Centre for Speech Technology Research

University of Victoria

****

FINAL REPORT

INVESTIGATION OF A SPEAKER VERIFICATION SYSTEM USING PARAMETERS DERIVED FROM THE CRC VOCODER

****

Prepared by:

B. Craig Dickson and Roy C. Snell
Centre for Speech Technology Research

for the

Communications Research Centre
Department of Communications

Contract No. 21ST.36001-3-3038

December 31, 1983

Research Director—Department of Linguistics, (604) 721-7426

Administration Director—JEM Research, (604) 477-7246

Marketing—Software Research Corp., (604) 477-7246

Centre for Speech Technology Research
University of Victoria
P.O. Box 1700
Victoria, B.C. V8W 2Y2
Telex 049-7222

Centre for Speech Technology Research
UNIVERSITY OF VICTORIA P.O. Box 1700, Victoria, B.C. V8W 2Y2

REPORT DOCUMENTATION PAGE

PROJECT TITLE: Investigation of a Speaker Verification System Using Parameters Derived from the CRC Vocoder

TYPE OF REPORT AND PERIOD COVERED: Final Report
May-December, 1983

DATE: Dec 31, 1983

CONTRACTING AGENCY: Communications Research Centre

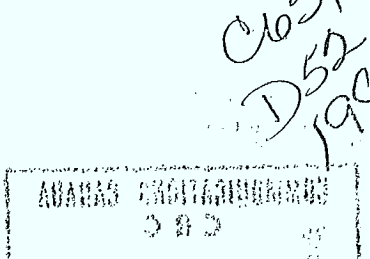
ADDRESS: Department of Communications
P.O. Box 11490, Station 'H',
Shirley Bay, Ontario
K2H 8S2

CONTRACT NUMBER: (DSS) 21ST.36001-3-3038

PROJECT CONTROL NUMBER: 83-01

PROJECT MANAGER: B. Craig Dickson

AUTHOR(S): B. Craig Dickson



SUMMARY

This paper is the final report on the Speaker Verification Research Project to investigate the feasibility of performing automatic speaker verification using parameters generated or derived from a vocoder system that is based on the ten-pole linear prediction model of speech. The project was undertaken through Supply and Services Canada Contract No. 21ST.36001-3-3038 for the Space Systems Directorate, Communications Research Centre (CRC), Department of Communications.

The project was contracted through Mr. Brian Bryden of the CRC, who was recently involved in the development of a real-time digital vocoder that produces speech output of sufficiently high quality to make it attractive as a general purpose narrow band communications device. The speaker verification research was intended to be an initial study to determine the feasibility of incorporating an automatic verification routine into the vocoder system, making use of values generated by the encoding parameters as they are transmitted by the system. A speaker recognition procedure developed in an earlier project at the Centre for Speech Technology Research was modified to operate in conjunction with the vocoder.

The results of the preliminary study indicate that, when using a preprocessing phase that isolates components of the signal according to general phonetic classes, sufficient information about the speaker can be determined from the encoding parameters to make the development of a verification system possible. Various strategies for obtaining speech samples and for the adequate manipulation of statistical comparisons are discussed.

TABLE OF CONTENTS

1. INTRODUCTION

The success of automatic speaker recognition systems designed to be used with communications channels has been hampered by two general sources of error. Primarily, signal interferences in the message path, which cause linear and non-linear distortion, have been shown to mask the speaker-discriminating characteristics of the signal so that successful procedures developed in the laboratory can not be replicated in the field. This has been compounded by the second source of error, intra-speaker differences in the speech samples used for recognition. These differences tend to interfere with the task of extracting acoustic parameters from the signal that describe the distinctive characteristics of individual speakers.

The Communications Research Centre in Ottawa has recently completed the development of a new vocoder system for digital transmission of speech in real time, using new signal processing technology to provide a low cost digital communications system. For automatic speaker recognition, the use of vocoders alleviates the majority of problems associated with signal interference generated by the message path. Because vocoders involve the digital transmission of an encoded signal (followed by decoding upon receipt) any distortion that occurs is a result of the encoding function, and any transmission errors that do occur can be detected and/or corrected by standard digital techniques (parity, error detecting and correcting codes).

The object of this research is to find solutions to the problem of intra-speaker differences between speech samples, in order to develop a recognition system that will operate in connection with the newly developed vocoder system. Errors produced by this source of interference stem from the kinds of information in the signal that are focused upon during the analysis phase, during which parameters for speaker recognition are extracted. A high amount of variability in the speech signal produced by an individual can be attributed to natural phonetic processes that are mistakenly analyzed as being speaker-specific. Improvement of the parameter extraction phase would therefore contribute to improvement of the overall task, both in the laboratory and in the field.

To achieve the goal of improved parameter extraction, the Centre for Speech Technology Research in Victoria, B.C. has been involved in research to develop procedures that

automatically isolate phonetic events containing a high degree of speaker-specific information for recognition purposes, and that manipulate these events to reduce the influence of intra-speaker differences between speech samples used as test and reference data. Recently, work has been continued under contract with the Department of Communications, in an effort to develop a functional automatic speaker verification* procedure to be used in connection with the new vocoder system. This paper addresses the underlying theoretical issues that are involved with the research, describes the results obtained from the preliminary investigation, and comments on the expected success and applications of the proposed system.

---

* Speaker recognition is generally divided into two kinds of tasks: verification and identification. Verification, or authentication, involves the task of determining whether or not a speaker is if fact who he claims to be. A sample utterance produced by the speaker is compared with a file of accumulated reference data belonging to the person of claimed identity, and a binary decision of acceptance or rejection is made. This contrasts with speaker identification, in which an unknown sample of speech is compared with reference data of a (possibly large) group of known speakers, the possibility existing that the unknown speaker is not represented in the reference data. At some level of analysis, both tasks require that a decision be made about whether the sample data and the reference data are closely enough matched to constitute an acceptance. Errors of false acceptance of a speaker, versus errors of false rejection are determined by this measure.

2. THEORETICAL BACKGROUND

2.1 Sources of Intra-speaker Variability

Acoustic signals generated by the speech event are composed of a complex orchestration of information. Information about the linguistic message, the dialect, age, sex and respiratory health of the speaker are all transmitted simultaneously, and integrated with these are indicators of the speaker's intent and emotional state. Also, and of particular importance here, the speech event produces acoustic indicators of individual speech characteristics which, for our purposes, are defined as acoustic manifestations of anatomical aspects such as the size of the resonating cavities, or habitual aspects such as tendencies to produce certain phonation types and the timing of articulatory movements.

In theory, a number of these anatomical or habitual characteristics are expected to influence the signal in such a way as to provide information that is unique to any one speaker. However, the highly dynamic properties of the signal that are caused by differences in the linguistic message, together with the transient aspects of mood, intent etc., produce variations in the speech signal that are difficult to control while concentrating on those aspects that we expect to be unique.

In processing the signal, the human listener appears to be capable of focusing on one of the above kinds of information while ignoring, or at least accommodating, the other kinds. The isolation of information about the speaker would then be performed on the basis of the listener's knowledge about the language, regional and social indicators, and how for example, the speaker's mood might influence the quality of the signal. In using this prior knowledge, the listener also appears to be capable of shifting from one set of auditory parameters to another in order to effect recognition. However, while a large source of prior knowledge is not available to the the automatic system, it is also not limited by the effects of short-term memory. It is the limitations of short-term memory that have been cited [1] as the reason why automatic speaker recognition experiments conducted under laboratory conditions have produced more successful results than parallel auditory tasks.

Because of the dynamic properties of speech, the capability of an automatic recognition system to selectively shift from one set of parameters to another according to other information about the signal is therefore expected to improve the system's performance. Segmentation of the signal on the basis of the occurrence of phonetic events may be of interest for this purpose, but their selection is not a simple task.

For example, it has been found in our earlier research [2,3] that the effects of coarticulation (the articulatory influence of one phonetic segment upon another) are a major influence on the acoustic quality of speech segments that may be used for recognition, even when these segments are well separated by other segments. To compound this, a slight change in the dynamic content of the utterance, such as a difference in speech rate or sentence stress, will alter the coarticulation characteristics in two otherwise identical utterances. In speaker recognition experiments, coarticulation effects have been observed in context-dependent tasks (the same series of words repeated to presumably produce a similar acoustic pattern) as well as for context-independent tasks, although the influences of the phonetic environment are reduced in the former.

2.2 Selection of Parameters

The principal task of speaker recognition research, then, is to develop a set of parameters that account for such intra-speaker variability. Effects such as those described above must be controlled to avoid the acceptance of information that is not consistently unique to the speaker, but care must be taken that the elimination of certain elements of speech does not lead to the elimination of useful information about the speaker. The division of the signal into categories according to general acoustic characteristics is seen as a means of reducing variability that is related to the linguistic component, so that these categories can be examined for habitual and anatomical influences that are thought to be distinctively representative of the speaker.

2.3 Long-Term Parameters

In the search for useful parameters, several attempts have been made to make use of parameters in the signal that are not greatly influenced by modifications in the linguistic message. It has been theorized that the averaging of information over the duration of an utterance would smooth the effects of linguistic variation, so that individual characteristics are more accessible. For example, the mean and range of the fundamental frequency (F0) have been used effectively in the laboratory, [4,5,6] as have accumulated values derived from LPC such as reflection coefficients [4,6,7] and cepstrum coefficients [4,8,9], and other processes such as the long-term power spectrum [10,11] and even zero-crossings [12]. However, parameters gathered in this manner appear to be susceptible to shifts in speech patterns that occur over a period of time, perhaps brought about by changes in vocal tract settings or phonation type [3]. These kinds of parameters may therefore be more representative of the transient aspects spoken of earlier, such as changes in emotional state, than the kind of information that is anatomically or habitually based and unique to a speaker.

2.4 Phonetically-Based Approaches

An earlier approach to speaker recognition, and one that has recently been revisited (eg. [13]), involves the location of phonetic segments within an utterance, and an examination of the segments for speaker-specific information. The theory of this phonetically-based approach is that the acoustic manifestation of the phonetic unit will reflect the manner in which the speaker habitually produces the segment, at least within the context of a previously specified utterance. For a speaker verification task in which the speaker's evironment is strictly controlled, such an approach has been highly successful [14] as long as several tests are run in parallel, and the reference samples are updated regularly to account for changes in voice quality over time. However, if there is no control of the context, as is usually the case for speaker identification for forensic purposes, and which is often desired in less strictly controlled situations involving speaker verification, the acoustic qualities of a single phonetic unit are far too variable to be of use.

Research in the phonetically-based approach has shown that certain phonetic segments are more constistent in providing speaker-specific information than others [15]. Phonetic units such as [n], [m], [u], [ɵ] and [i] were found to outrank other units in Fisher discriminant analysis tests, indicating that the production of these events reflects more information about the anatomy of the subject's speech mechanism than do other events. However, for these units to be useful, a means of controlling the acoustical effects of coarticulation must be developed.

2.5 Phonetic Classes

In an attempt to control coarticulation while concentrating upon phonetic units, the current research has incorporated a preliminary stage of analysis in which the speech signal is categorized into a series of phonetic classes, each class being characterized by its broad acoustical qualities. A phonetic class is defined as a class of acoustical events that are the result of a particular set of articulatory and phonatory gestures, and are not necessarily related to the phonemic intent of the gestures. Acoustic events thus isolated are accumulated over the duration of a speech sample and, with other similarly accumulated events belonging to different phonetic classes, are used to describe an acoustic profile of the speaker.

As an example of a phonetic class, a nasal murmur is an acoustical event that occurs when the majority of the pulmonary energy is deflected through the nasal cavity during voicing. Generally this is the principal acoustic indicator of a nasal phoneme, but the event is often masked by the predominance of oral energy. If this occurs, other acoustic indicators of the nasal phoneme will remain, such as the amount of nasalization in the vowel and the transitional effects of tongue body movement to a target in the oral cavity. On the other hand, a nasal murmur can occur during any articulation that involves increased oral impedence when the velar port is relaxed or open. Thus a voiced stop consonant or a resonant such as [w] may create sufficient impedence in the oral cavity to give rise to a nasal murmur.

The separation of the speech signal into acoustic classes is therefore not directly related to the phonetically-based approach described above. However, the new approach avoids the difficulties that arise from attempts to derive parameters from the signal without

accounting for the different kinds of acoustic events within the utterance. Because the categorization of the signal into classes according to broad acoustic patterns avoids the comparison of highly dissimilar acoustic events, a large proportion of the variability within the signal is controlled.

The accumulation of a series of events belonging to a common phonetic class appears to reduce the effects of coarticulation when the appropriate statistical procedures are applied to determine their predominant parametric values. The expected result is that information in the acoustic signal that reflects the less variable habitual or anatomical characteristics of the speaker are enhanced while the highly variable influences of coarticulation are reduced.

2.6 Phonetic Classes Selected

Three classes of phonetic events have been focused upon in the theoretical stages of the research: nasal murmurs, high front vowels, and high back vowels. As was mentioned earlier, research in the phonetically-based approaches has indicated that phonetic units belonging to these classes contain a higher proportion of speaker-specific information than other phonetic units. Also, these events may be more descriptive of the physiological characteristics of the organs of speech than other articulatory gestures. The nasal murmurs reflect information about the resonant characteristics of the nasal cavity, which is a relatively fixed body, and the two categories of high front and high back vowels represent positions of extreme oral articulation. Thus the accumulation of acoustic descriptions of these events as they occur in an utterance will not only reduce the effects of coarticulation, but will also provide an acoustic profile that is more specific than one that results from the accumulation of acoustic information from an unsegmented signal.

3. TECHNICAL APPROACH

3.1 Results of Previous Work

In our earlier research in speaker recognition, the class of nasal murmurs was selected for examination to determine the feasibility of the procedure. An automatic procedure for locating the nasal murmurs was developed on a small computer, using parameters derived from energy calculations and profiles taken from the power spectra. A data base of 880 short sentences containing nasal murmurs in a range of phonetic environments produced by 10 subjects was used to develop the phonetic class extraction routine, and the context-independent data of 50 subjects were used in a closed-set identification experiment. Twenty higher order cepstrum coefficients were derived from the extracted data, these parameters yielding 100% accuracy when a pooled covariance matching procedure was employed. A minimum of fifty 20 ms. frames of nasal data for each subject's covariance matrix was used.

These high success rates appear to have been influenced by the fact that the test and reference samples were recorded on the same day. We were encouraged, however, with the results of closed-set experiments involving test samples taken 17 months later from groups of ten and five subjects. The success rates of 70% and 85% respectively were achieved, which is better than other reported results in which extensive periods of time have elapsed between the collection of test and reference samples.

3.2 Current Approach

The purpose of the current research project has been to determine whether speech that has been encoded using the vocoder system developed at the Communications Research Centre contains sufficient speaker-specific information to allow automatic speaker verification. In order to do this, the recognition model involving the classification of phonetic events was modified to operate on speech that has been processed through the encoding stage of the vocoder. Because of the preliminary aspects of the work, the phonetic class of nasal murmurs was again examined, facilitating the rapid comparison of results with the earlier work. Also, the concern has been expressed that an all-pole model of LPC would mask the nasal segments because of the presence of

zeros (anti-resonances) during these events. It is therefore expected that, if this class of events can be automatically extracted and used for verification, there will be little difficulty in achieving success with other classes.

3.3 Isolating the Class of Nasal Murmurs

The 880-sentence data base of the ten subjects used previously was re-sampled and analysis files were created, using the parameters taken from the modified vocoder routines. The segment extraction routine that was subsequently developed proved to be more accurate in isolating nasal events than the one used in the previous research. This success has also given a strong indication that the other phonetic classes of interest, these being high front vowels and high back vowels, can be identified with relative ease using strategies developed for extracting the nasal murmurs.

Parameters available through the encoding stage of the vocoder include 10 reflection coefficients (K's), the residual energy and the fundamental frequency (including the voiced/voiceless decision). Values for these parameters are computed at intervals of 20.4 ms from a signal that has been digitized at approximately 8K samples per second.

It was found that the reflection coefficients are too speaker-dependent to be useful for extracting the segments. An initial attempt that relied on the autocorrelation (or filter) coefficients (A's) was more promising, but further examination revealed that these coefficients were influenced by the shape of the waveform, such that phonation types that involved harsh or creaky voice failed to produce values that were within tolerances assigned to nasal murmurs produced with modal voice. As harsh or creaky voice may be habitual, or may be the result of a respiratory disorder or smoking, other parameters that were independent of laryngeal characteristics had to be determined.

The frequency of the resonant peaks (formants), derived from the filter coefficients by means of a peak-picking routine operating on the filter response curve, proved to be sufficiently speaker-independent to be used for identifying many aspects of the nasals, as long as variability was allowed for a formant peak in the 800 to 1900 Hz range. The bandwidths and the relative magnitudes of the formants were not as useful as the formant frequencies for identifying the nasals, as they were easily influenced by the vocalic environment.

The nasal murmur extraction routine takes the following values from parameters that are either available in or derived from each frame of encoded data produced by the vocoder:

a) voicing occurs (pitch value reported);

b) zero-crossing (0x) count set at maximum of 13 per frame;*

c) normalized residual energy value (RMSN) set at a maximum of 0.8;

d) filter coefficient A2 set at a value of less than -0.2;

e) a formant peak occurs below 375 Hz and a peak does not occur between 375 Hz and 800 Hz;

f) one or two peaks occur in the 1900-2800 Hz region and no more than one peak occurs between 800 and 1900 Hz; and

g) all peaks in the 1900-2800 Hz region are separated by more than 400 Hz and any peak in that range is separated from peaks outside that range by more than 700 Hz.

Each parameter was selected on the basis of its computational simplicity and its power to eliminate non-nasal events. As can be seen in the decision logic presented in Table 1, the order of access to each parameter is determined by the level of computational complexity required to eliminate the frame. The individual parameters each contribute in the following manners:

---

* The number of zero crossings per frame is not one of the parameters that is transmitted by the vocoder, although it is calculated to allow voice/voiceless decisions in the pitch extraction routine. Because the narrow bandwidth used for transmission of the encoded data is not expected to support the extra four bits that would be required for 0x, it is expected that this parameter can be calculated from the decoded signal. The comparison of a number of samples of original and synthesized data revealed that, in the frames in which this parameter is required to distinguish the nasals, the 0x values in the synthesized frames are not significantly different from those in the original sample.

- The use of a maximum 0x value provides an inexpensive means of eliminating signals with dominant high frequency energy, avoiding the need to compute the spectrum for all voiced frames.

- Because RMSN (with a maximum possible value of 1.0) indicates the ratio of the excitation energy to the total signal energy, it is being used experimentally as a means of avoiding the false capture of signals in which voicing is very weak or absent.

- Although the filter coefficients were found to be influenced by laryngeal characteristics, it was also found that the single filter coefficient A2 reliably exhibited a value above -0.2 during the occurrence of most vowels. As this parameter involves a simple derivation from the K's, the expense of computing the spectrum for all voiced frames that meet the above criteria is avoided. The parameter's value may reflect the sharpness of amplitude fluxations in the waveform, as nasals generally exhibit a smooth shape and most vowels contain more rapid shifts in amplitude over time.

- Categories e, f and g above provide general spectral characteristics that are descriptive of nasal murmurs but that are flexible enough to accommodate both speaker-dependent and context-dependent variations in the signal. It was found that a careful description of the separation of formant frequencies would avoid the false capture of the majority of the resonant consonants.

- Because the above parameters are likely to exhibit values within the stated tolerances during transitions, and nasal murmurs are generally steady state events lasting 60 ms or more, the segment isolation decision also requires two or more consecutive frames to have values within the tolerances to effect the capture of the event.

In the previous project, gain information was used as a primary indicator of the temporal location of possible nasal murmurs in order to avoid expensive computational decisions. However, because this approach required a two-pass analysis, gain has not been employed in the current routine, other than as a threshold check and to compute normalized residual energy. (The previous strategy required prior information regarding the possible location of the events, making use of the gain computed from overlapping frames to locate rapid shifts in energy as the signal was damped or released from damping caused by nasal coupling.)

```
INPUT PARAMETERS FOR 1 FRAME

IF   (VOICED  = YES)    AND
     (0X (= 13)         AND
     (RMSN (= 0.8)

     THEN   BEGIN

     COMPUTE FILTER COEFFS

     IF   (A2  ) -0.2)

          THEN   REJECT FRAME
     ELSE
          COMPUTE SPECTRUM/APPLY PEAK PICKING

          IF   (LOWEST FREQUENCY PEAK ) 375 HZ)

               THEN   REJECT FRAME
          ELSE
               IF   (A PEAK OCCURS IN 375-800 HZ RANGE)

                    THEN   REJECT FRAME
               ELSE
                    IF   (# OF PEAKS IN 800-1900 RANGE ) 1)

                         THEN   REJECT FRAME
               ELSE
                    IF   (# OF PEAKS IN 1900-2800 HZ RANGE = 1)     AND
                         (DIFF BETWEEN THIS PEAK AND A PEAK
                         OUTSIDE 1900-2800 HZ RANGE ( 700 HZ)

                         THEN   REJECT FRAME
               ELSE
                    IF   (# OF PEAKS IN 1900-2800 HZ RANGE = 2)     AND
                         (DIFF BETWEEN PEAKS IN 1900-2800 HZ
                         RANGE ( 400 HZ OR DIFF BETWEEN ONE
                         OF THE 2 PEAKS IN THIS RANGE AND A
                         PEAK OUTSIDE THIS RANGE ( 700 HZ)

                         THEN   REJECT FRAME
               ELSE
                    IF   (# OF PEAKS IN 1900-2800 HZE RANGE ) 2)

                         THEN   REJECT FRAME
               ELSE
                    IF   (PREVIOUS FRAME ACCEPTED)
                      OR   (NEXT FRAME ACCEPTED)

                                   THEN   ACCEPT FRAME

               END
               ELSE
                    REJECT FRAME
```

Table 1: Decision logic for isolating consecutive frames with acoustic parameter characteristics of nasal murmurs.

3.4 Results of Segment Isolation

The corpus of data used for segment isolation consisted of 88 sentences, each averaging 1.25 sec in duration for a total duration of 110 sec per subject. For the development of the nasal murmur isolation routine, a total of 1100 sec (18.3 min) of speech data produced by 10 male subjects was therefore processed. This amounted to an estimated total of 54,000 frames of speech data, each frame representing an approximate 20.4 ms interval of the signal (see Table 3).

Each sentence in the corpus contained one nasal phoneme in a unique phonetic environment. However, the occurrence of the acoustic event described as a nasal murmur was not guaranteed, due to the environmental influences of coarticulation. As was found in our previous research, the acoustic manifestation of some phonemic nasals may only be the presence of transitional information, or of nasal activity in coarticulation with a vowel. It was observed from the examination of our corpus of utterances that the murmur was likely to be absent or of very short duration if the nasal phoneme was at the beginning of an utterance, if the signal energy was too low, or if the phoneme was followed by a voiceless consonant such as a fricative, which would require that the velar port remain closed to effect articulation.

In addition to the nasal phonemes in the corpus, some occurrences of voiced stops and resonants produced with a highly impeded oral air flow resulted in the occurrence of nasal murmurs. Subjective auditory examinations of some of these events showed that when they were isolated from the neighbouring vowels and transitional influences, their auditory effect was exactly that of a nasal murmur.

The combined effects of (1) coarticulation during the production of nasals, and (2) the occurrence of murmurs associated with non-nasal phonemes, made it difficult to predict the number or positions of these acoustic events in the data without the aid of careful auditory and spectral analysis. Examination of the data in this manner showed that an average of 570 frames of data per subject were attributed to nasal murmurs that were 40.8 ms in duration or longer and of sufficient energy to have a distinctive spectral shape and auditory quality. Of these, an average of 400 frames were associated with the nasal phonemes in the corpus, and 170 were the result of some other phonetic event such as a voiced stop producing a murmur.

A total of 5700 frames of data associated with nasal murmurs was therefore expected in the 880-sentence data base produced by the ten subjects. As summarized in Table 2, the segment isolation routine located a total of 7000 frames, with an estimated 5400, or 94.7% of the expected nasal murmur frames being included. An estimated 300 frames of nasal data were not captured, generally because of unexpected irregularities in the murmurs produced either by perturbations or irregular formant structure.

	Est. Number of Frames	Duration (sec.)	% to Known Murmurs	% to Total Captured
Known Nasal Murmurs:	5,700	116.3	100.0%	81.4%
Known Murmurs Captured:	5,400	110.2	94.7%	77.1%
Murmurs Not Captured:	300	6.1	5.3%	4.3%
Apparent False Captures:	1,600	32.6	28.1%	22.8%
Total Events Captured:	7,000	142.8	122.8%	100.0%

TABLE 2: Estimated percentages of nasals and non-nasals captured by the segment isolation routine, relative to the total known nasal murmurs, and the total events captured, for the data base of 880 short sentences spoken by 10 male subjects.

As can be seen from Table 2, an estimated 1600 frames were the result of apparent false captures. Upon examination, it was found that all these frames were associated with events that contained some degree of velarization. Velarization occurs when oral airflow is restricted by placing the tongue body near the velum, and is common during the production of [w], [u], [l] and [r]. (The velum is the articulatory organ that is also responsible for controlling the passage of energy into the nasal cavity.) Because no adjustment to the parameters was found that

rejected these events without also rejecting some nasals, tests were conducted to determine their phonetic characteristics.

Observation of the resonant peaks after autoregressive analysis showed that the apparent false captures were similar in spectral shape to the general spectral distribution for nasal murmurs, ie. a dominant first formant peak generally below 350 Hz, the absence or attenuation of F2 between 800 and 1900 Hz, and the presence of a weak and broadened F3 in the 1900 to 2800 Hz range. To illustrate these similarities, Figure 1 provides spectral representations of some nasal segments that were isolated by the routine, compared with some non-nasal segments that were isolated.

	Est. Number of Frames	Duration (sec.)	Percentage of Signal
Total Signal Analyzed:	54,000	1,100.0	100.0%
Known Nasals Captured:	5,400	112.2	10.0%
Apparent False Captures:	1,600	32.6	3.0%
Total Events Captured:	7,000	142.8	13.0%

TABLE 3: Accumulated results of the segment isolation routine, showing percentage of signal isolated from data base of 880 short sentences spoken by 10 male subjects.

Auditory tests were also conducted, and it was found that, before synthesis, the events did not always give the impression of being predominately nasalized, although they all involved a highly impeded oral air flow. The error is therefore assumed to be related to the limitations of the analysis procedure, or to the process of dividing the sampled data into frames to effect analysis. However, this is speculative in that auditory examination of the synthetic speech produced by the vocoder usually showed that the auditory distinction between nasal and velar articulation was maintained. It is expected that a 12-pole model of LPC would be more effective than the 10-pole model in providing parameters that make this distinction.

This apparent failure is not expected to interfere with the verification task however, as the falsely captured events and the nasal murmurs are similar enough in their parametric values to be regarded as belonging to the same class of phonetic events. Although further tests are required, there is some speculation that when the oral airflow is impeded and glottal energy is sufficiently high, resonance may be transferred to the nasal mask where it would radiate in much the same manner as a nasal murmur. Table 3 summarizes the percentage of frames captured in relation to the total signal.

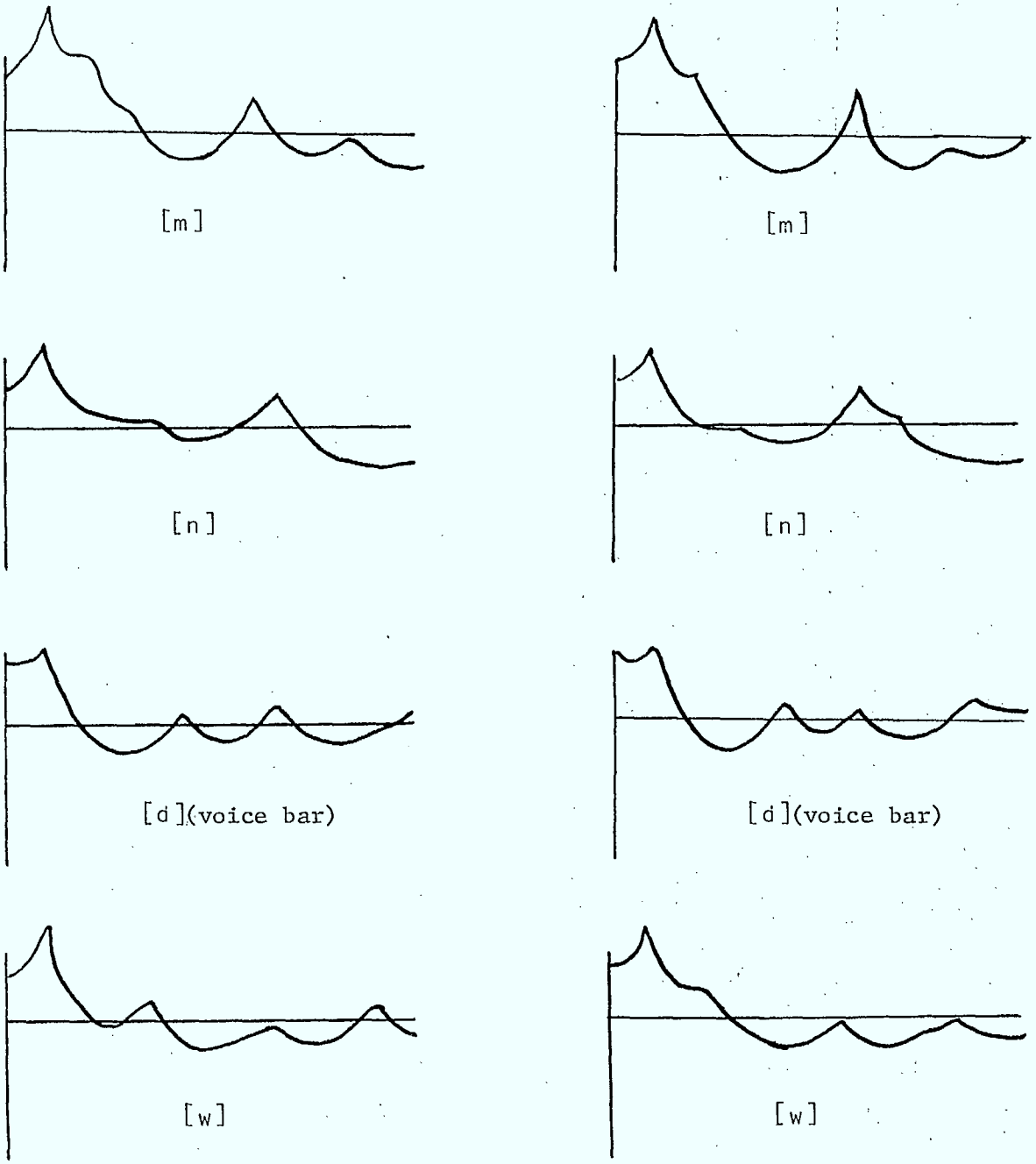


FIGURE 1a: Spectra of some nasal and non-nasal segments taken from Subject 3 using autoregressive analysis. Horizontal scale is frequency (0-4000) and vertical scale is log magnitude.

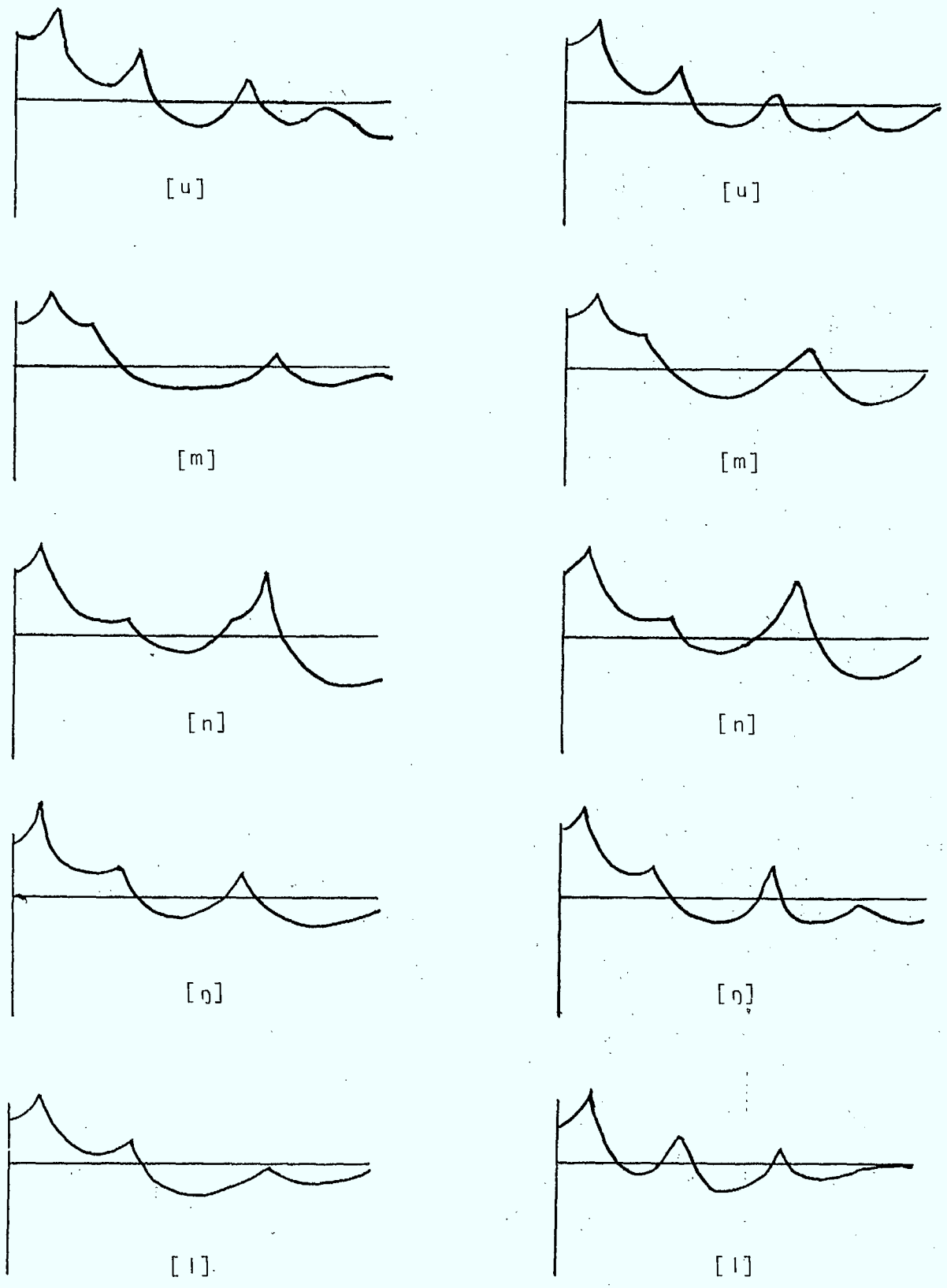


FIGURE 1b: Spectra of nasal and non-nasal segments taken from Subject 4 using autoregressive analysis. Horizontal scale is frequency (0-4000 Hz) and vertical scale is log magnitude.

4. PATTERN MATCHING EXPERIMENTS

4.1 Data Collection for Matching Experiments

The corpus of data used for identifying and isolating the class of nasal murmurs was collected during single sittings for each of the ten subjects. This corpus was used in our previous research, the results of which were summarized in Section 3.1 above. In order to make use of data from this corpus for the pattern matching experiments, test and reference files were created for each subject through random selection of data frames from larger files. The larger files were created earlier by the nasal murmur isolation routine, after processing the data through the vocoder system. The new files were then used in the preliminary selection of parameters for the pattern matching experiments.

In order to provide more conclusive evidence regarding the efficiency and application of the proposed speaker verification system under conditions of elapsed time, a second set of data was also recorded and processed through the vocoder and segment isolation routines. The new data set was devised to test the accuracy of our approach, using context-dependent data taken from 10 male and one female subjects. The eleven subjects were each given the following code:

"My number is BCM-199-379D."

In a practical application, this code could be different for each subject, but for experimental purposes, ten of the eleven subjects must be regarded as imposters while the eleventh is being tested, so all subjects were given the same code. The sample contains nine positions where nasal murmurs might be detected. For future work, the sample also contains four high front vowels, with four more possible, depending upon the subject's tendency to diphthongize.

The average duration of each speech sample was just over five seconds. A context-dependent scenario, rather than a context-independent one, was chosen to simplify the test, as it was expected that the influences of coarticulation would require the accumulation of a larger sample of data for the context-independent task.

To allow for the accumulation of reference data over time, the subjects were recalled on three consecutive days, all three recordings contributing to the reference data. The subjects were then recalled one week following the last reference sample recording session, and the test sample was recorded. The accumulation of reference data by taking samples over time has been shown in other research (eg.[9]) to improve success rates. Successfully matched test samples are generally added to the reference data to update them, thus reducing the elapsed time between the collection of samples being matched.

4.2 Statistical Techniques for Feature Evaluation

Our earlier research [2,3] supports the usefulness of the pre-processing phase that isolates nasal murmurs for speaker recognition. For the adaptation of this approach to the vocoder system, a number of parameters are available from the vocoder that are expected to be useful for speaker verification. These include:

a) the excitation energy
b) the fundamental frequency (Fo)
c) the vocal tract response as determined by ten reflection coefficients (K's).

Two alternative representations of the vocal tract response may be obtained from the above parameters. These are the autocorrelation coefficients (A's) and the cepstral coefficients (C's). For the selection for the nasal data the A's are used directly and indirectly (the indirect use being for the determination of vocal tract resonances for formant estimation). The success attained in the nasal extraction procedure indicates that vocal tract data expressed in terms of the formant frequencies tends to be speaker-independent and thus of no use in speaker recognition. The indication that the A's are influenced by shifts in laryngeal activity also shows that these coefficients are of little use in speaker recognition.

It might be expected that the other two representations, that is, the K's and C's would suffer from the same problem of speaker independence, but investigations have shown that this is definitely not the case. In fact, the data gathered to date indicate a very high degree of speaker dependence in the statistical behavior of the reflection coefficients taken from the isolated nasal frames. Because of time restrictions associated with this preliminary investigation, the cepstrum coefficients have

not yet been examined. However, the work of other investigators [4,8] indicates that if the reflection coefficients show a high degree of speaker dependence, then the cepstra will be even more efficient. In the anticipated further development of the verification system, the C's will therefore be automatically derived from the K's transmitted by the vocoder system (see Figures 2 and 3).

A variety of mathematical techniques are available for application to the pattern recognition problems arising in speaker identification and speaker verification. In the discussion that follows we will frequently refer to "data vectors". Each data vector consists of an n-tuple of parameter values extracted from a single 20 ms data frame. For most of the discussion the data vectors examined will consist of the fundamental frequency as one component and the ten K's as the others.

For the pattern recognition problem at hand we have a set of data vectors forming a reference set for each subject who is to be identified by the system. It is expected that the reference data available will exhibit sufficient subject-dependent behavior to allow us to carry out the desired identification or verification task at hand.

The actual operation of the system will involve the capture of additional data vectors from a speaker who is either completely unknown or who claims to be one of the reference subjects. In the first instance, the system is required to determine whether or not the unknown speaker is one of the reference subjects; in the second the system must judge whether or not the speaker is indeed the person he claims to be.

The mathematical techniques that must be used in a problem of this nature involve an examination of the statistical behavior of the individual components of the data vectors for each reference subject and analysis of any interactions between components. One such technique involves the extraction of what are known as the Principal Components (cf. [16], chapter 8) of the data vectors. This method recognizes that all of the parameters that make up a data vector may not be of equal importance in explaining the variation of values observed for a particular speaker. The computations that take place in principal components extraction are designed to determine a linear transformation that may be applied to the data vectors for a specific set of reference data that will produce a new set of data vectors. These transformed vectors have the following key properties:

a) the individual components may be ranked in order of importance related to the amount of the variation in the original data explained by each.

b) the individual components are statistically uncorrelated.

The first property is particularly useful if there is a requirement to identify certain data components as being more significant than others. Additionally, if the length of the data vectors in a particular application is longer than desired, the original data may be replaced by transformed data in which only the first few principal components have been retained.

Another standard technique used in statistical pattern recognition is Linear Discriminant Analysis(cf. [16], chapter 6, and [17], chapter 5). By analyzing the statistical properties of the reference data base, a set of linear functionals is determined (a different functional corresponding to each reference subject). When an unknown test subject is to be identified, the set of linear functionals is applied to the data vectors captured from the subject and the largest functional value obtained is used to classify the subject as one of the reference subjects. This technique is more suited to the case of "closed set" identification where the test subject is known to be one of the reference subjects and all that must be determined is which one.

One major problem that arises from the application of discriminant analysis techniques is the strict requirements that must be satisfied by the data being analyzed. The common assumption that data vectors from all the reference populations possess a common covariance matrix is often satisfied to a sufficient degree that the method is appropriate. However, the other standard requirement is that the data vectors for each subject possess a multi-normal probability distribution. It is very common for researchers to completely ignore this requirement and proceed to use the method on data for which it is not suited. The major problem that arises in this instance is the determination of error rates that have no justification from probability theory.

4.3 Statistical Properties of Nasal Murmurs

A number of researchers have made use of the reflection coefficients and fundamental frequency values taken from the

voiced component of the speech signal as the basis for speaker recognition. The focus of our investigation has been the selection of acoustic events belonging to a phonetic class as the basis for recognition, with the expectation that more speaker-dependent information is available during these types of vocalic events.

An indication of the expected effectiveness of the pre-processing phase that isolates the nasal murmurs can be obtained through a comparison of the statistical properties of the variables extracted from the nasal murmurs with those extracted from voiced events that have not been separated into acoustic classes. In particular, most pattern recognition techniques are based on variations, from subject to subject, of the average values of the selected parameters. The general behaviour of the parameters under discussion may be seen from the data in Tables 4a through 4e, in which the means and standard deviations of the parameters are listed for large samples of nasal murmur data and voiced data.

TABLE 4: Comparison of extracted nasal murmur data and voiced data of five subjects, showing means and standard deviations of each category. The variables 1 to 10 are the values of the 10 K's and the 11th variable is Fo.

	NASAL		VOICED	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
1	-0.925	0.062	-0.701	0.355
2	0.466	0.211	0.141	0.432
3	0.014	0.272	-0.123	0.348
4	0.263	0.243	0.151	0.286
5	-0.101	0.189	0.096	0.266
6	0.225	0.219	0.272	0.229
7	0.178	0.261	0.123	0.238
8	0.114	0.198	-0.016	0.245
9	0.028	0.212	-0.089	0.215
10	0.066	0.208	0.064	0.187
11	119.984	17.560	112.486	26.785

TABLE 4a: Subject 1, Data Set 2.
No. of Nasal Frames: 129
No. of Voiced Frames: 611

	NASAL		VOICED	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
1	-0.840	0.096	-0.423	0.415
2	0.025	0.185	0.032	0.333
3	-0.264	0.232	-0.233	0.350
4	0.386	0.206	0.057	0.256
5	0.212	0.198	-0.023	0.274
6	0.058	0.157	0.218	0.222
7	0.001	0.137	0.096	0.214
8	0.178	0.141	0.253	0.185
9	-0.053	0.152	-0.049	0.179
10	0.131	0.132	-0.010	0.170
11	104.717	9.037	103.407	14.399

TABLE 4b: Subject 2, Data Set 2.
No. of Nasal Frames: 53
No. of Voiced Frames: 551

	NASAL		VOICED	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
1	-0.842	0.101	-0.304	0.501
2	0.121	0.277	-0.023	0.332
3	-0.352	0.198	-0.037	0.387
4	0.303	0.263	0.092	0.310
5	0.167	0.250	-0.054	0.285
6	0.069	0.148	0.209	0.220
7	0.254	0.216	0.186	0.189
8	0.095	0.157	0.352	0.213
9	-0.033	0.223	-0.012	0.194
10	0.047	0.148	-0.021	0.136
11	148.180	16.568	141.446	24.716

TABLE 4c: Subject 3, Data Set 2.
No. of Nasal Frames: 50
No. of Voiced Frames: 581

	NASAL		VOICED	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
1	-0.837	0.107	-0.569	0.326
2	0.140	0.161	0.290	0.329
3	-0.284	0.280	-0.247	0.317
4	0.548	0.166	0.247	0.249
5	0.153	0.248	0.005	0.253
6	-0.048	0.234	0.208	0.260
7	0.081	0.153	0.182	0.200
8	0.199	0.172	0.283	0.214
9	0.173	0.214	0.006	0.202
10	0.042	0.210	-0.070	0.157
11	109.776	10.274	105.383	14.557

TABLE 4d: Subject 3, Data Set 1.
No. of Nasal Frames: 161
No. of Voiced Frames: 2056

	NASAL		VOICED	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
1	-0.883	0.077	-0.723	0.303
2	0.204	0.216	0.222	0.363
3	-0.156	0.242	-0.156	0.310
4	0.297	0.289	0.113	0.264
5	0.094	0.251	0.048	0.285
6	0.061	0.239	0.186	0.242
7	0.170	0.157	0.173	0.187
8	0.139	0.183	0.054	0.223
9	0.017	0.240	-0.132	0.223
10	-0.028	0.192	-0.027	0.158
11	131.146	21.345	131.571	20.296

TABLE 4e: Subject 4, Data Set 1.
No of Nasal Frames: 219
No. of Voiced Frames: 1661

Several patterns of behaviour are apparent from the data listed in Table 4, which may be confirmed by similar analysis of the data available for the other subjects:

1) There is a significant difference in the mean values of the first few reflection coefficients for the nasal data as opposed to the general voiced data.

2) More importantly, the standard deviation of these variables is much lower for the nasal data. These smaller values produce very tight confidence limits about the means for samples drawn from the population of nasal frames produced by an individual speaker.

3) The average values of the fundamental frequency vary a good deal from speaker to speaker, again with a much smaller standard deviation for the nasal frames.

Although the fundamental frequency values are often cited as being valuable for speaker recognition, the ease with which these parameters may be disguised eliminates them from serious consideration for forensic identification purposes. For verification, there is also concern that they may vary excessively from the reference sample values under conditions of emotional stress, and that they may be consciously imitated by imposters. The values of the first few reflection coefficients, however, are believed to be related to the actual vocal tract geometry of the speaker (configured in this case for the production of a nasal or nasal-like event) and should prove more difficult for an imposter to replicate. The lower standard deviation values for the nasal data indicate that we have been able to isolate a particular acoustic event during which a subject displays less variation in the form of his speech production.

The basic statistical test to decide whether or not a sample has been drawn from a given population involves the comparison of the sample mean to that of the population. With the population standard deviation known and with knowledge of the statistical distribution of the population, decisions can be made with a known level of accuracy. A statistical principle known as the "Law of Large Numbers" provides some relief from the requirement that the population distribution be known if one is dealing with large samples (typically more than 30 values). In either case, the ability of the statistical tests to discriminate between two populations is directly proportional to the population standard deviation and inversely proportional to the size of the sample. The advantage of the smaller standard deviations exhibited by the reflection coefficients

taken from the nasal murmur data is the ability to discriminate between speakers using considerably smaller sets of sample data.

An example of this may be seen in the data in Tables 4a to 4e. For Variable no. 1 (the parameter K1), the standard deviations in the nasal data are approximately 1/4 the values of the equivalent parameter in the voiced data. This means that with 1/16 the sample size of the voiced information, we can obtain the same level of discrimination between mean values taken from the nasal information.

It is evident from the data provided that no single parameter is sufficient to distinguish between the five subjects listed. However, the variations in a number of the parameter values from one subject to the next indicate a good chance of achieving appropriate separation by using a combination of values.

4.4 Selection of Parameters for Discrimination

The previous sections have indicated the strong possibility that successful subject recognition can be performed using the mean values of a number of the LPC parameters obtained from frames of speech data that have been identified as nasal or nasal-like information. The problems that must be examined are:

a) which parameters should be chosen for optimal discrimination?

b) is it possible to set detection thresholds that will allow us to make the decision that the current subject is not one of the reference subjects, (ie. to perform open set verification)?

A number of techniques have been used during the course of investigations related to this project in order to answer the first question. These will be described briefly here, with some indications being provided of the level of success achieved by each technique.

The major data analysis was carried out on a subset of the two data bases. (The first set of data was the sentence list spoken by 10 subjects and used for determining methods for isolating the class of nasal murmurs. The second set of data consisted of the four separate recordings with unvarying contexts made by 11 subjects, as described in section 4.1 above.) The nasal data taken from five of the

subjects in the first group and the nasal data from eight of the subjects in the second group were used.* Test and reference files were taken from the first set by random selection to assure context-independence. For the second set, the results of the first three recordings made one day apart were used as the reference data and the results of the fourth recording one week later were used as the test data.

The data vectors for each 20 ms nasal frame consisted of the 10 reflection coefficients and the fundamental frequency produced by the vocoder system. The number of frames for each set of data is given in Table 5.

SUBJECT	REF. FRAMES	TEST FRAMES
SET 1, SUB 1	78	80
SET 1, SUB 2	65	40
SET 1, SUB 3	70	76
SET 1, SUB 4	114	113
SET 1, SUB 5	74	69
SET 2, SUB 1	129	52
SET 2, SUB 2	53	16
SET 2, SUB 3	50	20
SET 2, SUB 4	134	44
SET 2, SUB 5	116	37
SET 2, SUB 6	116	42
SET 2, SUB 8	163	45
SET 2, SUB 11	70	33

TABLE 5: Number of 20 ms reference frames and test frames used in the data base for pattern matching experiments.

---

* Because of time restrictions, the final version of the nasal murmur extraction routine was not implemented in time to begin the collection of data for the pattern matching experiments. The original version which relied heavily on the values of the A's instead of the spectral peaks failed to produce a sufficient number of test frames from the 5-second utterances of three of the subjects for adequate analysis. However, the new routine was tested separately on these and all other subjects, and no failures to produce adequate samples resulted. The frames isolated by the original routine are essentially a subset of the frames isolated by the final version.

4.5 Statistical Techniques for Discrimination Tests

- Linear Discriminant Analysis

The statistical technique of Linear Discriminant Analysis referred to above was applied to the data in the preliminary stages of the investigation. The method was found to function adequately for purposes of closed set identification but was later discarded for the following reasons:

1) The magnitudes of the discriminant functions when they are evaluated have no absolute significance, but must be interpreted across the reference data being used. This makes the method inadequate for open-set identification.

2) The discriminant functions produce a single real value that is used to classify the test data. This means that no method is available to directly determine the effect of a single aberrant component of the data vector - and a single such value may seriously affect the classification outcome. (The problem of severe effects on the classification results stemming from aberrant behaviour of one or two components of the data vectors will also affect the performance of other common classification techniques such as the weighted Euclidean distance and the Mahalonobis distance.)

3) The statistical properties that are demanded of the data in the standard discriminant analysis techniques (multi-normal distributions with common covariance matrix) are not satisified closely enough by the data vectors available.

- Principal Components Analysis

The other technique mentioned above, Principal Components Analysis, was applied to the data before developing the classification procedures. In order to carry out an analysis that is based on the deviations of individual test sample means from the corresponding reference data means, it is desirable to be able to identify and apply higher weights to the most significant variables. Also, if the individual components were known to possess normal distributions, this would imply that the principal

components were independent, making it possible to make meaningful statements concerning the setting of detection thresholds and false alarm rates for open-set classification.

- Classification Using Z-values

In Section 4.3 above it was indicated that a reasonable inter-speaker separation of mean values exists for many of the components, and that quite small values of standard deviation accompanied these values. For this reason, an attempt was made to measure the degree of separation of the mean values of the test data components from the corresponding components of the reference data. This was undertaken for both the original data and the derived principal components. The technique employed is outlined below.

For each speaker in the set of those to be identified, the reference data vectors that have been obtained are processed to extract the means μ_i and standard deviations σ_i ($i=1,\ldots N$) where N represents the number of components in the data vector. Typical values used were N=11 (all components used) and N=6 (only the first 5 K's and Fo used). In the case where Principal Components Analysis has been applied, the data vectors are obtained from the original reference data by applying the appropriate principal components transformation.

When an unknown speaker is to be identified, the data from a number of nasal frames is extracted from the vocoder system while he is speaking (see Figures 2 and 3). These data, referred to as the test data are then compared to the stored information from the reference data. The comparison takes place against the reference data for each of the original speakers and is carried out in the following manner:

1) If the original data are obtained by extraction of the principal components, the test data are converted using the same transformation that was originally applied to the reference data of the speaker in question. Then the mean value is calculated for each component of the transformed data vectors, yielding a set of mean values $\overline{x_i}$ ($i=1,\ldots N$).

2) If the original data are analyzed without using the principal components, the mean values $\overline{x_i}$ ($i=1,\ldots N$) represent the averages of the components of the test data vectors with no modifications.

3) For each component, a value is calculated (which will be referred to as the Z-value) using the formula

$$Z = \frac{(x_i - \mu_i)\sqrt{M}}{\sigma_i} \qquad (i=1,\ldots N)$$

where M represents the size of the test sample (ie. the number of nasal data frames captured from the speech of the unknown subject).

A few comments are appropriate concerning the behaviour and interpretation of the Z-values:

1) If each component of the data vectors was normally distributed, the Z-values would possess a normal distribution with mean 0 and standard deviation 1;

2) If, in addition, the individual components were uncorrelated (hence independent) then the probabilities of normal values greater than the Z-values obtained could be multiplied to obtain the overall probability that the test data came from the speaker to whose reference data the comparison was being made.

Although the requirement for normality is partially satisfied for the data obtained and the Principal Component Analysis produces uncorrelated variables, it is not possible to use the standard analysis techniques with the Z-values with any real justification. A detailed study of the statistical properties of the K's in the data should allow decisions to be made concerning the significance of the Z-values. This should permit the setting of accurate thresholds for open-set matching of test samples with claimed reference samples for verification purposes.

The separation of mean values for the components of the data vectors for the reference speakers is, however, large enough that unsophisticated methods can be used to analyze the Z-values in order to perform classification of the unknown speakers. In Section 5 below, we will describe two straightforward implementations that involve the use of the Z-value approach to classify the thirteen speakers in the data base described in Section 4.4.

5. CLASSIFICATION RESULTS

In order to adequately assess the effectiveness of any pattern matching techniques, a number of applications must be carried out using different test and reference data sets. Random subsets of the test and reference data files of each subject were generated for this purpose. For example, a particular simulation might involve the choice of random samples of 50 frames from the reference data of each subject and the use of those data to attempt to identify the speaker when random samples of 30 frames are chosen from the test data.

Two key parameters that must be determined if the identification system is to function efficiently are (1) the minimum number of frames required in the reference data base for each subject, and (2) the minimum number of frames required to be extracted from the sample utterance in order to obtain a high degree of recognition accuracy. As stated earlier, the larger the test sample, the more closely the parameter means will cluster around the corresponding population means for the speaker in question. However, for efficiency, it is desirable to minimize the the number of frames that are required for the task. The simulation technique involving random selection of a number of frames allows us to examine the error rates obtained for various test sample sizes and reference data base sizes.

5.1 Method 1: Classification with K's and Fo

The classification algorithm involves the summing of all Z-values computed in relating the test data means to the means for one of the reference subjects. The sums S_j ($j=1,\ldots$) obtained for each of the reference groups are compared and the test speaker is identified as the reference subject for whom S_j is the minimum. No attempt is made in this method to discard certain Z-values that appear to represent aberrant behaviour for the particular test sample being classified.

This method was applied to the data from the 13 subjects. All of the reference frames for each subject were used to calculate the mean and standard deviation values required for the analysis. In all cases, with the exception of Subjects 2 and 3 of Data Set 2, for whom an insufficient

number of test frames were available (see Table 5 and previous footnote), a sequence of 200 randomly selected sets of 30 test frames was chosen from those available and matched with one of the 13 reference subjects using the technique just described. As summarized in Table 6, the results yielded an error rate of 2.14% A very encouraging aspect of these results is that only one incorrect match was reported for the second set of data in which the reference samples were collected over a period of three days and the test samples were collected one week later.

TEST SUBJECT	TRIALS	CORRECT	INCORRECT	ERRORS(%)
SET 1, SUB 1	200	167	33	16.5
SET 1, SUB 2	200	200	0	0.0
SET 1, SUB 3	200	200	0	0.0
SET 1, SUB 4	200	194	6	3.0
SET 1, SUB 5	200	193	7	3.5
SET 2, SUB 1	200	199	1	0.5
SET 2, SUB 4	200	200	0	0.0
SET 2, SUB 5	200	200	0	0.0
SET 2, SUB 6	200	200	0	0.0
SET 2, SUB 8	200	200	0	0.0
SET 2, SUB 11	200	200	0	0.0
OVERALL	2200	2153	47	2.14%

TABLE 6: Results of classification experiment involving all parameters (10 K's and Fo), and comparing test files from 11 subjects with reference files of 13 subjects. The reference files consisted of all available frames, and the test files consisted of 30 randomly selected frames. A total of 2200 trials were conducted, involving random selection of test data 200 times per subject.

In order to determine the optimum sizes of the test and reference data when this technique is applied, a number of further tests were conducted. A total of 1100 classification attempts was conducted for each examination of a different test or reference size, that is, 100 randomly selected test files generated from each of the 11 larger test files used. These tests are summarized in Table 7. As can be seen from Table 7, all 11 available parameters were used except in two tests cases, in which only the first 5 K's and Fo were used. The results indicate a slight improvement in performance obtained from the use of all 11 parameters.

NUMBER OF VARIABLES	REFERENCE DATA SIZE	TEST DATA SIZE	ERRORS(%)
11	40	20	8.73
11	40	30	6.27
11	40	30	1.27
6	40	30	2.00
11	60	20	1.27
11	60	30	0.82
11	60	40	0.00
6	60	30	1.82
11	80	20	2.73
11	80	40	0.18
11	80	60	0.00

TABLE 7: Results of classification experiments involving K's and Ro, and different test and reference sample sizes. Each result involved 1100 separate classification attempts, or 100 attempts for each of 11 test subjects.

In terms of the general effectiveness of the method, it appears from the results summarized in Table 7 that reference data bases consisting of data vectors from 80 nasal frames or more are required, and test samples should be capable of providing 40 or more nasal frames to provide the best results. Test and reference files of these sizes are easily obtained using a method of elicitation such as that outlined in Section 4.1. If context-independent speech samples are to be used, a typical sample length of 16 seconds or more would be required to obtain 80 nasal frames, with an expected 10% of the signal being identified as nasal or nasal-like (see Table 3).

5.2 Method 2: Principal Components Classification

In this method, the selected number of principal components is extracted from the reference data and the means and standard deviations are calculated. Before the test data are compared with a specific reference file, the test data are first transformed using the same principal components transformation as was applied to the reference file being examined for the match. The Z-values are then calculated as they where in Method 1 above.

To take advantage of the power of the principal components, the Z-values that are derived from the analysis

are weighted according to the theoretical significance of the principal components in the data. The weight w, ranging from M for the first component,, (M-1) for the second, and down to 1 for the component of the least theoretical importance, is first applied to the Z-values. The Z-values are then summed to obtain the selection values S_j, as was done previously. Table 8 summarizes the results of a series of trials using this method. The trials were conducted in the same manner as those shown in Table 7.

NUMBER OF VARIABLES	REFERENCE DATA SIZE	TEST DATA SIZE	ERRORS(%)
11	40	30	0.82
6	40	30	11.72
11	60	20	1.23
11	80	20	0.36
11	80	30	0.45
11	80	40	0.73
6	80	30	3.55

TABLE 8: Results of classification experiments involving principal components and different test and reference sample sizes. Each result involved 1100 separate classification attempts, or 100 attempts for each of 11 test subjects.

A comparison of Tables 7 and 8 shows that in general there is a slight improvement in the effectiveness of using the principal components and the weighting function, when the same test and reference sample sizes are referred to. However, this slight improvement in performance does not appear to warrent the increased computation required to obtain the principal components, suggesting that efforts should be concentrated upon the development of a verification routine that uses the original data.

6. DISCUSSION

Our success in determining parameters in the vocoded data for phonetic classification, and the preliminary results of the speaker-discriminating efficiency of parameters that are available in the vocoder system, lead us to anticipate success in designing a speaker verification system that will operate in conjunction with this digital communications device. The results of the experiments carried out to date suggest that the reflection coefficients obtained from the nasal frames contain sufficient speaker-dependent information to be used for speaker verification. A thorough study of the statistical behaviour of the reflection coefficients produced during nasal activity should result in the development of a more sophisticated analysis technique for examining the Z-values. Such as study is also expected to lead to the determination of detection thresholds for open-set verification.

However, higher success rates are expected to result through the use of cepstrum coefficients in place of the K's. As the implementation of cepstrum coefficients involves a simple transformation, the examination of their statistical properties may be done in parallel with further analysis of the K's.

The phonetic class of nasal murmurs was chosen for this stage of the experiments because of projected difficulties in isolating these events when restricted to the selection of parameters produced by the 10-pole model of LPC. However, our success indicates that the isolation of other classes of phonetic events will be trivial in comparison, making it possible to implement several parallel levels of analysis for speaker verification purposes. As stated in the theoretical approach above, sets of parameters in the different classes of events that are isolated may be appealed to selectively, This way, any failures in one set of parameters can be offset by successes in others.

Such a scheme has more practical applications in context-independent verification than in context-dependent, as temporally-based methods could be implemented in order to find the phonetic events of interest in the latter task (although the task would then be weakened by the failure of a speaker to produce a phonetic segment at the appropriate position in the utterance). However, in an application such as the verification of an aircraft pilot's identity, the use of a code may not always be feasible, as it may interfere

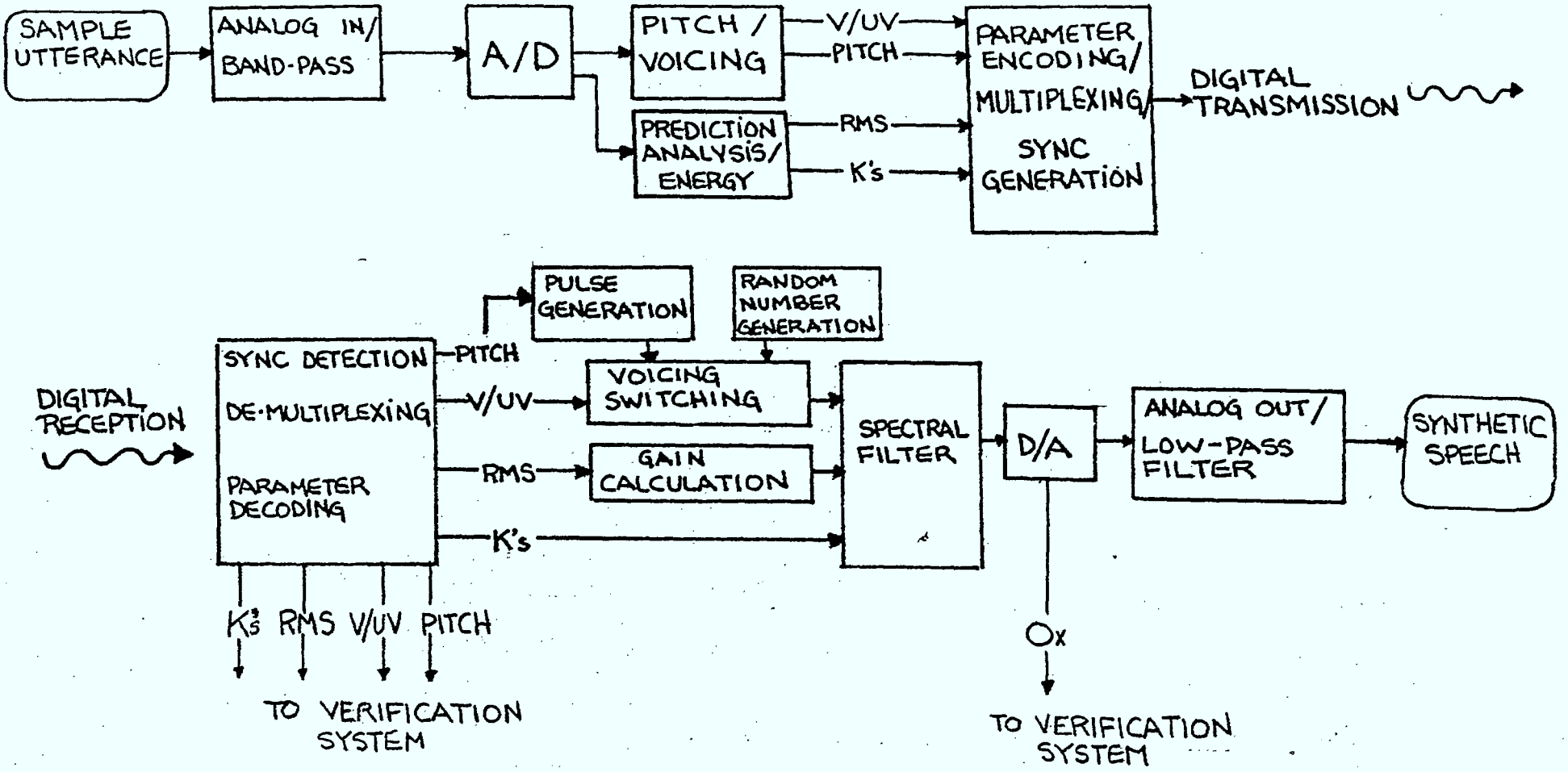


FIGURE 2: BLOCK DIAGRAM OF VOCODER SHOWING TRANSMIT AND RECEIVE STAGES, WITH POINTS OF INTERFACE TO SPEAKER VERIFICATION SYSTEM

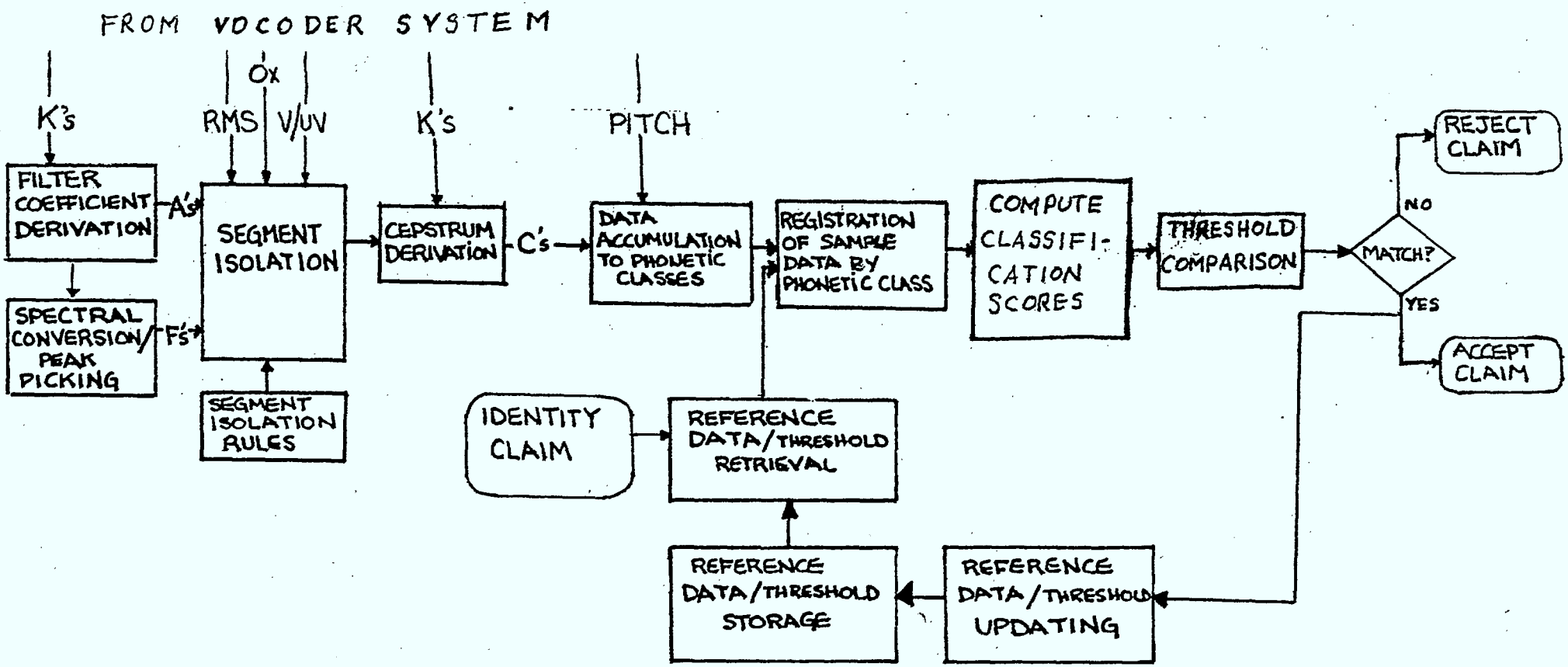


FIGURE 3: BLOCK DIAGRAM OF SPEAKER VERIFICATION SYSTEM, SHOWING PARAMETERS INPUT FROM VOCODER

with other tasks involving communication. The selection of phonetic events from free context would therefore mean that verfication could be performed from any stretch of the speech signal, as long as sufficient information can be gathered to perform the task. As illustrated in Section 5.1 above, a total duration of 16 seconds of free speech is the minimum requirement if a context-independent approach must be used.

The block diagrams of Figures 2 and 3 are schematic representations of the Communication Research Centre's vocoder system interfaced with the proposed speaker verification system. In Figure 2, the vocoder system is represented, showing the encoding and decoding stages. After the parameters are decoded they are fed to the verification system (except for 0x, which is picked up after D/A conversion) to be operated on independently of the processing phases that produce the synthetic speech.

As shown in Figure 3, the values are then fed into the various stages of preprocessing for segment isolation, and rules are applied to control the selection of data frames. Two or three segment isolation operations will be conducted in parallel at this stage, depending upon the number of phonetic classes to be isolated in the final version. For computational efficiency, gating procedures will be used to control the various stages of analysis (cf. Table 1) and cepstrum derivation will be performed only on frames that match the segment isolation rules.

For speaker verification, a critical component of the system is the capability to update reference files and thresholds, using information received through successful matches. This essentially converts successfully matched test samples to reference data by pooling them with the reference data held in storage.

A procedure for registering an identity claim will be implemented, in order to retrieve the reference file belonging to the claimed speaker. For verification this claim enables a single comparison to be made, that being the comparison of the test sample being received through the system with the reference file retrieved by the individual wishing to verify the speaker's identity. This accentuates the need for an efficient threshold determination procedure as discussed in the previous section.

In further development of the verification routine, factors that have been previously controlled, but that may interfere with the task must be investigated before accurate predictions can be made regarding the extent of applications

for the system. Such factors include the effects of emotional stress, respiratory disorders and background noise. Also, hardware requirements and processing time must be taken into account in the design of the system, with due consideration being given to the complexity of statistical measures that are to be implemented. These problems can only be solved by the examination of a large population of speakers under a variety of conditions.

REFERENCES

[1] Rosenberg, A.E., Listener performance in speaker verification tests, IEEE Transactions in Audio and Electroacoustics AU-21:221-225, 1973.

[2] Dickson, B. Craig, and Warkentyne, H.J, SPIDENT Project Final Report, research and development of a computerized speaker identification system, unpublished final report to the RCMP, Ottawa, contract 1SU80-0024, May, 1982.

[3] Dickson, B. Craig, An Investigation of Theories and Parameters Pertaining to Speaker Recognition, Unpublished Thesis, University of Victoria, 1980.

[4] Atal, B.S., Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification, Journal of the Acoustic Society of America 55:1304-1312, 1974.

[5] Sambur, M.R., Selection of acoustic features for speaker identification, IEEE Transactions in Acoustics, Speech, and Signal Processing ASSP-23:169-176, 1975.

[6] Markel, J.D., and Davis, S.B., Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base, IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-27:74-82, 1979.

[7] Pfeifer, L.L., New techniques for text-independent speaker identification, 1978 IEEE International Conference Record on Acoustics, Speech, and Signal Processing 3:283-286, 1978.

[8] Hunt, M.J., Yates, J.W., and Bridle, J.S., Automatic speaker recognition for use over communication channels, 1977 IEEE International Conference Record on Acoustics, Speech, and Signal Processing 2,:764-767, 1977.

[9] Furui, S., Cepstral analysis technique for automatic speaker verification, IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-29:254-272, 1981.

[10] Geppert, R., Kuhn, M.H., Piotrowsky, H., and Tomaschewski, H., An operational system for personnel authentication using long-term averaged voice spectrum, 1979 Carnehan Conference on Crime Countermeasures, University of Kentucky, Lexington, pp. 77-80, 1979.

[11] Hollien, H., Childers, D.G, and Dohorty, E.T., Semi-automatic system for speaker identification (SAUSI), 1977 International Conference Record on Acoustics, Speech, and Signal Processing 2:768-771. 1977.

[12] Basztura, C, and Majewski, W.: The application of long-term analysis of the zero-crossing as a speech signal in automatic speaker identification, Arhives of Acoustics 3:3-15, 1978.

[13] Hoefker, U., Jesorsky, P., Kreiner, B., and Wesseling, D., A new system for authentication of voice, 1979 Carnehan Conference on Crime Countermeasures, University of Kentucky, Lexington, pp. 47-51, 1979.

[14] Doddington, G., Voice authentication gets the go-ahead for security systems, Speech Technology 2:14-24, 1983.

[15] Paul, J.E., Rabinowitz, A.S., Rigonati, J.P., and Richardson, J.M., Development of analytical methods for a semi-automatic speaker identification system, 1975 Carnehan Conference on Crime Countermeasures, University of Kentucky, Lexington, pp. 52-64, 1975.

[16] Morrison, D.F., Multivariate Statistical Methods (2nd Edition), McGraw-Hill, 1976.

[17] Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis Wiley-Interscience, 1973.