Automatic Speaker Recognition by Linear Prediction: a study of the parametric sensitivity of the model.

by

Anthony McLaren Collins

Thesis for Degree of Master of Arts
In
Information Sciences
of
The Canberra College of Advanced Education

submitted
August 1982

(Replacement copy of title page)

Synopsis

The application of the linear prediction model for speech waveform analysis to context-independent automatic speaker recognition is explored, primarily in terms of the parametric sensitivity of the model. Feature vectors to characterize speakers are formed from linear prediction speech parameters computed as inverse filter coefficients, reflection coefficients or cepstral coefficients, and also power spectrum parameters via Fast Fourier Transform coefficients. The comparative performance of these parameters is investigated in speaker recognition experiments. The stability of the linear prediction parameters is tested over a range of model order from p=6 to p=30. Two independent speech databases are used to substantiate the experimental results.

The quality of the automatic recognition technique is assessed in a novel experiment based on a direct performance comparison with the human skill of aural recognition. Correlation is sought between the performance of the aural and automatic recognition methods, for each of the four parameter sets. Although the recognition accuracy of the automatic system is superior to that of the direct aural technique, the error distributions are highly variable. The performance of the automatic system is shown to be empirically based and unlike the intuitive human process.

An extended preamble to the description of the experiments reviews the current art of automatic speaker recognition, with a critical consideration of the performance of linear prediction techniques. As supported by our experimental results, it is concluded that success in the laboratory rests upon a rather fragile foundation. Application to problems beyond the controlled laboratory environment is seen, therefore, to be still more precarious.

Acknowledgement

The experiments have been performed while the author was working in the Information Science Laboratory at the Department of Engineering Physics, Australian National University, Canberra. The encouragement and critical interest of Professor Steven Kaneff is gratefully acknowledged. Generous support for this project was provided by the Australian Federal Police Force, with the cooperation of Dr. Malcolm C. Hall. A French Government Professional Scholarship to work at Ecole Nationale Superieure des Telecommunications, Paris, and the cooperation of Professor C. Gueguen and Mr. Y. Grenier, catalysed an intensive exploration of the performance of their scheme for automatic speaker recognition based on linear prediction parameters.

The continuing interest, encouragement and patience of my supervisor, Dr. Brian Stone, was instrumental in the culmination of the work described in this thesis.

Table of contents

CHAPTER 1

Speech and speaker recognition

1.1 Introduction

The use of a person's voice as a reliable cue for unique identification has been the subject of extensive research. It has long been known that we can identify speakers from their voices quite reliably. During the past decade, there have been many attempts to design a computer system to perform this task automatically and accurately. Such an automatic speaker recognition system could be employed for many purposes including voice data banks for security controlled access, military communications and the forensic sciences.

Many recent studies have convincingly demonstrated the feasibility of automatic speaker recognition systems while revealing some problems and limitations in a practical application of the techniques. However, as most experiments have been performed upon relatively small and independent

speech databases, it is difficult to compare the efficacy of the different techniques. This is particularly true of techniques based upon the linear prediction model of speech production: several experiments have produced contentious and even conflicting results. Moreover, it is remarkable that (to the author's knowledge) no recent experiment has compared directly the relationship between the intuitive human skill of aural recognition and the actual performance of an automatic speaker recognition system. Thus our experiments focus on a critical consideration of these two points. Due to the rather fragmentary nature of this domain of research, a substantial prologue to the experimental work introduces and broadly defines the principal facets of the problem of automatic speaker recognition. This prologue occupies the latter part of chapter 1.

1.2 Speaker recognition experiments: human v machine.

In subsequent chapters this thesis presents a study of the performance of an automatic speaker recognition system based upon speech parameters derived from the linear prediction (LP) model of the acoustic waveform. A database of digitized speech samples recorded from 25 speakers is used for experiments which explore the stability and efficacy of different LP parameter sets.

Secondly, the same database is analysed by the Fast Fourier Transform (FFT) to facilitate a direct comparison of the efficiency of speaker recognition schemes based upon Fourier spectral parameters with those derived by LP.

Finally, a subsidiary (and smaller) speech database is used to perform a speaker recognition experiment testing human aural skills, and also to independently assess the consistency and validity of the LP and FFT results. Thus a comparison of the results of subjective (human) recognitions is performed upon the same speech database as used for automatic (computer) recognition experiments. Several of the voices in the database are found to be highly susceptible to confusion of identity by the group of listeners performing the subjective recognition experiment. This occurrence is exploited. It affords the opportunity to critically assess the relevance of LP parameters as efficient features for speaker characterization, from the viewpoint of a meaningful perceptual criterion.

1.3 The nature of speech and speaker recognition

The principal aim of speech is communication between human beings. For this reason, each language employs a common semantic code and a common set of phonemes. A given message uttered by different speakers contains basically the

same sequence of sounds. Nevertheless, when two speakers utter the same phrase or even when the same speaker utters the same phrase on different occasions, only certain aspects of the sounds produced are the same.

There are several reasons why aspects of the acoustic patterns of a word differ on separate occasions. The anatomical features of each individual vocal tract are unique, and even for one person, the act of speaking is directly influenced by the current physical and mental state. In fact, in addition to the specific semantic information of the message being uttered, a person's speech includes information about physical identity, mood, the manner of speaking and state of health. Thus, when we attempt to recognise a person from voice alone, we are trying to perceive information specifically about identity but borne by a complex and multi-faceted speech signal.

Speech consists of vowels and consonants produced by air passing through the vocal cords and into the pharyngeal, oral and nasal cavities. The articulation of vowels is determined by the particular position of the tongue in the mouth (high/low,front/back etc.), while consonants are formed by interaction between the lips, teeth, palate and velum. There are individual anatomical variations in the size and the shape of the tongue, the teeth and the three

cavities. They combine with idiosyncratic styles of speaking to produce the differences evident when two speakers utter what are allegedly the same sounds.

It is this difference between speakers, known as inter-speaker variability, which admits the possibility of personal identification by voice alone. Significant differences have been found, however, even when the same sound is uttered by the same speaker on separate occasions. These differences result from such factors as phonetic context, psychological stress, age and illness. They are known as intra-speaker variability.

The validity of any speaker recognition technique must therefore be founded upon the assumption that inter-speaker variability is always greater than intra-speaker variability. If this premise is not upheld, the natural differences in the manner of speaking of a particular person will render the technique unreliable.

1.4 Terminology

Speaker recognition is a term used to describe both the verification and identification of voices. To perform IDENTIFICATIONS, reference samples for N speakers are first established. Given a new and unknown speech sample, the goal

of the system is to identify the speaker as one of these N reference speakers. In speaker VERIFICATION, a person claiming to be speaker X offers a voice sample. The system is designed to compare the patterns from this test utterance with the stored reference pattern of X to accept or reject this claim. Thus the system is required to distinguish between a "friend" and an "imposter". The essential difference between the two situations is that there are N possible decisions for identification, but only two are possible for verification. An optional "no match" category may be permitted for identification, allowing for the situation when the unknown sample differs excessively from all of the stored reference samples.

According to these definitions, our speaker recognition experiments are specifically situations of identification rather than verification.

1.5 Survey of the literature

1.5.1 An overview

A comprehensive review of the field of automatic speaker recognition was published by Atal (1976), with an accompanying paper by Rosenberg (1976) addressing the particular

aspect of automatic speaker verification. The U.S. National Academy of Sciences (NAS,1979) has published "an evaluation of the use of sound spectrograms for identifying speakers from the sound of their voices". Although specifically commissioned by the F.B.I. to review forensic applications, this report also concisely summarizes the general state-of-the-art of the technology of automatic speaker recognition.

It is apparent from these overviews that techniques based upon LP parameters, following the first exposition by Atal (1974), have attracted a great deal of attention. Numerous subsequent papers have considered at length the finer points of various methods of implementation. As some controversial issues have arisen, the literature in this area is considered in greater detail.

1.5.2 LP techniques under scrutiny

The attraction of LP parameters for digital speech analysis lies greatly in the simplicity and speed of the computational algorithms, particularly as they are amenable to efficient implementation by digital hardware. Thus Atal's paper (1974) demonstrating successful automatic speaker recognition catalysed extensive interest in this field of research. Studies by Sambur (1975) and Bunge et al. (1977)

firmly established the value of LP parameters in this application. Another paper by Sambur (1976) described a scheme, called "Orthogonal linear prediction for context-independent speaker recognition", claiming in addition considerable immunity to distortion factors such as those incurred by telephone transmission.

More recently, the results of the latter paper have been subject to critical scrutiny by Fasolo and Mian (1978) and Bogner (1981). It is now apparent that Sambur's results were presented in an unduly optimistic manner. The small size of the database (21 male speakers) biased the performance. A comparison of the Atal and Sambur techniques performed by Fasolo and Mian, using a common database (10 male speakers), in fact favoured the earlier technique. Furthermore, the Bogner experiment revealed minor discrepancies between the published description and the actual execution of Sambur's orthogonal linear prediction technique, and proceeded to show that the claims of immunity from artifacts of telephone transmission are not fulfilled with a larger database (some 50 male speakers). Both Fasolo and Bogner remarked that the small database of test samples would be expected to confer a low statistical significance upon Sambur's claims. Yet, Sambur's paper has been one of the most widely cited in the late 1970s.

This unfortunate sequence of events highlights a general problem in the field of automatic speaker recognition. It is the difficulty of objectively evaluating and comparing the performance of the diverse published schemes. Most of the experiments are based upon relatively small and quite different databases. Thus to assess or to extrapolate the domain of applicability is extremely hazardous, when various acoustical, statistical and socio-linguistic factors of unknown gravity intervene.

1.6 Automatic speaker recognition: a critical comment

There has been a rapid growth of interest in the application of digital computers to speech analysis during the 1970s, and in particular to the dual problems of automatic speech understanding and speaker recognition. This is apparent from the voluminous published literature. However, it is interesting to observe that research on automatic speaker recognition techniques in the USA appears to have peaked in the mid-1970s and is now in a state of relative decline. Evidence for this view is given by the instance of several entire sessions at the Acoustical Society of America conferences being devoted to papers on aspects of this field then, while only a few papers have been presented at recent conferences. The current exigencies of funding and the perhaps undue optimism of the recent past appear to have ef-

fected their toll.

Although the widespread availability of these powerful new tools stimulated the field of research, it is notable that as early as 1970 a well-known paper by Bricker et al. (1971) remarked on the pitfalls of a superficial experimental approach:

> "These results with small populations suggest that the speech signal contains so much information about the talker that one can be distinguished from among 30 or so by a variety of procedures, and that we cannot learn from these studies the relative merits of various ways of representing the signal and reaching a decision."

This profound statement remains the key to the relative lack of general agreement or applicability being achieved by any one of a large number of recently published schemes for automatic speaker recognition, each claiming outstanding performance in the laboratory. Nevertheless, many meritorious and interesting techniques have been elaborated. The difficulty lies in trying to reconcile the varying experimental conditions and so to discern the true basis of the claims, to future profit. It is in this context that our experiments study in detail the stability and efficacy of LP parameters as applied to automatic speaker recognition.

Although the general feasibility of automatic speaker

recognition is clearly demonstrable in laboratory experiments, it is evident that great caution is necessary in applying the techniques to real-world situations. The statistical implications of an inadequate database have been discussed. Careful experimental design can alleviate them, but a more serious problem is outstanding. Most experiments have been performed upon speech databases of some 10 to 20 speakers with voices drawn from a professional or student community such as a university. These people are usually of diverse background, but we do not yet know if other communities are equally likely to contain a similar proportion of confusable voices. In fact, it is not unreasonable to postulate that people of similar socio-economic background and physiological characteristics may have similar voices. Everyday aural experience suggests this. Thus it is difficult to confidently extrapolate the performance of experimental systems to many typical real-world situations.

This is especially true for the applications of interest to forensic scientists (NAS,1979) where various limiting factors usually intrude: the subjects may be uncooperative or attempting mimicry, the recording may be noisy or low quality telephone channels may distort the voices. For this reason, the scientific community is reluctant to endorse the reliability of any automatic or even semi-automatic technique. There are the broadly-based reservations expressed in

the National Academy of Sciences report, "On the Theory and Practice of Voice Identification" (1979). Other papers which reveal the vigorous and sometimes even bitter polemic enveloping the field include those on the controversial machine-aided technique of so-called "voice-prints", Stevens (1973), Tosi (1973) and Hollein (1974). As a practical implementation of any known technique for automatic speaker recognition technique in the real world faces many unsolved problems, including the afore-mentioned, it is evident that this domain of research remains fertile.

As to a suitable database for statistically meaningful tests, note that the scope of our experiments is also bounded. The LP experiments to be described have been performed on two different databases, comprising 25 and 21 speakers respectively, drawn from an Australian digitized speech database of some 100 speakers in all. The speakers are selected of necessity from a varied professional and student community. It is beyond the available resources to compile a database of voices with screening of the subjects for socio-economic, physiological or specific linguistic attributes. However, within this constraint, we specifically address the question of the consistency and relevance of LP and FFT techniques, with reference to a real-world, perceptually meaningful benchmark.

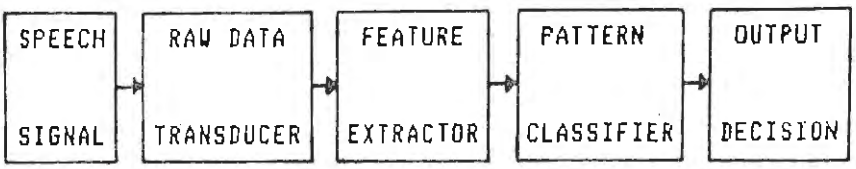
CHAPTER 2

Automatic speaker recognition schemes

2.1 Automatic speaker recognition

A generalised pattern recognition system must be capable of observing a sample of data, pre-processing and transforming it into a meaningful representation space, and then classifying the resultant pattern correctly. The basic configuration of a speech pattern recognition system is shown in figure 1. It consists of three inter-related components: a transducer, a feature extractor, and a classifier. Although these sub-units are highly interdependent in any implementation of a pattern recognition system, it is convenient to consider them separately.

The transducer stage includes analogue-to-digital conversion of the sampled acoustic signals, then pre-processing of the raw data to produce a more compact representation. The feature extraction methods are empirical-



A speech pattern recognition system.

Figure 1.

ly determined, while the pattern classifier design is usually based on statistical procedures. For automatic speaker recognition, all schemes may be categorized broadly on the basis of the speech features used, the computational techniques to extract these features and the nature of the recognition algorithms. The next three sections of this chapter address these points in turn.

2.2 Speaker dependent acoustic feature selection

The quest for stable and reliable speaker dependent features to characterize voices continues. Extensive research in the past decade has yielded steady progress but no fundamental breakthrough. Automatic machine methods of speaker recognition, although sophisticated in terms of computer technology and decision theory, are probably still simple in comparison to the intuitive process involved in human perception. Human speaker recognition can implicitly exploit the entire knowledge accrued in speech communication experiences by an individual. Nevertheless, it should not be overlooked that a machine might be able to make use of factors that a human listener cannot assimilate.

It is not easy to define and then to extract from the speech waveform a set of acoustic features which is invariant for an individual speaker but greatly variable from per-

son to person. Moreover, acoustic features referring to the various sub-messages contained in the speech signal do not fall into distinct and separate sets. Yet recovering one of the sub-messages, viz identity, is the essence of the task of speaker recognition. As it is not possible to isolate the specifically speaker dependent features of speech, the strategy adopted by necessity is to extract and compare some general characteristics bearing speaker identity together with other irrevelant information. The individuality of the voice often manifests itself in utterances of particular phonetic context. In fact, it seems that peculiarities in the pronunciation of known sounds are powerful clues to identity. This trait clearly links the problem of speaker recognition with that of speech understanding and also indicates the difficulties involved in identifying voices speaking in an unknown language. Similarly, the recognition of a person uttering senseless or isolated sounds is not a usual experience. (But note that laughter is frequently distinctive and recognisable, thus should not be categorized as such a "senseless or isolated sound". In this context, its role is often similar to that of the more specific expletives .)

The goal of context-independent, automatic speaker recognition is thus fraught with complexity. Recognition as performed by human intuition is significantly

context-dependent, and so it seems reasonable to try to emulate the practice with an automatic system. This implies that comprehension of the speech message is a desirable pre-requisite to assist in the successful recognition of the speaker identity. But automatic speech understanding is a complementary and equally difficult problem for computers, and it also lacks a simple solution, so the general thrust of research has persevered with the goal of context-independent speaker recognition. This constraint has motivated extensive experimentation with spectral, cepstral and LP parameters to derive recognition features from global measurements of speech parameters rather than from local analyses of the acoustic waveform.

There are other important aspects to consider in the selection of features. The desirable properties for features to be exploited for speaker recognition are (Wolf,1972): (1) a natural and frequent occurrence in normal speech, (2) ease of measurement, (3) time invariance and immunity from the speaker's health, (4) large variation between different speakers but consistency for a specific speaker, (5) tolerance to background noise and resilience with respect to specific transmission distorsions, (6) resistance to conscious effort of the speaker to disguise the voice and (7) resistance to mimicry by other speakers.

Both the unique properties of a speaker's articulation (including physical defects and lisping) and some aspects of the learned pattern of speaking may be suitable candidates. However, no one feature is known to combine all the desirable properties listed above. For this reason, there have been some efforts to evaluate and rank features in terms of their efficiency and practicality (Wolf,1972;Sambur,1975).

Feature extraction and selection have received the greatest attention from the point of view of efficient speech communication. As the semantic information content of speech is of the order of only 50 bits/sec while faithful transmission of the acoustic waveform requires the order of 50000 bits/sec, the incentive for efficient data reduction is obvious. Thus stimulated, highly efficient speech data reduction techniques have been developed. LP encoding of the speech waveform is one of these. Although such techniques are superficially attractive for the generation and storage of compact features for speaker recognition, it is to be noted that speech communication of high intelligibility does not necessarily preserve personal voice characteristics. A coding technique optimal for communication applications therefore may be less appropriate to generate features for automatic speaker recognition systems.

For example, an efficient LP-based speech vocoder system using an all-pole model cannot accurately encode nasal sounds. As these sounds often strongly characterize individual voices, important clues to identity are suppressed in transmission even though the overall speech intelligibility is high. Thus we must carefully distinguish intelligibility as measured by the accuracy of speech message communication from the fidelity of representation of all of the information in the acoustic waveform, including perhaps even sounds other than speech.

This brief digression considering aspects of the problem of speech encoding reveals a dilemma inherent in the process of feature extraction and selection. The choice of a data reduction/parameterization technique, as necessarily applied to condense the speech waveform data during the preliminary stage of processing, substantially constrains the scope of these features. It is the opinion of the author that the failure of LP to meet the optimistic claims made in the mid-1970s, heralding a breakthrough in speaker recognition techniques, is significantly explicable on these grounds. LP is an excellent tool but not a panacea.

2.3 Feature extraction by computer

2.3.1 The speech signal

Two different types of sounds can be distinguished in the speech signal. They are voiced sounds, created by a quasi-periodic impulsive excitation of the vocal tract generated by the vibration of the vocal cords, and unvoiced sounds, created by a noise-like excitation generated by a partial obstruction of the vocal tract. Typical voiced sounds are vowels, while fricatives such as 's' and plosives such as 'p' belong to the second category. Voiced sounds have a fundamental frequency which is the inverse of the period between two successive excitation impulses of the vocal cords. Depending on the vocal-tract configuration, voiced sounds also exhibit several formants (or resonant frequencies) which can be measured by spectral analysis of the speech signal. During the articulation of speech, the shape of the vocal tract is varied almost continuously. However, as the rate of change is limited by the dynamics of the articulatory system, the assumption of a short-time stationary linear model of the vocal tract is a reasonable premise.

Spectral analysis of a brief segment (or time-window) of voiced speech signal reveals a frequency distribution (or power spectrum) containing peaks corresponding to the formants. The spectrum is in fact a line spectrum with a ser-

ies of harmonics of the excitation frequency shaped according to the formant envelope. The noise-like signal of unvoiced sounds displays no such fine line structure. Lengthening the time-window for analysis increases the frequency resolution of the harmonic structure but blurs the temporal variations of the formant-dictated envelope. Conversely, a shorter time-window reduces the frequency resolution but more clearly reveals the temporal variations of the formants, viz of the frequency distribution. This property of the time-window highlights a compromise which is central to an understanding of the concepts of digital speech processing.

2.3.2 Digital speech processing

The advent of digital computers has facilitated the analysis and modelling of speech signals with a variety of powerful and flexible mathematical techniques. The usual procedure is to process a series of contiguous (or partially overlapping) time-windows of sampled and digitized speech data, to produce a more compact and perceptually meaningful representation of the signal. This representation may in turn be used to efficiently encode the speech for storage or transmission, or to investigate diverse aspects of the mechanism of speech communication, eg. the role of pitch as a measure of psychological stress in the speaker.

Fourier analysis and LP (or inverse filtering) of digitized speech signals form the basis of many methods of spectral analysis, pitch extraction and filtering. Our study investigates aspects of the performance of these techniques for speaker characterization. The general objective is to demonstrate a transformation to some representational space from which it is readily possible to derive efficient, compact and reliable features for speaker recognition.

2.3.3 Global features from LP and FFT

We have considered (section 2.2) the desirable attributes of features for automatic speaker recognition systems, noting that context-dependent features are potentially attractive but difficult to derive in an automatic system. For this reason, many experiments have been performed to investigate the properties of global features, viz those which may be computed without A PRIORI knowledge of the speech signal. To this end, the transducer, as shown in figure 1, accepts raw data of the speech utterances to be classified. It is required to implement blind signal processing to transform each window of the raw data into an n-dimensional vector in an n-dimensional Euclidean pattern space. Typical digital signal processing algorithms, such as the FFT and LP, are executed in this phase of the computations to generate the feature vectors of condensed raw data.

2.3.4 LP analysis of speech

The theory of the LP model of speech production is expounded in great detail by Makhoul (1975). A complementary view appears in Markel and Gray (1976), covering both theoretical and computational considerations including practical FORTRAN algorithms. There are several alternative formulations of the LP model of the speech waveform which lead to different sets of coefficients. As our experiments compare three LP coefficient sets, we briefly discuss their inter-relationships.

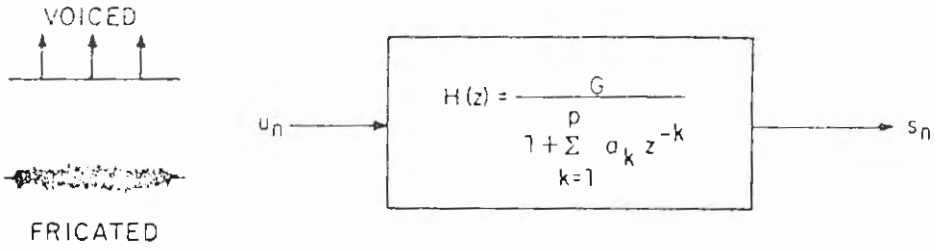
Linear prediction is a method of designing a time-varying filter, optimal according to a least mean squares error criterion, with a transfer function approximating the spectrum of a given signal. The result of applying some uniform excitation to the filter's input should be the given signal as output. A compact set of coefficients is sufficient to specify a digital implementation of the filter (as hardware or software) in the form of a simple iterative structure or algorithm. The alternative name for this technique, inverse filtering, arises from the design criterion, viz if the filter is estimated in the form of an all-pole filter with a transfer function for sampled data

$H(z) = G / A(z)$, then

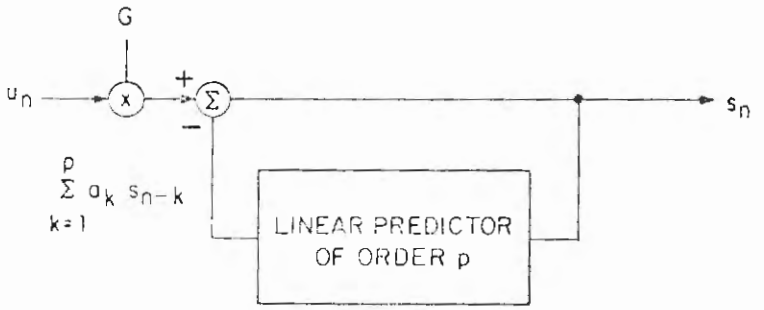
$$A(z) = 1 + \sum_{K=1}^{P} a_K z^{-K}$$

is an inverse filter where the a_Ks are the filter coefficients (the set referred to as the Ai). The result of subjecting the given signal to this inverse filter should be the removal of all BUT RANDOM variations in the output. The order of the model, p, is the number of poles or predictor coefficients. G is the gain factor. The alternative frequency-domain and time-domain models are illustrated in figures 2a and 2b respectively, where the input sequence U(n) excites the models to yield a corresponding output sequence S(n) of speech data samples. For a time-window of N samples of speech data, the computational complexity is of the order of $N.p+p^2$ multiplications.

If H(z) describes a stable, all-pole filter structure (minimum phase), then A(z) may be realised in either the direct form of a filter or as an equivalent lattice form. The reflection (or partial correlation (PARCOR)) coefficients, k_i, of this lattice are related to the predictor coefficients, a_K, by the backward recursion



(a) FREQUENCY-DOMAIN MODEL



(b) TIME-DOMAIN MODEL

Alternative discrete models of speech production
by Linear Prediction (after Makhoul, 1974)

Figure 2.

$$k_i = a_i^{(i)}$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2} \quad , \quad 1 \leq j \leq i-1 \quad , \text{ where}$$

index i takes in turn values p,p-1,...,1
and $a_j^{(p)} = a_j$, $1 \leq j \leq p$ initially.

This transformation of the LP coefficients is of particular interest because the set of lattice filter coefficients, Ki, has been successfully interpreted by Atal (1971) and Wakita (1973) as the reflection coefficients of an acoustic tube that directly models the vocal tract.

Another transformation, following Oppenheim (1968), derives the set of cepstral coefficients, Ci, corresponding to the spectrum of the signal, by the recursive relationship, viz

$$c_1 = a_1$$

$$c_j = a_j - 1/j \sum_{k=1}^{j-1} (kc_k) a_{j-k} \quad , \text{ for } 2 \leq j \leq p.$$

The simplicity of this derivation of the cepstral coefficients from the LP inverse filter coefficients is in contrast to the direct method of cepstral analysis via the FFT (see section 2.3.5). Further details of the interesting relationship between the cepstral and predictor coefficients

are given by Schroeder (1981).

2.3.5 FFT analysis of speech

The Cooley-Tukey algorithm (1967) for the efficient computation of Fourier coefficients has made practicable the widespread application of spectral analysis techniques, performed either by digital hardware or by software. Known as the fast Fourier transform (ie. the FFT), it is usually applied to a time-window of digitized speech data comprising a number of samples N which is an exact binary power. In this case, the computational complexity is of the order of $N.\log_2 N$ multiplications: not dissimilar to LP (for typical values of say N=256 and p=12) providing that the requisite sine and cosine values are tabulated.

The application of the FFT algorithm yields a spectral vector of N/2 complex frequency domain elements. To form a power spectrum the squared moduli of these N/2 complex numbers are calculated. A corresponding cepstral vector is generated by computing the spectrum of the logarithm of the power spectrum. Thus two FFT operations with an interposed logarithmic scaling factor are necessary, making this the most complicated of the commonly used speech analysis algorithms.

2.4 Classification techniques

After extracting appropriate features (at least as candidates) to characterize a speaker it is necessary to apply a pattern classification scheme. The assessment of the suitability of a particular classification technique is by no means an easy task. Most speech features for automatic speaker recognition are not subject to a precise or quantitative definition, so statistically based techniques of pattern classification are usually favoured. An iterative design process may be necessary to develop a viable pattern recognition system which optimises the many variables. This is especially true for a system based on global (context-independent) features, which may in turn be derived from the speech signal by an empirically determined procedure.

A lucid and comprehensive review of the field of statistical pattern recognition was published recently by Chen (1978). Bricker et al.(1971) surveyed statistical techniques specifically used in the domain of speaker recognition. Although some years old relative to the rapid developments during the seventies in this field, the paper is still a valuable reference. Here, we briefly summarize the issues of specific relevance to our experiments.

The reduction of dimensionality of the chosen feature vectors is an issue central to the practical implementation of automatic pattern classification techniques. It is usually necessary, on grounds of computational complexity, to reduce the dimensionality of the feature vectors before submission to the classification algorithm. We use the method of principal component analysis as a preliminary step towards classification. This device, also known as the Karhunen-Loeve transform and closely related to Factor Analysis (Watanabe,1965; Fukunaga,1972), is widely used for speech data reduction.

Given that the mean value of a feature vector of order n and its corresponding matrix of covariance, both computed over a total of M windows of speech data, completely characterize the data of a sample set assumed to follow Gaussian statistics: then, a single new direction in the vector space is determined to explain as much of the variance of the data as possible. This direction in the vector space is a linear combination of the n original directions, with direction cosines defined by the principal eigen vector of the matrix of covariance. The remaining eigen vectors in turn define mutually orthogonal dimensions, ordered in decreasing contribution of variance to the total variance of the dataset.

A powerful statistical tool may be both convenient and

necessary for the reduction of dimensionality to generate concise features for classification. Nevertheless, it must be emphasised that a careful selection of the raw features is still fundamental to success. As Chen (1978) states:

> "It is a misconception that feature extraction is nothing more than dimensionality reduction and that the Karhunen-Loeve expansion solves all mathematical feature extraction problems."
>
> "As a final remark ... feature extraction will remain a key problem in pattern recognition. Experimental methods should be relied on whenever the theoretical mechanism is inadequate."
>
> "A major weakness of statistical pattern recognition is in the difficulty to take the contextual relations into account in the recognition process."

The dimensionality of feature vectors must also be considered in relation to the test dataset. Due to the usually limited sample sizes for recognition experiments in laboratory environments, misleading or optimistic classification results may be obtained. Foley (1972) and Sarma and Venugopal (1977) have commented upon this problem. It is recommended that the ratio of the sample size per class (N) to the dimensionality of the feature vectors (L) should be at least 3 for stable results. Smaller values of N generally lead to an optimistic bias. According to this criterion, the results of many prominently quoted experiments are of dubious value. Bogner (1981) also demonstrates the statist-

ical limitations of experiments such as Sambur's (1976). Certainly the trends observed are illustrative, but it is extremely hazardous to extrapolate and compare results of low statistical significance.

Finally, after the generation of an appropriate feature vector, a distance measure is required to distinguish between and so to classify the vectors. The comparative performance of various metrics is presented by Bricker et al. (1971), Bunge (1977) and Grenier (1978). As the Mahalanobis metric is shown to be both powerful and independent of arbitrary linear transformations of the feature vectors, it has been chosen for our experiments. According to this metric, the distance between an utterance vector x and the reference vector u is

$$d = [\ (x-u)^T W^{-1} (x-u)\ ]$$

where W is the covariance matrix averaged over all speakers. Although in fact speaker-dependent, a good estimate of the covariance matrix is often difficult to compute (due to small sample sizes). Thus the averaged matrix is usually taken as a compromise which in practice is found to perform satisfactorily.

CHAPTER 3

A Linear Prediction recognition experiment

3.1 Automatic speaker recognition using LP parameters

An automatic speaker recognition experiment based on LP parameters is described, using the scheme of Grenier (1978) as a point of departure. The success of this scheme, with quite short phrases (=1.5 secs) and achieving up to 98% recognition accuracy, is impressive. However, the small database of 11 speakers and 10 reference phrases is insufficient to establish a high statistical significance. Also, although not highlighted in the published description of the experimental conditions, the database comprised 3 women and 8 men: the implicit partitioning into high and low pitched voices probably simplified the classification task and so positively biased the success scores. (Note : the speaker recognition scores in the 1978 paper by Grenier are drawn from his 1977 thesis but are corrected for a systematic error extant in the thesis. I diagnosed the problem and recomputed all results while working in collaboration with

Grenier at ENST, Paris.)

It is in this general context that our experiment is designed to further validate and extend the performance tests of the Grenier scheme. As a more stringent condition, our reference phrases are chosen to be different: a set of 14 is used instead of the single key phrase, repeated 10 times. The Australian test database comprises 25 male speakers, each contributing 14 different, brief phrases for recognition and also to generate the references.

For each speaker, the 14 phrases are individually modelled in terms of a mean vector of LP coefficients of order p, typically p=18, and the corresponding matrix of covariance. This mean vector is in fact the average value taken over each set of time windows for analysis. For a typical 1.6 second phrase comprising some 100X 256-sample analysis windows, 200 coefficient vectors are averaged (with a 50% overlap per advance of the analysis window). Following Grenier, after principal components analysis, a composite feature vector consisting of a concatenation of the original mean vector and the 1st eigen vector (or principal axis) is formed to model each phrase. This was found empirically by Grenier to be the most successful feature vector of several combinations presented in his 1978 paper. The ensemble of feature vectors for each speaker is then averaged to gen-

erate a reference vector to characterize this speaker. Thus a total of 25 reference vectors is produced.

Finally, each of the phrases of each speaker is tested in turn against the set of references and the nearest reference is selected as the speaker's identity, according to a Mahalanobis metric. The references are modified on a "leave-one-out" basis to compensate for the bias arising from the inclusion of the specific test phrase. Using the Australian database for example, the recognition score is computed as the overall percentage of correct matches made upon individually testing each of the 14 test phrases per speaker against the 25 references in turn, a total of 350 tests eg. a 98% score arises from 7 errors of recognition .

3.2 The design of the LP experiment

The goals formulated for our extended tests of the Grenier scheme are:

(1) to explore the performance with an expanded speech database, establishing greater statistical significance for the original conclusions and employing context-independent data.

(2) to explore the relative performance of the princi-

pal LP parameters, viz Ai, Ki, Ci, with respect to the order of the LP model, over a range of p=6 to p=30.

(3) to explore the stability of the performance of these LP parameters with respect to different test databases of speech.

(4) to extend the comparison of LP parameters to those of the FFT, under otherwise identical experimental conditions.

(5) to explore the capability of an automatic system based upon LP and FFT parameters to perform recognitions in a direct comparison with the human skill of aural recognition.

The experimental conditions for each of (1)-(4) are elaborated, together with the results and comments, in sections 3.4.1 to 3.4.4 respectively. A discussion of (5) is deferred until section 5.1, following a description in chapter 4 of the corresponding experiments on speaker recognition as performed aurally by people.

3.3 The speech database and processing parameters

The principal database used for the LP and FFT experi-

ments is a subset of a large digitized, speech database assembled by the author at the Department of Engineering Physics of the Australian National University (Collins, 1978). For this experiment, a set of 14 sentences is chosen from the whole. Firstly, six sentences as used by Wolf (1972) and Sambur (1975), and then a further eight sentences composed in consultation with Dr. D. Bradly (of the Department of Linguistics, Australian National University) to produce a still small but more phonetically balanced speech sample set:

1. Cool shirts please me.
2. Pay the man first please.
3. I cannot remember it.
4. Papa needs two singers.
5. A few boys bought them.
6. Cash this bond please.
7. Today I auction beer.
8. How do you know?
9. There she sits.
10. A boy played a tune.
11. The bear chews on his paw.
12. Are you poor?
13. June danced hard.
14. We are firm.

The speech comprising (inter alia) the above set of utterances is recorded for each of the contributing speakers with a high fidelity reel-reel taperecorder (Revox A77), prior to digitization with an analogue-to-digital conversion accuracy of 12 bits and a sampling frequency of 16kHz. A 48dB/octave anti-aliasing lowpass filter is employed. All LP and FFT analysis of the digitized speech data is performed on Hanning windows of 256 samples, ie. 16msec windows, computed every 8msecs. For each phrase in turn a parametric model comprising the mean vector and the matrix of covariance is computed, usually of order n=18 for LP coefficients. Via FFT spectral analysis of 62.5 Hz resolution an analogous spectral vector of the same order is generated by averaging spectral samples (in groups of 6) over the range of 62.5 Hz to 6.75 kHz.

3.4 Discussion of the results

3.4.1 LP models ranked in context-independent recognition

As a foundation for our experiments, the performance of the Grenier scheme (corrected as noted above) is tested with Australian speech data comprising 14 brief phrases. Apart from the change of language from French to English, the use of context-independent data is the primary variable of in-

terest. All other experimental parameters are identical, excepting an increase in the sampling frequency from 10 to 16kHz which suggests a corresponding increase in the LP model order from p=12 to 18. This point is further studied in section 3.4.2. Our results shown in table 1 validate the Grenier technique for automatic speaker recognition. Applied to the larger database with the more stringent condition of context-independence, the increased error rate is to be expected. A clear hierarchy is apparent from the recognition scores for the 3 LP coefficients, Ki, Ai and Ci. The cepstral coefficients (Ci) most successfully characterize the individual speakers, confirming results by Atal (1974), Bunge (1977) and Grenier (1978).

3.4.2 Sensitivity to the LP model order

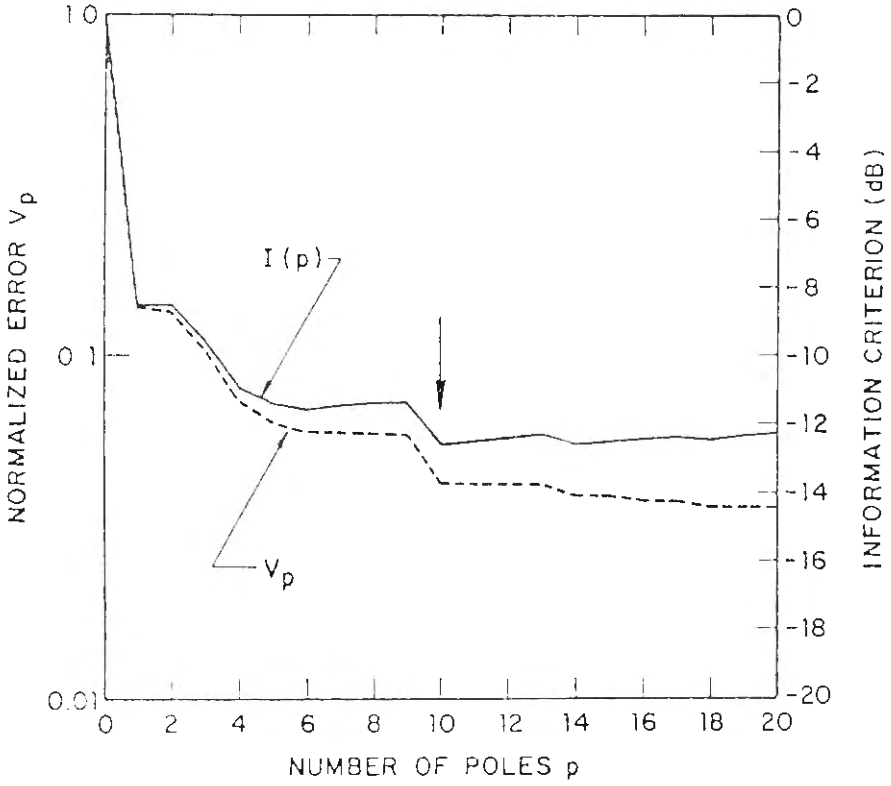
We remark in 3.4.1 that the LP model order p has been increased in proportion to the sampling frequency. Like many other experimenters, Grenier used a "rule-of-thumb" value for p, viz sampling frequency expressed in kiloHertz plus 2. The selection of a suitable value for p has been studied from various viewpoints in the literature: Makhoul (1974) presents both analytical and experimental arguments. The former, based on Akaike's (1974) information criterion I(p), is shown from figures 3 and 4 to recommend p=3 for unvoiced and p=10 for voiced sounds. Here,

Experiment	Phrases	Speakers	Ki	Ai	Ci
Grenier	11	10	97.3%	98.2%	98.2%
Collins main database	14	25	87.7%	92.0%	97.1%

Table 1.
Comparison of Grenier and Collins recognition scores.

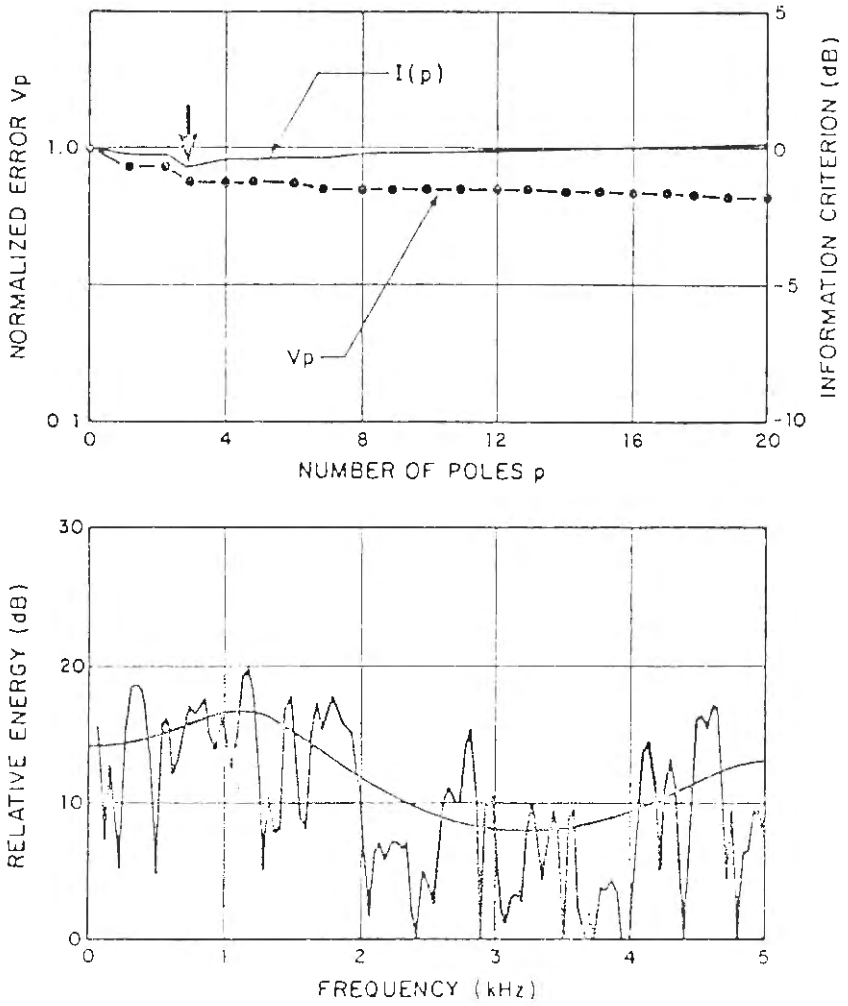
Experiment	Phrases	Speakers	Ki	Ai	Ci
Collins main database	14	25	87.7%	92.0%	97.1%
Collins ave of 11 partitions	14	15	91.8%	95.0%	98.6%
Subsidiary database aural expt.	6x12	21	93.7%	96.8%	98.4%

Table 2.
LP parameter stability experiments.



A plot of Akaike's information criterion versus order p of the predictor for voiced sound. The optimal value of p occurs at the minimum of I(p), shown by the arrow at p=10 (after Makhoul, 1974).

Figure 3.



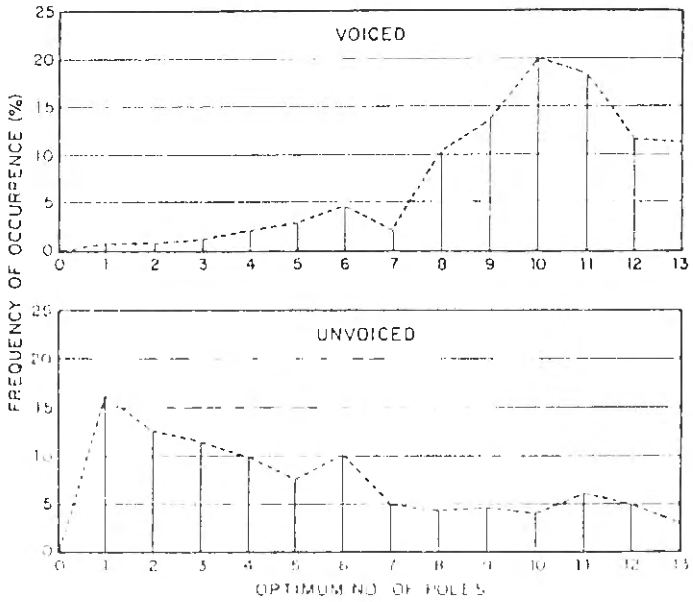
Application of Akaike's information criterion to a fricative sound (after Makhoul, 1974).

Figure 4.

$$p = (I(p) - LOG(Vp)) *0.5 *N *c ,$$

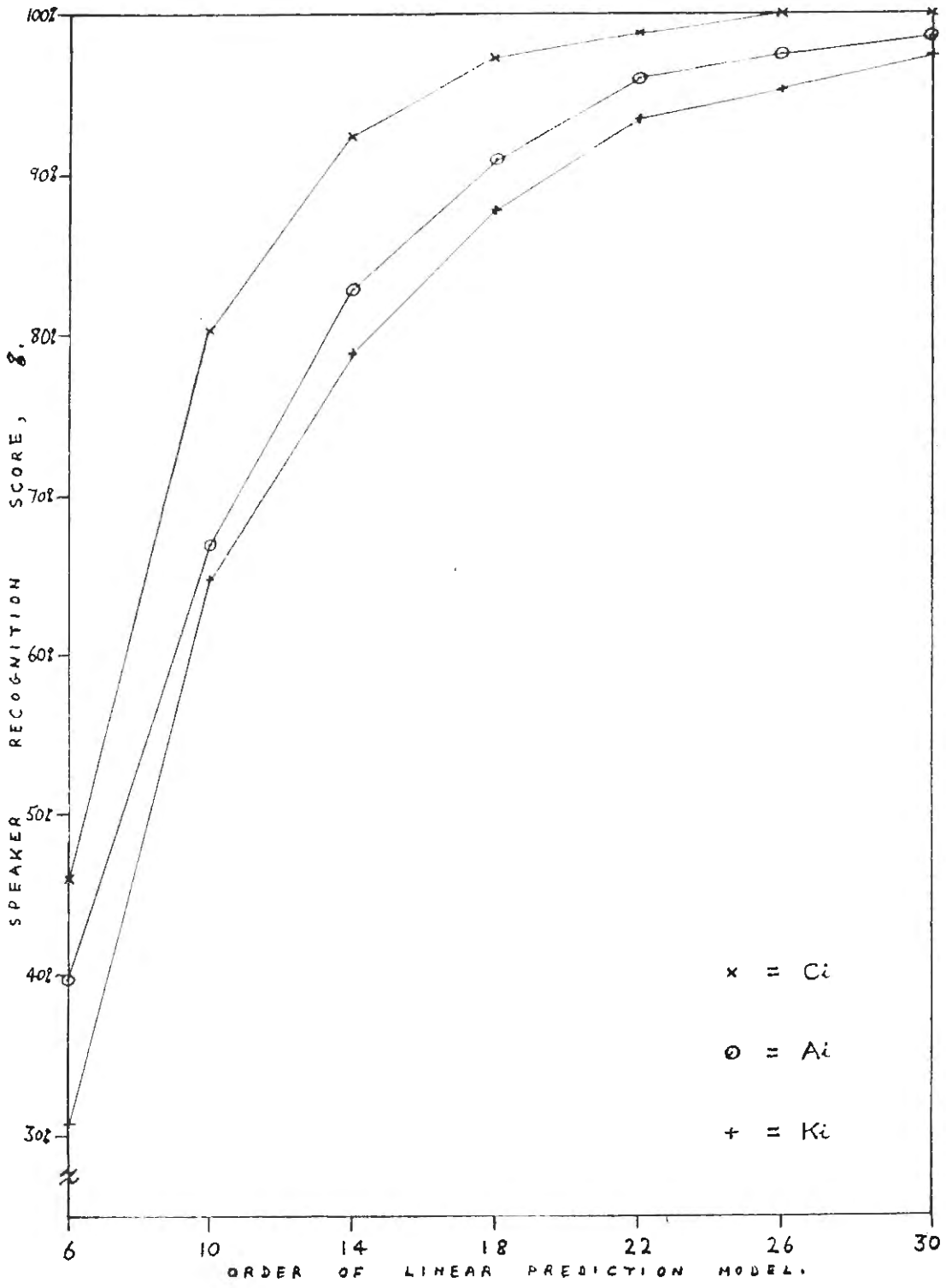
where Vp is the normalised error of the LP model and c is a function of the shape of the analysis time window (of N=200 points). On this basis with Grenier's window of 24 msec, for voiced sounds at 10kHz p=12 while at 16kHz p=18. On the other hand, Makhoul's experimental results are plotted as histograms in figure 5: the AVERAGE optimal values of p under the same conditions as for figures 3 and 4 are seen as p=5.2 for unvoiced and p=9.6 for voiced sounds.

Context-independent LP analysis of speech for recognition purposes is performed with a fixed model order, so the larger value appropriate for voiced sounds is chosen by default. As a reliable voiced/unvoiced decision for an individual window of speech data is often both ambiguous and computationally laborious, it is deemed preferable to bypass the question. Therefore the "rule-of-thumb" value is clearly a practical compromise, but we know of no quantitive study of its effect upon the performance of an actual LP-based speech understanding or speaker recognition system. Figure 6 presents the results of our study of the relative efficiency of the LP parameters Ai, Ki and Ci versus the order p of the model. A range of p=6 to 30 in steps of 4 is plotted. Observe that the cepstral coefficients Ci, converge most rapidly to produce a zero recognition error rate while



Histograms of Makhoul's (1974) experimentally determined number of poles to model voiced and unvoiced sounds.

Figure 5.



LP parameter efficiency versus the order p of the model.

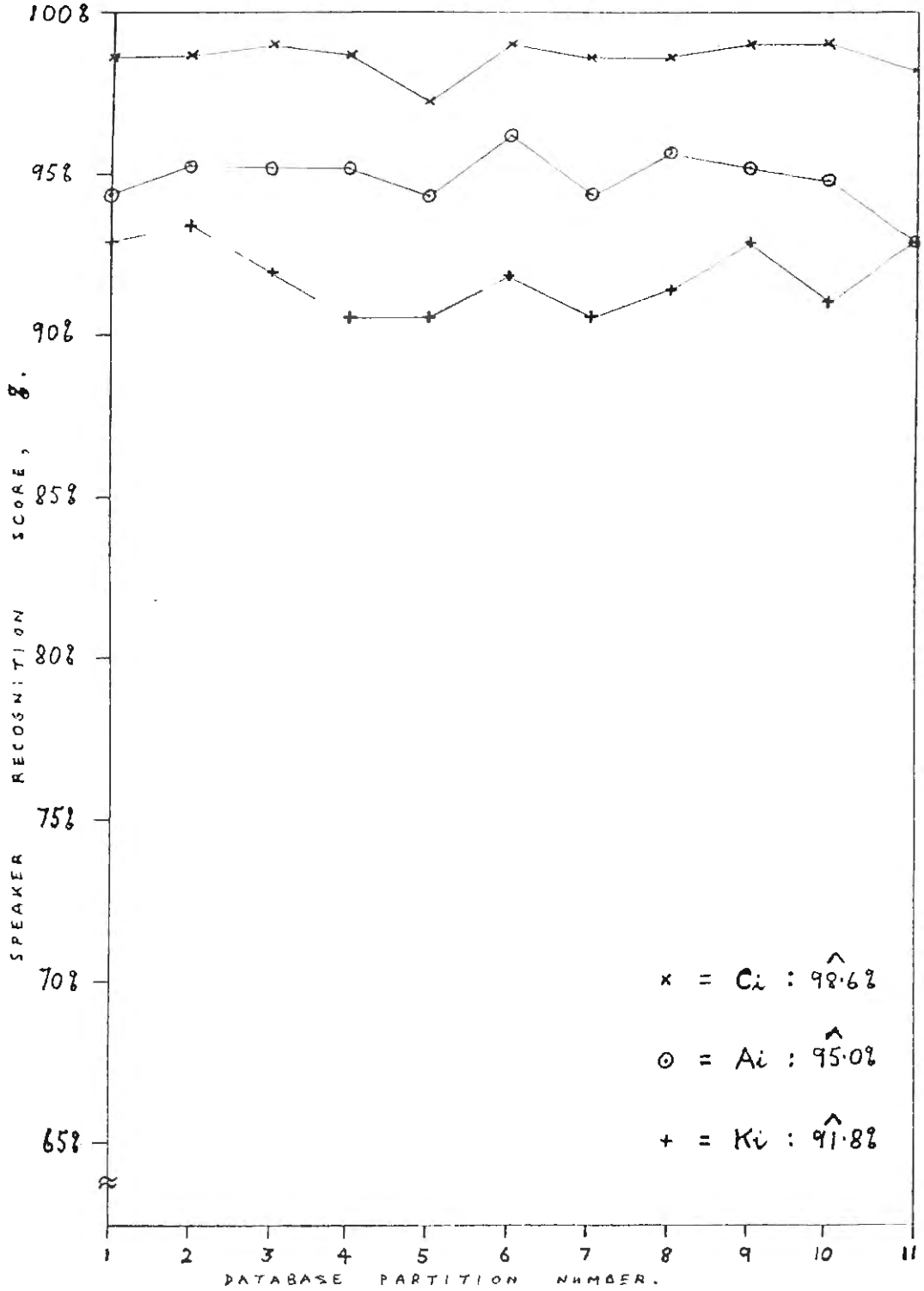
Figure 6.

the Ai and Ki plots become essentially asymptotic at high orders, with but a slow decrease in the error rate. As the computational complexity for all LP coefficients is of the order of $N.p+p^2$, the Ai and Ki are relatively less efficient at high orders compared with the Ci. While the generally chosen "optimal" order (viz p=18 for our experiments) is reasonable, it is apparent that the order may be increased to advantage.

3.4.3 Stability of different LP parameters

In 1.5.2 and 1.6 we have critically commented upon the relative fragility of many notable experiments in LP-based automatic speaker recognition. It is difficult to compose a sufficiently large and unbiased database for statistically sound results. As our database is also subject to size and homogeneity limitations, two measures are taken to confirm the veracity of our deductions.

The results of the first approach are presented in figure 7, using the same experimental techniques and database of 25 speakers as in 3.4.1 and 3.4.2, with a model of order p=18. Eleven separate recognition experiments are performed, no.1 ... no.11, each for a different set of 15 speakers selected by partitioning of the database. In fact, test no. 1 used speakers 1 through 15 of the total of 25,



LP parameter efficiency versus database changes,
model order p=18.

Figure 7.

no.2 used 2 through 16, and finally no.11 used 11 through 25. Thus for each of the LP parameters, Ai, Ki and Ci, results are obtained for each of the 11 different database partitions. By this means, it is possible to investigate the stability of the experimental results using a still small database. It is apparent that partition no.11, for example, displays a serious departure from the average trend. The average recognition scores taken over the 11 sub-partitions is entirely consistent with the results obtained with the whole database. Thus the hierarchy of the LP parameters observed in Grenier's original experiment and also in our extended experiment (as reported in 3.4.1) is positively confirmed.

The second approach is to use a database recorded under completely different circumstances. Further details of this subsidiary database are to be found in the following chapter, here we merely note that it consists of 6 phrases recorded 12 times from each of 21 speakers.

Table 2 summarizes all of these results: line 1 is for the principal database of 25 speakers, line 2 is the averaged result for the 11 sub-partitions of 15 speakers each (sliding through the 25), and line 3 shows the scores for the subsidiary database. The three sets of results consistently display the hierarchy of the LP parameters Ki, Ai and

Ci, while the absolute values of the recognition scores increase as the database size decreases - to be expected due to the simplified pattern recognition task with smaller databases.

3.4.4 Comparison of LP and FFT parameters

In sections 2.3.4 and 2.3.5 we have discussed the LP and FFT digital speech analysis techniques. Here we compare their speaker recognition performance when applied to both the principal and subsidiary databases. Table 3 shows the recognition scores for an FFT-derived spectrum versus the three LP parameters. The feature vectors used are as described in 3.3. As the LP Ai coefficients encode essentially the spectral information of the speech signal, it is noteworthy that the FFT recognition scores are comparable to those of the Ai, and likewise fall between those of the Ki and the Ci.

3.4.5 Summary of LP experiments

A complete summary of the four experiments involving LP parameters is presented in Table 4. We deduce that

1. there is established a strong hierarchy of effici-

Experiment	Phrases	Speakers	Ki	Ai	Ci	FFT
Collins main database	14	25	87.7%	92.0%	97.1%	90.0%
Subsidiary database aural expt.	6x12	21	93.7%	96.8%	98.4%	96.0%

Table 3.
Scores for LP versus FFT recognition.

Experiment	Phrases	Speakers	Ki	Ai	Ci	FFT
Grenier	11	10	97.3%	98.2%	98.2%	-
Collins main database	14	25	87.7%	92.0%	97.1%	90.0%
Collins Ave of 11 partitions	14	15	91.8%	95.0%	98.6%	-
Subsidiary database aural expt.	6x12	21	93.7%	96.8%	98.4%	96.0%

Table 4.
Complete summary of Grenier and Collins recognition scores.

ency in the recognition performance of the LP parameters. Note that Grenier's original experiment used a common key phrase in French while all of our experiments are context-independent, using a set of brief but different phrases in English. The technique is thus validated for context-independent speaker recognition, a more robust demonstration of the original claims. The increased overall error trend reflects the greater difficulty of this task, but for the LP cepstral coefficients, the Ci, extremely low error rates are achieved at the higher LP model orders. "Ma position est indeterminable", the French key phrase, is strongly voiced for most of its duration. This would be expected to favour stability of convergence of the LP analysis algorithms, and may account for Grenier's Ai score being comparable to the Ci one. Nevertheless, our more varied set of test phrases performs successfully.

2. FFT-derived spectral parameters are comparable in performance to the LP inverse filter coefficients, the Ai, but significantly weaker than the Ci under otherwise identical experimental conditions. As the derivation of the Ci is simpler than FFT parameters, LP is clearly advantageous.

3. for all experiments the LP reflection coefficients, the Ki, are substantially less efficient than both the Ai and the Ci.

CHAPTER 4

An aural speaker recognition experiment

4.1 Speaker recognition performed by people

The ability to effect rapid and positive identification of voices is a daily experience, yet the intuitive process executed by the ear-brain system is virtually unknown. A broad range of knowledge and linguistic experience is apparently applied to the task, which usually concerns either a relatively idiosyncratic utterance of a person or else a contextual or social situation familiar to the listener. Thus an isolated act of speaker recognition is less common. However, in order to conduct an objective investigation of aural recognition skills, it is the latter act which we aimed specifically to assess.

A subsidiary database has been assembled for this purpose. This database also enables an independent test of the consistency of the LP and FFT recognition techniques for comparison with the principal database (see 3.4.3).

4.2 The design of an aural recognition experiment

An audience of fifteen persons was selected to attempt aural identifications of the speakers of utterances of unrelated and infrequently used words, viz "spectrogram", "identification", and "dovetail, sidewalk: dovetail sidewalk". The taperecorded utterances were presented in random order, spoken by randomly selected colleagues of the members of the audience (60%) or else by unknown persons (40%). The utterances were chosen to minimise the occurrence of a bias favouring identification through familiar speech patterns or mannerisms.

4.3 Experimental procedure

Recordings of the utterances comprising the database were made in an office with low background noise using a pair of Beyer M69 microphones and a REVOX A77 stereophonic taperecorder. The recording conditions were essentially uniform for all of the 21 speakers included in the database. Each speaker read several sentences of introductory text as a preamble. Apart from labelling the individual donor, this text was intended to establish a relaxed and normal manner of expression before the utterances of specific interest were recorded.

After editing and ordering, the utterances were reproduced from the tape-recordings at 15 second intervals via an extremely high quality three channel electrostatic loudspeaker system which recreated an illusion of the presence of a speaker on the podium of a lecture room. The use of a centre-fill loudspeaker stabilised the stereo image of the voices for listeners seated throughout the room.

The audience, all members of the Department of Engineering Physics, ANU, was assembled in the room. They were informed that the known voices to be presented for identification were of colleagues and that others of unknown identity were interspersed. A list of names of all members of the department headed the questionaire papers as a memory aid. Otherwise the inability to immediately write the name corresponding to an identifiable voice would have biased the scores: it is not uncommon to recognise a person by voice but to momentarily forget the name.

4.4 Discussion of the results

39 utterances were presented from 21 speakers (12 known, 9 unknown) as "open" trials: in the sense that a positive identification was not necessarily possible. A summary of the scores is presented in table 5. Note that 3 voices, denoted as #1, #8 and #9, were responsible for a large

Known voice			Unknown voice	
Correct Ident.	Incorrect Ident.	No decision Made	No decision Possible	Mistaken Ident.
Score I	Error A	Error B	Score J	Error C
311	8	48	179	34

NOTES:

(1) Total of 580 respondent trials, 5 inconsistencies were eliminated.
(2) A + C = Errors of identity = 42/580 = 7.2%.
B = Errors of indecision = 48/580 = 8.3%.
A+B+C = Total of all errors = 90/580 = 15.5%.
(3) Analysis of errors in relation to 3 voices: viz #9 known, #1 and #8 unknown.

a. 4/8 = 50% cases of incorrect Ident. involved #9, i.e. #9 ascribed to 50% of the incorrectly identified known voices.
b. 16/48 = 33% of misdecisions involved #9, ie. #9 was not identified as a known voice.
c. 6/34 = 18% instances of #1 mis-identified as #9.
6/34 = 18% instances of #8 mis-identified as #9.

(4) Errors A,B,C have the same denotation as in the open trials of Tosi (1972).

SUMMARY:

16/42 = 38% of all errors of identity involved #9,
16/48 = 33% of all errors of indecision involved #9,

32/90 = 36% of all errors involved #9.

Table 5.

Recognition scores for aural speaker recognition.

number of the errors and, in particular, errors and confusions arising from the voice of #9 accounted for 35.6% of all errors from a total of 580 trials. #9 has an educated Australian accent with a strongly nasal but otherwise bland quality. It is probable that the latter characteristic gives rise to the high incidence of errors involving #9.

The experiment differs from those of Pollack (1954) and Stevens (1968) in employing a group of unskilled respondents simultaneously attempting a "one-off" identification. This is a more difficult situation in which to form a decision than repeated listenings and inter-comparisons of voices on a filecard principle. Nevertheless, the scores achieved (≈15% errors) are qualitatively similar to those of comparable experiments in this field, viz Pollack (5-80% errors, depending on phrase length) and Stevens (≈10% errors on average).

CHAPTER 5

Automatic v. subjective experimental results

5.1 Evaluation of LP and FFT versus aural recognition

The subsidiary database and the experiment in subjective speaker recognition performed with it affords the opportunity to investigate

1. consistency of LP and FFT recognition techniques as demonstrated with the principal database, and

2. the power of LP and FFT techniques to resolve confusions of identity as perceived by human listeners.

As discussed in section 3.4.4 and shown in table 3, both LP and FFT techniques perform consistently over the principal and subsidiary databases - the efficiency heirarchy for parameters Ki, FFT, Ai and Ci is indeed maintained.

Tables 6-9 respectively present a more detailed ana-

TYPE THRESHOLD FOR PRINT : 1 E- LP : Ai

TYPE NO. OF HIGH/LOW SCORES REQUIRED (MAX=10) : 10

RECOGNISED AS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	69								.		.		.								
2		70						.									.				
3			65	2	.		.								.					2	
S 4				69	2											.					
5			.		70											.					
P 6						66		6													
7							66				.	2	.				.				.
E 8	.				4			65				2									
9	2					.			58		7		2			2					
A 10										72											
11	.								.		68				.	.					
K 12	4										.	65	.	.							
13	3					.					.		66			.					
E 14			.		.							.		69							
15	.												.		70						
R 16							.	4			2		3			62					
17		.															71	.			
18				.												2	.	68			
19																			72		
20																				70	
21																					72

LOWEST SCORES

i: score	9: 58	16: 62	3: 65	8: 65	12: 65	6: 66
7: 66	13: 66	11: 68	18: 68			
% of total error	0.93%	0.66%	0.46%	0.46%	0.46%	0.40%
0.40%	0.40%	0.26%	0.26%			

HIGHEST SCORES

i: score	10: 72	19: 72	21: 72	17: 71	2: 70	5: 70
15: 70	20: 70	1: 69	4: 69			
% of correct ID	100%	100%	100%	99%	97%	97%
97%	97%	96%	96%			

MAX CONFUSIONS

i/j: score	9/11: 7	6/ 8: 6	8/ 6: 4	12/ 1: 4	16/ 8: 4	13/ 1: 3
16/13: 3	3/ 4: 2	3/20: 2	4/ 5: 2			

Table 6.

Recognition scores for LP Ai on the subsidiary database, presented as a confusion matrix.

TYPE THRESHOLD FOR PRINT : 1 ≡· LP : Ki

TYPE NO. OF HIGH/LOW SCORES REQUIRED (MAX=10) : 10

		RECOGNISED AS																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	1	57							.			.	5	5		.	.					.
	2		70										.	.								
	3			61		3	.	.			.			.				.			.	2
S	4			.	70																	.
	5			2	.	66			.			.						.				
P	6	3					56		4				.	.			4					3
	7	3				.		64						2			2					
E	8	6				2	.		61				.									.
	9	.					.			61		2		.			5	.				
A	10			.			2	.			67						.					
	11			.					.		66				.		3					
K	12	3			.				3			.	60	2	.		.					
	13								.	.		2	.	61	2		3					.
E	14			7				.							61		3					
	15	2		.										2		67						
R	16	.					3					3		2	.		61					.
	17											2		.				69				
	18				3							3							65			.
	19					.													.	68	2	
	20													3	.			3		65		
	21	2			.		2	.	.					2							63	

LOWEST SCORES

i: score 6: 56 1: 57 12: 60 3: 61 8: 61 9: 61
13: 61 14: 61 16: 61 21: 63

% of total error 1.06% 0.99% 0.79% 0.73% 0.73% 0.73%
0.73% 0.73% 0.73% 0.60%

HIGHEST SCORES

i: score 2: 70 4: 70 17: 69 19: 68 10: 67 15: 67
5: 66 11: 66 18: 65 20: 65

% of correct ID 97% 97% 96% 94% 93% 93%
92% 92% 90% 90%

MAX CONFUSIONS

i/j: score 14/ 3: 7 8/ 1: 6 1/12: 5 1/13: 5 9/16: 5 6/ 8: 4
6/16: 4 3/ 5: 3 6/ 1: 3 6/21: 3

Table 7.

Recognition scores for LP Ki on the subsidiary database, presented as a confusion matrix.

TYPE THRESHOLD FOR PRINT : 1E- LP t Ci

TYPE NO. OF HIGH/LOW SCORES REQUIRED (MAX=10) : 10

R E C O G N I S E D A S

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	1	69										2										
	2		69																			
	3			67												2						
S	4				72																	
	5					70																
P	6						71															
	7							72														
E	8								67			3										
	9									65		2					3					
A	10										71											
	11											71										
K	12	3						2					65									
	13													71								
E	14			3											67	2						
	15															70						
R	16																71					
	17																	71				
	18																		70			
	19																			68		
	20																				71	
	21																					72

LOWEST SCORES

1: score	9: 65	12: 65	3: 67	8: 67	14: 67	19: 68
1: 69	2: 69	5: 70	15: 70			
% of total error	0.46%	0.46%	0.33%	0.33%	0.33%	0.26%
0.20%	0.20%	0.13%	0.13%			

HIGHEST SCORES

1: score	4: 72	7: 72	21: 72	6: 71	10: 71	11: 71
13: 71	16: 71	17: 71	20: 71			
% of correct ID	100%	100%	100%	99%	99%	99%
99%	99%	99%	99%			

MAX CONFUSIONS

i/j: score	8/11: 3	9/16: 3	12/ 1: 3	14/ 3: 3	1/11: 2	3/15: 2
9/11: 2	12/ 7: 2	14/15: 2	1/ 6: 1			

Table 8.

Recognition scores for LP Ci on the subsidiary database, presented as a confusion matrix.

TYPE THRESHOLD FOR PRINT : 1■. FFT

TYPF NO. OF HIGH/LOW SCORES REQUIRED (MAX=10) : 10

		R E C O G N I S E D A S																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	1	63						4				5										
	2		70	.	.																	
	3			68		.								.	.		.					
S	4				70			.						.								
	5			3		68		.														
P	6						72					.										
	7			2		.		67				2										
E	8								71	.												
	9					2		8		60			.	.								
A	10			.	.						69			.								
	11	.										71										
K	12	5								.		.	65									
	13	.		.										70								
E	14					.									69		.		.			
	15	2		.	2	8				2	.				.	54			.			
R	16						.		2								68			.		
	17			.	2		.											68					
	18			4	.	.												2	64			
	19	2			.						2				.			.	2	62	.	
	20				.						.	.			.						68	
	21	.			3																	68

LOWEST SCORES

i: score	15: 54	9: 60	19: 62	1: 63	18: 64	12: 65
7: 67	3: 68	5: 68	16: 68			
% of total error	1.19%	0.79%	0.66%	0.60%	0.53%	0.46%
0.33%	0.26%	0.26%	0.26%			

HIGHEST SCORES

i: score	6: 72	8: 71	11: 71	2: 70	4: 70	13: 70
10: 69	14: 69	3: 68	5: 68			
% of correct ID	100%	99%	99%	97%	97%	97%
96%	96%	94%	94%			

MAX CONFUSIONS

i/j: score	9/ 7: 8	15/ 5: 8	1/11: 5	12/ 1: 5	1/ 7: 4	18/ 3: 4
5/ 3: 3	21/ 4: 3	7/ 3: 2	7/11: 2			

Table 9.

Recognition scores for FFT on the subsidiary database, presented as a confusion matrix.

lysis of the speaker recognition scores for the subsidiary database with these parameters. Shown as confusion matrices, the off-diagonal terms signify recognition errors. Error statistics are tabulated for the speakers with the lowest and highest recognition scores and also for the most common confusions made between speaker identities. Similiary, table 10 recasts the results of the aural recognition experiment of chapter 4 as a confusion matrix. The most frequent real-world (aural) confusions are of speaker #1 as #9: 6 times, #6 as #13: 2, #8 as #9: 6, #9 as #11: 3, and #21 as #4: 4 times. There is a conspicuous occurrence of speaker #9 in the error scores. Examination of the corresponding confusion matrices for the computed recognitions reveals an overall poor performance for speaker #9. However the confusions of identity show no consistent overall trend for any LP parameter set or for the FFT parameters, considered individually or in combination, comparable with the aural results.

It appears that the computer analyses, through a powerful application of statistical pattern recognition techniques, are capable of highly reliable speaker categorization. However, the resultant "characterization" of speakers is neither consistent across individual parameters nor similar to human judgements. From this viewpoint, the success of the computer techniques is seen to be essentially of an empiri-

		RECOGNISED AS																						
	NULL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	OTHER	TOTAL
	1	24								6														30
	2		30																					30
	3	3		27																				30
S	4	22			8																			30
	5	15																						15
P	6	26												2		•							•	30
	7	•						29																30
F	8	23								6		•												30
	9	18								8		3											•	30
A	10	8		•							19												2	30
	11											45												45
F	12												30											30
	13	•												29										30
E	14	•													14									15
	15									•						29								30
R	16	9															5						•	15
	17	8																21					•	30
	18	22																		•	•		6	30
	19	3																		12				15
	20	24																			4		2	30
	21	23			4									•									2	30

Table 10.

Recognition scores for the aural experiment on the subsidiary database, presented as a confusion matrix.

cal nature rather than revealing particular insight into the solution of the problem. In other words, they do not indicate those specific voice attributes which strongly characterize individuals or equally those which are prone to cause confusion for human listeners. The LP cepstral parameters, the Ci, are the most efficient in all experiments, but no clear association between any particular parameter and the recognition process in humans is supported by comparison with the perceptually-based experiment.

5.2 Conclusion

Having demonstrated the relative fragility of context independent automatic speaker recognition, when based primarily upon statistical pattern recognition techniques, it is apparent that greater insight at the feature extraction stage is necessary. Many recent experiments have attempted to exploit raw computational power and an empirical approach to feature extraction to overcome the paucity of knowledge about the real nature of the human skill of speaker recognition. Thus we see an armoury of techniques successful within the confines of a laboratory but failing to perform well in real-world situations. In 1978, Jesorsky et al. succinctly commented upon this dilemma.

> "In the past, research in speaker-recognition was primarily aimed at pointing out the efficiency of a cer-

> tain algorithm for feature-extraction or classification.
> In this sense, the speech input seemed to be of less importance and served only as a vehicle for the investigations. In our opinion, the speech input ... is an essential part of the system design.
> Generally, the computational effort of the recognition system can be reduced significantly without decreasing the performance of the system."

Our experiments have been performed upon a subset of a large Australian speech database (Collins, 1979). Using the whole database, there is the opportunity to extend the scale and scope of the experiments, including context-dependent speaker recognition since the spoken text is uniform over all speakers. Experiments studying the effect of telephone transmission distortions upon the stability of LP and FFT parameters are a logical further step. Thus the experiments to date are exploratory, forming a systematic foundation for controlled investigation of a variety of techniques. The general aim is to span the gulf between successful but highly constrained laboratory results and potential real-world applications of automatic recognition techniques in varying and antagonistic environments.

5.3 References

AKAIKE, H., "A new look at statistical model identification", IEEE Trans. on Automatic Control, vol. AC-19, pp. 716-723, Dec. 1974.

ATAL, B.S. and Hanauer, S.L., "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Amer., vol. 50, no. 2, pp.637-655, 1971.

ATAL, B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Amer., vol. 55, no. 6, pp.1304-1312, 1974.

ATAL, B.S., "Automatic recognition of speakers from their voices", Proc. IEEE, vol. 64., no. 4, pp.460-475, April 1976.

BOGNER, R.E., "On talker verification via orthogonal parameters", IEEE Trans. ASSP-29, no. 1, pp.1-12, February 1981.

BOLT, R.H., Cooper, F.S., David, E.E., Denes, P.B., Pickett, J.M. and Stevens, K.N., "Speaker identification by speech spectrograms: some further observations", J. Acoust. Soc. Amer., vol. 54, pp.531-534, 1973.

BRICKER, P.D. et al., "Statistical techniques for talker identification", Bell System Tech. J., vol. 50, pp.1427-1454, April 1971.

BUNGE, E., et al., "Statistical techniques for automatic speaker recognition", Conf. Rec. ICASSP, pp.772-775, Hartford 1977.

BUNGE, E., "Speaker recognition by computer", Philips Tech. Rev., vol. 37, no. 8, pp.207-219, 1977.

CHEN, C.H., "A review of statistical pattern recognition", NATO Advanced Study Institute Conference on Pattern Recognition and Signal Processing, pp.117-132, 1978.

COLLINS, A.M., "A database for digital speech processing research", Proc. IREECON, pp.565-568, September 1979.

COOLEY, J.W. and TUKEY, J.W., "An algorithm for the machine calculation of complex Fourier series", Math. of Comp., vol. 19, pp.297-301, 1965.

CRICHTON, R.G. and FALLSIDE, M.A., "Linear prediction model of speech production with applications to deaf speech training", Proc. IEE, vol. 121, no. 8, 1974.

FASOLO, L. and MIAN, G.A., "A comparison between two approaches to automatic speaker recognition", Conf. Rec.

ICASSP, 1979.

FOLEY, D.H., "Considerations of sample and feature size", IEEE Trans. Inf. Theory, vol. IT-18, no. 5, pp.618-626, September 1972.

FUKUNAGA, K., "Introduction to statistical pattern recognition", Academic Press, New York, 1972.

GRENIER, Y., "Identification du locuteur et adaptation au locuteur d'un systeme de reconnaissance phonemique", Rapport ENST-E-77005, Ecole Nationale Superieure des Telecommunications, Paris, 1977.

GRENIER, Y., "Speaker identification from linear prediction", Proc. 4th IJCPR, pp.1019-1021, November 1978.

HOLLIEN, H., "The peculiar case of voiceprints", J. Acoust. Soc. Amer., vol. 56, pp.210-213, 1973.

JESORSKY, P., HOFKER, U. and TALMI, M., "Extraction of speaker-dependent features from spoken code sentences", Conf. Rec. ICASSP, pp.279-282, 1978.

MARKEL, J. and GRAY, A.H., "Linear prediction of speech", Springer, Berlin-Heidelberg, 1976.

MAKHOUL, J. et al., "Natural communications with computers", Report 2976, Bolt Beranek and Newman, Cambridge, 1974.

MAKHOUL, J., "Linear prediction: a tutorial review", Proc. IEEE, vol. 63, pp.561 -580, April 1975.

MIAN, G.A., "Some factors influencing the performances of a speaker recognition system based on LPC", Conf. Rec. ICASSP, 1978.

NATIONAL ACADEMY OF SCIENCES (NAS),
NATIONAL RESEARCH COUNCIL, "On the theory and practice of voice identification", Committee on evaluation of sound spectrograms, Assembly of Behavioral and Social Sciences, Washington, D.C., 1979.

NOLL, A.M., "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection",J. Acoust. Soc. Amer., vol. 36, pp.296-302, 1964.

OPPENHEIM, A.V., et al., "Non-linear filtering of multiplied and convolved signals", Proc. IEEE, vol. 56, pp.1264-1291, Aug. 1968.

OPPENHEIM, A.V., "Speech spectrograms using the fast Fourier transform", IEEE Spectrum, pp.57-62, Aug. 1970.

POLLACK, J. et al., "On the identification of speakers by voice", J. Acoust. Soc. Amer., vol. 26, pp.403-406, 1954.

POLS, L.C.W., TROMP, H.R.C., and PLOMP, R., "Frequency analysis of Dutch vowels from 50 male speakers", J. Acoust. Soc. amer., vol. 53, no. 4, pp.1093-1101, April 1973.

ROSENBERG, A.E., "Automatic speaker verification: a review", Proc. IEEE, vol. 64, pp.475-487, April 1976.

SAMBUR, M.R., "Acoustic features for speaker identification", IEEE Trans. ASSP-23, no. 2, pp.176-182, April 1975.

SAMBUR, M.R., "Speaker recognition using orthogonal linear prediction", IEEE Trans. ASSP-24, no. 4, pp.283-289, August 1976.

SARMA, V.V.S. and VENUGOPAL, D., "Performance evaluation of automatic speaker verification systems", IEEE Trans. ASSP-25, no. 3, pp.264-266, June 1977.

SCHROEDER, M.R., "Direct (nonrecursive) relations between cepstrum and predictor coefficients", IEEE Trans. ASSP-29, no. 2, pp. 297-301, April 1981.

STEVENS, K.N. et al., "Speaker authentication and identification", J. Acoust. Soc. Amer., vol. 44, pp.1596-1607, 1968.

STEVENS, K.N. et al., "Speaker identification by speech spectrograms: some further observations", J. Acoust. Soc. Amer., vol. 54, pp.531-534, 1973.

TOHKURA, Y., and ITAKURA, F., "Spectral sensitivity analysis of PARCOR parameter for speech data compression", Conf. Rec. IEEE ICASSP, pp.1059-1063, 1978.

TOSI, O.I. et al., "Experiment on voice identification", J. Acoust. Soc. Amer., vol. 54, pp.2030-2043, 1972.

TOSI, O.I., et al., "Reply to 'Speaker identification by speech spectrograms: some further observations'", J. Acoust. Soc. Amer., vol. 54, pp.535-537, 1973.

WAKITA, H., "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. on Audio Electroacoust., vol. AU21, pp.417-427, Oct. 1973.

WATANABE, S., "Karhunen-Loeve expansion and factor ana-

lysis: theoretical remarks and applications", Trans. 4th Prague Conf. Inform. Theory, pp.635-660, Prague 1965.

WOLF, J.J., "Efficient acoustic parameters for speaker recognition", J. Acoust. Soc. Amer., vol. 54, no. 6, pp.2044-2056, December 1972.