IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

COMPARITIVE STUDY OF VARIOUS DISTANCE MEASURES FOR ISOLATED SPEECH RECOGNITION APPLICATION

Akash Prakash*, Divyankitha M.Urs, Lahari S, Preetish H.S, M. A. Anusuya

[*] Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India

Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India

Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India

Student, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India

Associate Professor, Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India

ABSTRACT

Speech recognition applications are becoming common these days as many of the modern devices designed are user-friendly for the convenience of general public. Speaking or communicating directly with the machine to achieve desired objectives make usage of modern devices easier and convenient. Although many interactive software applications are available, the uses of these applications are limited due to language barriers. Hence, development of speech recognition systems in local languages will help anyone to make use of this technological advancement. In India, speech recognition systems have been developed for many indigenous languages [7]. In this project, we present the performance analysis of acoustic template method using different distance measures for Kannada language. The main objective of the paper is to show that City Block distance measure has been identified equivalence to Euclidean distance measure. This analysis shows that for simple speech recognition applications like character or isolated speech recognition, City Block distance measure can also be used.

KEYWORDS: Speech recognition, acoustic method, Vector Quantization, Euclidean Distance, City Block Distance, Spearman Distance

INTRODUCTION

Speech is the vocalized form of human communication. Speech is natural, easy, fast, hands free and does not require technical knowledge. Human beings are comfortable with speaking directly with computers rather than depending on primitive interfaces such as keyboards and pointing devices. The primitive interfaces like keyboard and pointing devices require certain amount of skill for effective usage. Use of mouse requires good hand-eye coordination. Physically challenged people find it difficult to use computer. It is difficult for blind people to read from monitor. Moreover, current computer interface assumes a certain level of literacy from the user. It expects the user to have certain level of proficiency in English apart from typing skill. Speech interface helps to resolve these issues.

Interaction with computer through a convenient and user-friendly interface has always been an important technological issue. Machine-oriented interfaces restrict the computer usage to a small fraction of the population, who are both computer literate and conversant with written English. Computers which can recognize speech in native languages enable common man to make use of the benefits of information technology. Speech recognition system keeps elderly, physically challenged especially blind people closer to the Information Technology revolution. Speech recognition benefits a lot in manufacturing and to control applications where hands or eyes are otherwise occupied. It has large application for use, over the telephone, including automated dialing, telephone directory assistance, spoken database querying for inexperienced users, and voice dictation systems. Section 2 describes the collection of related works.

Section 3 describes the methodology used for isolated speech recognition. Section 4 describes the system implementation process. Section 5 discusses the results and conclusions along with the future enhancement.

The Speech Signal

A brief introduction to how the speech signal is produced and perceived by the human system can be regarded as a starting point in order to go into the field of speech recognition. The process from human speech production to human speech perception, between the speaker and the listener, is shown in Figure 1 [8].

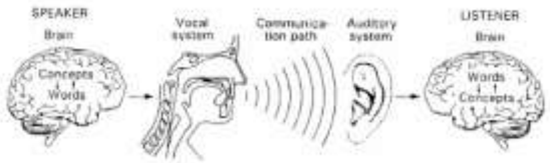


Figure 1: Human speech communication Speech recognition

In computer science and electrical engineering, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). Some SR systems use "speaker independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system. Systems that do not use training are called "speaker-independent" systems. Systems that use training are called "speaker-dependent" systems. Speech recognition systems try to establish a similarity to the human speech communication system. A source-channel model for a speech recognition system is illustrated in Figure 2.

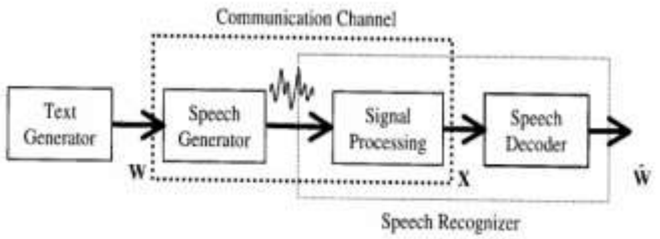


Figure 2: Source channel model

The different elements from the human communication system are related to the modules or components of the source-channel model, giving a short explication of how human speech communication and speech recognition systems are performed. The aim of human speech communication is to transfer ideas. They are made within the speaker's brain and then, the source word sequence 'W' is performed to be delivered through her/his text generator. The human vocal system, which is modeled by the speech generator component, turns the source into the speech signal waveform that is transferred via air (a noisy communication channel) to the listener, being able to be affected by external noise sources. When the acoustical signal is perceived by the human auditory system, the listener's brain starts processing this waveform to understand its content and then, the communication has been completed. This perception process is modeled by the signal processing and the speech decoder components of the speech recognizer, whose aim is to process and decode the acoustic signal X into a word sequence Ŵ, which is hopefully close to the original word sequence W. Thus, speech production and speech perception can be seen as inverse processes in the speech recognition system [8].

RELATED WORKS

Human computer interactions as defined in the background is concerned about ways users (humans) interact with the computers. Some users can interact with the computer using the traditional methods of a keyboard and mouse as the main input devices and the monitor as the main output device. Speech recognition systems help users who in one way or the other cannot be able to use the traditional Input and Output devices. For about four decades human beings have been dreaming of an "intelligent machine" which can master the natural speech. In its simplest form, this machine should consist of two subsystems, namely automatic speech recognition (ASR) and speech understanding. The goal of ASR is to transcribe natural speech while SU is to understand the meaning of the transcription. Recognizing and understanding a spoken sentence is obviously a knowledge-intensive process, which must take into account all variable information about the speech communication process, from acoustics to semantics and pragmatics [9].

Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following:

Isolated Words

Isolated word recognizers require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. It recognises one word uttered at a time. Dictation systems are the best examples of it.

Connected Words

Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.

Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries. It also depends on the different types of accents used by the user.

Spontaneous Speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters [2].

Current State of ASR Technology and its Implications for Design

The design of user interfaces for speech-based applications is dominated by the underlying ASR technology. More often than not, design decisions are based more on the kind of recognition the technology can support rather than on the best dialogue for the user. The type of design will depend, broadly, on the answer to this question: What type of speech input can the system handle, and when can it handle it? When isolated words are all the recognizer can handle, then the success of the application will depend on the ability of designers to construct dialogues that lead the user to respond using single words. Word spotting and the ability to support more complex grammars opens up additional flexibility in the design, but can make the design more difficult by allowing a more diverse set of responses from the user. Some current systems allow a limited form of natural language input, but only within a very specific domain at any particular point in the interaction.

Even in these cases, the prompts must constrain the natural language within acceptable bounds. No systems allow unconstrained natural language interaction, and it’s important to note that most human-human transactions over the phone do not permit unconstrained natural language either. Typically, a customer service representative will structure the conversation by asking a series of questions. With “barge-in” (also called “cut-through”), a caller can interrupt prompts and the system will still be able to process the speech, although recognition performance will generally be lower. This obviously has a dramatic influence on the prompt design, because when barge-in is available it’s possible to write longer more informative prompts and let experienced users barge-in. Interruptions are very common in human conversations, and in many applications, designers have found that without barge-in people often have problems. There are a variety of situations, however, in which it may not be possible to implement barge-in. In these cases, it is still usually possible to implement successful applications, but particular care must be taken in the dialogue design and error messages. Another situation in which technology influences design involves error recovery. It is especially frustrating when a system makes the same mistake twice, but when the active vocabulary can be updated dynamically, recognizer choices that have not been confirmed can be eliminated, and the recognizer will never make the same mistake twice. Also, when more than one choice is available (this is not always the case, as some recognizers return only the top choice), then after the top choice is disconfirmed, the second choice can be presented [9].

Problems in Designing Speech Recognition Systems

ASR has been proved to be a not easy task. The main challenge in the implementation of ASR on desktops is the current existence of mature and efficient alternatives, the keyboard and mouse. In the past years, speech researchers have found several difficulties that contrast with the optimism of the first speech technology pioneers. According to Ray Reddy, in his review of speech recognition by machines says that the problems in designing ASR are due to the fact that it is related to so many other fields such as acoustics, signal processing, pattern recognition, phonetics, linguistics, psychology, neuroscience, and computer science. And all these problems can be described according to the tasks to be performed.

- **Number of speakers**: With more than one speaker, an ASR system must cope with the difficult problem of speech variability from one speaker to another. This is usually achieved through the use of large speech database as training data (Huang et al., 2004)
- **Nature of the utterance**: Isolated word recognition imposes on the speaker the need to insert artificial pause between successive utterances. Continuous speech recognition systems are able to cope with natural speech utterances in which words may be tied together and may at times be strongly affected by co articulation.
- **Vocabulary size**: In general, increasing the size of the vocabulary decrease the recognition scores.
- Differences between speakers due to sex, age, accent and so on.
- **Language complexity**: The task of continuous speech recognizers is simplified by limiting the number of possible utterances through the imposition of syntactic and semantic constraints.
- **Environment conditions**: The sites for real applications often present adverse conditions (such as noise, distorted signal, and transmission line variability) which can drastically degrade the system performance [9].

METHODLOGY

The following figure 3 shows the different phases of speech recognition application.

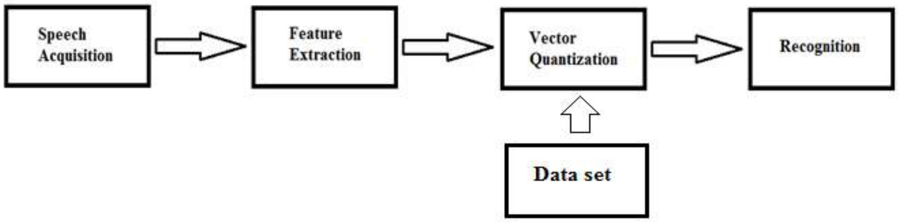


Figure 3: Basic system architecture

Speech Acquisition

The input to the algorithm is recorded using software called Praat. The samples are recorded in .wav format. The sampling frequency of each of the samples is 16 KHz. The signals are acquired using mono channel. Initially, recordings of each word are supplied to the algorithm to train the system.

Feature Extraction

For each of the sample recordings, the MFCC features are extracted. The resulting features are called the feature vector. The MFCC features are extracted up to 13 coefficients. One coefficient is energy feature coefficient and the remaining 12 coefficients are features of the signal.

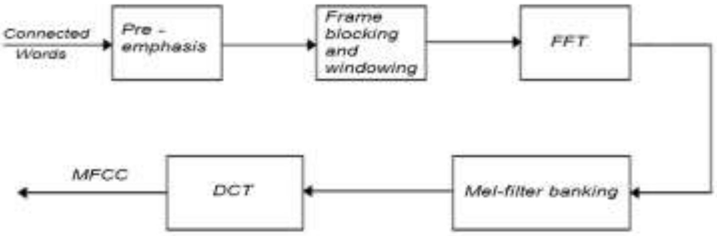


Figure 4: Steps in MFCC feature extraction

The above figure 4 shows the steps involved in the feature extraction phase using Mel Frequency Cepstral Coefficients. A pre-emphasis filter is used to compensate the high-frequency part of the speech signal, which was suppressed during the human sound production mechanism. The speech signal is divided into a sequence of frames where each frame of 10-12ms. These frames are analyzed independently and represented by a single feature vector. In order to reduce the discontinuities of the speech signal at the edges of each frame overlapping of 3ms frame size is applied. A tapered Hamming window is applied. To convert each frame of N samples from time domain into frequency domain, Fourier Transform is applied. A set of triangular filters called the Mel Filter Bank is used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Discrete Cosine Transform (DCT) is used to convert the log Mel spectrum into time domain. Using the above procedure the MFCC coefficients are calculated [10] [11].

Vector Quantization

Vector quantization (VQ) is a classical quantization technique from signal processing which allows the modelling of probability density functions by the distribution of prototype vectors. VQ works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms. The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error.

A simple training algorithm for vector quantization is:

- Pick a sample point at random
- Move the nearest quantization vector centroid towards this sample point, by a small fraction of the distance
- Repeat

The algorithm can be iteratively updated with 'live' data, rather than by picking random points from a data set, but this will introduce some bias if the data are temporally correlated over many samples. A vector is represented either geometrically by an arrow whose length corresponds to its magnitude and points in an appropriate direction, or by two or three numbers representing the magnitude of its components.

Classical K-means Algorithm

The K-means algorithm is proposed by MacQueen in 1967. It is a well-known iterative procedure for solving the clustering problems. It is also known as the C-means algorithm or basic ISODATA clustering algorithm. It is an unsupervised learning procedure which classifies the objects automatically based on the criteria that minimum distance to the centroid. In the K-means algorithm, the initial centroids are selected randomly from the training vectors and the training vectors are added to the training procedure one at a time. The training procedure terminates when the last vector is incorporated. The K-means algorithm is used to group data and the groups can change with time. The algorithm can be applied to VQ codebook design.

The K-means algorithm can be described as follows:

1. Randomly select N training data vectors as the initial code vectors Ci, i = 2, 1,..., N from T training data vectors.
2. For each training data vector X j, j = 2,1 ,...,T , assign X j to the partitioned set i S
3. Compute the centroid of the partitioned set that is code vector
   where Si denotes the number of training data vectors in the partitioned set i S.

If there is no change in the clustering centroids, then terminate the algorithm; otherwise, go to step 2.
There are various limitations of K-means algorithm. Firstly, it requires large data to determine the cluster. Secondly, the number of cluster, K, must be determined beforehand. Thirdly, if the number of data is a small it difficult to find real cluster and lastly, as per assumption each attribute has the same weight and it quite difficult to knows which attribute contributes more to the grouping process.

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a centroid the collection of all code words is called a codebook. Inside the kmeans algorithm we have proposed the recognition accuracies using different distance measures and the results are tabulated. Commonly Euclidian distance measure is used whereas this paper proposes other two distance measures.

Speech Identification

In this step, the speaker's voice is represented by a sequence of feature vectors which is then compared with the database i.e. the trained data set. The distance between the test sample and the centroid for each of the clusters is measured. The test sample belongs to the cluster with which the test sample has the minimum distance, for that the signal is matched. And this is identified as speech recognised. Two other types of distance measures and their performance measures are proposed in this paper along with the Euclidean distance. Cityblock and Spearman, distance measures are used for comparing the test the trained signals.

Distance Measures

In the speech recognition phase, an unknown speaker's voice is represented by a sequence of feature vectors {x_1, x_2 ….x_i), and then it is compared with the codebooks from the database. In order to identify the unknown speech, this can be done by measuring the distortion distance of two vector sets based on minimizing distance.

Euclidean distance

The Euclidean distance is the "ordinary" (Straight line) distance between two points in Euclidean space. The Euclidean distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity. The Euclidean distance between points p and q is the length of the line segment connecting them.

The formula used to calculate the Euclidean distance can be defined as following:
The Euclidean distance between two points p and q in Cartesian plane is,

$$P = (p_1, p_2 \ldots p_n) \text{ and } Q = (q_1, q_2 \ldots q_n),$$

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\textstyle\sum_{i=1}^{n} (p_i - q_i)^2}$$

City block Distance

City block distance, considered by Hermann Minkowski in 19th century Germany, is a form of geometry in which the usual distance function of metric or Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. The city block metric is also known as rectilinear distance.

The distance, d1, between two vectors p, q in an n-dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

where (p,q)are vectors

$$\mathbf{p} = (p_1, p_2, \ldots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \ldots, q_n)$$

For example, in the plane, the taxicab distance between (p_1, p_2) and (q_1, q_2) is $|p_1 - q_1| + |p_2 - q_2|$.

Spearman Distance

As in the case of the Pearson correlation, a distance measure corresponding to the Spearman rank correlation can be defined as

$$d_s \equiv 1 - r_S$$

where *rs* is the Spearman rank correlation.

The Spearman rank correlation is an example of a non - parametric similarity measure. To calculate the Spearman rank correlation, each data value is replaced by their rank if the data in each vector is ordered by their value. Then the Pearson correlation between the two rank vectors instead of the data vectors is calculated. Weights cannot be suitably applied to the data if the Spearman rank correlation is used, especially since the weights are not necessarily integers.

Data Set

The speech signals for both training and testing phases are collected from different speakers of the same age. It consists of male and female speakers of age 21-25 years who knows Kannada language well.

Dataset for Training Phase

Our database consists of male speaker and female speaker, who utters three Kannada digits (ondhu, eradu, mooru) 10 times. These signals are for isolated words. All samples are stored in wave format files with 16000Hz sampling rate. The signals are captured in the normal environment of having normal noise. The training and testing database is created and tested on noisy speech utterances. Totally 60 signals are collected - 30 signals for single word utterance and 30 signals for connected three word utterance (ondhu eradu mooru, naalku aidu aaru, elu entu ombathu). These are also treated as single word where the length of the signal is more.60 signals for male and female speakers are collected for training purpose. Two male and two female speakers were used for the database creation.

Dataset for Testing Phase

For testing purpose, again the same male speaker and same female speaker are considered, who utters isolated and connected Kannada digits another 10 times. This is repeated for both the male and female speakers. Totally 40 signals are used for testing purpose (isolated and connected signals). These signals are not included in the trained system.

SYSTEM IMPLEMENTATION

The parameters considered are shown in table 1 and the procedure is explained in this section.

Speech Acquisition

Recording Software	PRAAT software
Sample format	File stored as .wav
Frequency	16000Hz
Number of channels	1
Noise cancellation	No

Table 1: details of acquired speech

Feature Extraction

The samples are read using the audioread() function. The feature extraction technique used is MFCC.

[y,Fs] = audioread(filename)

This reads data from the file named filename, and returns sampled data, y, and a sample rate for that data, Fs.
[FMatrix]=mfccf(num,s,Fs)

This computes and returns the MFCC coefficients for a speech signal s where num is the required number of MFCC coefficients. In our experiment num is set to 13. First 13 coefficients are tested for different distance measures.
M = mean(A)

This returns the mean of the elements of A along the first array dimension whose size does not equal 1.

Vector Quantization

In this step, the feature vector of all the samples are sent to a function called kmeans() and they are automatically clustered. [idx,C] = kmeans(signal,k)

This performs k-means clustering to partition the observations of the n X p data matrix 'signal' into k clusters, and returns an n X 1 vector (idx) containing cluster indices of each observation.

Speech Recognition

In this step of the implementation, the feature vector of the test sample is extracted and it is compared with the centroids of the three clusters created. The MATLAB function ***pdist2 ( )*** is used to compare the two matrices.

D = pdist2(X,Y,distance)

This returns a matrix D containing the Euclidean distances between each pair of observations in the *mx*-by-*n* data matrix X and *my*-by-*n* data matrix Y. Rows of X and Y correspond to observations, columns correspond to variables. D is an *mx*-by-*my* matrix, with the (*i*,*j*) entry equal to distance between observation *I i*n X and observation *j i*n Y.

X: The feature vector of the test sample.

Y: The feature vector containing centroids of the clusters of training sample.

Distance measures can be:

'euclidean': Euclidean distance

D = pdist2(X,Y,'euclidean')

'cityblock': City block metric

D = pdist2(X,Y,'cityblock')

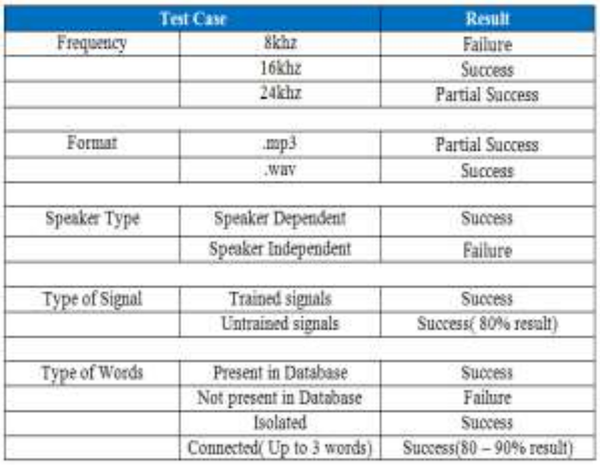
'spearman': One minus the sample Spearman's rank correlation between observations, treated as sequences of values.

D = pdist2(X,Y,'spearman')

RESULTS AND CONCLUSIONS

Testing

The following are the few test cases and their results for which the system was tested:



Test Case		Result
Frequency	8khz	Failure
	16khz	Success
	24khz	Partial Success
Format	.mp3	Partial Success
	.wav	Success
Speaker Type	Speaker Dependent	Success
	Speaker Independent	Failure
Type of Signal	Trained signals	Success
	Untrained signals	Success(80% result)
Type of Words	Present in Database	Success
	Not present in Database	Failure
	Isolated	Success
	Connected(Up to 3 words)	Success(80 – 90% result)

Table 2: Test cases and their result

Testing single word recognition

The following tables and graphs show the recognition rate for trained and untrained signals for single word recognition for three types of distance measures using Vector quantization Technique.

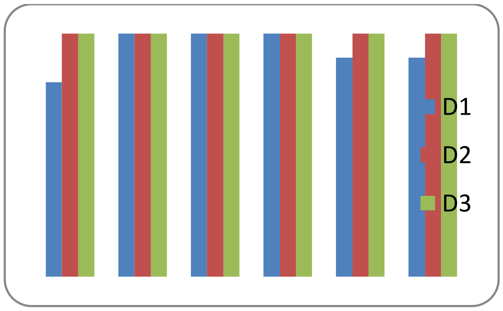
D1 = Euclidean Distance measure
D2 = Cityblock Distance measure
D3 = Spearman Distance measure

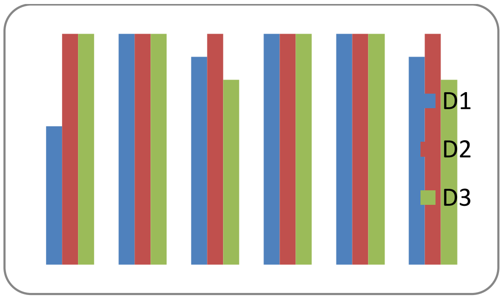
TRAINED SIGNALS	*NO. of SIGNALS*	*MALE*			*FEMALE*		
Distance Measures		*Isolated words*			*Isolated Words*		
		Ondhu	*Eradu*	*Mooru*	*Ondhu*	*Eradu*	*Mooru*
D1	*10*	*80%*	*100%*	*100%*	*100%*	*90%*	*90%*
D2	*10*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*
D3	*10*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

Table 3: Single Word Recognition (Trained)

UNTRAINED SIGNALS	NO. of SIGNALS	MALE			FEMALE		
Distance Measures		Isolated words			Isolated Words		
		Ondhu	Eradu	Mooru	Ondhu	Eradu	Mooru
D1	10	60%	100%	90%	100%	100%	90%
D2	10	100%	100%	100%	100%	100%	100%
D3	10	100%	100%	80%	100%	100%	80%

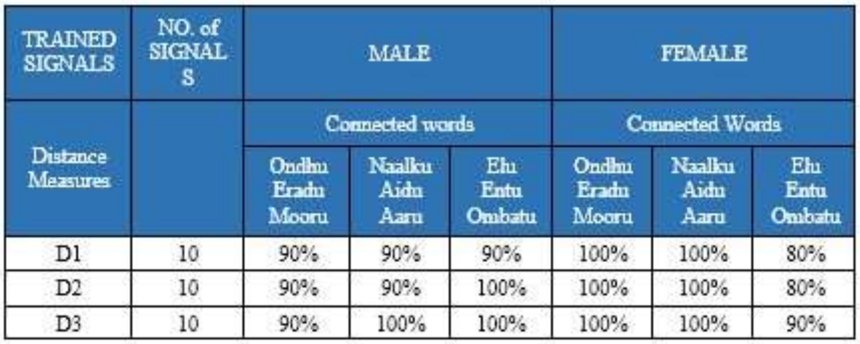
Table 4: Single Word Recognition (Untrained)





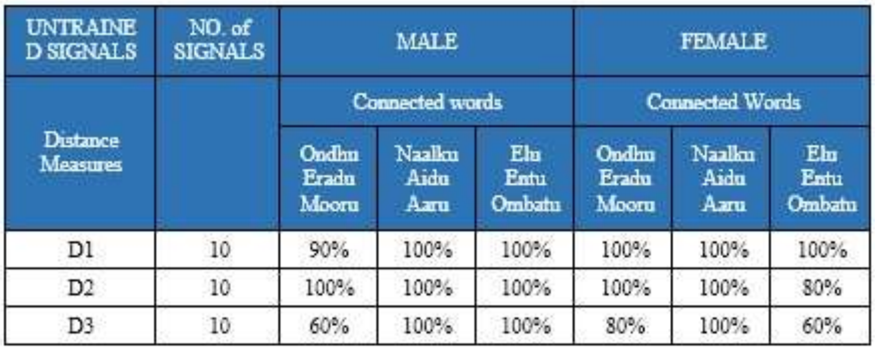
Testing connected word recognition

The following tables show the recognition rate for trained and untrained signals for connected words recognition for three types of distance measures using Vector quantization Technique



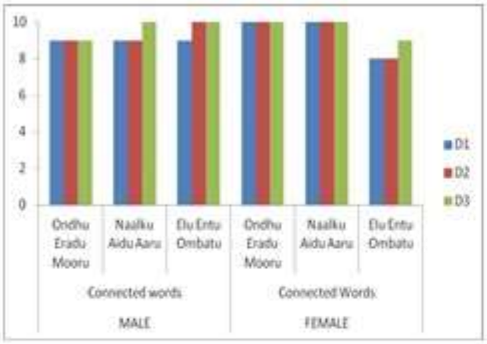
TRAINED SIGNALS	NO. of SIGNAL S	MALE			FEMALE		
Distance Measures		Connected words			Connected Words		
		Ondhu Eradu Mooru	Naalku Aidu Aaru	Elu Entu Ombatu	Ondhu Eradu Mooru	Naalku Aidu Aaru	Elu Entu Ombatu
D1	10	90%	90%	90%	100%	100%	80%
D2	10	90%	90%	100%	100%	100%	80%
D3	10	90%	100%	100%	100%	100%	90%

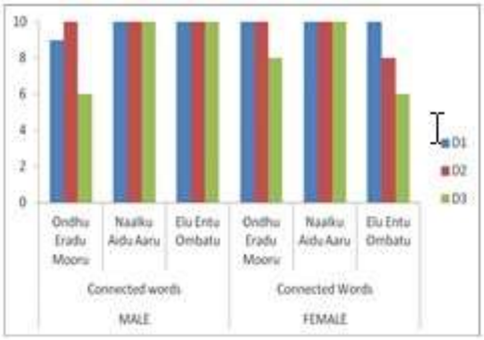
Table 5: Connected Word Recognition (Trained)



UNTRAINE D SIGNALS	NO. of SIGNALS	MALE			FEMALE		
Distance Measures		Connected words			Connected Words		
		Ondhu Eradu Mooru	Naalku Aidu Aaru	Elu Entu Ombatu	Ondhu Eradu Mooru	Naalku Aidu Aaru	Elu Entu Ombatu
D1	10	90%	100%	100%	100%	100%	100%
D2	10	100%	100%	100%	100%	100%	80%
D3	10	60%	100%	100%	80%	100%	60%

Table 6: Connected Word Recognition (Untrained)





CONCLUSIONS AND FUTURE ENHANCEMENT

Three different distance measures namely, Euclidean Distance, Cityblock Distance and Spearman Distance, were applied for male and female voice for both trained and untrained data set. It is found that "CITYBLOCK" distance measure obtains better accuracy in recognizing the signals in the data set in case of Connected Speech Recognition when compared to other distance measures. For trained both Euclidean and Cityblock distance measures gives the same performance. From this analysis, it is observed that City block distance and Euclidean measure can be used alternatively. For the limited lengthy signals where grammar builder is not used, city block distance measure can be proposed for template matching method.

The system can be further enhanced for different lengths of the speech signal, different speakers, for different languages and for confusable datasets also. The system can also be tested with more number of MFCC coefficients taking 26 or 39 features of the speech signal. For the limited lengthy words (connected), the performance the speech recognition system can be analyzed.

ACKNOWLEDGEMENTS

We would like to thank Dr. Syed Shakeeb Ur Rehman, Principal, Sri Jayachamarajendra College of Engineering and Dr. C. N. Ravi Kumar, HOD, Department of Computer Science and Engineering for their support. We are extremely grateful and thankful to our beloved professor M.A. Anusuya for her guidance, and providing us this opportunity. Lastly, we are thankful to our parents and friends for their motivation and encouragement.

REFERENCES

[1] Patel, Kashyap, and R.K.Prasad, "Speech Recognition and Verification Using MFCC & VQ", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013

[2] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.

[3] Santosh k. Gaikwad, Bharti W. Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", international journal of computer applications, November 2010.

[4] Rabiner, Lawrence R., and Biing-Hwang Juang, "Fundamentals of speech recognition" Vol. 14. Englewood Cliffs: PTR Prentice Hall, 1993

[5] Reddy, D. Raj. "Speech recognition by machine: A review", Proceedings of the IEEE 64.4: 501-531, 1976

[6] Gaikwad, Santosh K., Bharti W. Gawali, and Pravin Yannawar, "A review on speech recognition technique", International Journal of Computer Applications, vol.10, no. 3, pp. 16-24, 2001

[7] Kurian, Cini "A Survey on Speech Recognition in Indian Languages, "International Journal of Computer Science & Information Technologies 5.5, 2014.

[8] Meseguer, Noelia Alcaraz, "Speech analysis for automatic speech recognition, "Norwegian University of Science and Technology, Master's Thesis 109 (2009).

[9] Jackson, Muhirwe. "Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language." A Project Report Submitted in Partial Fulfillment of the Requirements for the Award of the Degree Master of Science in Computer Science of Makerere University August (2005).

[10] http://practicalcryptography.com/miscellaneous/machinelearning/guide-mel-frequency-cepstral-coefficients-mfccs/

[11] http://en.wikipedia.org/wiki/Mel-frequency-cepstrum

[12] http://mirlab.org/jang/matlab/toolbox/asr/

[13] http://in.mathworks.com/