**ABSTRACT**

One of the problems faced in speech recognition is that the spoken word can be vastly altered by accents, dialects and mannerisms. In South Africa, there is a large variety of languages and dialects. Even the most basic speech recognition systems perform poorly when trying to recognise words spoken by English second language speakers. The motivation behind this survey is to investigate speech recognition and more specifically what research has been around dealing with the problem of large variations in dialects.

1. **INTRODUCTION**

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Many speech recognition applications, such as voice dialing, simple data entry and speech-to-text are in existence today.

Automatic speech recognition systems involve numerous separate components drawn from many different disciplines such as statistical pattern recognition, communication theory, signal processing, combinatorial mathematics, and linguistics.

Attempts to build automatic speech recognition (ASR) systems were first made in the 1950s. These early speech recognition systems tried to apply a set of grammatical and syntactical rules to identify speech. If the spoken words adhered to a certain rule set, the system could recognise the words. However, human language has numerous exceptions to its own rules. The way words and phrases are spoken can be vastly altered by accents, dialects and mannerisms [3]. Therefore, today's speech recognition systems increasingly rely on statistical methodology, moving away from approaches such as template matching, dynamic time warping, and non-probabilistically motivated distortion measures that were initially proposed [5]. The statistical model that dominates the field today is the hidden Markov Model.

2. **OVERVIEW**

The remainder of this literature survey is structured as follows: Section 3 discusses the different approaches to speech recognition. This is followed in Section 4 by a summary of the speech recognition process. Section 5 briefly discusses hidden Markov Models (HMM) and their application to speech processing. Section 6 looks at some research that has been done on speaker independent systems to handle various dialects. Finally, Section 7 summarises and concludes the paper.

3. **VARIOUS APPROACHES TO SPEECH RECOGNITION**

The three broad approaches to automatic speech recognition are the acoustic-phonetic, pattern recognition and artificial intelligence (AI) approaches [2]. The acoustic phonetic approach to speech recognition has not been very successful in practical speech recognition systems. Both the pattern
recognition and AI approach to speech recognition have achieved higher success rates than the acoustic-phonetic approach.

3.1 Acoustic-Phonetic Approach

In this speech recognition algorithm, the system tries to decode the speech signal in a sequential manner based on the observed acoustic features of the speech waveform and the known relations between acoustic features and phonetic symbols. Figure 1 shows a block diagram of the acoustic-phonetic approach to speech recognition. The first step in the process is the parameter measurement process, which provides an appropriate spectral representation of the speech signal. The next step in the processing is the feature detection stage where the spectral measurements are converted to a set of features that describe the acoustic properties of the various phonetic units. Finally, the recogniser tries to determine the best matching word or sequence of words.

![Fig 1. Acoustic Phonetic Approach to Speech Recognition](image)

3.2 Pattern Recognition Approach

In this approach, the speech patterns are used directly without explicit feature determination and segmentation. The method has two steps-namely, training of speech patterns, and recognition of patterns by way of pattern comparison. Figure 2 shows a block diagram of the pattern-recognition approach. In the parameter measurement phase, a sequence of measurements is made on the input signal to define the “test pattern”. The unknown test pattern is then compared with each sound reference pattern and a measure of similarity between the test pattern and reference pattern is computed. Finally, the decision rule decides which reference pattern best matches the unknown test pattern based on the similarity scores from the pattern classification phase.

![Fig 2. Pattern Recognition Approach to Speech Recognition](image)
3.3 Artificial Intelligence Recognition Approach

This approach is a hybrid of the acoustic-phonetic approach and the pattern recognition approach. In the artificial intelligence approach (AI), an expert system or self-organising (learning) system, implemented by neural networks is used to classify sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem at hand.

4. SPEECH RECOGNITION PROCESS

In essence, the basic task involved in speech recognition is that of going from speech recordings to word labels. As the pattern recognition approach to speech recognition is the most widely used approach, this approach will be discussed in more detail. There are two main variants of the basic speech recognition task, namely isolated word recognition and connected word recognition.

4.1 Variants of the Speech Recognition Task

4.1.1 Isolated word recognition

Isolated word recognition refers to the task of recognizing a single spoken word where the choice of words is not constrained to task syntax or semantics. As described in [4], HMMs can be used to build an isolated word recogniser. Briefly, the HMM approach is a well-known and widely used statistical method of characterising the spectral properties of the frames of a pattern. HMMs are particularly suitable for speech recognition as the speech signal can be well characterised as a parametric random process and the parameters of the stochastic process can be determined in a precise, well-defined manner. [2]

4.1.2 Fluent speech Recognition

Fluent speech recognition is a more complicated task than isolated word recognition. In this case the task is to recognize a continuous string of words from the vocabulary.

4.2 Feature Extraction and Pattern Recognition

The input into an automatic speech recognition system is the speech signal. The two major tasks involved in speech recognition are feature extraction and pattern recognition.

4.2.1 Feature Extraction

In all speech recognition systems the first step in the process is signal processing. Initially a spectral and temporal analysis of the speech signal is performed to give observation vectors which can be used to train the HMMs [4]. One way to obtain observation vectors from speech samples is to perform spectral analysis. A type of spectral analysis that is often used is linear predictive coding (LPC) [4].

4.2.2 Pattern Recognition

Pattern recognition refers to the matching of features. The pattern recognition process consists of training and testing. During training, a model of each vocabulary word must be created. Each model consists of a set of features extracted from the speech signal. The exact form of the model depends on the type of pattern-recognition algorithm used. During testing, a similar model is created for the unknown word. The pattern-recognition algorithm compares the model of the unknown word with the models of known words and selects the word whose model score is highest [7]. There are many different pattern matching techniques. These include templates, Dynamic Time Warping and HMMs.
5. HIDDEN MARKOV MODELS IN SPEECH RECOGNITION

Despite huge amounts of research trying to create an intelligent speech recognition machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments [2]. To date, the best results in speech recognition systems have been achieved by those based on hidden Markov models. Hence, most current automatic speech recognition systems are based on HMMs.

5.1 Three Fundamental Problems of HMM design

HMM design is characterised by three fundamental problems [4] namely:

1. The evaluation of the probability of a sequence of observations given a specific HMM.
2. The determination of a best sequence of model states.
3. The adjustment of model parameters to best account for the observed signal.

There are various methods for solving the above problems discussed in the literature. The most popular technique used to solve problem 1, the Forward-Backward procedure, is an algorithm for computing the probability of a particular observation sequence [8].

The Viterbi algorithm [9] is a dynamic programming algorithm for finding the most likely sequence of hidden states that results in a sequence of observed events. This algorithm is a popular technique for solving problem 2, that of finding the best state sequence for the given observation sequence.

The third and most difficult problem in the design of HMMs is the problem of determining a method to maximise the probability of the observation sequence given the model. As Rabiner mentions, there is no known way to analytically solve this problem, neither is there an optimal way of estimating the model parameters. There are various iterative procedures such as the Baum-Welch method, and expectation modification method [10] or gradient techniques [11] that can be used to choose model parameters. The standard criterion for estimation of HMM parameters is maximum likelihood.

5.2 Types of HMMs

There are many different types of hidden Markov models. In the ergodic or fully connected HMM, every state of the model can be reached (in a single step) from every other state of the model [4]. In speech processing, the left-right model or Bakis [12] model has been used. The benefit of this model is that it can model signals whose properties change over time [4].

![Fig. 3. Illustration of 2 types of HMMs. (a) A 4-state ergodic model. (b) A 4-state left-right model.](image-url)
5.3 Limitation of HMMs

HMMs have been successfully applied to problems in both isolated and connected word recognition. There are however some limitations of this type of statistical model for speech. One major limitation is the assumption made in the model that successive observations are independent. A second limitation is the assumption that the distributions of individual observation parameters can be well represented as a mixture of normal or autoregressive densities. The final assumption limiting the HMM model is the assumption that the probability of being in a given state at a certain time only depends on the previous state. This assumption is inappropriate as speech sound dependencies often extend through several states [4].

6. SPEAKER INDEPENDENT SPEECH SYSTEMS

There are many different languages and dialects throughout the world. Even the most basic (isolated word) speech recognition systems perform poorly when trying to recognise the words spoken by English second language speakers. The following section discusses some research that has been done into various ways of handling difficult situations with large variations in dialects.

Improved performance for speaker independent speech recognitions systems requires better modelling of different dialects of the target language. Previous work that has been carried out on this topic, suggests that separate modelling of dialects is needed to accurately capture the many pronunciation differences that occur. Regardless of how much dialect data is included in training, some speakers will not be covered by the resulting model. Those speakers not covered by the model include nonnative speakers of the language and speakers whose speech patterns have been affected by surgery [14].

The BBN BYBLOS system [15] is a continuous speech recognition system that has been used to develop a method of speaker adaptation from limited training. The authors show that the system performs poorly for speakers with strong dialects. They also show how the degradation can be overcome by using speaker adaptation from multiple reference speakers. Their results obtained from testing showed that their current state-of-the-art (SI - speaker independent) models perform poorly when a test speaker's characteristics differ markedly from those of the training speakers. Their SI models have difficulty with a native speaker of English with an African-American dialect for instance. Moreover, nonnative speakers of American English nearly always suffer significantly degraded SI performance.

To try to overcome this degradation, they tried to adapt the training models to the new dialects by estimating a probabilistic spectral mapping between each of the training speakers and the test speaker. [15] They found that the overall average word error rate after speaker adaptation was 5 times better than SI recognition for these speakers. These results are evidence of the need and usefulness of speaker adaptation to be able to recognise the speech of speakers whose dialects differ from those found in the training data.

The results achieved by the BYBLOS system [15] agree with the testing done on the Cambridge University HTK (CU-HTK) System for the automatic transcription of conversational telephone speech [16]. Testing on the HTK system showed that adaptation to the test speaker and the acoustic environment greatly improves the performance of automatic speech recognisers. [16]
7. CONCLUSION

There has been much progress in the field of automatic speech recognition since its humble beginnings in the 1950s. Various approaches to ASR have been mentioned. Current speech recognition systems are generally based on hidden Markov models as these models have lead to the best results in speech recognition systems thus far. Although HMMs have been very successful, there are a few limitations of the models that were mentioned. The need for, and usefulness of speaker adaptation in speaker independent systems was highlighted. We are a long way from achieving perfect speech recognition and there is much research still to be done in the field of automatic speech recognition.
8. References

Last accessed 28 April 2009.


