100	105

W16	1	59.8	2.2	56	62
	2	64.6	7.1	56	74
	3	65.4	3.2	60	68
	4	77.8	5.8	69	85
	5	76.6	4.7	72	83
	6	81.6	7.1	71	90
W18	1	37.6	5.8	23	43
	2	38	6.2	30	45
	3	55.2	5.9	45	60
	4	57.4	11.4	40	70
	5	73.2	7.7	60	80
	6	109.4	.5	109	110

W19	1	53	5.7	44	59
	2	61.4	12.5	44	75
	3	65.4	3.2	60	68
	4	66	12	52	84
	5	84	7	75	92
	6	75	7	70	85
W20	1	83.6	3.6	19	105
	2	88.2	2.1	74	94
	3	59.2	7	52	71
	4	57.2	7.3	48	65
	5	75.4	6.9	70	87
	6	85	5	80	90

Table 14 : Mean, standard deviation, minimum and maximum closure duration (msec)

Word	1	3	4	5	6	8	9	10	11	13	14	15	16	18	19	20
1 Vs 2	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
I Vs 3	+	+	+	-	-	-	-	-	-	-	+	+	-	+	+	+
1 Vs 4	+	+	+	+	-	+	+	-		+	+	-	+	+	+	+
1 Vs 5	+	-	+	-	+	+	-	+	-	+	+	-	+	+	+	-
1 Vs 6	+	-	+	-	+	-	+	-	+	+	-	+	+	+	+	-
2 Vs 3	-	+	-	+	+	-	-	-	-	-	+	+	-	+	-	+
2 Vs 4	+	+	-	+	-	+	+	-	-		-	-	+	+	-	+
2 Vs 5	+	-	-	-	+	-	-	+	-	+	-	-	+	+	+	-
2 Vs 6	+	-	-	+	-	+	+	-	-	+	-	+	+	+	+	-
3 Vs 4	+	-	-	+	+	-	-	+	-	+	-	+	+	-	-	-
3 Vs 5	+	+	-	-	+	-	-	+	-	+	-	+	+	+	+	-
3 Vs 6	+	+	-	-	+	-	-	-	-	+	+	+	+	+	+	+
4 Vs 5	-	+	-	+	+	+	+	+	-	-	-	-	-		+	-
4 Vs 6	-	+	-	+	+	+	+	-	-	+	-	+	-	+	-	+
5 Vs 6	-	-	-	-	+	-	-	+	-	-	-	+	-	+	-	-

Table 15 : Significant difference between closure duration

Table 16 shows the mean, standard deviation, minimum and maximum values of F_2 transition duration for all the 29 words. Table 17 shows the significant difference between TDF_2 of the six subjects. It was observed that the TDF_2 (except in word 4, 19 and 23) was significantly different between subjects.

TDF_2 was considered to be same when there was a difference of 1msec between two TDF_2 of a word. A total of 41 same and 133 different TDF_2 were obtained. TDF_2 was same in 23.5% of measurements and different in 76.5% of measurements.

Words	Subject	Mean	S.D.	Max	Min
W ₁	1	50.6	0.89	50	52
	2	52	1.9	50	55
	3	43	2.6	40	46
	4	58	5.7	52	57
	5	55	1.4	54	56
	6	75	3.3	70	77
W ₂	1	53	1.5	57	55
	2	48	5.27	43	56
	3	52	5.4	46	62
	4	54	3.6	50	58
	5	43	6	41	57
	6	43	5.7	84	99

Words	Subject	Mean	S.D.	Max	Min
W ₃	1	63	4.2	59	68
	2	73	0.2	71	76
	3	61	1.9	59	69
	4	73	5.3	67	81
	5	59	11.4	50	79
	6	88	9.3	76	100
W ₄	1	46	3.6	43	51
	2	50	9.6	42	64
	3	41	9.9	30	49
	4	45	5.16	37	50
	5	51	10.10	37	59
	6	86	6.54	79	95
	7	53	16.87	30	95

W ₅	1	41	3.3	40	48
	2	53	6.1	43	63
	3	45	4.8	41	53
	4	64	6.2	54	70
	5	56	10.13	45	70
	6	72	2.58	70	73
	7	56	11.79	40	73
W ₆	1	51	13	50	53
	2	56	3.7	51	60
	3	52	9.3	44	65
	4	72	8.5	67	75
	5	92	4.9	88	96
	6	92	5.7	85	100
W ₇	1	49	15	47	51
	2	49	7.2	41	57
	3	42	3.4	37	46
	4	80	3.5	68	91
	5	52	9.9	40	62
	6	80	5.06	75	86
W ₈	1	80	3.1	73	83
	2	75	4.7	70	81
	3	70	6.1	61	78
	4	90	8.04	80	101

	5	79	4.6	74	85
	6	103	2.9	100	106
W ₉	1	107	2.3	103	110
	2	107	6.8	102	119
	3	80	11.0	68	93
	4	99	4.02	93	106
	5	86	4.43	79	90
	6	91	2.49	90	96
W ₁₀	1	45	3.2	40	48
	2	44	3.4	39	48
	3	41	3.5	40	48
	4	57	10.7	45	70
	5	81	6.4	75	89
	6	70	5.7	65	79
W ₁₁	1	49	12	42	72
	2	55	3.6	49	52
	3	45	6.9	36	53
	4	80	13.3	71	104
	5	82	12.1	72	101
	6	96	5.9	90	105
	1	109.60	4.7	102	114
	2	105.60	70.9	98	116
	3	82.60	9.1	72	93

	4	112.40	2.9	100	116
	5	89.40	16.0	62	100
	6	106.40	5.9	100	62
W ₁₃	1	35.4	3.2	30	36
	2	47.6	13.0	34	69
	1	68	1.3	66	69
	4	73	17.1	49	94
	5	97	1.6	95	99
	6	104	3.7	100	112
W ₁₄	1	50	1.2	49	62
	2	46	4.7	43	54
	3	37	9.0	25	46
	4	49	8.01	40	51
	5	45	3.96	39	49
	6	82	70.79	73	97
W ₁₅	1	46.4	1.6	45	49
	2	42.8	2.2	39	46
	3	47	7.7	39	66
	4	60	6.3	50	65
	5	60	6.4	55	70
	6	70	4.9	39	76
W ₁₆	1	120	4.1	117	127
	2	104	3.0	103	110

	3	106	3.5	103	110
	4	104	8.7	69	109
	5	91	15.07	80	118
	6	96	4.1	90	100
W ₁₇	1	54	2.3	50	56
	2	100	6.8	90	109
	3	74	6.3	63	78
	4	84	9.3	72	98
	5	106	7.9	93	114
	6	90	4.7	82	94
W ₁₈	1	61		59	64
	2	34	2.1 2.6	50	57
	3	51	2.9	50	56
	4	92	9.2	89	102
	5	74	4.1	68	78
	6	92	6.5	85	100
W ₁₉	1	105.2	13.5	31	113
	2	106.8	3.6	103	112
	3	117	3.8	113	121
	4	76	14.7	50	86
	5	91	5.1	87	100
	6	85	6.9	80	93
W ₂₀	1	47	2.7	45	51

	2	81	7.1	70	87
	3	82	8.3	71	90
	4	91	4.8	86	97
	5	98	1.9	46	51
	6	110	4.1	106	116
W ₂₁	1	75	2.4	73	79
	2	67	3.3	64	72
	3	66	6.22	62	77
	4	85	9.27	76	99
	5	67	4.4	62	72
	6	78	0.54	78	79
W ₂₂	1	93.2	3.5	90	99
	2	81.2	8.4	71	90
	3	84.6	3.6	79	88
	4	84.8	7.5	76	95
	5	73.2	11.7	64	93
	6	83.4	7.3	73	90
W ₂₃	1	84	5.4	80	90
	2	82	2.7	79	85
	3	88	13.7	65	99
	4	83	2.7	80	86
	5	90	8.3	76	96
	6	96	11.6	70	98

W ₂₄	1	69	6.8	60	76
	2	55	6.6	49	65
	3	45	5.2	40	54
	4	77	15.88	50	90
	5	74	2.3	72	78
W ₂₅	6	76	1.3	75	78
	1	68	5.5	59	72
	2	83	6.4	73	93
	3	63	7.6	55	75
	4	73	11.2	58	86
	5	74	2.7	70	77
	6	96	3.9	90	100
W ₂₆	1	58	7.1	50	65
	2	69	4.3	65	75
	3	64	5.9	58	72
	4	68	2.7	64	71
	5	86	20.4	67	118
	6	84	4.1	79	88
W ₂₇	1	48	6.05	42	55
	2	29	0.8	38	40
	3	42	5.5	35	48
	4	55	3.9	49	59
	5	49	4.5	46	57

	6	79	5.7	72	86
W₂₈	1	55	3.7	49	58
	2	50	7.6	42	59
	3	51	6.4	46	59
	4	89	6.04	81	99
	5	64	6.1	60	75
	6	67	1.9	65	70

W₂₉	1	35	3.3	30	38
	2	40	2.5	37	43
	3	48	3.7	42	32
	4	64	2.3	62	68
	5	78	3.3	72	80
	6	91	2.2	90	95

Table 16. Mean, S.D., minimum and maximum TDF2 for 29 words (msec)

Word	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1 Vs 2	-	-	+	-	+	-	-	-	-	-	-	-	+	-	-	+	+	+	-	+	+	+	-	+	+	-	+	-	+
1 Vs 3	+	-	+	-	-	-	-	+	-	-	+	+	+	+	-	+	+	+	-	+	+	-	+	+	-	-	-	-	+
1 Vs 4	+	-	+	-	+	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+	-	-	-	-	-	+	+	+
1 Vs 5	+	-	-	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	+	+	-	-	-	-	+	+	+	+
1 Vs 6	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	-	-	-	-	+	+	+	+	+
2 Vs 3	+	-	+	-	+	-	-	-	+	+	-	+	+	+	-	-	+	-	-	+	-	-	-	+	+	-	-	-	+
2 Vs 4	+	+	-	-	+	+	+	+	-	-	+	-	+	-	+	-	+	+	-	+	+	-	-	+	-	-	+	+	+
2 Vs 5	-	-	+	-	-	+	+	-	+	+	+	+	+	-	+	+	-	+	-	+	-	-	-	+	+	+	+	+	+
2 Vs 6	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	-	+	+	-	-	+	+	+	+	+	+
3 Vs 4	+	-	+	-	+	+	+	+	+	+	+	+	-	-	+	-	+	+	-	+	+	-	-	+	+	+	+	+	+
3 Vs 5	+	-	-	-	+	+	+	+	-	+	+	-	+	-	+	+	+	+	-	+	-	+	-	+	+	+	+	+	+
3 Vs 6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	+	+	-	-	+	+	+	+	+	+
4 Vs 5	-	-	+	-	-	+	+	+	+	+	-	+	+	-	-	+	+	+	-	+	+	+	-	-	-	+	+	+	+
4 Vs 6	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	+	-	-	+	+	-	-	-	+	+	+	+	+
5 Vs 6	+	+	+	+	+	-	-	+	-	+	+	+	-	+	+	-	+	+	-	+	+	-	-	-	-	+	-	+	+

Table 17 : Significant Difference between TDF₂

Table 18 shows a summary of inter-intra subjects differences across all the words.

It appears that more than 67% of the measurements were different between the subjects. However, within the subject not more than 61% of measurements were different. The total duration of the phoneme was the most similar and the frequency of onset of burst was the least similar among the parameters.

Subjects	F ₂		Onset of Frication		Onset of burst		Closure duration		Total duration		TDF ₂	
	% same	% Diff	% same	% Diff	% same	% Diff	% same	% Diff	% same	% Diff	% same	% Diff
S ₁	54	46	65	35	25	75	85	15	89	11	79	21
S ₂	50	50	50	50	28	72	30	70	30	70	41	59
S ₃	51	49	40	60	33	67	71	29	71	29	31	69
S ₄	40	60	45	55	44	56	26	74	31	69	38	62
S ₅	40	60	50	50	28	72	28	72	35	65	45	55
S ₆	55	45	50	50	78	92	50	50	45	55	54	46
Average	43	52	43	52	39	61	48	52	52	48	47	53
Inter subject	20.6	79.4	25	75	12.5	87.5	33	67	22	78	23.5	76.5

Table 18 : Summary of percentage same and percentage difference within and across subjects.

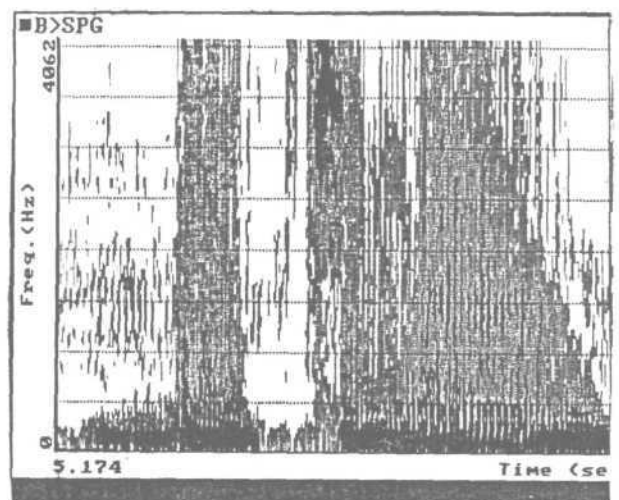
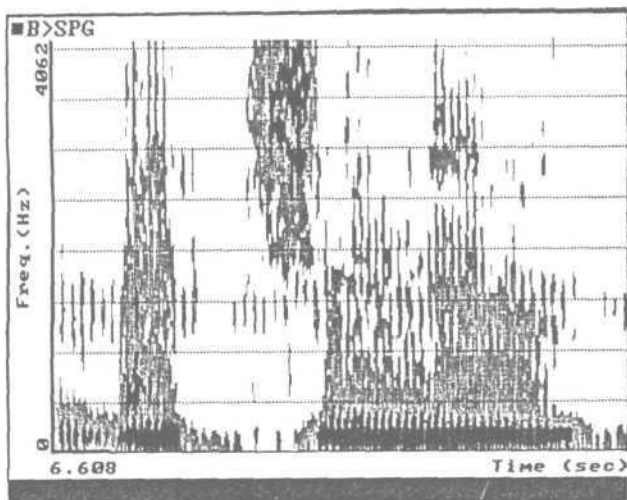
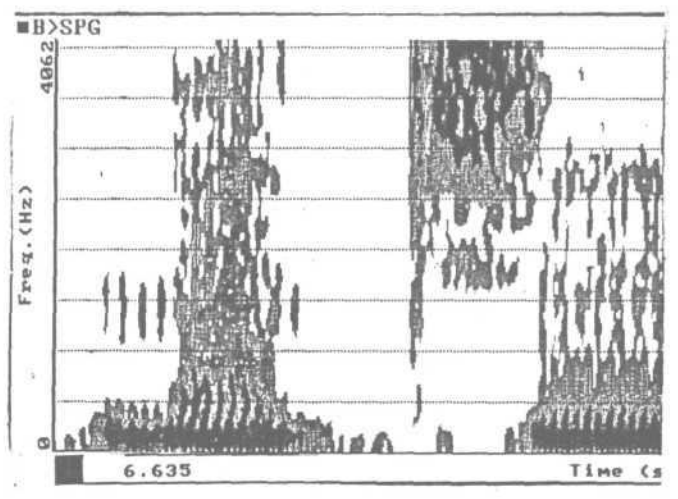
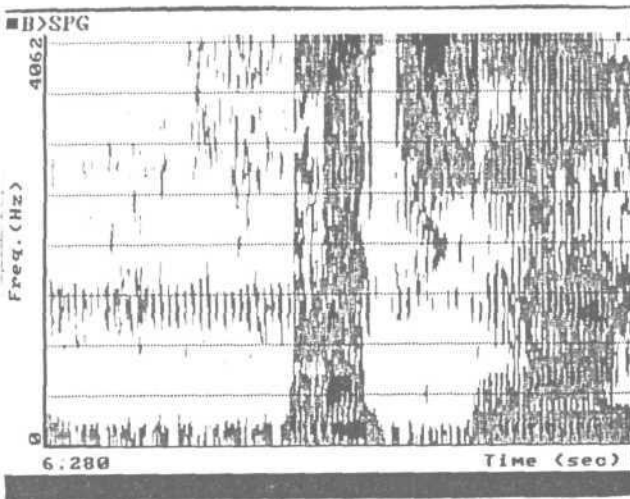
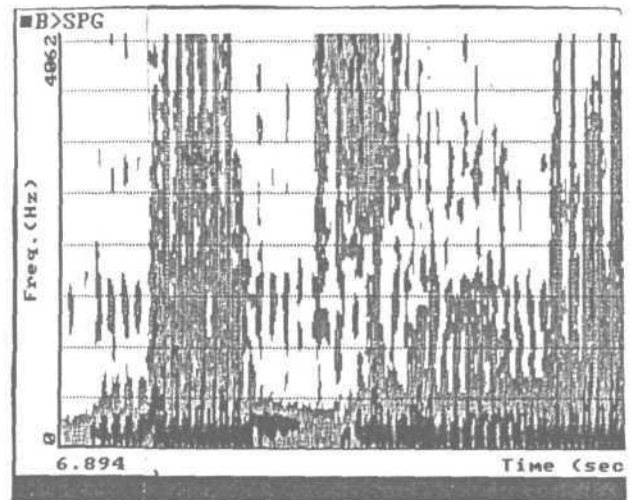
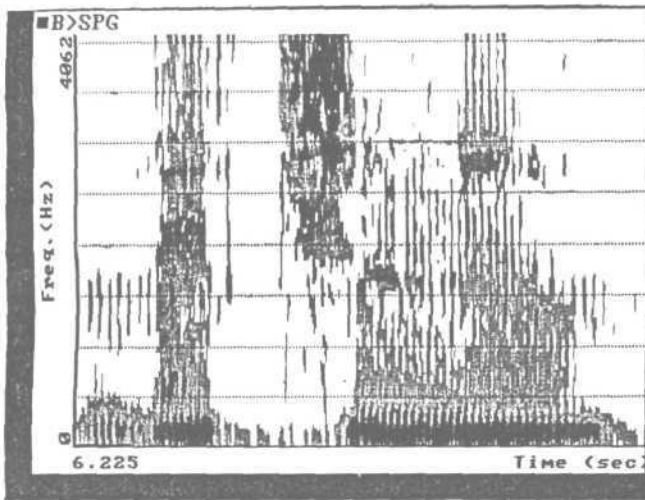


Figure 7 shows the spectrograms of the word 'bacna' for all the six subjects.

Figure8: Variations in closure durations.

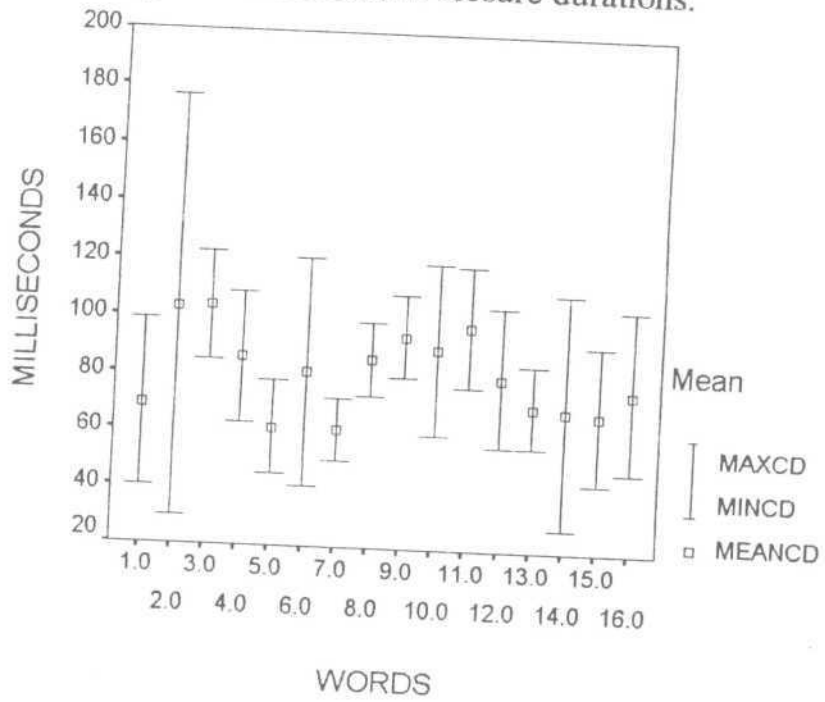


Figure9: Range of F2 values of different vowels.

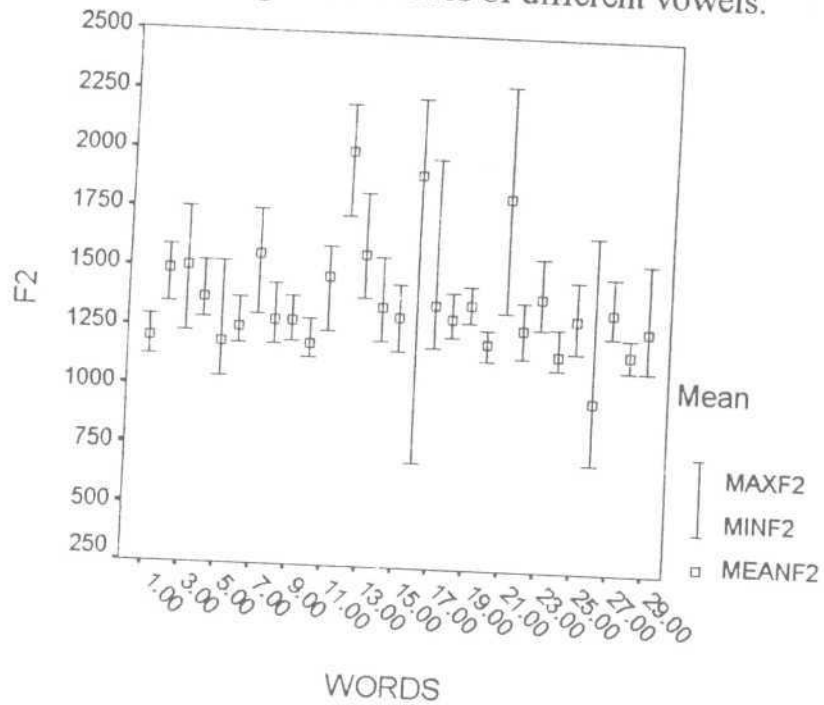


Figure 10: Variations in total durations of phonemes across words.

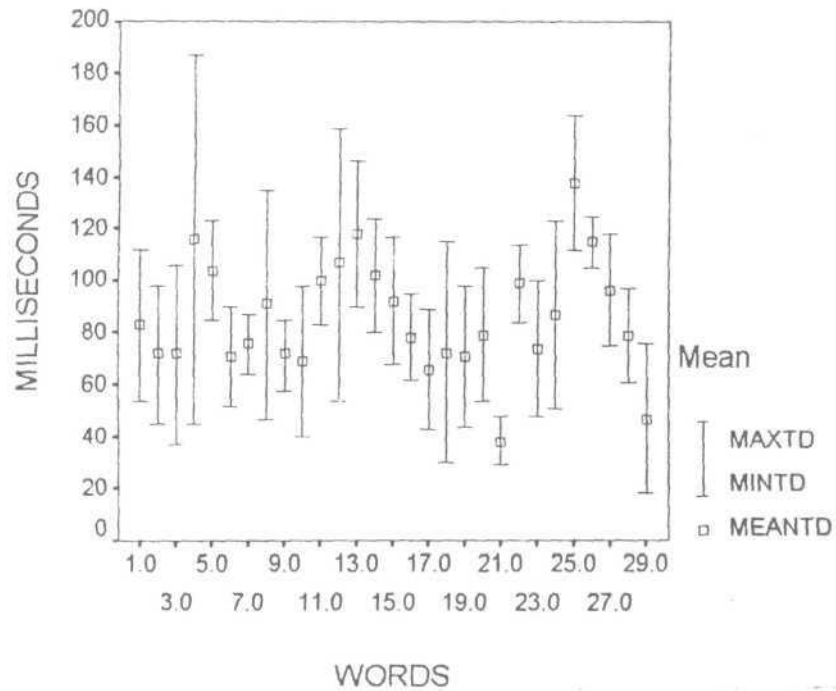


Figure 11: Variations in burst frequency range across plosives.

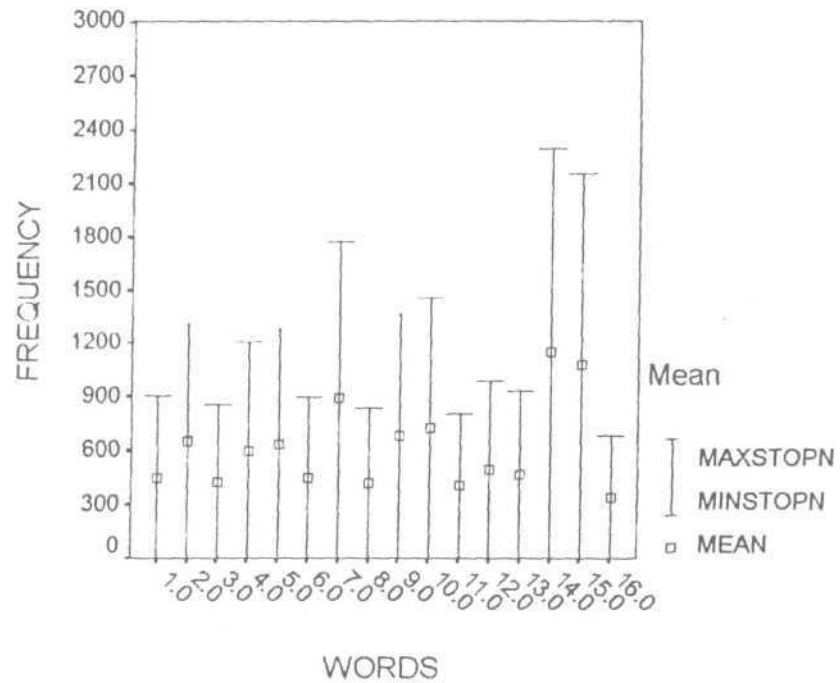


Figure 12: Variations in F2 transition duration across words.

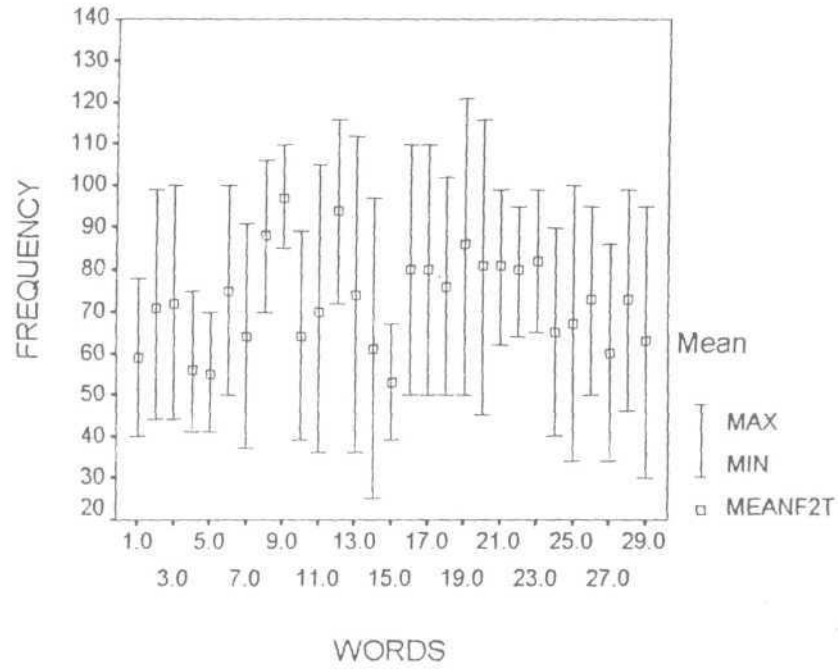
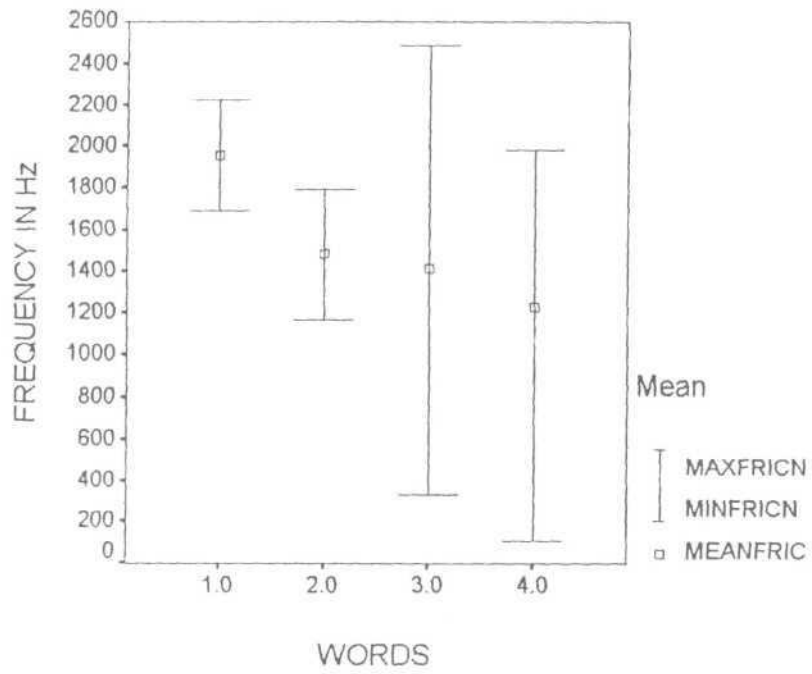


Figure 13: Variations in frication noise across words.



III Intra and Intersubject differences for individual words:

Figure 8 shows the range of closure durations of plosives in all the 16 words. It was observed that the closure durations varies maximally for the retroflexes /t/, /th/, /d/ and /dh/ (as in words 2,5,10 and 11) and variations were least for /d/ as in word 7. Figure 9 shows the range of F2 of vowels as in all the 29 words. The range F2 was maximum in vowels /e/, /i/ and /u/ (as in words 16, 21 and 26) and minimum for /a/ as in word 20.

The range of total duration of phonemes was maximum for /tt/ (word 4) and minimum for /n/ (word 21) (figure 10). Figure 11 shows the range of burst frequencies for all the plosives, the retroflex aspirated /dh/ was highly variable in the onset of the burst and the bilabial aspirated /bh/ was least variable (words 14 and 16 respectively). The F2 transition duration was more variable for /a/ (word 13) and least variable for /a/ (word 5) (figure 12). Similarly, frication noise was most variable for the palatal /J/ and least variable for the dental /s/ (figure 13).

To summarize, the results indicate

- (1) High intra-subject variability
- (2) Least intra-subject variability for total duration
- (3) High intra-subject variability for F2, transition duration of F2, onset of the burst and closure duration.

(4) High intra-subject variability for retroflex in closure duration, onset of the burst, and total duration

(5) High intra-subject variability for F2 of high vowels.

The results reveal several points of interest. First of all, least intersubject variability and high intra-subject variability was observed for total duration of phonemes indicating that this could be considered as one of the best parameters for speaker verification.

Second, the results indicate that more than 67% of measurements were different across subjects. This is in consonance with Table 18 which shows a summary of inter-intra subject differences across all the words. However, it does not confirm the results of Tosi (1979), in that voiceprint is not foolproof 100%. Some false positives and false negatives appear to occur.

Third, high intra-subject variability for the closure duration and onset of the burst was noticed for the retroflex plosives. These being the shortest among the plosive and the most difficult, appears to be uttered differently by different speakers.

Fourth, F2 of high vowels varied largely among subjects. It appears that the positioning of the articulator, the constriction made in the oral cavity, the lip rounding and the length of the oral tract differs among speakers which is more reflected in high closed vowels than in open low vowels.

In view of these results, within the perview of this experiment, it is suggested that two speech samples can be considered to be of the same speaker when not more than 61% of the measurements made are different and two speech samples can be considered to be from different speakers when more than 67% of the measurements are different. It is also suggested that whenever possible, retroflex plosives and high vowels could be considered to bring out differences between speakers. Also, nasal continuants and fricatives may not be considered for acoustic analysis unless a condition prevails their inclusion. It should be kept in mind that out of the arrays of source of variations between the speakers, only phonetic variations are accounted for in this experiment. Even the phonetic variations may be different in telephone, mimiced for disguised speech. Caution should be taken in applying these results to such speech for speaker verification.

It is suggested that methods be established for speaker verification in telephones, mimiced or disguised speech. Also, the other sources of intra-speaker variations, which would be especially more relevant in a multicultural/ multilingual country like India, should be the focus of future studies.

CHAPTER V

SUMMARY AND CONCLUSIONS

Speaker identification is a topic of interest since decades. After the invention of spectrograph, voiceprints have been used in court as a proof. However, while matching the voiceprints of the culprit with that of the suspect, it is not known as to.

- a) What percent of matching provides information that both the prints are of the same person and
- b) What percent of matching provides information that the voice prints are of different persons.

As the speech system is highly variable, it leads to intra and inter-subject variability and how much variation can be accommodated is not known. It is essential that this reliability be known before concluding two voice prints, to be identical or different. In this context, the present study was planned. The objective of the study was to find out the reliability of acoustic measurements in voice identification. The method of test-retest reliability had been used for the study. Twenty nine bisyllabic (CVC, CVCV, CVCVC, CVCCV) meaningful Hindi words with 16 plosives, five nasal continuants, four affricates and four fricatives in the initial, medial and final positions formed the material. Six normal Hindi speaking male subjects in the age range of 20 to 25 years participated in the study. The

subjects were instructed to read the words visually presented into a microphone (H-legend) kept at a distance of 10cm from the mouth. They were required to read each list (randomized) five times. All these were audio recorded using a Sony Tape Deck (TC-FX170). The words were digitized at 8 KHz sampling frequency using a 12 bit A/D converter and stored in the memory of the computer. Using the wave display, the closure duration and duration of speech sounds were measured and using spectrography F2, F2 transition duration, onset of stop bursts and frication noise were measured. All the measurements were done using SSL software of the Voice and Speech Systems, Bangalore. The ANOVA and the non-parametric statistics were used to find out the inter-subject and intra-subject variability.

The results of the study indicate:

- (1) High intra-subject variability
- (2) Least intra-subject variability for total duration
- (3) High intra-subject variability for F2, transition duration of F2, onset of the burst and closure duration
- (4) High intra-subject variability for retroflex in closure duration, onset of the burst, and total duration
- (5) High intra-subject variability for F2 and of high vowels.

Within the perview of the experiment, it is suggested that two speech samples can be considered to be of the same speaker when not more than 61% of

the measurements made are different and two speech samples can be considered to be from different speakers when more than 67% of the measurements are different.

It is also suggested that whenever possible retroflex plosives and high vowels could be considered to bring out differences between speakers. Also, nasal continuants and fricatives may not be considered for acoustic analysis. Unless a condition prevails their inclusion, it has to be kept in mind that out of the array of the sources of variation between the speaker, only phonetic variations are accounted for in this experiment. Even phonetic variations may be different in telephone, mimiced or disguised speech. Hence, caution should be taken in applying the results to such speech for speaker verification.

Further, it is suggested that methods as established for speaker verification in telephone, mimiced or disguised speech. Also the other sources of intra-speaker variations, which would be especially more relevant in a multicultural/multilingual country like India, should be the focus of future studies.

BIBLIOGRAPHY

- Abercrombie, D. (1967): Elements of General Phonetics. Edinburgh : Edinburgh University Press. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Atal, B.S (1976) : Automatic Speaker Recognition based on pitch contours. JASA 52,1687-97. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Argyle, M. (1967) : The Psychology of Interpersonal Behaviour. Harmondsworth: Penguin. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Baldwin, J. (1977) : The forensic application of phonetics: Police Review (18th November) 1609. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Bell-Berti., F (1975) : Control of pharyngeal cavity size for English voiced and voiceless stops. JASA 57, 456-61. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Bell-Berti, F., Raphael, L.J., Pisoni, D.B. & Sawusch J.R. (1978) : Some relationships between articulation and perception. Status Report of Speech Research SR-55/56, 21-32 New Haven, Com: Haskins Laboratories. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Black, Lashbrook, Nash, Oyer, Pedrey, Tosi, Truby: Reply to Speaker Identification by Speech Spectrograms Some further observation. Vol.54.

No 2 (1973) JASA cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.

Bladon, R.A.W and A. Al-Bamerni (1976) : Coarticulation resistance in English /I/. Journal of Phonetics, 4, 137-50. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.

Bladon, R.A.W and B.Lindblom (1981) : Modelling the judgement of vowel quality differences. JASA 69, 1414-22. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.

Bolt H.R., F.S.Cooper, E.E David, P.B.Denes, J.M. Pickett, K.N. Stevens. (1970) : Speaker Identification by Speech Spectrogram. A Scientists view of its reliability for Legal Purposes. JASA, Vol 47, No. 2 (Part 2). cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.

Bolt, R.H., F.S.Cooper, D.M.Crean, S.L. Hamlet, J.G.McKnight, J.M. Pickett, O.I.Tosi and B.D Underwood (1979) : On the theory and practice of Voice Identification. Washington : National Academy of Sciences. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.

Bolt H.R., F.S. Cooper, E.E David, P.B Denes, J.M. Pickett, K.N. Stevens. (1970) : Speaker Identification by Speech Spectrogram. A Scientists view of its reliability for Legal Purposes. JASA, Vol 47, 597-612. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.

- Bricker, P.D and S.Pruzansky (1966) : Effects of stimulus content and deviation on talker identification, *JASA*, 40, 1441-9. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Bricker, P.D and S Pruzansky (1976) : Effects of stimulus content and deviation on talker identification, *JASA*, 40, 1441-9. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Broderick, P.K, J.E. Paul and R.J. Rennick (1975) : Semiautomatic speaker identification system. *Proceedings of the 1975. Canadian Conference on Crime Countermeasures* 29-37, Lexington: University of Kentucky, cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Brown, P. and S. Levinson (1979) : Social structure groups and interaction. In : K.R Scherer and H Giles (eds). *Social Markers in Speech*. Cambridge University Press. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Brown, R (1982) : What is speaker recognition ?. *Journal of International Phonetic Association* 12, 13-34. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Chomsky, N and M Halle (1968) : *The Sound Pattern of English*. New York : Harper and Row. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Coleman, R.O (1971) *Male and Female Voice quality and its relationship to vowel formant frequencies*. *JSHR* 14, 565-77. cited In F. Nolan, 1983,

- (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Coleman, R. (1973) : Speaker Identification in the absence of intersubject differences in glottal source characteristics. *JASA*. Vol 53, No.6. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Coleman, R.O (1976) : A comparison of the contribution of two voice quality characteristics to the perception of maleness and femaleness in the voice. *JSHR* 19, 168-80. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Corsi, P. (1982) : *Automatic Speech Analysis and Recognition. Speaker Recognition: A survey*. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Crompton, A. (1981) *Phonetic representation*. Unpublished paper, University of Nottingham cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Crystal, D. (1969) : *Prosodic Systems and Intonation in English*. London : Cambridge University Press. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Crystal, D. (1975) : *The English Tone of Voice*. London: Edward Arnold. cited In F Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Dane, H.M., Sarma V.V.S. (1980) : *A Pattern Recognition Model of Voice Based Personal Verification System for Forensic Application*.

- Delattre, P. (1967) Acoustic or articulatory invariance? Glossary 1, 3-25. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Douglas - Cowie, E. (1979) : Linguistic code switching in a Northern Irish Village. In: P Trudgill (eds), Sociolinguistic Patterns in British English. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Duncun, S. (1973). Toward a grammar for dyadic conversation. *Semiotica*, 9, 29-46. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Endres, Bambach, Flosser (1971) : Voice Spectrograms as a function of Age, Voice Disguise and Voice Imitation. Vol. 49, No. 6 (Part 2).
- Esling, J. H (1978) : Voice quality in Edinburgh, a sociolinguistic and phonetic study. PhD. Dissertation, University of Edinburgh. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Fischer-Jorgensen, E. (1975) : Trends in phonological Theory. Copenhagen: Akademisk.
- Fry, D.B. (1947) : The Physics of Speech. Cambridge : University Press, cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Fowler, C.A (1980) : Coarticulation and theories of extrinsic timing. *Journal of Phonetic* 8, 113-33. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.

- Garvin, P.L and P.Ladefoged (1963) : Speaker identification and message identification in speech recognition. *Phonetica* 9, 193-9. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Giles, H, K.R. Scherer and D.M. Taylor. (1979) : Speech markers in social interaction. In: K.R. Scherer and H. Giles (eds.), *Social Markers in Speech*. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Gimson, A.C (1980) : *An Introduction to the Pronunciation of English* (3rd edition). London: Edward Arnold. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Glenn, J.W. and N. Kleiner (1968) : Speaker Identification based on nasal phonation, *JASA* 43, 368-72. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Goldsmith (1976) : An overview of autosegmental phonology. *Linguistic Analysis* 2, 23-68. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Harshman, R. P. Ladefoged and L.Goldstein (1977) : Factor analysis of tongue shapes, *JASA* 62, 693-707. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Hazen, B. (1973) : Effects of differing phonetic contexts on spectographic speaker identification, *JASA* 54, 650-9. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.

- Hecker, M.H.L (1971) : Speaker recognition, basic consideration and methodology, JASA 49, 138 (A), cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Hollien, H. (1974a) : Peculiar case of 'Voice Prints' JASA 56, 210-13. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Hollien, H and M.Majewski (1977) : Speaker identification by long term spectra under normal and distorted speech conditions. JASA Vol 62, No.4, 975-80.
- Hollien, Majewski and Doherty (1981) : Perceptual Identification of Voices under normal stress and disguise speaking conditions. JASA.
- Jackobson, R, G.Fant and M.Halle (1952) : Preliminaries to Speech Analysis. Cambridge, Mass. MIT Press, cited in F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Jones, D. (1975) : English Pronouncing Dictionary (13th edition). London: Dent. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Kersta, L.G (1962a) : Voice Identification Nature 196, 1253-7. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Kerswill, L.G. (1962a) Voiceprint identification Mature 196, 1253-7. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Knowles, G. (1978) : The nature of phonological variables in Scouse. In: P. Trudgill (eds), Sociolinguistic Patterns in British English. cited In F.

- Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Kratochvil, P. (1973) : *Tone in Chinese*. In : E.Fudge (eds), *Phonology*. Harmondsworth, Penguin, cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Labov, W. (1966) : *Social stratification of English in New York City*. Washington D.C: Centre for Applied Linguistics. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Labov, W. (1972) : *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Ladd, D.R. (1980) : *The structure of Intonational Meaning*. Bloomington: Indiana University Press. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Ladefoged, P and R.V. Vanderslice (1967) : *The voiceprint mystique*. Working Papers in Phonetics, 7, 126-72. LosAngeles: UCLA. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Ladefoged, P. (1975) : *A Course in Phonetics*. New York: Harcourt Brace Jovanovich. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Lass, N.J., A.S. Beverly., D.K.Nicosia and L.A. Simpson (1978) : *An investigation of speaker height and weight identification by means of direct estimation*.

- Journal of Phonetics 6, 69-76 cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Laver, J. and S. Hutchenson (eds) (1972) : Communication in Face to Face Interaction. Harmondsworth : Penguin. cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Laver, J. (1976): The semiotic nature of phonetic data. York papers in Linguistics 6, 55-62. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Laver, J. (1979) . Voice Quality: A classified Bibliography. Amsterdam . John Benjamin. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Laver, J. (1980) : The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Lehiste, I. (1970) . Suprasegmentals. Cambridge, Mass: MIT Press. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Lehiste (1975) : The phonetic structure of paragraphs. In : A. Cohen and S.G. Nooteboom (eds.), Structure and Process in Speech Perception. Berlin : Springer. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Liberman. M.Y (1978) . The Intonational System of English. Bloomington. Indiana University Linguistics Club, cited In F. Nolan, 1983, (eds), The

Phonetic bases of speaker recognition London: Cambridge University press.

Lindblom, B, J.Lubber and T.Gay (1979) Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics* 7, 147-61. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition* London. Cambridge University press.

Lyons, J. (1977) : *Semantics* Cambridge : Cambridge University Press, cited In F. Nolan, 1983, (eds.). *The Phonetic bases of speaker recognition*. London: Cambridge University press.

MacNeilage, P.F (1979) Status report on speech production. *Proceedings of the 9th International Congress of Phonetic Sciences*, Vol, 9-39. Copenhagen. Institute of Phonetics. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.

Ni Chasaide A (1977) *The laterals of Donegal Irish and Hiberno English. an acoustic study.* MA Dissertation, University College of North Wales, Bangor.

Noll (1973) : Session: Speech Communication VT Speaker Identification, *Speech Recognition and Synthesis*, JASA Vol 53, 1.

Nolan, F.J (1982a) : *The nature of phonetic representations.* *Cambridge Papers in Phonetics and Experimental Linguistics* 1, Cambridge University Linguistics Department. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.

- Nolan, F.J (1982c) : The role of action theory in models of speech production. *Linguistics* 20, 287-308. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London. Cambridge University press.
- O'Connor J.D (1973) : *Phonetics*. Harmondsworth : Penguin, cited In F. Nolan, 1983, (eds). *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- O'Connor , J.D and G.F Arnold (1973) : *Intonation of Colloquial English* (2nd edition). London . Longman. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London. Cambridge University press.
- Pellowe, J and V.Jones (1978) : On intonational variability in Tyneside speech. In P. Trudgill (eds.), *Sociolinguistic Pattern in British English*, cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Perkell, J.S. (1979) : On the nature of distinctive features: Implications of a preliminary vowels production study. In B.Lindblom and S.Ohman (eds.) *Frontiers of Speech Communication Research*, cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London. Cambridge University press.
- Riordan, C.J (1977). Control of vocal tract length in speech. *JASA* 62, 998-1002. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Rosenberg, A.E (1973) : Listeners performance in speaker verification tasks. *IEEE Trans. Aud. and Electro acoustics*. AU21-221-5. cited In F. Nolan, 1983, (eds), *The Phonetic bases of speaker recognition*. London: Cambridge University press.

- Sawashima, M, H Hirose and H. Yoshioka (1978) : Abductor (PCA) and adductor (INT) muscles in the larynx in voiceless sound production. Annual Bulletin of the Research Institute of Logopedics and Phoniatics 12, 53-60. Tokyo, cited In F. Nolan, 1983, (eds). The Phonetic bases of speaker recognition. London: Cambridge University press.
- Scherer, K.R (1979) : Personality markers in speech. In: K.R. Scherer and H.Giles (eds), Social Markers in Speech, cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Stevens, K.N., C.E Williams, J.R Carbonell and B.Woods (1968) : Speaker authentication and identification, a comparison of spectrographic and auditory presentation of speech material. JASA 44, 1959-1609. cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Su, L-S., K.P.Li and K.S.Fu (1974) . Identification of Speakers by use of nasal co-articulation. JASA 56, 1876-82 cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Tatham, MA.A (1969) : Classifying allophones; Occasional Papers 3, 14-22. Language Centre, University of Essex, cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Trim, J.L.M (1959) : Major and minor tone units in English. Le maitre phonetique 112, 26-9. cited In F. Nolan, 1983, (eds). The Phonetic bases of speaker recognition. London. Cambridge University press.

- Trudgill, P. (1974a) : *The Social Differentiation of English in Norwich*. London: Cambridge University Press. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Trudgill, P. (1974b): *Sociolinguistics*. Harmondsworth: Penguin. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Tosi, O., H.Oyer, W. Lashbrook, C.Pedrey, J.Nicol, and E.Nash (1972) : *Experiment on Voice Identification Vol 51, Number 6 (Part 2)*.
- Tosi, O.I (1975) : *Voice Print myth or miracle?* In: J.G.Cederbaums and S.Arnold (eds.), *Scientific and Expert Evidence in Criminal Advocacy*. New York Practising Law Institute. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Tosi, O.I (1975) : *Voice Identification : Theory and Legal Applications*. Baltimore University Park Press. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition* London: Cambridge University press.
- Tosi, O.I (1979) : *Voice Identification : Theory and Legal Applications*. Baltimore University Park Press. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University

press.

- Umeda, M (1977) : *Consonant duration in American English*, JASA 61, 846-58. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- (1969) *The 'Voice Print' myth*. Educational Resources Center, Document ED 028442. Washington D.C. cited In F.

- Trudgill, P. (1974a) : The Social Differentiation of English in Norwich. London: Cambridge University Press, cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Trudgill, P. (1974b): Sociolinguistics. Harmondsworth. Penguin, cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Tosi, O, H.Oyer, W. Lashbrook, C.Pedrey, J.Nicol, and E.Nash (1972) : Experiment on Voice Identification Vol 51, Number 6 (Part 2).
- Tosi, O.I (1975) . Voice Print myth or miracle? In: J G.Cederbaums and S.Arnold (eds), Scientific and Expert Evidence in Criminal Advocacy. New York Practising Law Institute cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition London: Cambridge University press.
- Tosi, O.I (1975) : Voice Identification : Theory and Legal Applications. Baltimore University Park Press, cited In F. Nolan, 1983, (eds), The Phonetic bases of speaker recognition London Cambridge University press.
- Tosi, O.I (1979) : Voice Identification : Theory and Legal Applications. Baltimore University Park Press, cited In F. Nolan, 1983, (eds.), The Phonetic bases of speaker recognition. London: Cambridge University press.
- Umeda, M (1977) Consonant duration in American English, JASA 61, 846-58. cited In F Nolan, 1983, (eds), The Phonetic bases of speaker recognition. London. Cambridge University press.
- Vanderslice, R (1969) : The 'Voice Print' myth. Educational Resources Information Center, Document ED 028442. Washington DC. cited In F.

- Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Wang, W.S.Y. and C.J.Fillrore (1961) : Intrinsic cues and consoneant perception. *JSHR4*, 130-6. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Wells, J.C. (1970) : Local accents in England and Wales. *Journal of Linguistics* 6, 231-52. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Wells, J.C (1982) : *Accents of English : An Introduction*. 3 volumes. Cambridge: Cambridge University Press, cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- Wolf, J.J (1972) : Efficient acoustic parameters for speaker recognition. *JASA* 51, 2044-56.
- Young, M.A and R.A. Campbell (1967) : Effects of context on talker identification, *JASA* 42, 1250-4. cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.
- * Lieberman,P(1967) : *Intonation, perception and Language* - Cambridge, Mars : MIT Press - 1977. *Speech Physiology and acoustic phonetics*. New York: Macmillan . cited In F. Nolan, 1983, (eds.), *The Phonetic bases of speaker recognition*. London: Cambridge University press.