AN INVESTIGATION INTO THE PRACTICAL IMPLEMENTATION OF SPEECH RECOGNITION FOR DATA CAPTURING

By Craig van der Walt

Thesis submitted in partial fulfilment of the requirements for the Masters Diploma in Technology to the Department of Electrical Engineering (Light current) at the Cape Technikon.

Telkom Development Institute (TDI)
Telkom SA

Cape Town
South Africa
October 1993

DECLARATION

I declare that the contents of this thesis represents my own work and the opinions contained herein are my own. It has not been submitted before for any examination at this or any other Institute.

C. Van Der Walt

(Signature)

ABSTRACT

A study into the practical implementation of Speech Recognition for the purposes of Data Capturing within Telkom SA. is described. As datacapturing is increasing in demand a more efficient method of capturing is sought. The technology relating to Speech recognition is herein examined and practical guidelines for selecting a Speech recognition system are described. These guidelines are used to show how commercially available systems can be evaluated. Specific tests on a selected speech recognition system are described, relating to the accuracy and adaptability of the system. The results obtained illustrate why at present speech recognition systems are not advisable for the purpose of Data capturing. The results also demonstrate how the selection of keywords words can affect system performance. Areas of further research are highlighted relating to recognition performance and vocabulary selection.

OPSOMMING

Hierdie verhandeling beskryf 'n ondersoek na die praktiese implimenteering van spraak herkenning vir die vaslegging van data binne Telkom SA. Met die toenamme van aangevraagde data moes daar 'n meer doeltreffende metode van vaslegging gevind word. Die tegnologie van spraak herkenning word geevalueer en praktiese rig-lynne word verskaf, wat gebruik kan word wanneer 'n spraak herkenning stelsel gekies word. Hierdie rig-lynne is gebruik om 'n kommersiel beskikbaar stelsel te evalueer. Spesifieke toetse met betrekking tot die akuraadheid en aanpasbaarheid van die stelsel word beskryf. Die toetse bewys hoekom spraak herkenning hydelik vir data vaslegging nie geskik is nie. Die resultate wys ook hoe die keuse van sleutelwoorde die stelsel kan beinvloed. Gebiede van verdere ondersoek met betrekken tot woord verkiesing en stelsel akuraatheid word ten einde beskryf.

ACKNOWLEDGEMENTS

I would like to thank the following people for their assistance in compiling this document:

Dr G.J. Prinsloo whose expertise and advice was invaluable in completing this document. The informative discussions and insight is greatly appreciated.

Mr B. Mortimer of Cape Technikon whose enthusiasm and encouragement was never lacking.

Mr G. van Wyk of Datafusion for readily providing technical assistance, despite having a busy schedule of his own.

Telkom Development Institute (TDI) Cape Town and its staff for providing the facilities and resources for completing this project.

Datafusion for supplying the Recognition System hardware.

Telkom Drawing Office for providing the staff for the tests as well as assisting in the training of the capturing software.

Telkom SA for providing me the opportunity to complete this project.

I would also like to thank all those who assisted me in completing this project.
I apologise for not mentioning you all by name.

Finally I would like to thank my wife, Janine and my son, Calvin for the constant encouragement and patience you have displayed while I was completing this Project.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

One of the greatest joys parents can have is to hear their child utter its first word. From this moment on a never ending learning cycle of speech begins. Although a child may take a few months to utter its first words, it is able to recognise the speech of others even earlier. This aspect of recognition and uttering of speech is part of communication. So much of our developing world relies on communication. Scientists therefore started looking at methods of improving communication.

One of the fields that was investigated involved the reproduction of speech. From this the synthesizer was invented. It involved the manipulation of sounds in such a way that eventually intelligable speech was reproduced. This synthesized speech could easily be recognized by even the youngest child. It was however a long time before any machine was able to do the same.

This aspect of speech recognition was another area investigated by scientists. Their investigations have resulted in a number of speech recognition systems being realeased in the commercial market. This document examines the feasibility of using one such recognition system within Telkom SA.

1.1 Speech Recognition within Telkom SA

In order to use current technology Telkom has been upgrading its working environment to cater for the advances that are being made in the computer industry. One area that has been given particular attention by Telkom is the computerisation of existing topographical and cadastral information. At present most of this information is found on large and often unmanageable maps. Due to the vastness of South Africa, computerisation of all this

information is a tremendous task. As this is labour intensive, the question was raised as to whether the present method of data capturing was the most practical. One of the possible solutions investigated involved speech recognition. The aim of this thesis was to establish the practicality of implementing speech recognition as a means of data capture. The study would also determine whether implementation of a recognition system would increase productivity.

This document consists of two main parts. Chapters 2 to 5 look at the speech recognition technology. Some of the advances that have been made in this field of study are discussed. The criteria and specifications involved in selecting a speech recognition system are identified.

The second part of this document is covered in chapters 6 to 8. This section uses the guidelines already presented and demonstrates how they are applied in selecting a recognition system for digital mapping (DM).

1.2 Detailed chapter description:

1.2.1. Section 1:

In **Chapter 2** the early history of speech recognition is summarised after which the principles of speech recognition are explained by looking at a brief description of a recognition system.

Chapter 3 takes this description a step further by examining the input stage of a speech system in greater detail.

Chapter 4 looks at the recognition stage of the system and discusses the modern approaches applied to this stage of speech recognition, highlighting two main areas of research namely, pattern matching and knowledge engineering.

Chapter 5 provides the reader with practical guidelines for choosing a speech recognition system.

1.2.2. Section 2:

Chapter 6 shows how the specified criteria is applied in an attempt to solve a problem relating to digital mapping in Telkom. This chapter also identifies a commercial system that was chosen to be tested.

Chapter 7 describes the practical tests that were applied to the chosen system and tabulates the results for examination.

Chapter 8 presents the conclusion and recommendations. In this chapter conclusions are drawn and recommendations for the furthering of this work are presented.

CHAPTER 2 : PRINCIPLES OF SPEECH RECOGNITION

This section discusses the early history of speech recognition and the basic principles involved in this field of research. Some of this work has been presented in investigations [1] into the practical use of speech recognition, as well as in other sources [2] dealing with speech recognition.

2.1. Historical Overview of Speech Recognition

2.1.1. Development in the 30's and 40's

Speech research started in the 1930s when the practical digital transmission method of pulse code modulation (PCM) was developed. The invention of the sound spectrograph in 1946 contributed to increased speech analysis research [3]. This was because it provided a simple practical two-dimensional display of the acoustic output of the speech signal in the frequency domain [2].

2.1.2. Development in the 50's and 60's

Viewing speech segments as composed of discrete distinctive features originated in the 1950s [4]. This spurred development of electronic speech synthesisers, such as pattern playback synthesisers [5]. Digital speech coding in the form of delta modulation was developed at the same time [2].

Work done on the acoustic theory of speech production [6] introduced a decade of further speech research during which speech was synthesised by computer [7]. The application of digital signal processing (DSP) techniques [8] in the analysis of speech signals became popular around this time. Among the most common techniques used were the fourier transform, linear prediction spectrum analysis and cepstrum analysis.

2.1.3. Development in the 70's

Research in the early 1970's resulted in the development of time-adaptive speech coding [9] [10], speech understanding and the use of dynamic programming in matching templates of different speech signals [11]. In the late 1970's more complex speech systems such as sub-band and adaptive transform speech coders appeared. The development of large scale integrated circuits at this time helped to make the appearance of one-chip speech synthesisers.

2.1.4. Development in the 80's and early 90's

Significant developments during the 1980's have included,

- Single-chip digital signal processors,
- The use of vector quantization [12] for low-rate speech coding [13],
- Excitation for speech synthesis [14]
- The use of human auditory [15][16] models
- Neural models [17][18] in speech applications.

The knowledge engineering approach [19][53] applied to speech recognition in the late 1980's caused the revival of many feature based ideas investigated in the 1950's. The development of digital signal processing techniques [8][20] together with the use of artificial intelligence [21] played a major role in the re-introduction of the feature-based approach. Thus many of the early ideas that were overlooked due to the lack of computer power and memory were now given attention.

The latter approach is recently gaining more interest under speech researchers [22] as it has the potential to alleviate the inherent problem of most pattern matching speech recognition systems. Thus the previous limitation of recognising only unnatural discrete speech or word segments, instead of fluent conversational continuous speech with an unlimited vocabulary could be realised.

This would indicate that technology and hardware are at a level where real-time speech processing is practical. The trends in research indicate that speech recognition will play an important role in future technological advances. In this study a commercially available advanced speech recognition card is evaluated. This card makes use of the latest advances in speech recognition and is produced locally. Details of tests performed are recorded in latter chapters.

2.2. Speech Recognition : A Basic Description

To understand the various features of a speech recognition system, a description highlighting the main components of a generalised system, is given.

System components

Figure 2.1 illustrates the three main system components, namely :

- The input stage, consisting of the analogue to digital (A/D) conversion system and the Signal processing / Feature extracting component

- The recognition stage where matching takes place between the prototype storage and the input speech.

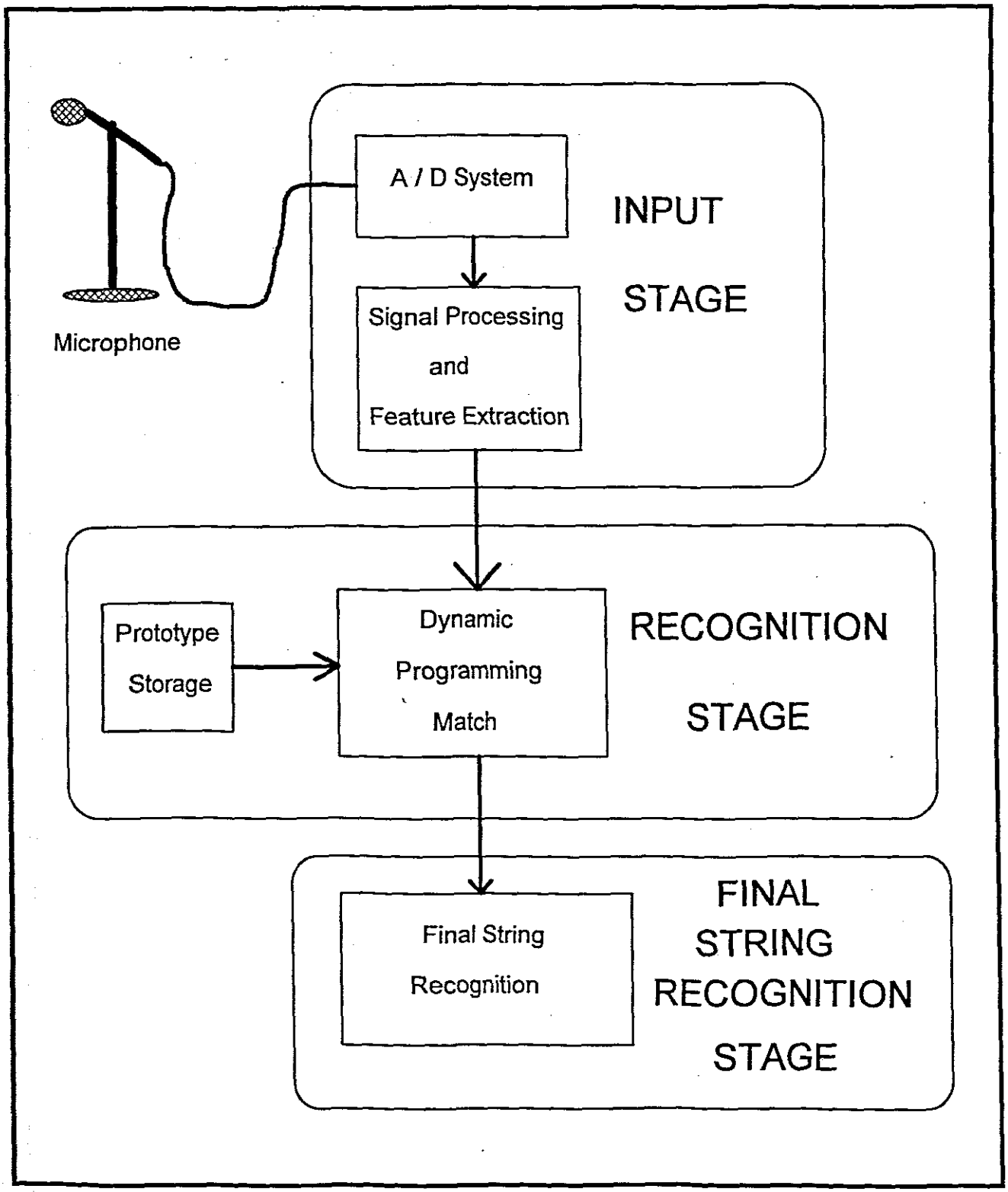


FIG 2.1 A Block diagram of a basic Speech Recognition System Model

- The final string recognition stage, where the recognition system interfaces with the user application.

2.2.2. Input Stage

The first component of the input stage is an A/D converter that is typically used in DSP applications. Enhancements such as automatic gain control (AGC), pre-emphasis equalisation, and or some form of noise cancellation are possible. The more expensive systems would possibly include all of the above.

The performance of this section will also be dependent on the bit resolution of the A / D converter. The higher the resolution the better the base from which to extract features. The disadvantage of higher bit resolution is that greater processing power is now required to still maintain the same real time recognition rates.

The second component of the input stage is the feature extraction component. This is where the each system differs from another. Every system incorporates its own unique method of extracting the necessary speech features. These extracted features are then used in the recognition stage. A more detailed discussion of how this is done is presented in chapter 3.

2.2.3. The Recognition Stage

The voice features that were extracted in the previous stage are used in the operation of this stage. This stage has the task of recognising or matching these features with a group or set of "reference" speech patterns. These patterns could have been trained by the user or the could be pre-trained by the manufacturer of the recognition system.

Systems trained by the user are obviously better adapted to the user wheras systems pre-trained by the manufacturer are faster to implement as no prolonged system training is required. The advantages/ disadvantages of the two options are further discussed in chapter 5.

For real time operation a match in as short a time as possible is required. Various methods and algorithms have been and are being developed to improve this search process. Common methods in use are:

- Dynamic Time Warping (DTW)

- Hidden Markov Models (HMM)

- Neural Nets (NN).

Each method is implemented differently and at different levels of complexity. Two main fields of research are being followed and these are described more fully in chapter 4.

Once the correct pattern match is found the result is transferred to the final stage of the recognition system.

2.2.4. The Final String Recognition Stage.

The main function of this stage is to interface to the user's operating system. The function performed by this stage will depend largely on the user requirements. Most systems however interface to the operating system via the standard system input device. The input speech is recognised and placed into the keyboard buffer as if the user had physically typed it in themself.

The input speech can be likened to a string of characters being entered via the keyboard. The function of this stage is to ensure that the voiced "input string" is associated with the equivalent keystrokes programmed into the system.

All these stages are normally incorporated on a separate plug-in board. Most of the DSP is done on this board so as to avoid taxing the main processor. This ensures that the user is still able to run his own applications whilst benefiting from voice recognition as an alternate input to the main process.

CHAPTER 3 : ACOUSTIC FEATURE EXTRACTION

In this chapter a more comprehensive description of the input stage is given. Characteristics of human speech applicable to speech recognition will also be described. Before examining the first stage, information regarding the spectral form of the various speech features is discussed.

3.1. Speech Input Features

Human speech can be described as a syntaxical pattern combination of numerous frequencies combined to form a complex, usually informative, wave form. Most human speech lies within the bandwidth of 0 to 8 kHz although certain higher frequencies are obtainable. (Clicking, hissing and certain consonants especially reach high frequency levels.)

The majority of speech recognition systems operate in the bandwidth 0.3 kHz to 4 kHz which is close to the standard bandwidth of normal telephone speech (0.3 kHz to 3.4 kHz). Every subject's voice is different. There are however common features or characteristics that can be extracted by the recognition system. They relate to the energy present in the speech, the pitch and tone of the speech.

Figure 3.1 [23] shows a graphical representation of some of these features. The X-axis represents the time domain whereas the Y-axis represents the frequency domain. The density is an indication of the power content. This diagram represents a sonogram or spectrogram of the speech. This has the following characteristics :

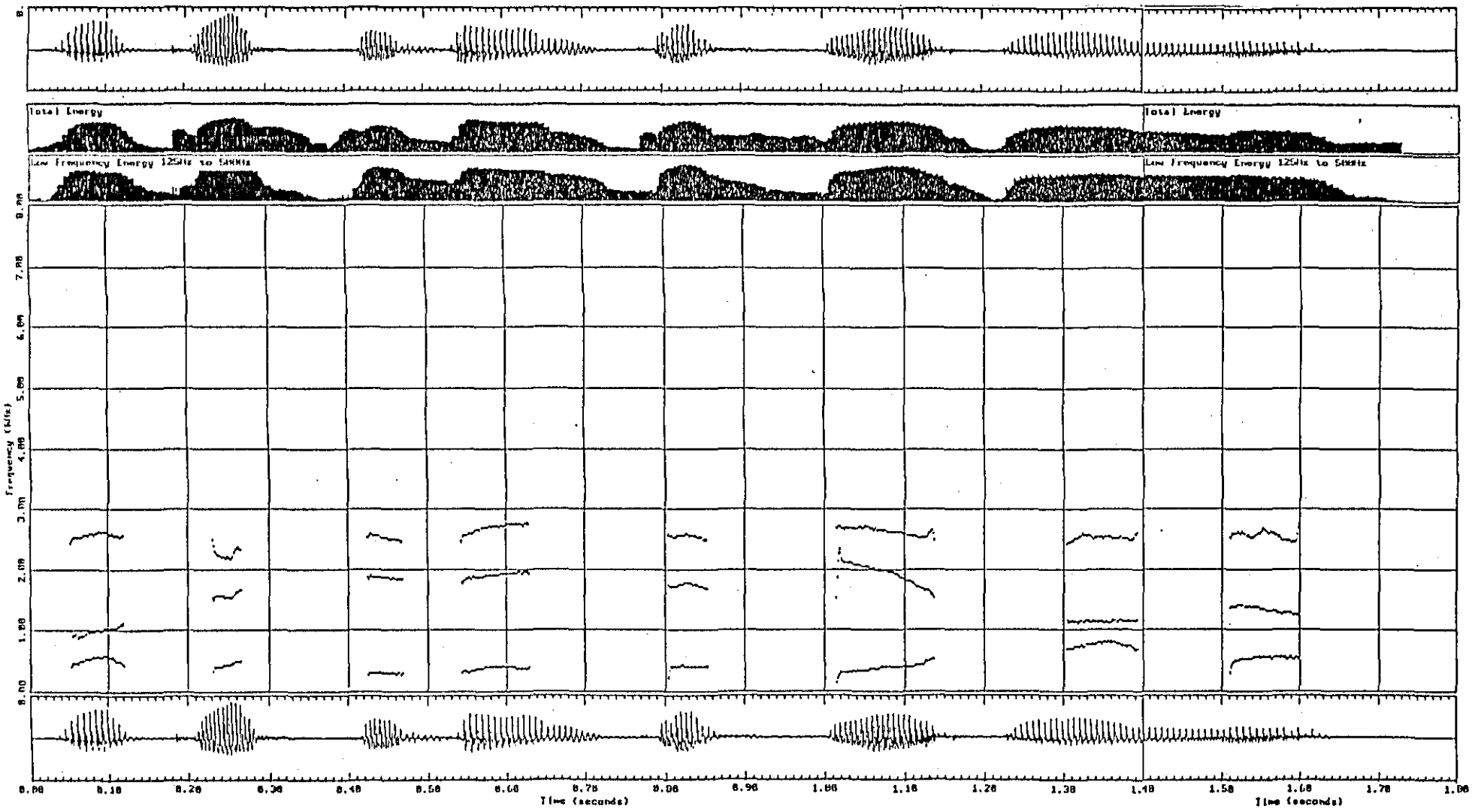

Total Energy
Total Energy
Low Frequency Energy 125Hz to 5000Hz
Low Frequency Energy 125Hz to 5000Hz
Frequency (kHz)
Time (seconds)
Time (seconds)

Fig 3.2

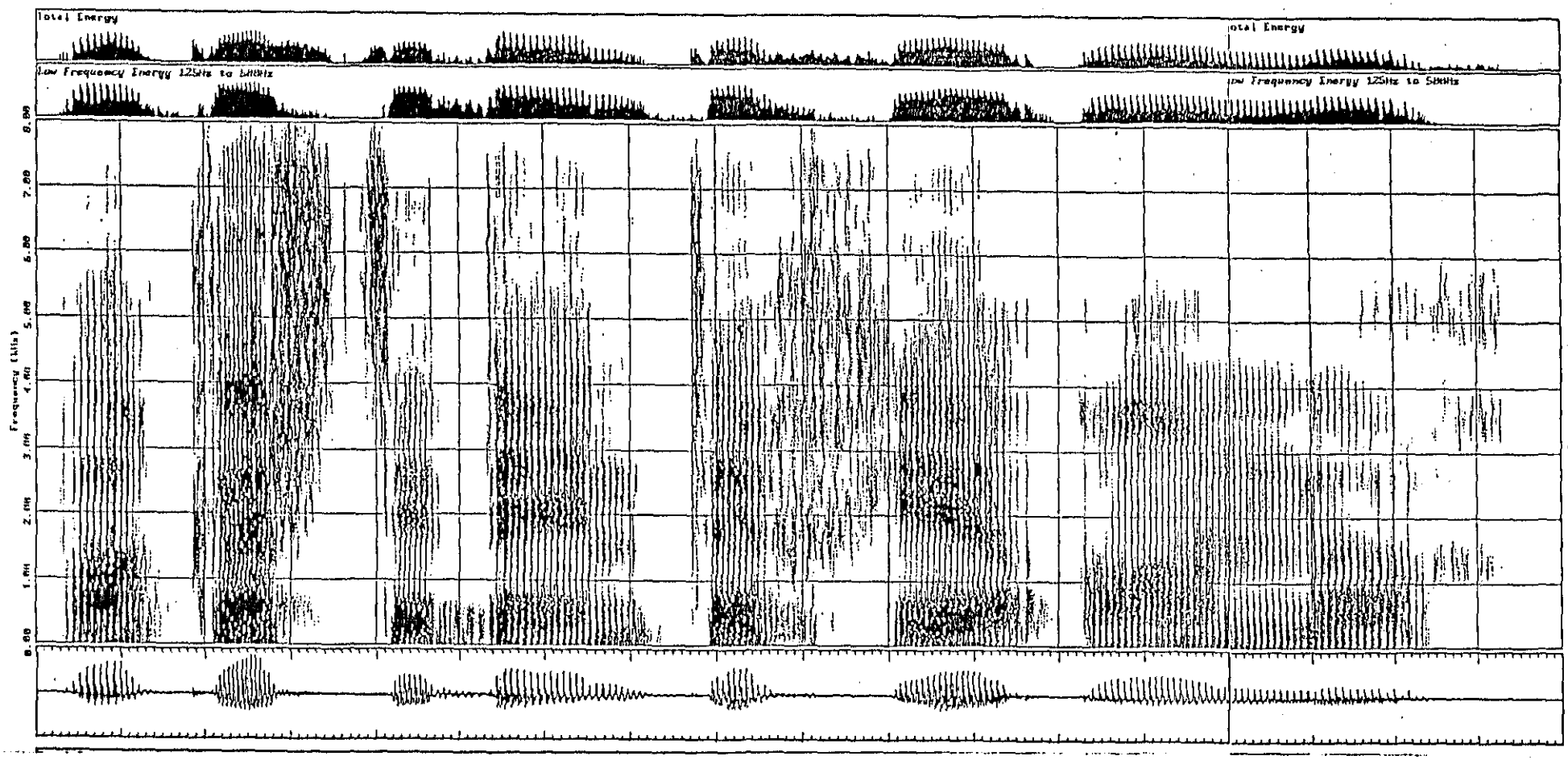

Total Energy
Low Frequency Energy 125Hz to 5kHz
Frequency (kHz)

Fig 3.1

- The darker areas correspond to the higher energy levels.
- Lighter areas represent low or no energy content.
- Most of the speech content lies below the 4 kHz mark.

Fig 3.2 [23]Shows the same speech patterns with only the higher energy areas or formants being displayed. This could be likened to a contour map showing only the higher threshold areas. As the computer system is to extract features out of these acoustic signals it is vital that these signals are as different as possible. To illustrate:

In Fig 3.3 and Fig 3.4 [23] the two words "conventional" and "potential" are represented. On examining the two spectrograms the reader will be able to identify certain features that are very similar. These similarities can also be heard when voicing the relevant words.

In comparison if any of the words "management" or "chimpanzees" in figures 3.5 & 3.6 [23] are compared a clear difference can be seen. These differences result in more accurate recognition as distinct differences can now be detected.

Thus it can be seen that the choice of words to be trained in a speech recognition system should not be random but rather be carefully selected in order to obtain the best feature extraction. The greater the difference in the spectrograms of the keywords, the harder it will be for the system to recognise false patterns.

Likewise the longer the speech pattern trained into a system, the more the distinctive features that are recognisable. The greater the number of features contained in the word, the less likely it will be for the pattern to be duplicated in another word. This factor

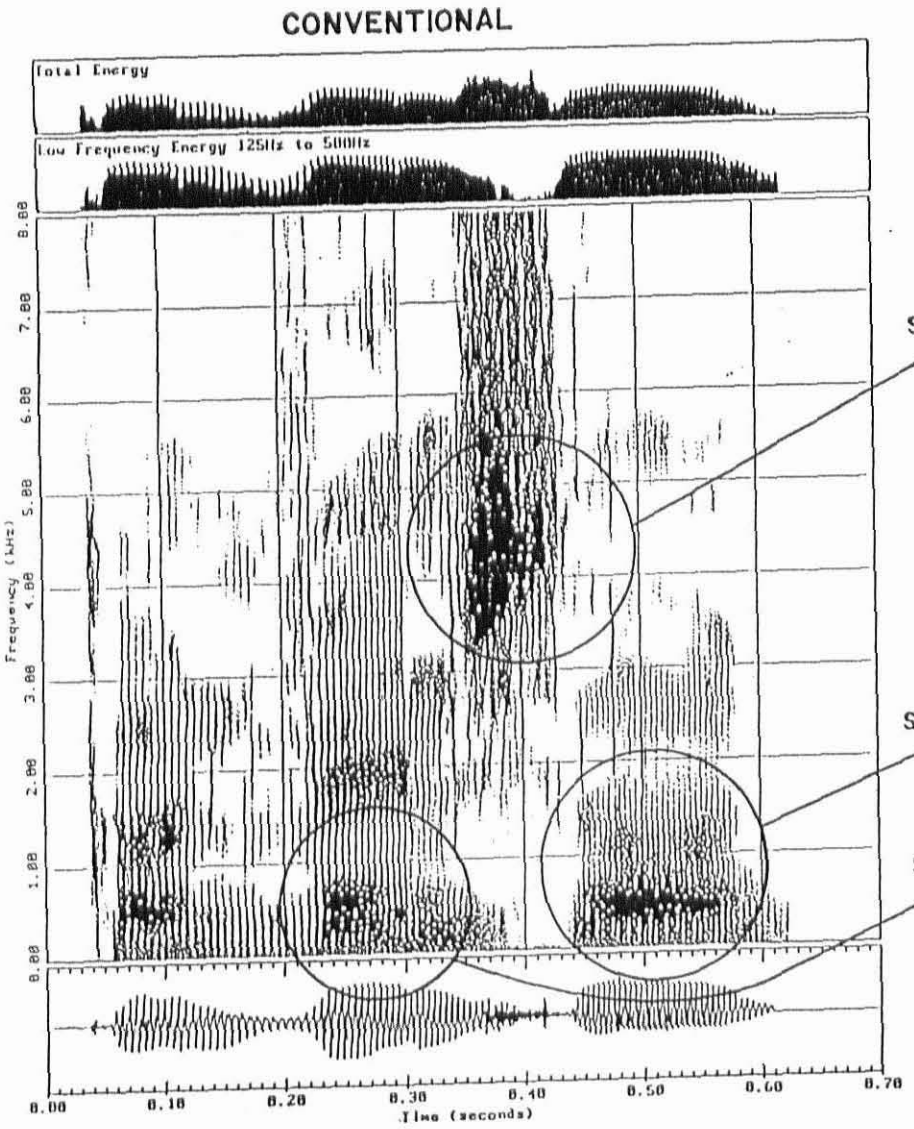


Fig 3.3

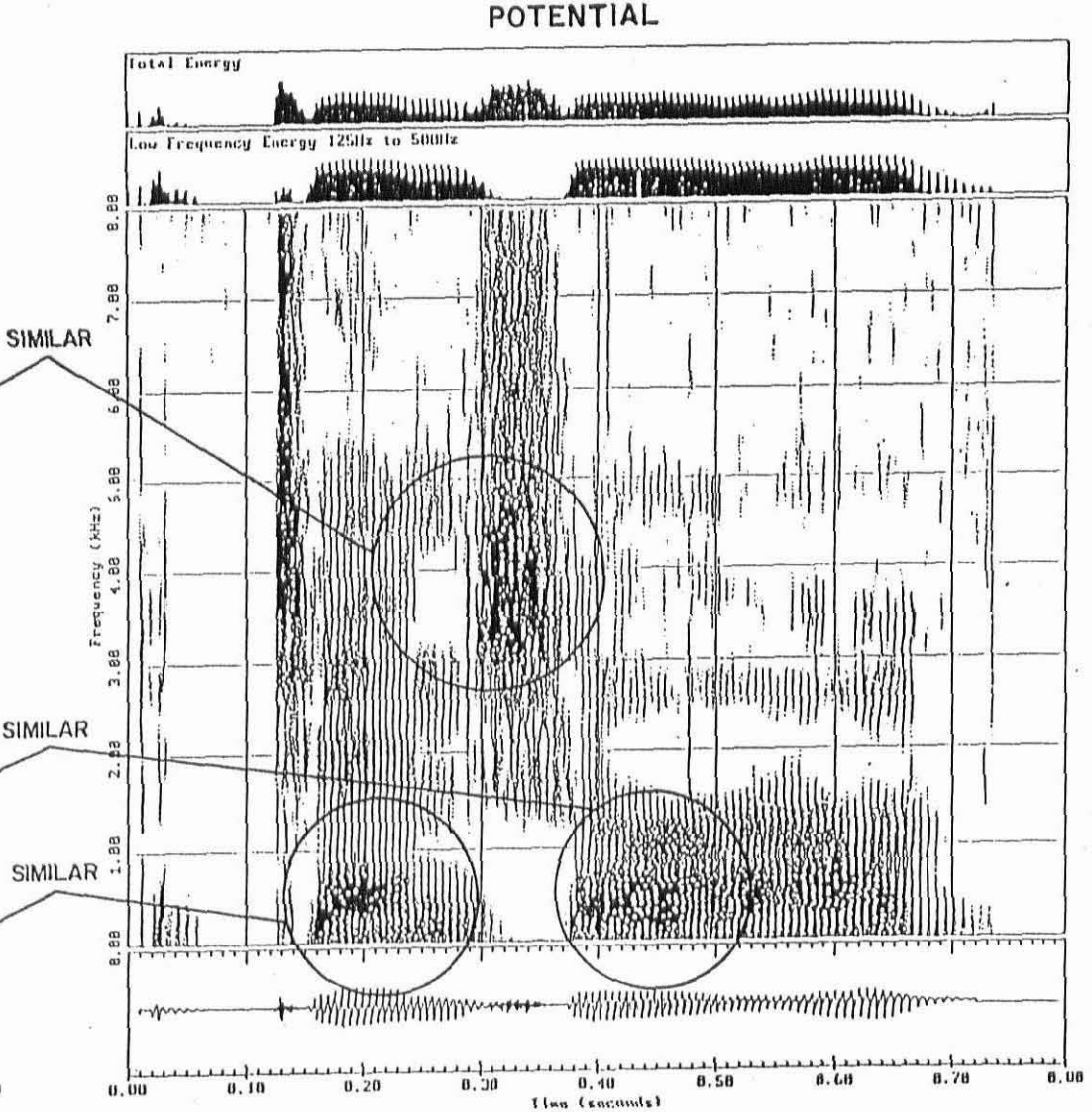


Fig 3.4

SIMILAR

SIMILAR

SIMILAR

CHIMPANZEES

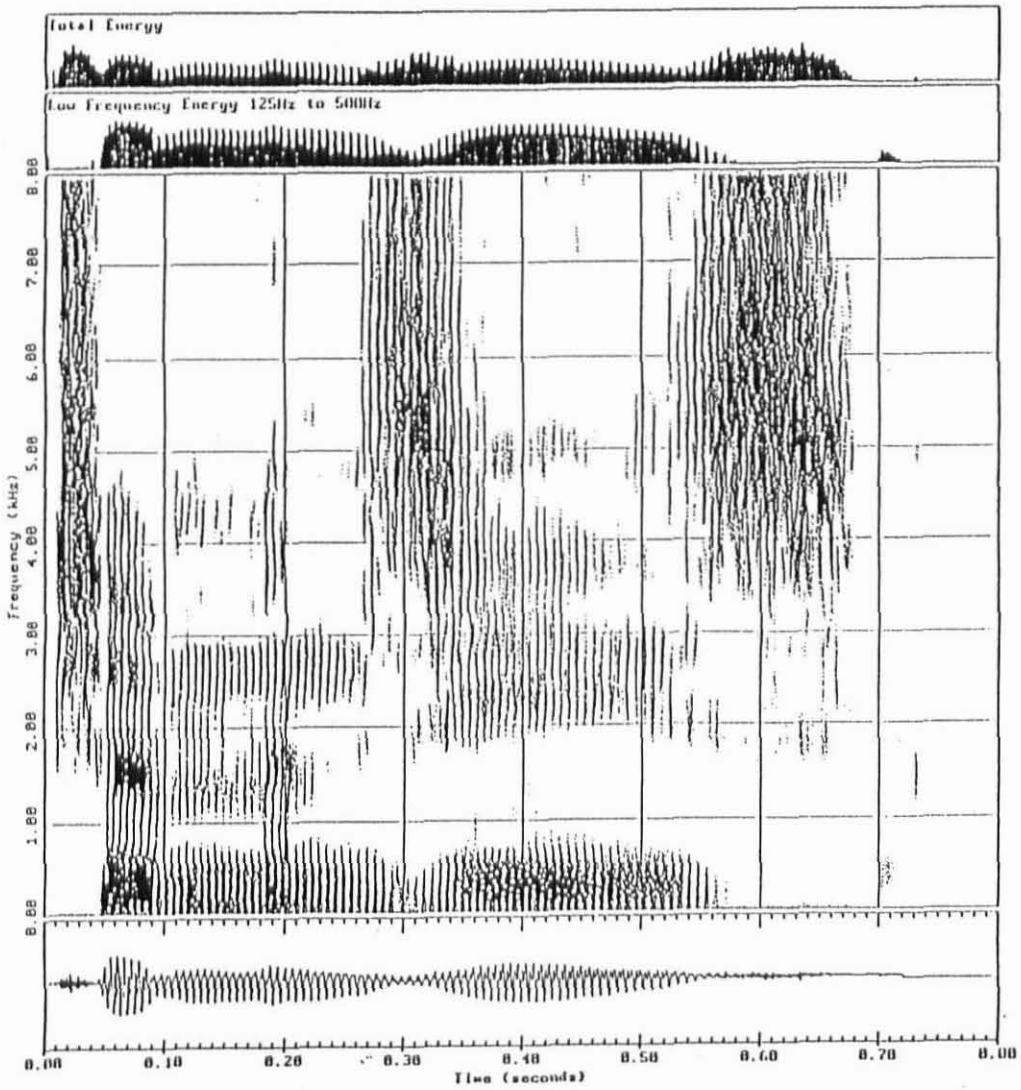


Fig 3.5

MANAGEMENT

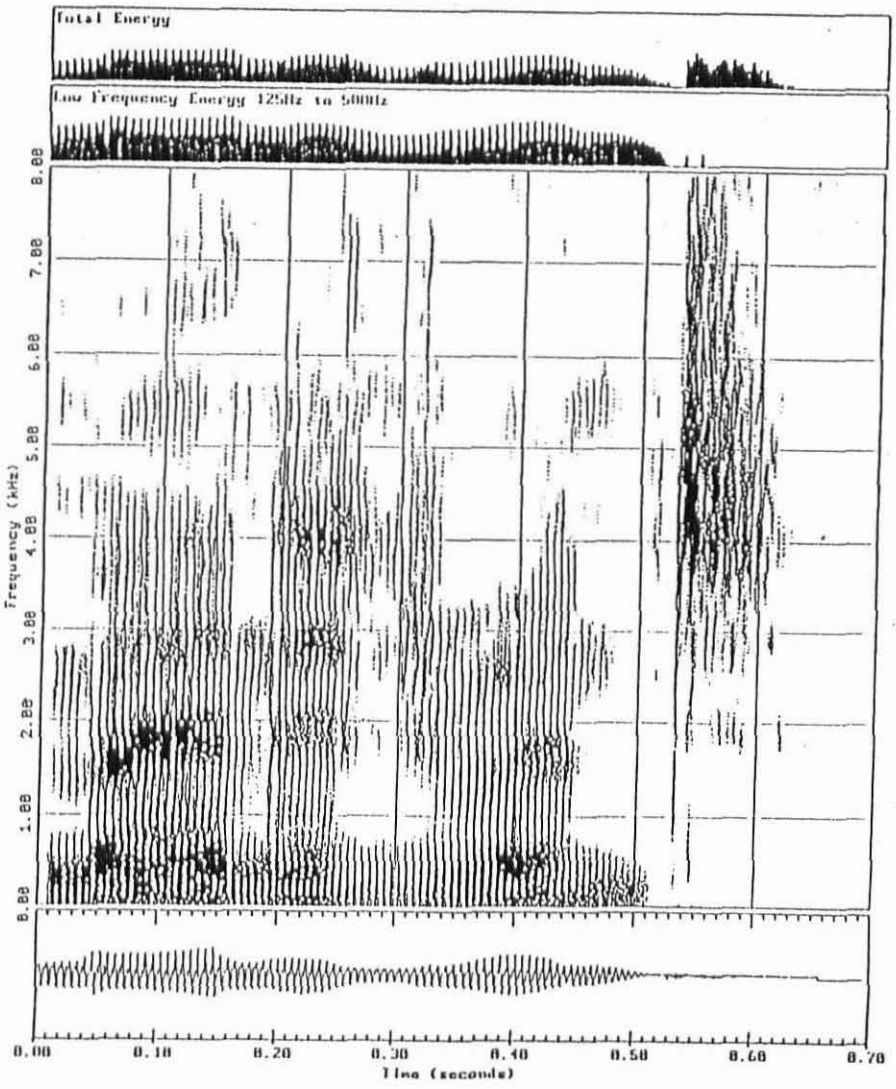


Fig 3.6

explains why recognition systems that are trained with very short keywords will often have a lower accuracy than systems that have longer keywords.

3.2. DSP used in Speech Features

By using DSP the input analogue speech is converted to a digital format that can be manipulated by the recognition system. Digital signal processing (DSP) involves first obtaining a representation of the input signal, based on a given model, and then the application of some higher level transformation in order to structure the signal into a convenient form. The last step in the process is the extraction and utilisation of the parametric representation and related information.

To paramatize the speech signal for speech recognition purposes, most recognisers follow the speech model exploited in voice coding (vocoding) [24][25]. This processing separates the excitation from the vocal tract response. Thus higher prosodic influences (intonation and word-stress) are eliminated from the speech signal. This ensures that the speech pattern to be recognised will virtually always be the same. It could be likened to removing all word stress and intonation from a public speaker in such a way that his speech now becomes a single monotonous string of words. Although not pleasing to the ear, it ensures that all speech patterns will appear the same at the recognition stage.

Current development of electronic models of the human ear have assisted in the understanding of how speech is recognised and understood. Thus researchers are more able to mimic the functions by hardware or software. These are referred to as peripheral auditory models. The models are then used as input stages to the speech recognition system [26][27]. Systems employing these types of models have been tested and promising results have been obtained [18][28].

3.3. Applying DSP in Recognition System

A number of speech recognition systems apply DSP in the following manner: The input speech is typically divided into time slices ranging between 10 and 40 ms[2]. These time slices overlap each other. As the input bandwidth of speech is normally limited to the 0 - 4 kHz range an adequate sampling frequency of 8 kHz [2] is usually selected. Some systems however attempt to improve this range by selecting a sampling rate of 16 kHz. This over sampling usually allows for a better signal to noise ratio for the same fixed bandwidth.

Once the conversion has been performed the signal is now in a form where speech features can be extracted.

3.4. Methods of Feature Extraction

Various methods of feature extraction are in use such as linear predictive coding (LPC) coefficients, mel-based cepstral coefficients, digital Fourier transform (DFT), zero crossing rates, adaptive filter energies, cross- and auto correlation [2][20][29]. As each of these principles represent a study on their own they will not be discussed further.

After applying one of the various methods of feature extraction, the input speech can now be represented as a multi-dimensional feature vector. A feature vector can be described as a point in a spacial matrix that represents the relative position of the uttered sound.

To illustrate: If one considers the surrounding universe, us the relative position of a particular planet or star can always be referenced by some mathematical equation representing its position. A speech sound will always be created within the mouth, throat or nasal cavity. As such any sound can be identified as occurring at certain location within this space. A feature vector describes this space mathematically.

As speech is a combination of sound it can be represented as a multi-dimensional feature vector. This feature vector can then be used as an input to the next stage where the pattern matching or recognition is performed.

CHAPTER 4 : SPEECH RESEARCH - MODERN APPROACHES

4.1. Introduction

Chapter 2 highlighted the advancement of speech recognition technology over the last few decades. During this time speech research has become a technological field of its own. The development of computer technology stimulates the research in this field continuously. The present-day developments in the speech field is divided into two main approaches:

- **Knowledge Engineering Approach**
- **Pattern Matching Approach**

Section [4.2] overviews the knowledge engineering approach while Section [4.4] deals with the pattern matching approach. Section [4.5] then compares these two approaches.

4.2. Knowledge Engineering Approach

Success with spectrogram reading in the last few years encouraged a substantial interest in the problem of computer recognition of continuous speech on a knowledge engineering approach [19]. The knowledge engineering approach makes use of knowledge base expert systems. In an expert system, expert information pertaining to the system function is programmed into the system. Decisions will then be made by the system based on this stored information. The system will try and obtain the best solution from its given rules and knowledge base.

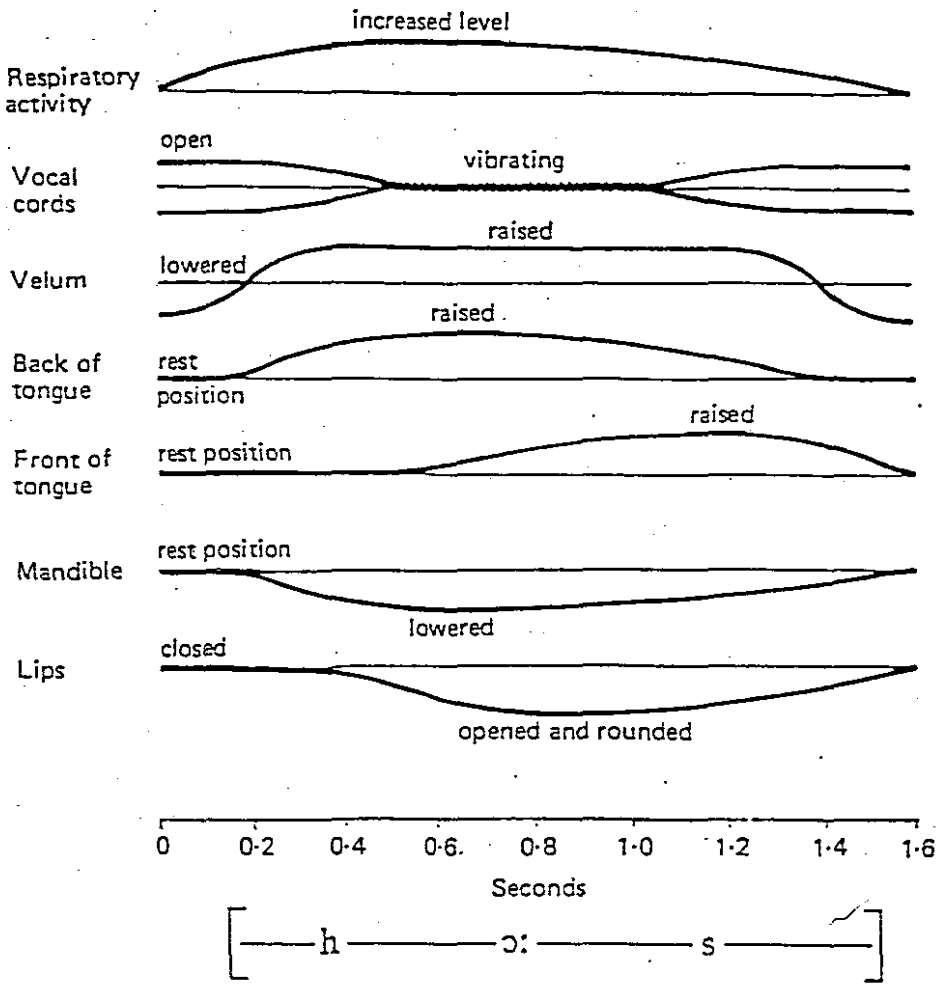
The expert system used in the knowledge engineering approach is trained with distinctive features found in acoustic speech patterns. Hence the pattern itself is not recognised, but

rather the features that make up the pattern. The spoken sentence is thus recognised by mapping the speech signal to a set of phonological units [30]. It involves a deep insight into the mechanics of speech. The phonological units refer to the phoneme sounds that combine to form intelligible speech.

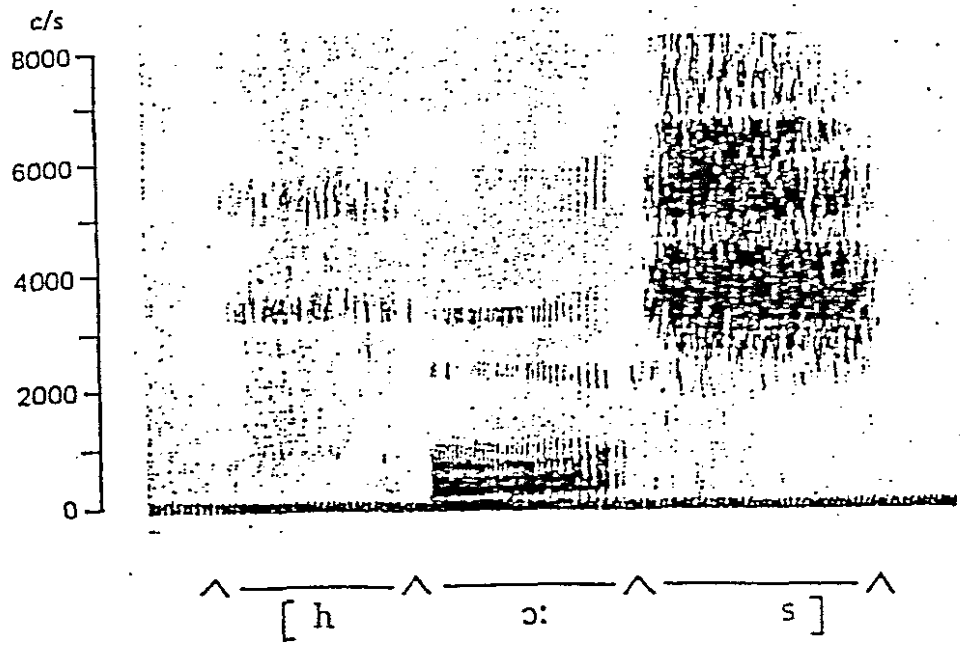
Principles involving the human vocal tract [6], constraints on the articulators and the grammatical structure [31][32] of the speech wave form [33] must be clearly understood. Speech articulators refer to the parts of the mouth and throat that affect the way speech is produced.(i.e. The tongue, teeth, lips etc.)

The knowledge engineering approach attempts to associate the various input signals with the actual position of the speech articulators. As each sound results in a different position of the articulators, the system is able to identify what was said. Fig 4.1[34] illustrates this principle by showing the relative position of the speech articulators when the word "horse" is pronounced . The way in which the movement of the articulators combine is called co-articulation. Notice how the tongue position changes during various stages of pronouncing the word. (Pronounce the word verbally and see if you can relate the movement of your tongue and lips to the diagram.)

Every phonetical sound can thus be represented by a combination of positions of the articulators. These positions are computed and represented in a mathematical model. This model is then used for training the expert system. The expert system would thus be trained with a vocabulary of phonetical sounds, associated with that language to be recognised. Rather than matching words, the knowledge engineering approach matches phonetical sounds. Adding the language constraints to the system now allows the system to thus build up an unlimited vocabulary.



Coarticulation.



Spectrogram of [hɔːs] *horse*
The arrowheads mark the transitions and the lines extend along the intervening segments

Fig 4.1

This is done in much the same way as a child uses various toy building blocks to make a boat or a car. Just as these blocks can only be joined in certain combinations so to will the speech sounds be joined in limited combinations. The extent of these combinations is dictated by the language that is to be recognised.

The Afrikaans language for example allows for only 64 000 possible phonetical combinations [27]. With these combinations the entire Afrikaans vocabulary can be created. Without the language constraints applied the possible combinations are approximately 7 million [27].

Whereas in one language certain combinations are not allowed, another language may allow these same combinations. These principles form the knowledge part of the system.

The engineering part of the approach is derived from the fact that engineering knowledge of the acoustics of the speech signal [35] are required. Knowledge of the spectral composition and spectral changes, are necessary in the training of computers to recognise speech patterns [36].

The engineering part of this approach thus requires that the system matches the spectral characteristics of speech with the respective movement of the articulators. As each sound results in a different position of the articulators, the recognition will occur at the phonetical level. Add to this the language constraints and the result is an unlimited vocabulary that can be recognised.

An attempt to mimic the knowledge-based approach in a computer speech recognition system includes an extraction module to deal with the automatic extraction of the discriminative acoustic features, and an expert or artificial intelligent system to deal with the linguistic interpretation of these features.

4.3. Advantages

The knowledge engineering approach is based on the way that visual characteristics are read from the spectral composition of speech. Combined with the knowledge of the articulatory dynamics and grammatical structure of speech, it is particularly suited for the recognition of fluent speech.

Fluent speech allows the speaker to use any combination of speech utterances such as words or syllables and to speak freely without having unnatural additional constraints such as silences between words.

The only constraint is that the speech should be from a fixed language. A fixed language refers to the language with which the expert system was trained.

In order to avoid recognition errors only one language will be trained into a system. Having several systems, trained with various languages, connected in parallel will allow the system to recognise these languages simultaneously.

4.4. Pattern Matching Approach

The second approach, namely Pattern matching, forms the basis of most commercial man-machine communication algorithms. The technique stems from the field of robotics (computer vision) and data communications. It also finds application in the fields of computer communications, image processing and computer interfaces.

Pattern matching basically consists of template matching in order to statistically find the nearest neighbour in a multi-dimensional feature space [37]. (This can be likened to matching different paint colours to get the closest match to the original.) The matching algorithms will try and find the shortest path through the space. This path will be

compared with existing prototypes already stored in memory. The closest match will thus be equivalent to the spoken speech. This approach provides solutions via pattern and / or statistical matching techniques such as neural networks and hidden Markov models [27][38][39].

In an automatic speech recognition application, the objective of the pattern matching algorithm is the following. To relate the speech input or pattern to a vocabulary of a fixed set of templates in order to find the closest match. It is essentially a mapping between speech and text so that each possible input wave form is identified with its corresponding text. Text here is used to mean the words, sentences or other linguistic units the speaker thinks of and verbalises in uttering speech.

Typical commercial systems use word templates. Such systems are characterised by speech constraints due to the limited number of word templates that are allowed in the vocabulary. In order to obtain the word units from the input speech in these systems, speakers are required to pause briefly after each word to facilitate automatic segmentation of the speech input. Hence most of the systems are described as discrete word recognition systems.

In general two types of pattern matching algorithms find application in the field of automatic speech recognition:

- **Dynamic Time Warping**

- **Parametric Feature Mapping**

4.4.1. Dynamic Time Warping

Dynamic time warping (DTW) in one of its various forms is at the heart of most commonly used decoding methods for speech recognition. The concepts of dynamic programming have been applied in the fields of data transmission [23][24] and digital error correction [40] for a number of decades. The application of dynamic programming methods to speech recognition [41] continues to be refined and enhanced at a rapid pace.

Most low cost (R1000 to R4000) commercially available recognisers use dynamic time warping. These systems address the problem of time alignment between a speech segment and the stored templates. This is achieved by using only digital time-domain samples as an input feature. The input speech can thus be "warped" in the time domain so that it matches the period of the stored templates. This technique is referred to as dynamic time warping.

4.4.2. Parametric Feature Mapping

The more complex Parametric Feature Mapping techniques such as hidden Markov models (HMM) [27] and neural networks on the other hand are used as feature based approaches. Stochastic digital signal processing techniques have contributed to this change in the pattern matching perspective. This technique allows for a statistical feature representation of the speech signal.

The hidden Markov model is basically a generalisation of the dynamic time warping algorithm, making it more versatile for speech recognition applications, but also more expensive (R5000 - R50 000).

4.5. Pattern Matching versus Knowledge Engineering

In order to get a better understanding of the differences between the knowledge engineering approach and the pattern matching approach the following example and illustration [42] will be examined.

Figure 4.2 illustrates the various levels into which speech can be divided. Starting from the bottom the following levels can be identified:

1. **The Acoustic level**

2. **The Phonetic level**

3. **The Phonemic level**

4. **The Orthographic level**

5. **The Lexical level**

6. **The Syntactic level**

7. **The Logical level**

This description will revolve around the first three levels as these levels are where the recognition process begins.

4.5.1. The Acoustic Level :

This is the level directly related to the speech signal. It consists of a graph with the X-axis representing time, the Y-axis representing the frequency spectrum and the Z-axis

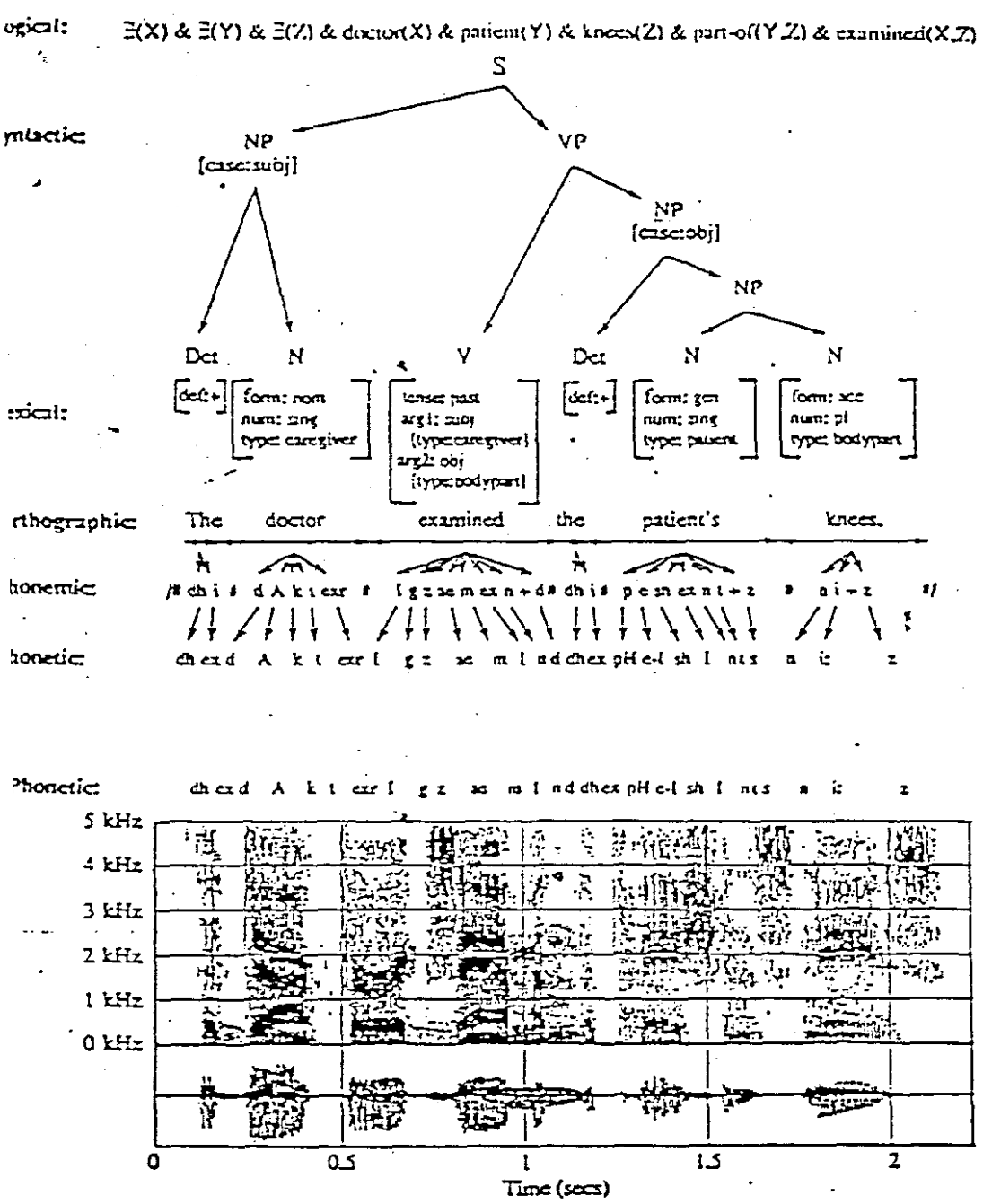

ogical:
∃(X) & ∃(Y) & ∃(Z) & doctor(X) & patient(Y) & knees(Z) & part-of(Y,Z) & examined(X,Z)
ntactic:
S
NP
[case:subj]
VP
NP
[case:obj]
NP
Det
N
V
Det
N
N
xical:
def:+
form: nom
num: sing
type: caregiver
tense: past
arg1: subj
{type:caregiver}
arg2: obj
{type:bodypart}
def:+
form: gen
num: sing
type: patient
form: acc
num: pl
type: bodypart
rthographic:
The doctor examined the patient's knees.
honemic:
honetic:
Phonetic:
5 kHz
4 kHz
3 kHz
2 kHz
1 kHz
0 kHz
0
0.5
1
1.5
2
Time (secs)

Fig 4.2

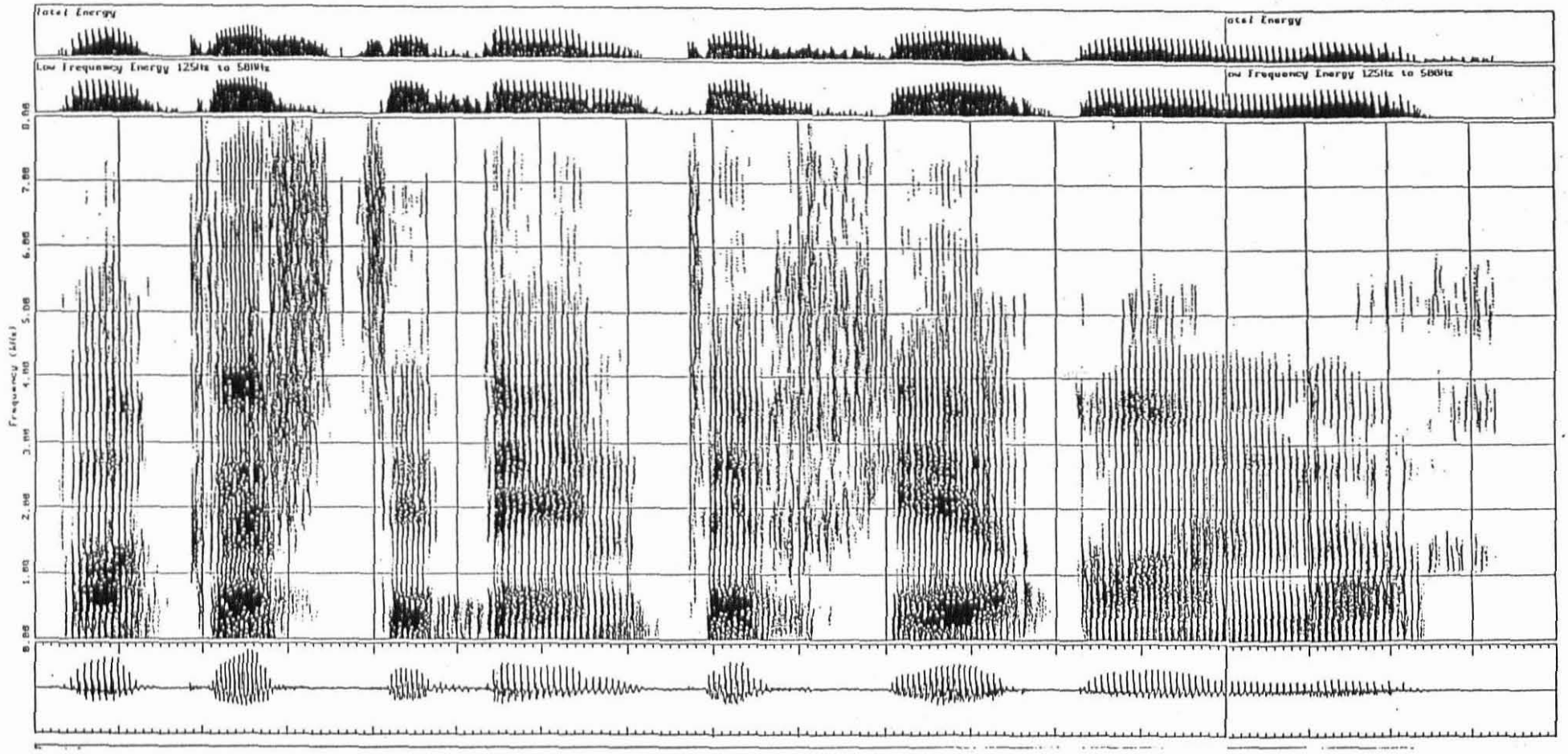

Total Energy
Low Frequency Energy 125Hz to 580Hz
Frequency (kHz)

Fig 4.3

representing the log-energy of the sound-wave. (The denser or darker the area the higher the acoustic energy in the signal.)

Figure 4.3 [23] graphically shows this spectrogram in more detail. The top graph details the energy levels present whereas the bottom graph shows the corresponding acoustical features. Notice that the denser areas represent a higher energy level.

Both the pattern matching approach and the knowledge engineering approach use this level to extract features from, in order to perform speech recognition.

4.5.2. Phonetic and Phonemic Level :

Although these two levels sound similar there is a vast difference between them. The phonetic level relates to the sound or movement of our speech articulators to produce a word whereas the phonemic level relates to the combination of the sounds or phones to form syllables or words.

4.5.3. Knowledge Engineering on these Levels

The knowledge engineering approach is said to be the ideal speech recognition system as it adapts to the speaker rather than the speaker adapting to the system. The knowledge engineering approach will start at the acoustic level. The system will extract various features from the acoustic level. The system will then use these features as an input to an expert system. This expert system will then decipher these features.

Depending on the language constraints that have been programmed into the system, the system will associate these input features with phonetic sounds that are characteristic to the relevant language.

The expert system will have an in-depth knowledge of the language that is to be recognised including the various language rules.

In general most languages can be represented by a fixed amount of phonetic sounds. (See Appendix F for International Phonetic Alphabet.) These phonetic sounds combine to form syllables and hence words are formed from these syllables. The phonetic sounds can only occur in certain combinations due to the constraints resulting from the movement of the speech articulators. The expert system thus attempts to reconstruct the input speech by using its knowledge of how our speech articulators work, as well as the language rules and constraints applicable to the language that is been recognised.

As this approach does not try and match pre-defined words, the vocabulary of a system using the knowledge engineering approach can be much larger than the one using the pattern matching approach.

4.5.4. Pattern Matching Approach

The pattern matching approach on the other hand will extract a feature vector from the acoustic level and try and match the feature with an existing template or prototype already stored in the speech vocabulary of the system. This matching is based purely on a comparison of the input features with an prototype from memory.

As only certain keywords have been trained into the system, only limited combinations of features relating to these keywords will be valid in the search path. The more keywords or prototypes trained into the system the greater the search path and also the longer the recognition time.

The user has to adapt his input speech in such a way that when the features are extracted from his speech, they will match or correspond to the prototype stored in the system.

Any words entered that are not already in the vocabulary will be rejected if no reasonable match is found. Depending on the quality of the system false recognition will occur if the input speech is very close to a stored prototype. The language is built into the stored vocabulary of words. Thus with the pattern matching approach one has larger constraints placed on the user. With the knowledge engineering approach the system will adapt to the user.

4.5.5. The Orthographic Level

This is the level which most recognition systems will output. It can be likened to the text equivalent of the spoken input to the system. The closer the system approaches this level the more accurate the system.

The recognition system can be configured so that this level is output to the user application. This would normally be the case if the system was used primarily as speech to text converter. If the system was configured for a specific user application then the output would correspond to the programmed instruction that is linked to this level.

4.6. Further Differences

Other than the aforementioned main differences there are further constraints that contribute to the differences between the two approaches. The following salient points will thus be considered:

4.6.1. Financial Constraints:

The knowledge engineering approaches has received less attention than the pattern matching approaches. One reason for this tendency is that the knowledge engineering approach has to date achieved less impressive practically implemented results [43] .

Virtually no commercial releases of speech recognition systems based on the knowledge engineering approach exist.

The main reason for the lack of interest in the knowledge engineering approach is that this type of research is expensive. This can clearly be seen by reviewing the explanation of the knowledge engineering approach. The person designing the system must have an expert knowledge of the language as well as the technology used to achieve this goal. In order to get a reliable expert system, expert knowledge is required and hence the financial burden of the design increases [43].

Pattern matching on the other hand has been much more successful due to the fact that researchers require only limited speech knowledge, some statistical knowledge and large amounts of computing power and memory [2]. Pattern matching is also used in a number of other applications such as fingerprint identification, voiceprint, voice mail etc. This factor further stimulates the commercialisation of this technology .

4.6.2. Segmentation Problem:

For both speech synthesis and recognition the input is divided up, typically into segments of some linguistic relevance, for efficient processing. In the case of speech recognition, the speech signal serves as the input which has to be divided into phonetic, syllable or word boundaries. The speech signal, however, gives only slight (hidden) indications of these boundaries which complicates the automatic segmentation process.

At present this is one of the most difficult problems in automatic speech recognition and is one of the main reasons for currently using pattern matching on discrete word or sub word segments. In order to obtain the word units from the input speech in these speech

recognition systems, speakers are required to pause briefly after each word to facilitate automatic segmentation of the speech input.

4.6.3. Recognition Problem :

In a pattern-matching approach, the objective is to relate the speech input to a vocabulary of templates in order to find the closest match. Typical mapping techniques used in the matching process are:

- **Hidden Markov Models (HMM)** [27][38],
- **Recurrent and Feed Forward Neural Networks (NN)** [39],
- **Bayes Classifiers** [37], **or**
- **One of the Dynamic Programming Techniques** [40][41].

The optimisation of the mapping search is a major topic in speech recognition research [2]. One simplification is to restrict the vocabulary a speaker may use, which limits the memory to be searched.

Knowledge engineering takes the recognition problem to another level by using expert engineering knowledge to extract speech features from the input speech and then using these features to recognise the speech.

CHAPTER 5: CONSIDERATIONS FOR CHOOSING A PRACTICAL SYSTEM

In this chapter various considerations will be highlighted which will assist in the decision to select an appropriate recognition system. These considerations will be discussed under the following headings:

- **COST**
- **RECOGNITION MODEL**
- **VOCABULARY**
- **PERFORMANCE EVALUATION**
- **SPEECH STYLE**
- **SPEAKER DEPENDENCY**
- **ROBUSTNESS**
- **LANGUAGE**
- **GENERAL**

5.1. Cost

In comparing the cost of automatic speech recognition systems, the first factor that should be considered is the type of system under investigation. The type of constraints that would be tolerated are also important. For example, most speaker-dependent systems that use

dynamic time warping recognition techniques simply utilise the time domain input in template matching and therefore require no pre-processing.

In contrast, speaker-independent parametric pattern matching approaches use computational intensive pre-processors for parametric feature extraction in order to provide more accurate and reliable systems. Since pre-processing has to be done in real time, the financial burden on the system increases. DSP and/or transputer cards may now be required.

Another factor to be considered is whether the implementation of a speech recognition system will be cost effective or not. (i.e. Will the, often marginal, increase in productivity or improvement in performance, warrant the purchase of a costly speech recognition system.)

In some instances it might be more advantageous to wait for the technology to improve or reach a plateau in quality before purchasing such a system.

5.2. Recognition Model

Of all the system features, the type of recognition model used in the system is the most difficult to identify. The reason for this is simply that this system feature is the unique part of the system, which the developers treat as their secret invention.

The recognition model is in most cases not mentioned in the marketing message of the system. As this is not always public knowledge the user will not really be able to compare systems at this level.

The one guideline that can be given is that the more complex the recognition model the more expensive the system. This does not necessarily mean that the more expensive

systems will perform better and so the user should compare the systems at other levels as described in this chapter.

5.3. Vocabulary

The training mechanism of the speech recognition system establishes a reference memory or dictionary of speech patterns, which are assigned to text labels. In speaker-independent systems, training is performed during system development using large speech databases and often combine manual and automatic methods.

Most low-cost, commercial, speaker-dependent speech recognition systems are trained by customers to suit their own needs at the expense of being able to recognise only one or two speakers.

The most commonly used vocabularies are the digit vocabulary consisting of the first ten digits (zero, one, two,..., nine) and the phonetically A-Z vocabulary which has 26 letters (alpha, bravo,..., Zulu).

Commercial systems typically use word templates. Such systems are characterised by speech constraints due to the limited number of word templates that are allowed in the vocabulary. Typical vocabularies consist of a set of 100-500 words. The size of the vocabulary is one of the primary considerations in the evaluation of a speech recognition system.

Smaller speech segments such as syllables, tri-syllables or demi-syllables, instead of using word segments have been used. This to a large extent relieves the limited vocabulary constraint (typically allows 2000 word vocabulary) with the trade-off of resultant poorer recognition rates.

One reason for poorer recognition rates is that syllable segments are typically much shorter than words, which from a pattern matching point of view complicates the matching process. Shorter speech patterns generally provide less discriminating power. (Refer to chapter 3 under the heading "Speech input features")

It is also noted that the poorer recognition rate in syllable based systems is as a result of neighbouring speech segments influencing one another, hence causing similar linguistic templates to have different acoustic patterns.

When selecting a practical system the user must realise that if considerable time is involved in training a system with a new vocabulary the cost of the system must be adjusted accordingly. Typically systems with a built in vocabulary are more expensive because of the costs incurred by the manufacturer in training the system. These systems however require only adaptation by the user and as such will present a saving to the user where extensive training was previously required.

So although the initial cost is greater, where a large number of users are going to utilise the system it might be more cost effective to select a system with a pre-trained database of keywords.

In a user trained system if the user requires a different vocabulary the user does the training. Considerable time is usually required to train the system. At the same time the user will need to have advanced knowledge of the system to perform this operation.

With systems where a database of words is supplied, the user can negotiate the training of a new vocabulary into the purchase price of the system. This will obviously require that time be spent deciding which words will be required for the new vocabulary, before having them trained, as any changes that need to be made at a later stage could be expensive

involving the manufacturer for such training. Thus if the user is going to continually change his vocabulary it might be more cost effective to opt for the user trained system.

5.4. Performance Evaluation

In most speech recognition applications, an a-priori decision has to be made as to the specific language which has to be recognised. The reason for this is that most speech recognition algorithms are language specific. As such they are trained on a fixed vocabulary of language dependent words.

Most automatic speech recognition systems quote accuracy of recognition in order to demonstrate how well they perform. These tests are usually subjective as most manufacturer will not advertise their systems weaknesses. These tests therefore, must not be used as the sole determining factor in evaluating performance.

The above measures, however, are quite variable across different applications. The performance of the system therefore depends on constraints such as :

- The recording environment (head-mounted, noise cancelling microphone in a quiet room vs. telephone speech with inherent noise).

- Size of the vocabulary of words the system can accept.

- Whether the system is speaker-trained or speaker-independent.

- Type of speech such as isolated-word or connected-speech.

Because the performance is dependent on the choice of the vocabulary, several word sets are commonly used to enable comparison of systems. The digit vocabulary uses the first

ten digits (zero, one, two,..., nine), while the phonetical alphabet vocabulary has 26 letters (alpha, bravo,..., zulu).

Several alternative performance measures have been used to directly account for vocabulary complexity. Some employ statistics of involving human speech perception while others simply measure the error rates. Care should therefore be taken as to the type of performance measure used where different speech recognition systems are to be compared. This study will evaluate the accuracy of the system as well as the speaker independency. The results will be based on the error rate of the system. The results of these tests are detailed in Chapter 7.

5.5. Speech Style

The segmentation problem, previously mentioned in the recognition model, can be partially overcome. This is achieved by the user compensating in his style of speech. Using the speech style to simplify the segmentation problem is currently the most widespread constraint enforced on the user.

In automatic speech recognition, three types of speech can be distinguished:

5.5.1. Continuous Speech Recognition :

This principle allows natural conversational speech with little or no adaptation of speaking style imposed on system users. Continuous speech allows most rapid input (50-150 words/min)[2] but is the most difficult to recognise.

No commercial systems have been developed that accept totally natural and continuous speech yet.

5.5.2. Isolated-word Recognition :

This method requires the speaker to pause for at least 150 ms after each word for segmentation purposes. This is unnatural for speakers and slows down the rate at which the speech can be input (40-75 words/min)[2].

5.5.3. Discrete Utterance Recognition :

Using this technique, a compromise between the two extremes described above is achieved. The speaker need not pause but must pronounce and stress each word clearly.

The vast majority of commercial speech recognition systems are based on isolated-word recognition and hence require that the speakers pause briefly after each word to facilitate segmentation.

The user must therefore determine which category his requirements fit into. Once this has been identified he will be able to narrow down his field of choices considerably for the best system.

5.6. Speaker Dependency

The differences between male and female speech, for example, are so profound that a number of systems choose to discriminate between these two types of voices. The systems typically use different recognition models for male and female speech. One of the reasons for this is because the pitch of a female voice is often higher than that of a male. This requires that the system adapt to the different frequency bands when applying DSP techniques in attempting recognition.

The user therefore has to determine whether the majority of the end users will be male or female so as to best choose a system to meets his needs.

Two terms are used to describe the speaker dependency feature in automatic speech recognition systems:

- **Speaker-Dependent Systems**
- **Speaker-Independent Systems**

5.6.1. Speaker-Dependent Systems :

Typically used in low-cost systems where the system will only recognise the voices of the one or two speakers who participated in the training of the memory templates. These have application in systems such as access control, where recognition is based on voiceprint.

5.6.2. Speaker-Independent Systems :

Typically used in high-cost speech recognition systems where voices from a wide variety of speakers can be accepted. These systems are usually trained using large speech databases, containing pre-recorded speech from a large number of speakers. Typical applications include public access to telephone enquiries, voice activated dialling etc. ("087" Number services employ systems such as these.)

The main difference between speaker-independent and speaker-dependent systems is that a large population of speakers is used in the training of the pre-defined memory templates.

Most commercial systems are speaker-dependent, demonstrating good performance only for speakers who have previously trained the system. These systems adapt to new users by requiring them to enter their speech patterns into the recogniser's memory. Every system will have its own method of training depending on the recognition model.

Since memory and training time in speaker-dependent systems grow linearly with the number of speakers, less accurate speaker-dependent recognisers are useful if a large population must be served.

The user must therefore clearly define how and who will use the system. By spending some time reviewing these needs the user will be able to identify which system best suits his needs.

5.7. Robustness

In order to understand the importance of considering this aspect of the recognition system, the term robustness will first be explained.

The robustness of the system, with respect to the error rate, refers to the ability of the system to cater for aspects of the input speech signal that reflect the recording environment. Variations in the manner of speaking, the noise and channel characteristics all affect the system robustness [44].

To a certain extent, the signal can be cleaned up or normalised in a pre-processing stage prior to parameterisation. Thus extraneous factors that may distort the automatic speech recognition process are eliminated.

If the environmental conditions are stationary and the variations can be determined, such effects can be removed from the signal. The simplest normalisation adjusts maximum signal amplitude to a standard level to account for variations in recording level, distance from the microphone, original speech intensity and loss in transmission. Automatic gain control, as employed in radio receivers, may be viewed as a simple algorithm to adjust amplitude variations.

The user must therefore determine in which environment the system will be utilised. The following factors must be considered :

- What will the level of environmental noise be where the system is to be implemented?

- Does the user have to purchase special noise cancellation microphones to eliminate false recognition?

- Will he need to incorporate special partitions or put people into separate offices in order to reduce office noise which can affect the system robustness?

Consideration of these factors is important. It would be pointless purchasing a cheaper system, that is not as robust, only to find that one now has to accommodate every user in a separate office because the noise level in an open plan office renders the recognition system useless.

5.8. Language

The language in which any commercial speech recognition system is trained is usually fixed. This means that to retrain the system for any other language is usually not very easy. Speech recognition systems based on the dynamic time warping approach are the only type of systems that are not language dependent. This is because these systems match patterns irrespective of the language. This is at disadvantage of being speaker-dependent mostly accepting only isolated words.

On the other hand, speech recognition systems based on the statistical feature mapping approach have a pre-defined structure for specific words. The input vocabulary thus consists of a fixed set of words and can only be retrained by the developers of the system.

In recent years, the development of technology in the parallel processing field has led to the use of parallel language systems in which the size of the vocabulary is extended to include words from a wide variety of languages. This technology is also used for language translation on international communication channels.

As most commercial systems have not been developed with the various South African languages in mind it is vital that the user considers this factor. An important aspect to be considered is that experience has shown that most systems developed in foreign countries do not recognise Afrikaans and even in some cases the South African English dialects. Locally produced cards that cater for the native South African tongues therefore are not affected by this problem.

5.9. General

Under this heading, general features of speech recognition systems that need to be considered will be described. Such features include factors like the type of hardware required, the number of input channels and system user friendliness.

These features are just as important as those described in the previous sections. It would be worthless spending all the time on choosing the right system only to find that the hardware where the system is to be installed cannot support the system. The user must therefore consider the following :

5.9.1 Hardware

The computer system that is required. Is a IBM (286 AT) or 386 required for the system to operate. (To be noted here is that although the system might be able to operate on a 286 machine, its recognition rate might no longer be in real time.)

The amount of random access memory (RAM) that is required by the device driver should also be examined in order to establish if the user can still run other application at the same time.

5.9.2. Social Impact

Another factor that must be considered is what will the social impact of the system be. Although the field of Social Psychology relating to the work environment is not the trained field of the author it is important that comments relating to this area are considered. It is the feeling of the author that in introducing a system, that so greatly changes the manner in which a person works, there must be some implications.

Although these factors represent a study in themselves the user would be wise to consider them before choosing a speech recognition system as the solution to his problem.

CHAPTER 6: GIS AND SPEECH RECOGNITION

The previous chapter described numerous guidelines for choosing the right speech recognition system for a given problem. In this section these guide lines will be applied to solve a problem relating to the capturing of numeric data co-ordinates for use in digital mapping within Telkom.

It would however be fitting to first introduce the existing system and then analyse where the speech interface can be applied. A commercial system meeting these guides will be scrutinised under the guidelines chosen to fulfil this interface.

6.1. Discussion of GIS

6.1.1. Introduction to GIS in Telkom

Geographical Information Systems (GIS) is a division of Telkom that is at present capturing details of topo-cadastral maps for internal use by the Telkom planning section. Topo-cadastral maps are maps that supply information relative to the registering and geographical layout of property in both urban and rural areas.

The information contained on maps of this nature can vary from erf numbers and street names to information regarding the actual location of the erf relative to a fixed geographical location. Telkom is using the features of GIS that allow information regarding the size of cables, as well as the location, cable type and route number etc. to be linked to these maps.

Due to the large amount of data capturing involved in the implementation of GIS it was essential to find a more efficient and practical way of capturing data. One of the methods sought out to achieve this efficiency was the implementation of speech recognition.

6.2. Where and how Speech can be Implemented

In order to best illustrate how the speech interface could be practically implemented a detailed description of the basic capturing software is provided. A step by step guide will first be given. Thereafter those steps where the speech recognition interface could be applied will be described in greater detail.

6.2.1 Steps to be taken for Data Capturing and Manipulation of Township Plans, Loose Erven and Farm Map

1) Check in the capture record program to ensure that job has not been previously captured

2) If job has not been previously captured fill in new information and obtain a line call number

3) Flag job using a dedicated number for every co-ordinated point, any intersection of lines and any change in direction

4) Make a photocopy of the flagged original for working purposes

5) Enter infodata program on PC

6) Capture co-ordinated points (data-base file)

7) Capture Co-ordinated points (verify file)

8) Make a comparison report of the data-base file and verify file

9) Print error report

10) Fix any errors on both files

11) Make comparison report again to ensure errors have been corrected

12) Process data using Infodata

13) Calculate traverses in Infodata using two given points and distances or using one given point with angles of direction and distances

14) Using Infodata make an output file (ASCII file)

15) Exit from Infodata

16) Enter mapper program on PC

17) Load points from output file

18) Enter required level and linestyle

i) Level 11 Linestyle 20 Stand lines

ii) Level 13 Linestyle 18 Block lines

iii) Level 18 Linestyle 15,16,50 Township lines

iv) level 25 Linestyle 38,39 Servitude lines

19) Join lines to captured points according to information on TP, erf map, farm map or servitude's

20) Save information at regular intervals and at the end of the job

21) Process data captured into new Informap drawing using a VMS program

6.2.2 Areas for Implementation

The description next presented will highlight those areas where the speech interface could be implemented. The description will start at step number 6.

Step 6.) This step involves doing the following:

After going through the preliminary set-up, the procedure for entering information is repetitive with various "x" and "y" co-ordinates being entered and processed. These co-ordinates correspond to actual surveyed reference points. These reference points are usually located at positions where the erven change direction (See fig C.1) Once all the co-ordinates have been entered a database file is created. The same procedure is then repeated to create a verification file. These files are then compared and a report is generated. If necessary any errors are corrected and another report is generated to ensure that errors have been corrected.

This process is repeated until all errors are removed in the data. The data is then processed using the software application and the final file is generated.

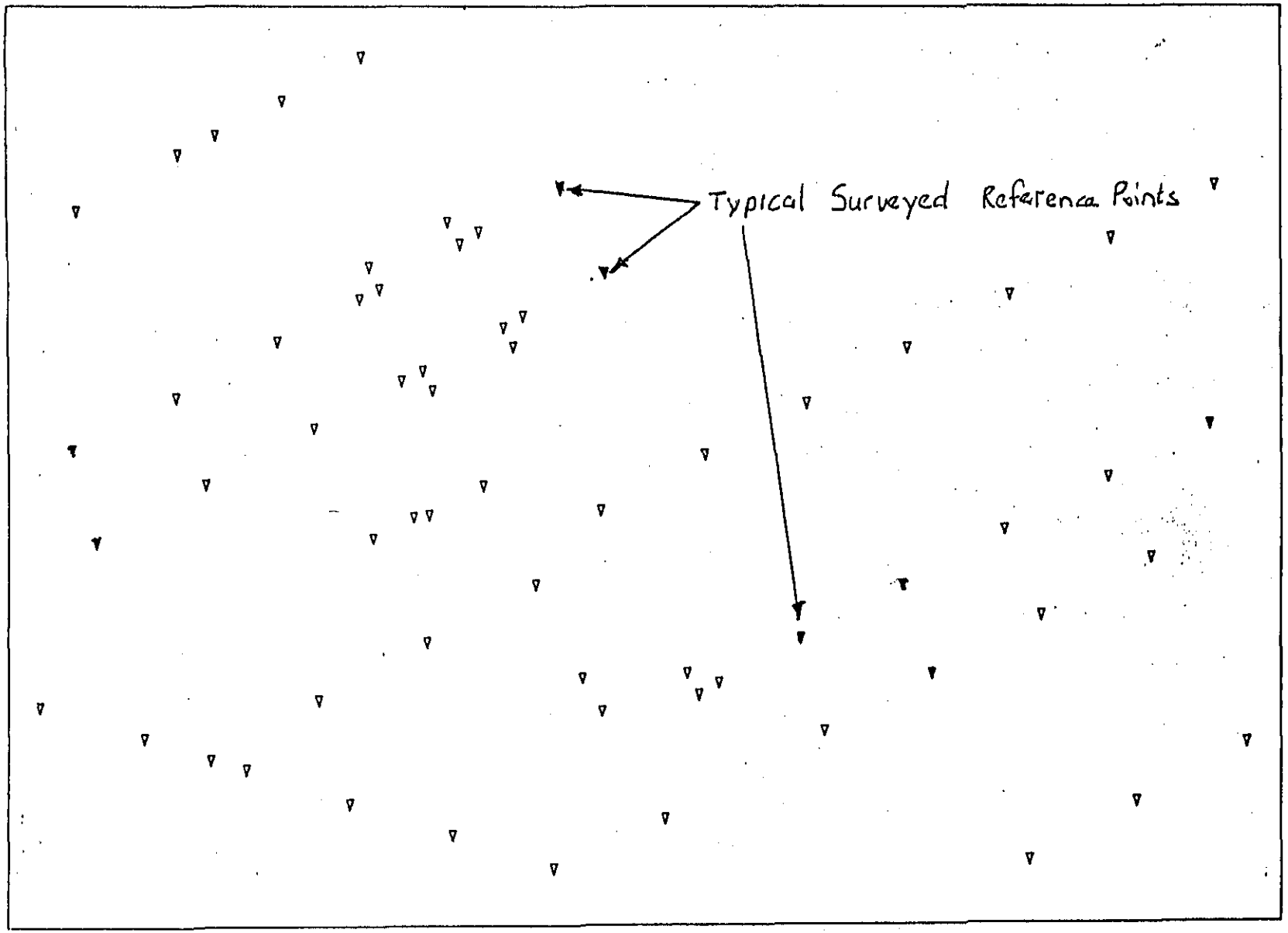


Fig C.1

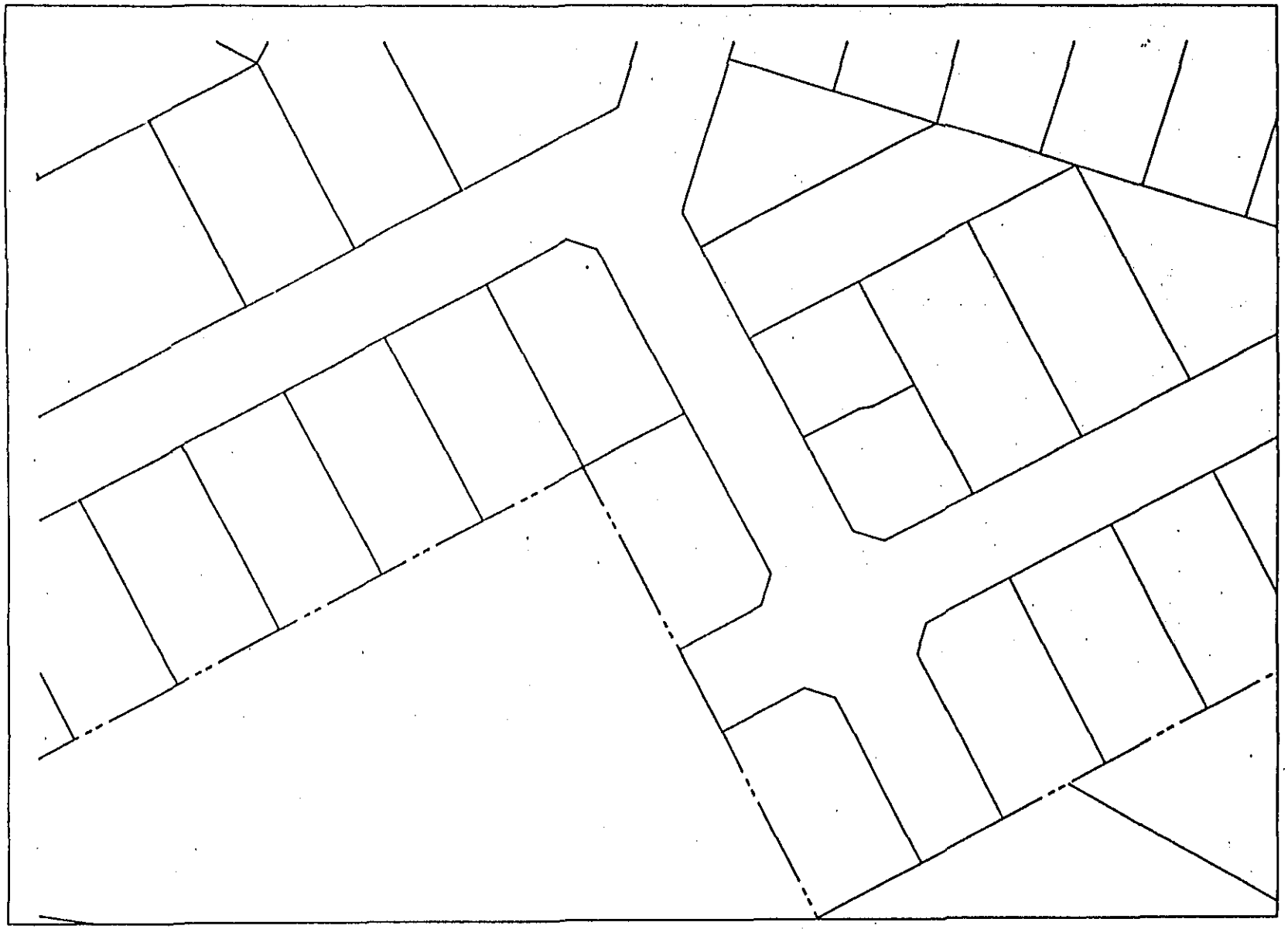


Fig C.2

The voice interface can be fully implemented in this section of the procedure. As the main form of input is via the keypad the only keywords required would be the numerical key keywords (i.e. Zero, One etc.)

The next step in the capturing cycle involves the capturing of traverses. These points are located between two surveyed reference points. They are usually situated at the corners of each erven. (See fig C.3) The method of capturing these points can vary although, in every method the voice interface can be utilised completely.

Once these points are all captured an output ASCII file is created which can be linked to the Mapper software package. The voice interface is also utilised in the Mapper software application. The steps that follow describe the correct procedure for mapping the co-ordinates. After loading the output ASCII file created by Infodata, the co-ordinate points are joined. (See fig C.2)

As this information is stored in different formats for each scale of drawing the capturer has to change to various linestyles or levels of input for each new facet of the drawing. This requires that various groups of keystroke combinations are performed during the capturing process. This is the type of operation where macro functions can be properly utilised. Since the software package does not cater for macro generation at present, the only means to get macro functions is from an external source. The cost of using a different package makes this option also impractical. Hence the speech interface software can be of assistance. Since the speech interface software can be adapted for Macros, it is ideally suited to cater for this software shortfall.

By using one keyword the capturer will for example, be able to change the level and line style. As the voice interface can be adapted without exiting the application, the capturer can cater for any repetitious task by just changing the keyword selection.

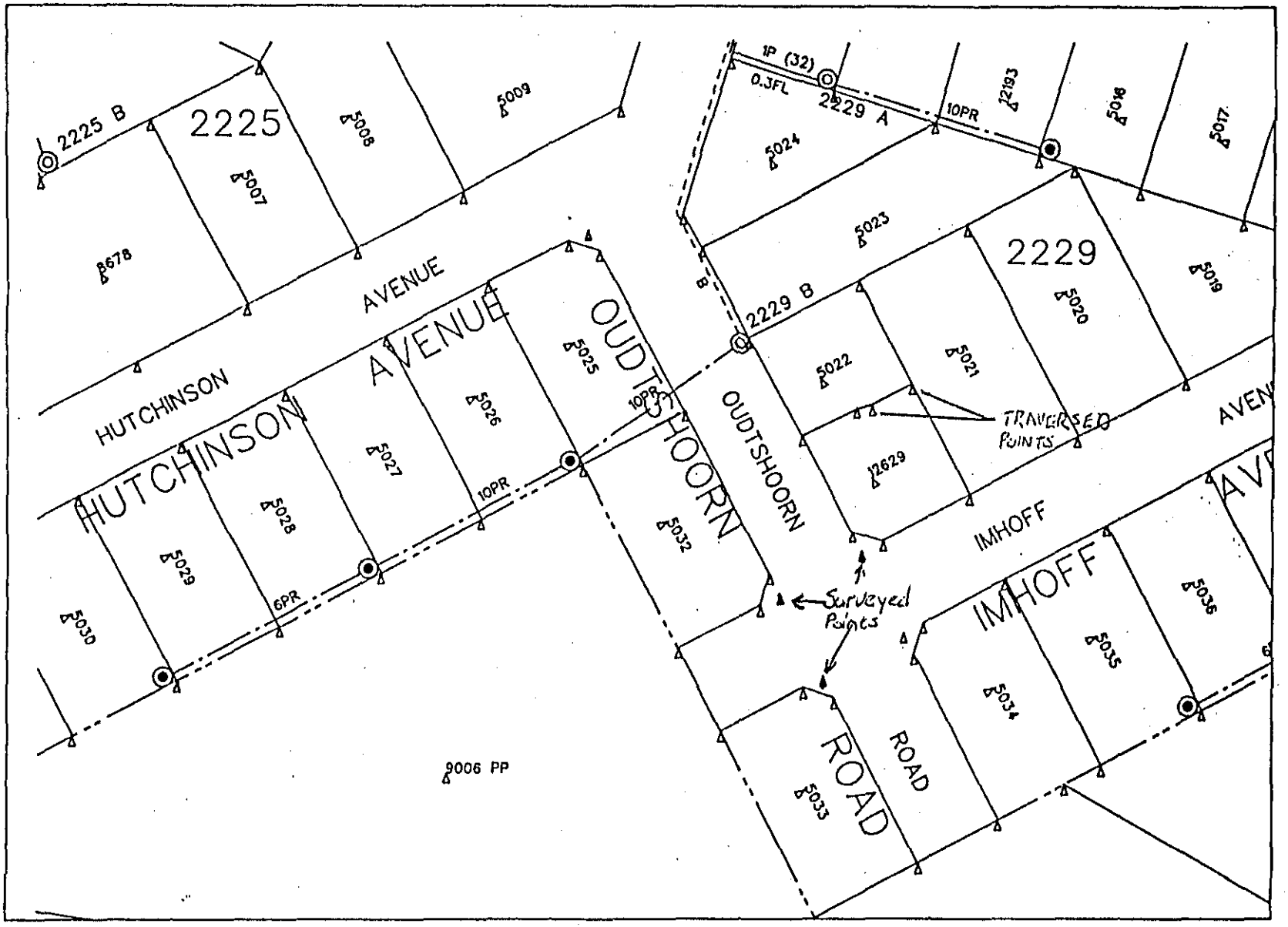

2225 B
2225
5007
5008
5009
8678
AVENUE
HUTCHINSON
HUTCHINSON AVENUE
5025
OUDTSHOORN
10PR
5026
5027
10PR
5028
5029
5030
6PR
5032
OUDTSHOORN
9006 PP
1P (32)
0.3FL
2229 A
10PR
12193
5018
5017
5024
5023
2229
5020
5019
2229 B
5022
5021
12629
TRAVERSED POINTS
IMHOFF
Surveyed Points
IMHOFF
5036
5035
5034
ROAD
ROAD
5033

Fig C.3

Once all the relevant lines have been joined the text must be entered onto the drawing. (See fig C.3). This also requires that certain font styles and types be selected. The macro facility of the speech interface can also be utilised in this section of the work.

The final operations that are performed involve the processing of the information into the "Informap" database using a VMS program.

Since the Mapper software package does not cater for macro generation the only means to get macro functions is from an external source. The chosen system would therefore have to provide this macro facility.

By using one keyword the capturer must be able to, for example, change the level and line style of a particular section of the drawing. If the speech recognition system can be adapted without exiting the application, the capturer can cater for any repetitious task by just redefining the keyword selection.

Once all the relevant information has been captured the text has to be entered onto the drawing. This also requires that certain font styles and types be selected. The system should also be able to cater for the changing of these line styles.

6.3. Problems Expected

When originally considering whether the speech interface would be able to improve the rate of data capture, certain possible areas requiring investigation were identified. These problems related to the following aspects.

- Would the system be able to easily recognise both male and female voices?

- It was not known how the card would perform if a person has a cold or other ailment which can affect his voice?

- As the application is for a South African environment will the card be able to recognise the key words if they are spoken with a Afrikaans accent? Other languages used in South Africa also needed to be considered.

- Could the system be trained for Afrikaans or other South African tongues?

- Other factors that played a major role in the decision to implement the speech interface were whether the card would be socially accepted by the end user. i.e Will the user be adept to the changes in the working environment as a result of the system being installed.

- Will the end users use the system if implemented?

- Will the working environment have to be changed to cater for the speech interface? (i.e. Partitioning screens to block off noise.)

- Is the technology advanced enough to inspire user confidence in using the speech interface?

- Would a speech recognition system provide any real performance benefit.

With these possible problems in mind the next section discusses which system was chosen and also explains the reasons for the choice.

6.4. Speech System Chosen to Test.

The speech recognition system that was chosen to be tested was designed by a local Electronics company called Datafusion Systems. The reasons for choosing this system, as well as certain features of the system, are outlined below. The reasons will be discussed under the same subheadings as described in chapter 5. The same subheadings have been chosen to show how the guidelines were practically applied.

6.4.1. Cost

The cost of the system was not finalised as the system was still under development. Early indications show however that the price would be compare favourably with existing cards that are commercially available. As the system would require that the recognition be continuous speech (see Section 6.4.5.), it had to be accepted that by implementing this type of system the initial costs would be fairly high.

Whether the system would be cost effective or not, was not examined as this would only be determined once the system had proven that it could be introduced. The preliminary investigation was not aimed at the costing of the card, rather it focused on the functionality and practicality of the card.

6.4.2. Recognition Model

The suppliers of the recognition system requested that the method employed in the recognising of speech not be revealed. The fact that the system is able to recognise continuous speech would however indicate that the system employs some form of hidden Markov model technique.

6.4.3. Vocabulary

The system was provided with a standard vocabulary consisting of the numbers, nul to nege (Afrikaans) as well as the Alpha, Bravo alphabet. An additional two words, Ster and Hekkie (also Afrikaans) was included. These pre-trained words were not ideally suited for the Telkom application and so it was realised that a new vocabulary would have to be trained into the system.

Although the method of training the system is not reviewed in this document a preliminary investigation was undertaken to determine the selection of the new vocabulary to suit the Digital mapping environment as well as a generalised vocabulary for Computer aided design (CAD) applications.

The selection of this list involved the conducting of a survey amongst a group of data captures to determine which were the most commonly used keywords in the existing application. (The Questionnaire for the survey is found in the Appendix note B.) The designers of the card were also approached to for advice in setting up the list of keywords.

Consideration was also given to the words that are common to CAD applications. A finalised list of words was drawn up consisting of words relating to the existing application as well as possible keywords applicable to CAD applications. The list is included in the Appendix B. The list was compiled using a combination of the results from the survey as well as consulting with the system manufacturer. Typical CAD function keywords also examined. This list was not trained into the system as the costs of such training were not yet justified.

As explained in chapter 3 under the heading "speech input features" the selection of the words could play a major role in the final stability of the system as certain words that are

too acoustically close can cause false recognition to occur. Limiting the list to as few words as possible would also help to improve the performance of the system.

6.4.4. Performance Evaluation

This facet of the systems performance is not definable until tests have been done. Although most manufacturers claim to have obtained various recognition accuracies, these tend to be subjective. The user must not accept these figures as final, rather a series of tests should be set up to confirm these figures. The next chapter will discuss the tests that were decided upon, along with the results obtained.

6.4.5. Speech Style

As the system was to be employed in a data capturing environment the system chosen had to cater for continuous speech. One of the main reasons for looking at the speech interface was to improve in the method used to capture data and hence slowing speech down to discrete utterances would be defeating the purpose. Tests done on the card chosen showed that it could handle continuous speech.

6.4.6. Speaker Dependency

As numerous data capturers would eventually have to use the system, it was decided that minimal time should be spent on training the system. Thus the chosen system had to meet the requirement of being speaker independent with as little training or adaptation as possible. The next chapter discusses tests that were done to determine whether the system met this requirement.

6.4.7. Robustness

The system was to be implemented in a drawing office environment. The card would therefore have to be able to handle the noise that is usually generated in a large open plan office. As the robustness of the card could also easily be improved by using a noise cancelling microphone consideration would have to be given as to the increased expense that would be incurred.

The chosen system does posses the ability to adjust its sensitivity to noise but at the expense of recognition and speaker independency. A good balance would therefore have to be reached when the system was implemented.

6.4.8. Language

As the applications requirements are such that the chosen system has to cater for speaker independency, it was necessary to choose a system that had a built-in database of keywords. Obviously these keywords would have to be trained into the system.

The language in which this training would be done has to be suited for South African standards and hence the chosen system should cater for South African English and Afrikaans as these are the languages most commonly used in the environment where the system is to be implemented.

The system chosen to be tested has a great advantage over other systems in this regard as it is locally manufactured and hence has already been adapted for South African conditions.

6.4.9. General

The chosen system has a number of favourable features which can be summarised as follows:

The ability to disable keywords individually. This can be very advantageous as the user can thus disable any keyword that is "miss-firing" due to current environmental noise being recognised as speech. This would tend to occur with keywords that contain a lot of high frequency speech components which can be wrongly interpreted due to noise.

The system also allows for speaker adaptation. This feature not only makes the system more speaker independent, but also allows the user to optimise the recognition performance for his speech under changing conditions. (e.g. If the user finds that one of the keywords is not being accurately recognised he can dynamically adapt or "retrain" the word and then continue with his application.)

The system also provides for redefinition of keywords. This is also very useful in applications where the user constantly performs various repetitious tasks. The user is thus able to define any keyword to represent a certain number of keystrokes and when these become invalid, the user can redefine the keyword to represent a different set of key strokes.

The recognition system card has its own on board RAM and processor and hence the system does not "tax" the user's personal computer(PC). Recognition occurs on-board and in real time. This means that the processing power of the user's PC is not sacrificed in order to accommodate speech recognition.

CHAPTER 7 : TESTS ON VOICE CARD

Tests that were performed using the speech recognition system are outlined in this chapter. At the outset it must be noted that all of these tests cannot be used as a standard for generalised testing of various speech recognition systems. Every application in which the speech recognition system is to be implemented will require that a different set of tests be decided upon.

Four main tests were performed, each serving to test a different feature of the recognition system. The objectives of each of these tests and methods followed are highlighted below.

It was envisaged that on the basis of these tests, a reliable indication could be obtained as to how well the speech recognition system operates. Although none of these tests were done using the existing data capturing software, the methods and principles tested are the same as those that exist under the GIS software environment. By not using the data capturing software the cards performance could be evaluated on its own. Before disscussing each of the tests a general disscussion will be given relating to the accuaracy versus throughput.

7.1 General discussion:

It is important to understand the terms accuracy and throughput are their relationship to each other in evaluating a system. The throughput is directly related to the stability of the recognition system.

To best understand throughput a comparison will be given between the formula for each of these terms.

The formula for calculating accuracy can be given as :

$$\frac{\text{Total entered elements - total corrupted elements}}{\text{Total entered elements}} \times 100\% = \%\ \text{Accuracy}$$

The equation for calculating the throughput is given as:

$$\frac{\text{Tot. entered sequences- Tot. corrupted sequences}}{\text{Total entered sequences}} \times 100\% = \%\text{Throughput}$$

Whereas the accuracy of a system is determined directly by the number of correctly recognised keywords the throughput is dependent on the amount of keywords which are combined to form a sequence.

To illustrate this the following example will be examined.

In a data capturing environment an individual has to input 10 digit numbers repetitively. If the recognition system has an accuracy of *90%*, only one error should occur for every 10 digits entered. This accuracy can be acceptable, but in the worst case scenario one error will occur with every 10 digit sequence entered. Although the accuracy would still be *90%* the throughput or number of correct sequences entered would approach *0%*. If however the numbers consisted of 5 digits only, the throughput would approach *50%*.

This subject will be disscussed further when comparing the results of the tests decsribed below:

7.2 Test 1

Objective : To establish the accuracy of the recognition system.

Discussion: This test was done using only one speaker. This would eliminate any possible errors due to speaker adaptation. The results obtained would therefore be directly related to the accuracy of the recognition system and not its ability to adapt to different speakers. Although this test would tend to be subjective, because it only uses a fixed combination of keywords, the results would represent a good indication of the system's performance.

Method : The chosen system has 38 keywords, comprising 26 letters (Alpha, Bravo ..... Zulu) 10 numbers (Nul, Een .... Nege) as well as 2 other words ("Ster" and "Hekkie".) The system was configured such that each alphabetic character was represented by its respective keyword.(i.e. "Romeo" = 'R') The numeric keywords represented the respective numerical values (i.e. "Sewe" = '7') The last two keywords were configured such that "Ster" = 'SPACEBAR' and "Hekkie" = 'CARRIAGE RETURN'.

In order to get a generalised even usage of the key words the Standard ASCII 'Fox' message was used. This message allows for the use of every keyword within one sentence.(See Appendix E) To read in the "Fox" message the user would, instead of reading in the letter "A", pronounce the equivalent keyword (In this case "Alpha" would be pronounced. As an example, to read in the word "back" the user would input into the speech system the words "Bravo" "Alpha" "Charlie" "Kilo". The system would then recognise these words and replace them in the text file with the letters "B" "A" "C" "K".

This fox message was repeated five times per test, using normal continuous flowing speech, without pausing between keywords, except for the taking of breath. The test was

repeated 5 times by the same individual. The card was adapted to the testers voice before the initial testing began.

The results were obtained by calculating the percentage of incorrect characters or character positions, as opposed to the total amount of characters read in. The user was not allowed to edit or correct any errors caused by incorrect system recognition. Whenever multiple characters were recognised instead of single characters only one error was counted. The results were then tabulated and the average recognition rate was then calculated.

Results:

The results obtained were tabulated and are represented in table 7.1 below.

READER	*TOTAL CHARACTERS*	*TOTAL ERRORS*	*% ACCURACY*	*% THROUGHPUT*
MALE 1	255	33	87,05	50.91
MALE 1	255	25	90.19	52.73
MALE 1	255	18	92.94	72.73
MALE 1	255	19	92.93	65.45
MALE 1	255	20	92.15	70.91

TABLE 7.1

Below is an extract from the actual results obtained in the first test. They will be discussed in the next section.

FOX TEST 1A

the quick brown fox jumps ovvr thv lgzy dogs bgck g1244567890

the qu41c brown fox jumps over thv lgzy dog7 bgck 01234567890

the quick brown fox jumps 8 venthv lgzy dogs bgck 01234547890

thv 3uick brown fox jumps ovvn the lgzy dogs bgc7 01234567890y

thv qu1ck brown fox jumps ovxr thv lazy dogs fgck 01234567890

FOX TEST 1 B

thv 64uick brown fox jumps 8 ver the lazy d8 gs bgck 0124436780

the quick brown fox jumps 8 ver thv lgzy dogs bgck 01234667890

the quick brown f8 x j4gmps 8 ver the lgzy dogs back 0123436789

0the 3uick brown f86x jumps over the lgzy dogs bgck 01j34567890

the quick brown fox jumps ver thv lgzy dogs bgck 0234567890

The complete list of all 5 test results can be found in the Appendix

Discussion of Results: The results indicate an accuracy of 90% for the full set of Keywords. This accuracy is less than that obtained using only the numerical keywords in test 3. These results confirm that a systems accuracy is directly related to the amount of keywords trained into the system. The actual results presented above show that errors occurred almost every time the same combination of keywords were spoken. By comparing the spoken words with those recognised by the system a clear pattern can be seen.

The word combinations forming the words "lazy" and "back" contained errors repetatively. This is due to the nature of speech where adjacent words affect each other. As a result of this the recognition of keywords is now more complex. The system now has difficulty in extracting distinct features from the uttered speech. Figure 7.1 shows a spectrogram of the keywords "Bravo" "Golf" and "Bravo" "Alpha".

By examining the spectrogram for the word "Golf" and the word "Alpha" it can be noted that there is a definite similarity between the two. They both contain a low frequency component and a slight high frequncy component. Although the second word "Alpha" has an additional component indicated on the spectrogram this portion could easily be lost if the user swallows the end of the word. This is easily done by placing more emphasis on the "ph" portion of Alpha thus overpowering the trailing "a". This factor will thus explain why the recognition system had difficulty in correctly recognizing the word "Alpha".

The problem is normally solved by the speaker pausing between each keyword, thus ensuring proper pronunciation. The net result is that continuous speech is now no longer possible. Thus if the user wishes to use continuos speech recognition, it would be advisable to limit the size of the vocabulary to maintain a good accuracy.

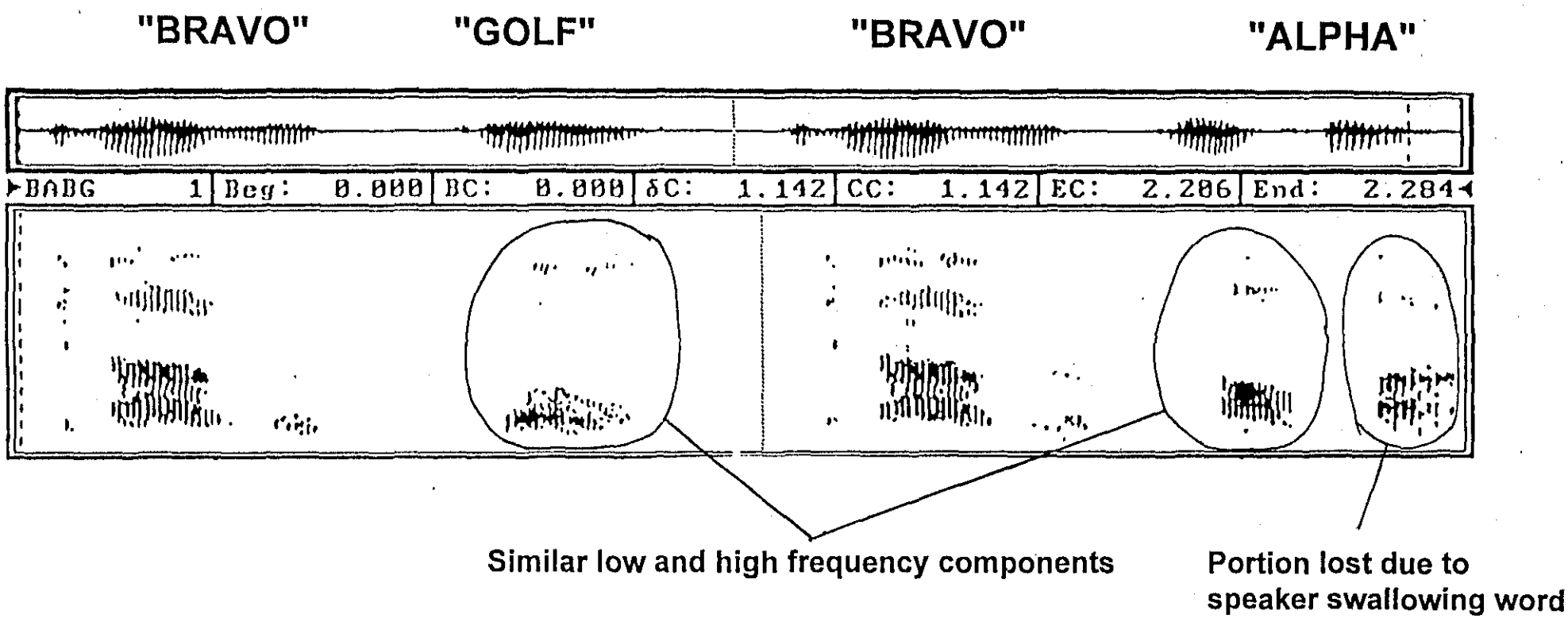
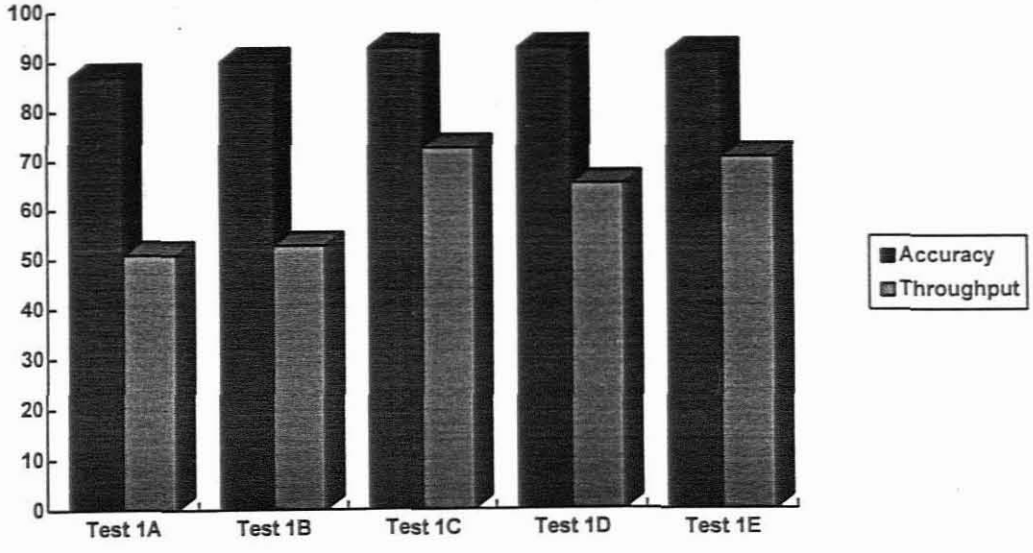


Fig 7.1 (Courtesy of Datafusion Systems, Stellenbosch)

These errors are also consistent between tests even though they were performed on different days. This leads to the conclusion that if the errors caused by the combination of these keywords were removed, the recognition rate would also be improved. Typically, for the tests performed in this instance, if the errors resulting from certain combinations of keywords were removed the accuracy could improve to approximately 96%.



GRAPH 7.1

By looking at graph 7.1 it can clearly be seen that although the accuracy of the system approached 90% the throughput remained at around 62.55%. This lower throughput is the figure that is used to determine the actual accuracy of the recognition system when determining its suitability for use in a data capturing environment. The throughput in this test is not based on a fixed number of keywords as would be the case in an actual data capturing environment but the next test will concentrate on this aspect.

7.3 Test 2

Objective: To determine the system accuracy with random selected speech

Discussion: Due to the possibility of recognition performance suffering as a result of certain combinations of keywords, it was necessary to test the system with randomly selected word combinations. As can be seen from test one certain errors occurred repeatedly when particular combinations of keywords were encountered.

This could result in the accuracy of the recognition system being reduced, because of only a few combinations of keywords. This being the case a better accuracy should be obtained with randomly selected speech. This test therefore attempts to determine if this assumption will be true. This test will also indicate what the general accuracy of the card is. It would also provide an indication of the throughput of the card.

Method : A randomly selected short article containing alphanumeric text was entered into the system. The input style was the same as test one, with the speaker using continuous normal speech. The text was read in twice. The keywords were set up the same as in test one. The system was also adapted beforehand to match the tester's speech. The results were calculated on the same basis as in test one.

Results:

A copy of the original text that was read is shown below. No emphasis was placed on the punctuation when checking for recognition accuracy. The speech system was not set up to represent Capital letters hence the lack of capital lettering in the results.

ORIGINAL TEXT : Scientists suspect their peers

How prevelant is scientific fraud ? The worlds largest general scientific society, the Ammerican Association for the Advancement of Science, recently sent surveys on this subject to 1500 members.

they suspect as phony. As to the causes of all the fraud, the scientists listed many,

Of the 469 scientists who responded 27 percent believed they have encountered or witnessed fabricated, falsified or plagiarized research over the past 10 years according to science magazine. only 2 percent believe fraud is on the decline; 37 percent feel that it is on the rise. Of those who had encountered fraud, 27 percent said they had done nothing about it and only 2 percent had publicly challenged the data such as the fierce competition to publish findings first and obtain government grants and public recognition.

TEST 1 : Icientists sigspect their peers

how prevalent is scivnt41fic fra3ud the worlds largest aeneral scient41fic society the american association for the advancement of science recxntly y sent siarveys on this sibject to 1500 members of thv 439 scientist7 who respondedy 27 percent bvlieve they have enc4ountened or witnessed fabricated faisifieh yyy or plajiarized research over thv past 10 years according to icixnce maaazine only 2 percxnt believe that fraua is on thv deciine 47 percvnt feel that it is on the rise of those who had vncountereh fraud 27 percent said thxy had done n8 thing about it and on13y 2 percent had publicly challengeh the data they suspecteh as phon3y aj6 to the causes of all the fraud the scivntists listed many such as the fierce c8 mpetition to publish findings first anh obtain government arants and pualic rec8 anition

TEST 2 : **scientists suipect their peens**

h8w prevglent is scientifi8 frg8ud the worlds largest general sc4ientific soc4iety the american associption fon the pdvanc4ement of science recently sent surveys on this subject to 1500 membens of the 469 scientisti who respondv8 27 per84ent believe they have encountered or witnessed fabricated falsified or plg8giarized neseg8nch over the past 10 yeg8rs ac8ording t8 science magg8zine only 2 percvnt believe that fng8d is on the dvlcline 37 per8ent fesl that it is on the rise of tho7e who hggd enc untered fraud 27 percent said they hg8d dons n th1ng about it an0 only 2 pergent hg8d publicly chg8llenged the dgtp they ligspected g8s phony as to the cg8uses of all the frpud the scientists li1ted mpny s1gch g8s the fierce c mpetition to publish findings first g8nd obtg8in government grants and public recognition

The results obtained are presentented in Table 7.2.

READER	*TOTAL CHARACTERS*	*TOTAL ERRORS*	*% ACCURACY*	*% THROUGHPUT*
MALE 1	*675*	*54*	*92*	*62.31*
MALE 1	*675*	*47*	*93,4*	*69.23*

TABLE 7.2

The average results are tabulated in Table 7.3

AVERAGE ACCURACY OF COMBINED TESTS	*92.7 %*
AVERAGE THROUGHPUT FOR COMBINED TESTS	*65.77 %*

TABLE 7.3

Discussion of Results: The results show that although the accuracy of the card was above 90% the throughput was below 70 %. These results are relatively similar to those obtained in test 1. Looking at the actual text results, it is quite clear that the throughput has a more noticable on the accuracy. At first glance it would not appear that the system is very accurate, yet the accuracy obtained was above 90%.

A point to noticed in these tests is that the problem associated with the words "Alpha" and "Golf" are repeated here as can be seen in the words "general" (Test1 line 1), "plagiarised" (Test 1 line 4) and "magazine" (Test 1 & 2 line 5.) Interestingly in test 2 the tendency was towards the word "golf" being recognised in most instances.

As their was only a small increase in accuracy and throughput between the two tests a reliable indication of the system accuracy can be obtained by using the ASCII fox message.

7.4 Test 3

Objective : To test the speaker-independency of the recognition system.

Discussion : The purpose of this test was to obtain an indication of the of how well the card was able to be adapted to different speakers. The test involved two phases. The

first phase was carried out without training or adapting the system to the user. The second phase was a repetition of the first, this time however the system was first adapted to the speaker. The first phase would give an indication of speaker independency of the system, whereas the second phase would test the adaptability of the system. The improvement obtained, if any, would thus give an indication of how well the system can be adapted. The tests were performed using three males and three females.

This was done to establish whether the system is gender dependent. As the prototype database was trained with male voices it was expected that test results would favour the male voices. To eliminate time consuming training or adaptation to the speakers it was decided to only use the numeric keywords.(i.e. Nul, een,.....nege) Using only the numeric keywords would also best simulate the actual data capturing environment. These numbers were tested in Afrikaans to determine how well the system was suited for South African languages.

Method: **Phase one :** Each speaker was given a list of numbers to be read into the system. The speakers were instructed to use normal continuous speech. These numbers were randomly generated by a computer and consisted of a nine character digit. No allowance was made for the speaker to edit or correct the file. Every speaker was given the same set of numbers to read in. The results are tabulated below and were calculated on the same basis as for Test 1. The throughput was based on a nine character keyword sequence. If any number was recognised incorrectly it was considered as a corrupted sequence in the formula below. The total entered sequences equaled the amount of number entered.

$$\frac{\text{Tot. entered sequences} - \text{Tot. corrupted sequences}}{\text{Total entered sequences}} \times 100\% = \%\,\text{Throughput}$$

This method would best represent a data capturing environment where the capturer has to enter a complete field before editing can take place. Hence one error in a field would result in the field being either edited or repeated.

Phase two : The procedure for this phase was exactly the same as for phase one, this time however the system was first adapted to the speakers speech. This adaptation was done by the speaker repeating the numeric keywords a number of times while the system adapted the recognition model to the specific speaker. The same set of numbers was used for this phase. The results are also tabulated below.

Results:

The The results before adaptation are tabulated in Table 7.4 below:

READER	TOTAL CHARACTERS	TOTAL ERRORS	% ACCURACY	% THROUGHPUT
MALE 1	369	25	93.22	53.66
MALE 2	369	120	67.48	2.44
MALE 3	369	37	89.97	41.4
FEMALE 1	369	116	65.85	2.44
FEMALE 2	369	253	31.44	0.0
FEMALE 3	369	172	53.39	0.0

TABLE 7.4

The results after adaptation are found in Table 7.5

READER	TOTAL CHARACTERS	TOTAL ERRORS	% ACCURACY	% THROUGHPUT
MALE 1	369	10	97.29	80.49
MALE 2	369	17	95.39	65.85
MALE 3	369	22	94.03	63.40
FEMALE 1	369	17	95.39	63.41
FEMALE 2	369	100	70.18	7.32
FEMALE 3	369	59	84.01	17.50

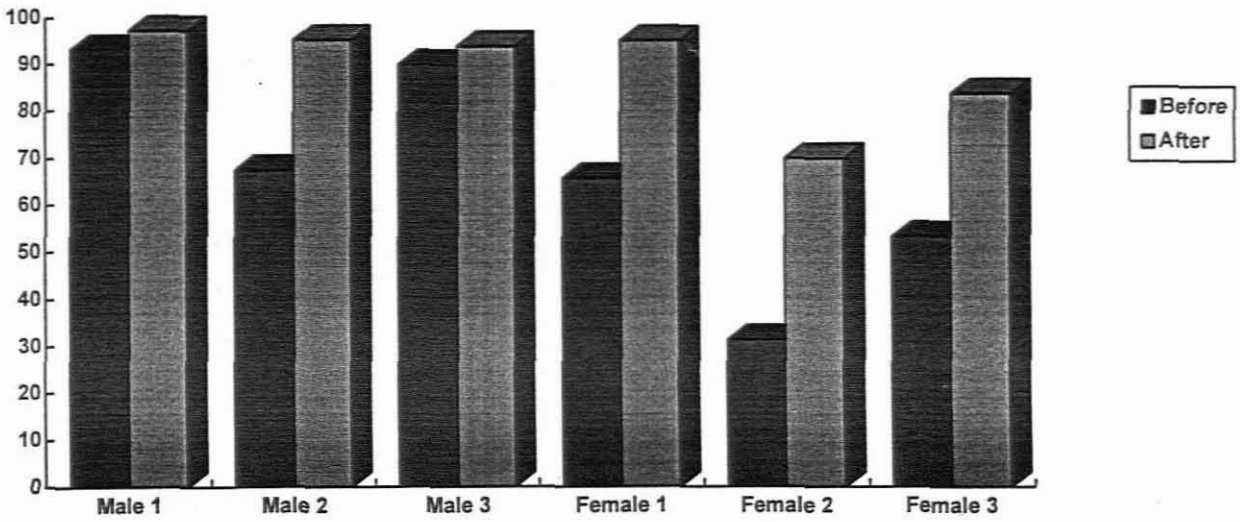
TABLE 7.5

The average of combined tests are tabulated below in Table 7.6:

Average of combined tests	**% Accuracy**	**% Throughput**
Males **Females**	**83.56** **50.23**	**32.41** **0.81**
Average after Adaptation	**% Accuracy**	**% Throughput**
Males **Females**	**95.57** **83.19**	**69.91** **29.41**

TABLE 7.6

Discussion of Results: A clear difference can be seen between the systems performance before and after adaptation is done. Graph 7.2 shows the difference between accuracy before and after adaptation. A difference is also seen between the performance of the system for male and female speakers. The results obtained prove that the speech database that is used to train the system should be based on the gender of the end user.



GRAPH 7.2

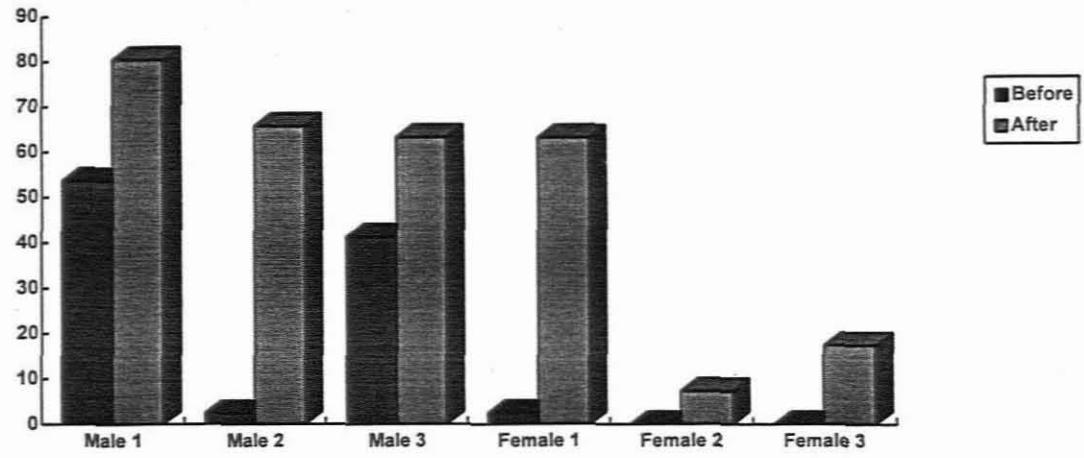
In the system under test, the recognition models were built up from a database of male voices, hence the better recognition rates for male speakers. The system's performance regarding Female no 2 was on the whole very poor. Even after adaptation the system was unable to recognise the numerical number 7 keyword. . This problem stems from the fact that the spectral components of female speakers differs from that of males. The frequency and pitch of females is normally higher than that of males and this causes differences in the recognition models. After examining the results of these tests it becomes apparent that the number "Sewe" is one examlpe of this irregularity.

Table 7. 7 presents the actual results of Female 2 and the results of Male 1 for comparison

	Female 2	*Female 2*	*Male 1*	*Male 1*
Original Text	*Before Adaptation*	*After Adaptation*	*Before Adaptation*	*After Adaptation*
653,712,560	0	3531312530	653712560	653712560
982,768,820	981380	98213880	982738820	982768820
197,261,942	19	191239421	1972619842	197261942
421,376,825	8	421338295	42136825	421376825
834,370,197	810191	8343101959	834370197	834370197
231,422,798	198	21314221398	231422798	231422798
484,280,456	8804	484280453	483280456	484280456
154,281,985	898	154281985	142819885	1542819885
839,349,113	8913	839349113	839349113	839349113
233,166,992	391	233133992	233166992	233166992
983,613,429	9819	9833134219	9883313429	983613429
224,473,699	199	22944133399	224473399	2244736998
663,869,287	898	3338392989	363869287	663869287
736,517,205	310	1333511205	736517205	736517205
164,877,695	8119	1348139395	16487695	164877695
922,794,878	919818	922994898	9227984808	9227948788
903,897,236	909	9038912133	103897236	903897236
992,693,836	9998	99393833	992693836	992693836
931,121,954	9	931121953	931121954	931121954
382,848,833	88880	382384833	382848833	382848833
618,059,886	80988	318059883	618059883	6180988
988,272,173	98811	988212113	988272173	988272173
724,436,571	31	1293433591	724436571	724436571
870,237,657	81011	8102315	870237657	878237657
235,215,612	1	23521312	235215612	235215612
951,478,386	1888	9513188384	985148386	985148386
911,851,916	999	91132931119	911851916	911851916
911,291,779	9199	0393938033	911291779	911291779
793,968,034	199803	292931292	793968034	793968034
985,285,557	98818	92985553	985285557	985285557
393,514,785	9118	393514188	393514085	393516785
758,626,625	18	1853243295	758626625	758626625
986,362,761	98811	983332131	986362731	986362761
867,160,421	81101	831304211	867160421	867160421
121,391,853	198	121391853	13139153	121391853
322,352,817	81	32235281	32252817	322352817
970,431,120	93010	904311210	970431120	978431120
847,697,470	89110	8413914150	847697470	847697470
714,578,822	1188	11459882929	714588822	714578822
764,932,765	13	13339321935	734932735	764932765

TABLE 7.7

Graph 7.3 shows the difference between throughput before and after adaptation.



GRAPH 7.3

The results indicate that the adaptation feature of the card is important when a large number of users are to be supported. Recognition performance increased by between 15 % and 30 % by just adding adaptation. Also to be noted is that the throughput, although increased after adaptation, is still below 70 % on average.

The average results for the males after adaptation is higher than those in test one. This is due to the size of the recognition database. As fewer keywords were to be recognised the system had less chance of false recognitions. This proves that system accuracy is dependant on the size of the database to be recognised.

The systems performance with Male number one showed the best results with throughput reaching just above 80 % for the nine character sequence. This however is still not accurate enough for the purposes of data capture. The accuracy of the system indicates that in certain applications speech recognition could be practical. The throughput of the system could be improved if the recogntion sequence of nine characters was reduced. This aspect will be disscussed further in chapter 8.

7.4 Test 4

Objective : To determine whether the recognition system could be practically implemented on the hardware platform for digital mapping.

Discussion:

It is very important that the user determines whether his computing system can support the voice recognition systems. As the recognition system is usually a "terminate and stay resident" (TSR) program the user must ensure that the systems driver does not clash with his applications. Some users will be using memory managers of networking drivers and hence it is vital that the introduced system does not clash with these types of applications. This type of examination will often require a reasonable knowledge of the operating system as well as the system hardware.

Method :

The environment in which the system would have to work could be outlined as follows: The application software will mostly operate on a Standard IBM AT(286) machine or compatible, with between 640 kbytes or more. The operating system is DOS. Using the smallest possible machine as a guide the hardware requirements are outlined as:

The user has to be connected to a network which uses approximately 100 to 120 kBytes of RAM for its network driver. This figure depends largely on whether the system has 640k RAM or more. The software application being run requires a further 182 kBytes of RAM. The driver for the speech recognition system takes up approximately 130 kBytes of RAM to operate. The results for this test are based on calculations and data retrieved using the DOS "mem" command.

Discussion of Results:

Doing a few calculations, it was found that only 151 kBytes of RAM was left for actual mapping data and this posed a possible problem in that the amount of data used at present is about 280 kBytes per map area, using the existing techniques.

The results indicate that by implementing the speech card the user would loose 130 bytes of RAM. This will result in smaller geographical areas being processed at one time. This would mean that the user would now have to adapt the size of this Geographical Data to cater for the reduction in RAM. This would thus affect the productivity negatively which is not acceptable and is contrary to the defined specifications.

The driver for the recognition system does appear to clash with the User application in that the dialogue box that is invoked for control of the speech recognition system operates in text mode whereas the application runs in Graphics mode. Incompatibility thus occurs when the user invokes this option. The problem can be overcome by exiting the application but this defeats the purpose of using the voice recognition system.

No noticable problems were found using the speech recognition system driver in conjunction with the networking software.

CHAPTER 8 : CONCLUSIONS AND RECOMENDATIONS

Before discussing the conclusion of the study into the feasibility of using voice for data capture the objectives will again be highlighted.

- Telkom SA needed a method of increasing the rate or speed of Data capture for Digital mapping.

- Irrespective of the method chosen, the current system had to be maintained and supported. (i.e. Present staff, software, hardware should not have to be significantly altered to support system introduced.)

- Telkom also required that the current status of speech recognition be examined to determine future possible applications.

These main objectives were used to decide whether Speech Recogniton could be used for the following:

8.1 Speech Recognitoin for Data capture?

After reviewing the test results in chapter 7 the following conclusion can be drawn :

Introducing the speech recognition system would not necessarily improve the rate or speed of data capture. The reason for this conclusion can be found in reviewing the following points:

At present the speech recognition systems available are not able to compete with the speed of Data capture by hand. Calculations based on statistics indicate that datacaptures are able to enter characters at speeds of between 3 and 5 characters per second. Even if

recognition performance could be improved to 100% the voice as a means of input would not be feasible. No individual would be able to talk as fast and continuously as an experienced Data capturer.

The study established that the speech interface was adaptable enough to be implemented with existing software applications. The speech recognition system was however not 100% compatible with the software package. Problems were experienced relating to video graphics modes. (The pop-up menus were in text mode whereas the software application was in graphics mode.) To remove this incompatibility would require engineering work on the part of the system manufacturers.

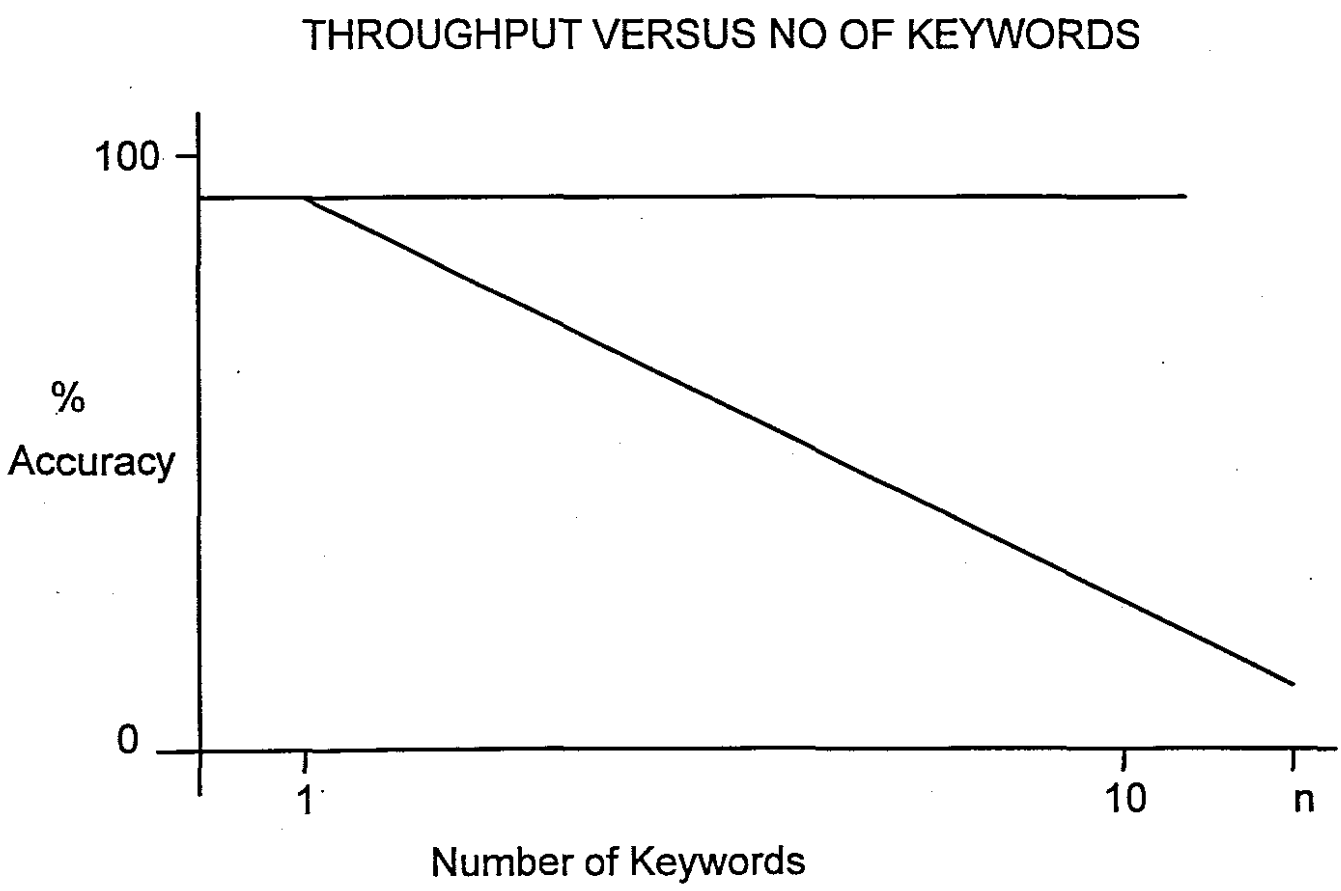
The tests also indicate that a separate database of female speakers would be required to make the system more robust. At present the accuracy and throughput for women was to low for practical implementation, irrespective of the application.

8.2 Speech Recognition Technology and Other Applications?

The tests performed do indicate that the technology of speech recognition has reached a stage where it can be introduced in certain applications. The tests show that the accuracy of the systems are at acceptable levels. Although the throughput in the tests was below the accepted limits, examining Graph 8.1 will identify the most suitable applications.

The graph shows the relationship between throughput and the number of keywords in a sequence. If for example the speaker used single keywords in a sequence then the throughput would equal the accuracy. As the number of keywords in a sequence increases the throughput decreases. In the worst case scenario the throughput would approach 0% as it did in certain of the tests described in chapter 7.

The obvious best solution would therefore be to use single keyword combinations to get the best throughput. This however results in isolated word speech recognition and not continuous speech recognition. The best use of a continuous speech recognition system would thus be in applications requiring short keyword sequences.



GRAPH 8.1

Applications such as the interrogation of Bulletin boards or databases would be ideal, where hotkeys could be replaced by uttered keyword sequences. In these types of applications the user would use short pre-defined phrases and hence only a limited number of keywords would be required.

Routing of calls or queries could also be handled by a voice recognition system. Account enquiries or voice mail applications are examples of such systems. The user is here given

a selection of menu options to chose from, in order to reach the desired option. These options could be verbally selected by the speaker.

Other applications are where the user requires a hands free interface and where speed is not the essence. Voice activated dialling for car phone, where hands free operation is essential is one example. The system however would have to allow the user to confirm the correct number as most phone numbers are relativevly long. (The system could also be designed to link actual names with phone numbers as an alternative.)

8.3. Summary

The overall conclusion that can be drawn from this investigation can be summarised as follows:

Using speech as a means of entering data within GIS is not advisable. The nature of data capturing is such that speech recognition would be impractical at present. The productivity would be negatively affected if implemented using existing hardware. The throughput of speech recognition systems is less than existing methods.

Speech recognition has however, reached a stage where the user can consider implementing it in other applications. The technology is advanced enough to be practically considered. Although recognition rates are acceptable further improvements are required before the systems can totally replace other input devices. The areas for implementation should be limited to applications that call for short phrases or menu driven options. In these applications user confidence can be established as the system performs satisfatorily.

8.3.1. Future Research

Areas of research could include the combination of male and female prototypes into one recognition system using parallel programming techniques. In this way the system will still be optomised for male or female speech. The system could use artificial intellegance techniques to distinguish the gender of the speaker and thus ensure the optimum recognition.

With the increasing popularity of multimedia it would also be advisable to ensure that future recognition systems are compatible with these environments. A windows version of the driver would be ideal as this could fully complement the multimedia scenario.

The question is also raised as to whether the throughput of a system could ever reach the same levels as the system accuracy. To optomise the throughput it would be advisable to limit the keywords of a recognition system so that all possible combinations can be catered for. A generic set of keywords could be establisehed as a standard for particular applications. This will ensure that accepted standards will be used by developers of speech recognition systems, as is the case in other areas of industry.

REFERENCES

[1] Prinsloo G.J., van der Walt C., *An investigation into Speech Input Systems*, Telkom Development Institute, Cape Town, 1992.

[2] O'Shaughnessy D, *Speech Communication: Human and Machine*, Addison-Wesley Publishing Company, 1987.

[3] Morse P.M, *Vibration and Sound*, MacCraw-Hill, New York, 1948.

[4] Fant G, *Acoustic Theory of Speech Production*, Mouton Publishers, 1960.

[5] Chomsky N, Halle M, *The sound pattern of English*, Harper & Row, New York, 1968.

[6] Fant G, Pauli S, *Spatial characteristics of vocal tract resonance modes*, Speech Communication Seminar, Stockholm, 1970

[7] Henke W, *Preliminaries to speech synthesis based upon an articulatory model*, Conference on Speech Communication and Processing, Aerospace Res. 1967.

[8] Oppenheim A.V, Schafer R.W, *Digital Signal Processing*, Prentice-Hall, 1975, 284-571.

[9] Atal B.S, Hanauer S.L, *Speech analysis and synthesis by linear prediction of the speech wave*, J. Acoust. Soc. Am. 50, 1971, 637-655.

[10] Markel J.D, Gray A.H, *Linear prediction of speech*, Springer Verslag, Berlin, 1976.

[11] Zue, V.W, Lamel L.F, *An expert spectrogram reader : A knowledge-based approach to speech recognition*, ICASSP, Vol I, 1986, 23.2.1-23.2.4.

[12] Makhoul J, Roucos S, Gish H, *Vector Quantization in speech coding*, IEEE Proceedings 73, 1985, 1551-1588.

[13] Nel I.J.N, *An adaptive homomorphic vocoder at 550 b/s*, Proc. COMSIG SA-90, Johannesburg, 1990.

[14] Flanagan J.L, Ishizaka K, Shipley K.L, *Signal models for low bit-rate coding of speech*, J. Acoust. Soc. Am. 68, 1980, 780-791.

[15] Hirahara T, Komakine T, *A Computational Cochlear Nonlinear Preprocessing Model with Adaptive Q Circuits*, ICASSP, Vol I, 1989, 496-499.

[16] Nezar L.H.T, *Model of the Cochlea using Adaptive Q Circuits and Lateral Inhibitory Neural Networks*, Thesis project, University of Stellenbosch, November 1989.

[17] Grossberg S, *Contour Enhancement, Short Term, Memory, and Constancies in Reverberating Neural Networks*, Studies of Mind and Brain, D. Reidel Publishing Company, 1982.

[18] Pickles J.O, *An introduction to the physiology of hearing*, Academic Press, 1988.

[19] Cole R.A, Zue V.W, *Speech as eyes see it*, Attention and Performance VIII, R.S. Nickerson, ed. Hillsdale, NJ, Lawrence Erlbaum, Chapter~2, 1980.

[20] Rabiner L.R, Schafer R.W, *Digital Processing of Speech Signals*, Prentice Hall Inc., Englewood Cliffs, 1978.

[21] Viterbi A.J, *Error bounds for convolutional codes and an asymtotically optimum decoding algorithm*, IEEE Transactions on Information Theory, Vol 13, 1967, 260-269.

[22] Mrayati M, Carre R, *New concept in acoustic-phonetic relations. Perspectives and applications*, ICASSP, Vol I, 1989, 231-234.

[23] Spectrograms courtesy of University of Stellenbosch ,SA , Speech Laboratory.

[24] Schroeter J, Larar J.N, Sondhi M.M, *Vector quantization of the articulatory space*, Trans. ASSP 36, 1988, 1812-1817.

[25] Schroeter J, Sondhi M.M, *Dynamic programming search of articulatory codebooks*, ICASSP, Vol I, 588-591.

[26] Moller A, *Auditory physiology*, Academic Press, New York, 1983.

[27] Prinsloo G.J., *Phoneme class recognition and Automatic Syllabification with a Phonological base Hidden Markov Model*, Masters thesis Stellenbosch University, 1988, 7-24.

[28] Miller M, Sacs M, *Representation of stop consonants in the discharge patterns of auditory-nerve fibers*, J. Acoust. Soc. Am. 74, 1983, 502-507.

[29] Kay M.S, Marple S.L, *Spectrum analysis- A modern perspective,* Proc. IEEE 69, No 11, 1981, 1380-1415.

[30] Jakobsen R, Fant C.G.M, Halle M, *Preliminaries in speech analysis*, 1952, MIT Press, Cambridge Mass, 1963.

[31] Joos M, *Acoustic Phonetics*, Supplement to Language 24, 1948

[32] Paivio A, Pegg A, *Psychology of Language*, Prentice-Hall,

[33] Borden G.J, Harris K.S, *Speech Science Primer: Physiology, acoustics and Perception of Speech*, Williams & Wilkens, London.

[34] Brosnahan L.F., Malmberg B, *Introduction to Phonetics*, Cambridge University Press, 1975

[35] Levinson S.E, Schmidt C.E, *Adaptive computation of articulatory parameters from the speech signal*, J. Acoust. Soc. Am. 74, 1983, 1145-1154.

[36] Carbonell N, Damestoy J.P, Fohr D, Haton J.P, Lonchamp F, *APHODEX, Design and implementation of an acoustic-phonetic decoding expert system*, ICASSP, Vol I, 1986, 1201-1204.

[37] Jelinek F, *Continuous speech recognition by statistical methods*, IEEE Proc. 64, 1976, 532-556.

[38] Rabiner L.R, Juang B.H, *An introduction to hidden Markov models*, IEEE ASSP 3, 1986, 4-14.

[39] Lippmann R.P, *An Introduction to Computing with Neural Nets*, IEEE ASSP Magazine, April 1987.

[40] Wakita H, *Direct estimation of the vocal-tract shape by inverse filtering of acoustic waveforms*, IEEE Trans. Audio Electroacoust., Vol 21, 1973, 417-427.

[41] Stern P.E, Eskenazi M, Memmi D, *An expert system for speech spectrogram reading*, ICASSP, Vol I, 1986, 1193-1196.

[42] Picone J, *Continuous Speech Recognition Using Hidden Markov Models*, IEEE ASSP Magazine July 1990

[43] Zue V.W., *The use of speech knowledge in automatic speech recognition"*, Proceedings IEEE, Vol 73 , pp 1602-1615, 1985

[44] Fissore L., Laface P., Micca G., *Comparison of Discrete and Continuous HMMs in a CSR Task over the Telephone*, IEEE,1991 ,253-256.

PPENDIX A: RECOMMENDED VOCABULARY FOR VOICE CARD

The following words are recommended in order to cover most commands that will be used in CAD applications. The list was compiled based on the results of a survey performed in the drawing office as well as disscusions with the speech recognition system developers.

ENGLISH WORDS

ACCEPT	AUTO	BREAK	CIRCLE
COMMAND	CONTROL	CURRENT	DELETE
DISPLAY	DOUBLE	DOWN	EIGHT
ENABLE	ENTER	ESCAPE	EXIT
FIVE	FOUR	FUNCTION	HASH
HELP	IN	JOIN	LEFT
LEVEL	LINES	LOAD	MARK
MOVE	NIL	NINE	NO
ONE	OUT	POINTS	PRINT
QUIT	REDRAW	REPEAT	RIGHT
ROTATE	SAVE	SCALE	SELECT
SET	SEVEN	SIX	SPACE
STAR	TEXT	THREE	TOGGLE
TRIPLE	TWO	UP	WINDOW
YES	ZERO	ZOOM	

AFRIKAANS WORDS

NUL	EEN	TWEE	DRIE
VIER	VYF	SES	SEWE
AGT	NEGE	JA	NEE
STER	HEKKIE		

ENDIX B WORD SELECTION SURVEY

Cover letter for survey done to establish keywords to be programmed into System.

ANALYSIS OF EXISTING MAPPER VERSION 3.2.1 SOFTWARE FOR RESEARCH PURPOSES.

I am doing a thesis project to establish if there is any way in which the existing method of data capture can be improved. The study is investigating an alternative method of data capture and any assistance in this regard will be appreciated.

Please note that these questions are not there to test your knowledge of the system so if you do not know what any of the commands mean please indicate this in the appropriate column named "unknown". (The reason why we want to know this is to determine which commands are unnecessary and no longer used.)

These questions are purely for statistical purposes and no action whether good or bad will necessarily result from them so please answer the questions honestly and objectively. Confidentiality will be maintained. Please indicate how often/seldom you use the hot keys of Mapper version 3.2.1 If you don't know what the hotkey does mark the column "unknown."

If you have any questions relating to this investigation please contact Craig van der Walt on XXXXXXXX. Thanking you for your co-operation and assistance.

SECTION 1: PLEASE MARK 1 COLUMN ONLY. PLEASE INDIATE HOW OFTEN YOU USE THE FOLLOWING COMMANDS

MENU OPTION	ABBREVIATION	USUAGE			
		UNKNOWN	OFTEN	AVERAGE	SELDOM
MARK POINT	P				
SELECT POINT	I				
ZOOM IN	Z				
ZOOM OUT	z				
SHOW WINDOW	W				
ZOOM WINDOW	w				
ENTER ANNOTATION	TS				
MOVE TEXT	TM				
ROTATE TEXT	TR				
DELETE TEXT	TD				
REDRAW	R				
LOAD TEXT	LT				
SAVE TEXT	ST				
LOAD LINES	LL				
OPTION MENU	O				
CLEAR DATA	C				
AUTO SCALE	A				
MOVE	M				
HELP	H				
TEXT ENABLE TOGGLE	ET				
LINES ENABLE TOGGLE	EL				
POINTS ENABLE TOGGLE	EP				
CURSOR SPEED	+/-				/
LOAD POINTS	LP				

MENU OPTION	ABBREVIATION	USUAGE			
		UNKNOWN	OFTEN	AVERAGE	SELDOM
JOIN POINTS	J				
SAVE LINES	SL				
LINE DELETE					
PRINT SCREEN	F1				
SET LINE TYPE	SY				
SET LEVEL	SV				
SET WINDOW	SW				
DISPLAY LEVEL	DV				
DISPLAY LINE TYPE	DY				
DISPLAY BEACON CODE	DB				
STORE WINDOW	F4				
RESTORE WINDOW	F5				
ABORT DRAW	ESC				
TEXT SIZE	TZ				
DOS COMMAND	DO				
REPEAT LAST COMMAND	F10				
MOD NG	NM				
DIGI MENU	DM				
DIGI SET DRAWING	DS				
DIGITISE	DP				
ONE POINT	[				
TWO POINT	]				
DIGI MENU 1 SET LINE TYPE					
DIGI MENU2 SET LEVEL					

APPENDIX C: TEST 1 RESULTS

THE QUICK BROWN FOX JUMPS OVER THE LAZY DOGS BACK 01234567890

FOX TEST 1 A

the quick brown fox jumps ovvr thv lgzy dogs bgck g1244567890

the qu41c brown fox jumps over thv lgzy dog7 bgck 01234567890

the quick brown fox jumps 8 venthv lgzy dogs bgck 01234547890

thv 3uick brown fox jumps ovvn the lgzy dogs bgc7 01234567890y

thv qu1ck brown fox jumps ovxr thv lazy dogs fgck 01234567890

FOX TEST 1 B

thv 64uick brown fox jumps 8 ver the lazy d8 gs bgck 0124436780

the quick brown fox jumps 8 ver thv lgzy dogs bgck 01234667890

the quick brown f8 x j4gmps 8 ver the lgzy dogs back 01234367890

the 3uick brown f86x jumps over the lgzy dogs bgck 01j34567890

the quick brown fox jumps ver thv lgzy dogs bgck 0234567890

FOX TEST 1 C

thv quick brown fox jumps 8 ver the lgzy dogs back g1234567890

the quick brown fox jumps 8 ver the la3y dogs back 01234567890

the quick brown f86x jumps 8 ver the lazy dogs back 0123456789g

the quick brown fox jumps over the lg3y dogs bgck 01234567890

the quick brown f8 x jumps 8 ver the lg3y dogs bgck 0124456789b

FOX TEST 1 D

the quick brown f8 x jumps over the lgzy dogs bgck 0123456789b

the quick brown fox jumps over the lazy dogs bgck 01234567890

the quick brown fox jumps 8 ver the lgzy dogs bgck 01234567890y

the qiick brown fox jumps 8 ver the lgzy dogs bgck 01234567890

the qgick brown fox j4gmps 8 ver the lgzy dogs bgck 01244567890

FOX TEST 1 E

the quick bnown fox jumps over the lazy dogs back 01234567890

thv quick brown fox jumps ovvn the lgzy dogs bgck 01234567890

the quick brown fox jumps ovvn thv lazy dogs bgck 01634567890

the 6uick brown fox jumps over the lazy dogs bgck 0123456789g

thv quick brown fox jumps ovxn thv lgzy dogs pgck 01234567890

APPENDIX D: TEST 2 RESULTS

ORIGINAL TEXT : Scientists suspect their peers

How prevelant is scientific fraud ? The worlds largest general scientific society, the Ammerican Association for the Advancement of Science, recently sent surveys on this subject to 1500 members.

Of the 469 scientists who responded 27 percent believed they have encountered or witnessed fabricated, falsified or plagiarized research over the past 10 years according to science magazine. only 2 percent believe fraud is on the decline; 37 percent feel that it is on the rise. Of those who had encountered fraud, 27 percent said they had done nothing about it and only 2 percent had publicly challenged the data they suspect as phony. As to the causes of all the fraud, the scientists listed many, such as the fierce competition to publish findings first and obtain government grants and public recognition.

TEST 2.1 : Icientists sigspect their peers

how prevalent is scivnt41fic fra3ud the worlds largest aeneral scient41fic society the american association for the advancement of science recxntly y sent siarveys on this sibject to 1500 members of thv 439 scientist7 who respondedy 27 percent bvlieve they have enc4ountened or witnessed fabricated faisifieh yyy or plajiarized research over thv past 10 years according to icixnce maaazine only 2 percxnt believe that fraua is on thv deciine 47 percvnt feel that it is on the rise of those who had vncountereh fraud 27 percent said thxy had done n8 thing about it and on13y 2 percent had publicly challengeh the data they suspecteh as phon3y aj6 to the causes of all the fraud the scivntists listed many such as the fierce c8 mpetition to publish findings first anh obtain government arants and pualic rec8 anition

TEST 2.2 : scientists suipect their peens

h8w prevglent is scientifi8 frg8ud the worlds largest general sc4ientific soc4iety the american associption fon the pdvanc4ement of science recently sent surveys on this subject to 1500 membens of the 469 scientisti who respondv8 27 per84ent believe they have encountered or witnessed fabricated falsified or plg8giarized neseg8nch over the past 10 yeg8rs ac8ording t8 science magg8zine only 2 percvnt believe that fng8d is on the dvlcline 37 per8ent fesl that it is on the rise of tho7e who hggd enc untered fraud 27 percent said they hg8d dons n th1ng about it an0 only 2 pergent hg8d publicly chg8llenged the dgtp they 1igspected g8s phony as to the cg8uses of all the frpud the scientists li1ted mpny s1gch g8s the fierce c mpetition to publish findings first g8nd obtg8in government grants and public recognition

APPENDIX E: TEST 3 RESULTS

Computer generated list of Random Numbers used in Tests to determine Speech system Adaptability

653 712 560
982 768 820
197 261 942
421 376 825
834 370 197
231 422 798
484 280 456
154 281 985
839 349 113
233 166 992
983 613 429
224 473 699
663 869 287
736 517 205
164 877 695
922 794 878
903 897 236
992 693 836
931 121 954
382 848 833
618 059 886
988 272 173
724 436 571
870 237 657
235 215 612
951 478 386
911 851 916
911 291 779
793 968 034
272 961 252
985 285 557
393 514 785
758 626 625
986 362 761
867 160 421
121 391 853
322 352 817
970 431 120
847 697 470
714 578 822
764 932 765

MALE 1 UNADAPTED	MALE 1 ADAPTED
653712560	653712560
982738820	982768820
1972619842	197261942
42136825	421376825
834370197	834370197
231422798	231422798
483280456	484280456
142819885	1542819885
839349113	839349113
233166992	233166992
9883313429	983613429
224473399	2244736998
363869287	663869287
736517205	736517205
16487695	164877695
9227984808	9227948788
103897236	903897236
992693836	992693836
931121954	931121954
382848833	382848833
618059883	6180988
988272173	988272173
724436571	724436571
870237657	878237657
235215612	235215612
985148386	985148386
911851916	911851916
911291779	911291779
793968034	793968034
272961252	272961252
985285557	985285557
393514085	393516785
758626625	758626625
986362731	986362761
867160421	867160421
13139153	121391853
32252817	322352817
970431120	978431120
847697470	847697470
714588822	714578822
734932735	764932765

MALE 2 UNADAPTED	MALE 2 ADAPTED
631130	653712560
913820	982768820
1911942	197261942
43113895	431376825
83440191	834370197
314198	261422798
48428045	484280456
154218198	154281985
839349113	839349113
33163992	233166991
98613429	983613469
24413499	224473699
663869981	663869281
13451120	766517205
18169	164877695
9221948	922194878
903891236	903897266
992693833	992693836
9112195	931121954
38848833	382848833
318088	618059886
988212113	988272173
8031667	870237657
3315612	235215612
95141838	951478386
911891	911851916
91191119	911291779
1931668034	976968034
21294152	272961251
98285551	985285557
39314185	393514185
158666615	758626615
9842141	986662761
9716041	867160421
1139185	121391853
235811	622652811
91043110	970461120
81691410	847697470
1141882	714578822
143932165	764932765

MALE 3 UNADAPTED	MALE 3 ADAPTED
65371250	64712560
98278820	982768820
192	1978261942
4385	421376835
33430197	834370197
331422798	2314227988
48280456154	484280456
15628105	154281985
839349113	839349113
233166992	233166992
983613729	983613429
224473691	224473699
663869287	6638692887
736517205	736517205
14877695	164897695
9227948788	922794898
9038912	9038978236
992693836	992693836
	931121954
	382848803
618059883	618059886
9888272173	988272173
724436571	7244365131
870237657	870367657
235215612	235215612
9514788386	9514788386
911851916	9511851916
911291779	9112951779
79398034	793968034
272961252	272961252
9852855500	985285557
3935147885	393514785
758626625	758626625
986362767	98662371
867160421	867160341
121391853	121139853
322352187	3223521818
970431120	970431120
847690470	847697470
7514578802	714578822
764932765	764932765

FEMALE 1 UNADAPTED	FEMALE 1 ADAPTED
53125	653712560
9827820	982768820
19292	197261942
42138	421376825
82097	834370197
231422798	231422798
828045	484180456
54281985	154281985
8939913	839349113
23316992	233166992
98212429	983612429
224299	229476699
663869280	663869287
517205	736517205
164870695	164877695
922798	922794878
902892	903897236
99269382	992692836
92112195	931121959
388388	382848863
6805988	618059886
988213	988272173
724365	724436571
80236	870237657
2325612	235215612
9540828	951478386
91185	96118519616
911299	911291779
2980	793968034
229252	272161252
9852855	985285557
	4935147885
58225	758626625
	985362761
8021	867160421
1219852	121391853
322528	322352810
021120	970431120
87970	847697370
71508822	716578822
79275	762932765

FEMALE 2 UNADAPTED	FEMALE 2 ADAPTED
0	3531312530
981380	98213880
19	191239421
8	421338295
810191	8343101959
198	21314221398
8804	484280453
898	154281985
8913	839349113
391	233133992
9819	9833134219
199	22944133399
898	3338392989
310	1333511205
8119	1348139395
919818	922994898
909	9038912133
9998	99393833
9	931121953
88880	382384833
0988	318059883
98811	988212113
31	1293433591
81011	8102315
1	23521312
1888	9513188384
999	91132931119
9199	0393938033
199803	292931292
98818	92985553
9118	393514188
18	1853243295
98811	983332131
81101	831304211
198	121391853
81	32235281
93010	904311210
89110	8413914150
1188	11459882929
13	13339321935

THE INTERNATIONAL PHONETIC ALPHABET.

(Revised to 1951.)

		Bi-labial	Labio-dental	Dental and Alveolar	Retroflex	Palato-alveolar	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngal	Glottal
CONSONANTS	Plosive	p b		t d	ʈ ɖ			c ɟ	k g	q ɢ		ʔ
	Nasal	m	ɱ	n	ɳ			ɲ	ŋ	ɴ		
	Lateral Fricative			ɬ ɮ								
	Lateral Non-fricative			l	ɭ			ʎ				
	Rolled			r						ʀ		
	Flapped			ɾ	ɽ					ʀ̆		
	Fricative	ɸ β	f v	θ ð | s z | ɹ	ʂ ʐ	ʃ ʒ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
	Frictionless Continuants and Semi-vowels	w | ɥ	ʋ	ɹ				j (ɰ)	(w)	ʁ		

VOWELS	Front	Central	Back
Close	i y	ɨ u	ɯ u
Half-close	e ø	ɘ	ɤ o
Half-open	ɛ œ	ɜ	ʌ ɔ
Open	a	ɐ	ɑ ɒ

(Secondary articulations are shown by symbols in brackets.)

APPENDIX G: PAPER PRESENTED AT COMSIG '93

The following is an extract from the Comsig '93 conference where this study was presented.

THE PRACTICAL APPLICATION OF A CONTINUOUS SPEECH RECOGNITION SYSTEM

C. van der Walt - Telkom Development Institute, Cape Town
B. Mortimer - Cape Technikon

ABSTRACT: *A study into the practical implementation of Speech Recognition for the purposes of Data Capturing within Telkom is described. Specific tests on a selected continuous speech recognition system are highlighted.*
The system chosen for the tests was locally produced and was trained using a male database of speech samples. The tests were aimed specifically at testing the accuracy and adaptability of the system. Typical results obtained, substantiate the need for training a recognition system with either male or female speech samples.
The results obtained using a 40 word vocabulary show that an accuracy of 92% is obtainable by the system. This figure can be increased to 96% by eliminating certain word combinations.
The system was also tested, using both male and female participants, on a numerical 10 word vocabulary. The results obtained for the male participants was 95% on average whereas the female average was 83%.
The difference between accuracy and system throughput is also discussed leading to conclusions as to the suitability of continuos speech recognition for this application.

1. INTRODUCTION

As the value of information continues to grow many companies realise that the capturing of this information should be done efficiently as possible. Telkom therefore decided to investigate the feasibility of using speech recognition as a means of data capturing.
This is an extract from an investigation [1] that was carried out by the Telkom Development Institute (TDI). The area that was chosen for investigation was that of data capturing of geographical information for use in planning and demand forecasting.
This report highlights the tests that were performed relating to the accuracy and adaptability of the recognition system. The tests were aimed at the practical implementation of the system and hence are directly related to the environment of numerical data capture.
Other factors such as the psychological and financial costs of implementing such a system are not discussed.

2 SYSTEM SELECTION

In choosing a system for tests it was decided that certain conditions would have to be met by the system. They are outlined as follows:

- Continuous speech must be catered for.
- The system must support bilingual speakers. (Afrikaans and English)
- The system must be speaker independent.
- The system must increase productivity.

Continuous Speech: Because the system was to be implemented in a data capturing environment the system had to recognise continuous speech. If an isolated word recognition system was chosen productivity would suffer.

Bilingual speakers: The present work force where the system was to be implemented consists of bilingual speakers. Most commercially available systems are trained using an English base of recognition templates. The question thus arose as to whether the system could be adapted to recognise Afrikaans speakers.

Speaker independent: As data capturing is usually performed by a large volume of capturer's it would have been impractical to select a system that would require individual training for each speaker.

Increase Productivity: This facet could only truly be gauged over a period of time. The initial performance of the system would however give a good indication of whether this method of Data capturing would be faster.

The system that was selected [2] matched most of the criteria and was made available by Datafusion Systems. The system had been trained using a database of ten male speech samples.

As indicated in [2] the Hidden Markov Model recognition technology was employed. This system has subsequently been used in further tests to improve iterative speaker adaptation for Speech Recognition [3].

3 TESTS PERFORMED

The tests that are herein described cannot be used as a standard for generalised testing of various speech recognition systems. Every application in which the speech recognition system is to be implemented will require that a different set of tests be decided upon.
Two main tests were performed, each serving to test a different feature of the recognition system.

Objective 1: To determine the system accuracy of the recognition system with repetitious speech.

Objective 2: To test the speaker independency of the recognition system.

It was envisaged that on the basis of these tests, a reliable indication could be obtained as to how well the speech recognition system performs. The working environment chosen for the tests are the same as those that exist within the data capturing environment. The speaker was equipped with a head-mounted microphone to facilitate hands-free operation.

3.1 TEST 1

Objective:
To establish the accuracy of the recognition system.

Discussion:
This test was performed using only one speaker. This would eliminate any possible errors due to speaker adaptation. The results obtained would therefore be directly related to the accuracy of the recognition system and not its ability to adapt to different speakers. Although this test would tend to be subjective, the results would represent a good indication of the systems performance.

Method :
The chosen system has 40 keywords (Alpha, Bravo ..... Zulu, Nul, Een .... Nege, as well as "Ster" and "Hekkie".)
The system was configured such that each Alphabetic character was represented by its respective keyword.(i.e. "Romeo" = 'R') The numeric keywords represented the respective numerical values (i.e. "Sewe" = '7')
In order to get a generalised even usage of the key words the Standard ASCII 'FOX' message was used. This message allows for the use of every keyword within one sentence.
This message was repeated five times per test, using normal continuous flowing speech, without pausing between keywords, except for the taking of breath. The test was repeated 5 times by the same individual. The card was adapted to the speakers voice before the initial testing began.
The results were obtained by calculating the percentage of incorrect characters or character positions, as opposed to the total amount of characters read in. The user was not allowed to edit or correct any errors caused by incorrect system recognition. Whenever multiple characters were recognised instead of single characters only one error was counted. The results were then tabulated and the average recognition rate was then calculated.

Summary of results for Testing of Card Accuracy.

READER	TOTAL CHARACTERS	TOTAL ERRORS	% ACCURACY
MALE 1	255	33	87.05
MALE 1	255	25	90.19
MALE 1	255	18	92.94
MALE 1	255	19	92.93
MALE 1	255	20	92.15

Table 3.1-Speaker accuracy

Discussion of Results:
The results indicate an accuracy of 90% for the full set of Keywords. This accuracy is less than those obtained using only the numerical keywords in the other tests described.

By comparing the Spoken words with those recognised by the system a clear pattern can be seen. In most cases the errors occurred almost every time when the same combination of keywords were said.

For example an extract from two of the tests is shown.

FOX TEST 1A
the quick brown fox jumps ovvr thv lgzy dogs bgck g1244567890
the qu41c brown fox jumps over thv lgzy dog7 bgck 01234567890
the quick brown fox jumps 8 venthv lgzy dogs bgck 01234547890
thv 3uick brown fox jumps ovvn the lgzy dogs bgc7 01234567890y
thv qu1ck brown fox jumps ovxr thv lazy dogs fgck 01234567890
FOX TEST 1 B
thv 64uick brown fox jumps 8 ver the lazy d8 gs bgck 0124436780
the quick brown fox jumps 8 ver thv lgzy dogs bgck 01234667890
the quick brown f8 x j4gmps 8 ver the lgzy dogs back 0123436789
0the 3uick brown f86x jumps over the lgzy dogs bgck 01j34567890
the quick brown fox jumps ver thv lgzy dogs bgck 0234567890

The same mistakes occur at the same word combinations within a given test. This is due to the nature of speech where adjacent words affect each other.
These errors are also consistent between tests even though they were performed on different days. This leads to the conclusion that if the errors caused by the combination of these keywords was removed, the recognition rate would also be improved.
Typically, for this test if the errors resulting from certain combinations of keywords were removed the accuracy could improve to approximately 96%. Tests performed on the same set of keywords but using randomly selected speech showed about a 2% higher recognition rate.
However if a reasonable standard is required to test various systems the ASCII FOX message could be used.

3.2 TEST 2

Objective :

To test the speaker independence of the recognition system.

Discussion :

The purpose of this test was to get a general indication of the of how well the card was able to be adapted to different speakers. The test involved two phases. The first phase was carried out without training or adapting the system to the user. The second phase was a repetition of the first, this time however the system was first adapted to the speaker.
The first phase would give an indication of speaker independency of the system, whereas the second phase would test the adaptability of the system. The improvement obtained, if any would thus give an indication of how well the system can be adapted.
The tests were performed using three male and three female bilingual speakers. This was done so as to establish the gender dependency of the system. As the prototype database was trained with male voices it was expected that test results would favour the male voices.
It was decided to only use the numeric keywords.(i.e. Nul, Een...Nege) These keywords were tested in Afrikaans to determine how well the system was suited for South African dialects.

Method:

Phase one :
Each speaker was given a list of numbers to be read into the system. The speakers were instructed to use normal continuous speech. These numbers were randomly generated by a computer. No allowance was made for the speaker to edit or correct the file.
Every speaker was given the same set of numbers to read in. The result were tabulated and calculated on the same basis as for Test 1 (Refer test 1)

Summary of results for tests performed in Phase 1:

READER	TOTAL CHARACTERS	TOTAL ERRORS	% ACCURACY
MALE 1	369	25	93.22
MALE 2	369	120	67.48
MALE 3	369	37	89.97
FEMALE 1	369	116	65.85
FEMALE 2	369	253	31.44
FEMALE 3	369	172	53.39

Table 3.2-Before adaptation

Phase two :
The procedure for this phase was exactly the same as for phase one, this time however the system was first adapted to the speakers speech style. The same set of numbers was used for this phase. The results are tabulated below.

The following results were obtained for Phase 2 after adaptation was performed.

READER	TOTAL CHARACTERS	TOTAL ERRORS	% ACCURACY
MALE 1	369	10	97.29
MALE 2	369	17	95.39
MALE 3	369	22	94.03
FEMALE 1	369	17	95.39
FEMALE 2	369	100	70.18
FEMALE 3	369	59	84.01

Table 3.3-After Adaptation

Discussion of Results:

A clear difference can be seen between the systems performance before and after adaptation is performed. The results indicate that the adaptation feature of the card is important when a large number of users are to be supported.
Recognition performance increased between 15 and 40 % after adaptation.
A major difference is also seen between the performance of the system for male and female speakers. The results obtained prove that the speech database that is used to train the system should be based on the gender of the end user.

In some instances with the female speakers, the system was unable to recognise certain numbers until after adaptation. An example can be seen by taking an extract from the test of female no 2:

Original Text	*Before Adaptation*	*After Adaptation*
653,712,560	0	3531312530
982,768,820	981380	98213880
197,261,942	19	191239421
421,376,825	8	421338295
834,370,197	810191	8343101959
231,422,798	198	21314221398
484,280,456	8804	484280453
154,281,985	898	154281985
839,349,113	8913	839349113
233,166,992	391	233133992
983,613,429	9819	9833134219
224,473,699	199	22944133399
663,869,287	898	3338392989
736,517,205	310	1333511205
164,877,695	8119	1348139395
922,794,878	919818	922994898
903,897,236	909	9038912133
992,693,836	9998	99393833
931,121,954	9	931121953
382,848,833	88880	382384833
618,059,886	80988	318059883
988,272,173	98811	988212113
724,436,571	31	1293433591
870,237,657	81011	8102315
235,215,612	1	23521312
951,478,386	1888	9513188384
911,851,916	999	91132931119
911,291,779	9199	0393938033
793,968,034	199803	292931292
985,285,557	98818	92985553
393,514,785	9118	393514188
758,626,625	18	1853243295
986,362,761	98811	983332131
867,160,421	81101	831304211
121,391,853	198	121391853
322,352,817	81	32235281
970,431,120	93010	904311210
847,697,470	89110	8413914150
714,578,822	1188	11459882929
764,932,765	13	13339321935

TABLE 3.4-Results Female 1

As indicated in the table, very poor recognition was achieved before adaptation was performed. However after adaptation the recognition performance was around 70 %, with an improvement of approximately 40 %.
Despite this improvement it is quite apparent that the system would have to be trained using a female database of speech samples if an accuracy above ***90 %*** is required.

The results for the Male participants appear to be satisfactory, but the question arises is this good enough for Data capturing purposes. To asses this, the system throughput must be considered.

3.3 ACCURACY VERSUS THROUGHPUT

An important concept must be realised when evaluating the accuracy of speech recognition systems. The concept pertains to Accuracy versus Throughput.
The throughput directly related to the stability of the recognition system. To best understand throughput a comparison will be given between the formula for each of these terms.
The formula for calculating accuracy can be given as :

$$\frac{\text{Total entered elements - total corrupted elements}}{\text{Total entered elements}} \times 100\% = \%\ \text{Accuracy}$$

The equation for calculating the throughput is given as:

$$\frac{\text{Tot. entered sequences - Tot. corrupted sequences}}{\text{Total entered sequences}} \times 100\% = \%\text{Throughput}$$

Whereas the accuracy of a system is determined directly by the number of correctly recognised keywords the throughput is dependent on the amount of keywords which are combined to form a sequence. To illustrate this the following example will be examined.

In a data capturing environment an individual has to input 10 digit numbers repetitively. If the recognition system has an accuracy of ***90%***, only one error should occur for every 10 digits entered. This accuracy can be acceptable, but in the worst case scenario one error will occur with every 10 digit sequence entered. Although the accuracy would still be ***90%*** the throughput or number of correct sequences entered would approach ***0%***. If however the numbers consisted of 5 digits only, the throughput would approach ***50%***.
In the tests described in section 3.2, nine digit word sequences were used. Using the respective formulas mentioned above, the average throughput of the recognition system approached ***70%*** for the male participants and ***30%*** for the females after adaptation.
From the test results presented in table 3.4 the throughput for the system was calculated at only ***7.32%*** even though the accuracy was above ***70%***.
As most requirements for data capturing consist of multiple digit sequences the throughput of the system will be a better criteria to use when determining system performance in data capturing or similar environments.

4 CONCLUSION

Tests performed on a Continuous Speech Recognition System were described. These tests were used to determine whether it was practical to employ this growing technology within Telkom, specifically in a data capturing environment.

The need for gender related training of speech databases in speaker independent systems was identified

The results obtained indicate that although accuracy's for continuous speech recognition systems approach 95% the final throughput is still too low for practical implementation in a data capturing environment.

These systems are at present better suited for environments that call for short phrases or command sequences, such as found in CAD applications. Other applications could include mobile car systems where hands-free dialling would be preferred

5 REFERENCES

[1] C van der Walt, "An investigation into the practical implementation of speech recognition for data capturing", Unpublished manuscript for Masters Diploma in Technology , Cape Technikon, July 1993.

[2] G. van Wyk, I.H.J. Nel, and W Coetzer, "A real time speaker independent speech recognition system", Proceedings of the IEEE South African Symposium on Communications and Signal Processing, 1991.

[3] F.J. Scholtz, J.A. du Preez, "Iterative Speaker Adaptation for Speech Recognition", Proceedings of the IEEE South African Symposium on Communications and Signal Processing, 1992.