Voice and Speech Recognition for Tamil Words and Numerals

V. S. Dharun, M. Karnan
Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.
Prof and Head, Department of Computer Science and Engineering, Tamilnadu College of Engineering, Coimbatore Tamilnadu, India.

***ABSTRACT*:** Voice Recognition is often confused with Voice and speech Recognition, which is the translation of spoken words (voice and speech) into text. But the identification of "who" is speaking is not the same activity as the recognition of what words are being spoken.

The main objective of this research is to develop a system for voice recognition in Tamil Word and Numeral using Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW).

Tamil language has 247 letters, but most of them are derived from the 12 vowels and 18 consonants. The other 216 letters are made by combining the sounds of a vowel and a consonant. Each letter is having unique sound.

Voice recognition is the system by which sounds, words or phrases spoken by humans are converted into electrical signals and these signals are transformed into coding patterns to which meaning has been assigned.

To extract valuable information from the voice and speech signal, make decisions on the process, and obtain results, the data needs to be manipulated and analyzed. This research work presents the feasibility of MFCC to extract features and DTW to compare the Tamil words and numerals test patterns. The extraction and matching Process is implemented right after the Pre Processing or filtering signal is performed.

I. INTRODUCTION

For the past forty years, voice and speech recognition research has been characterized by the steady accumulation of small incremental improvements. There has also been a trend to change focus towards more difficult tasks due both to progress in voice and speech recognition performance and to the availability of faster computers.

This research attempts to take advantage of the fact that in many applications there is a large quantity of speech data available, up to millions of hours. It is too expensive to have humans transcribe such large quantities of speech, so the research focus is on developing new methods of machine learning that can effectively utilize large quantities of unlabeled data. Another area of research is better understanding of human capabilities and to use this understanding to improve machine recognition performance.

Because of their limitations and high cost, voice recognition systems have traditionally been used only in a few specialized situations. For example, such systems are useful in instances when the user is unable to use a keyboard to enter data because his or her hands are occupied or disabled. Instead of typing commands, the user can simply speak into a headset. Increasingly, however, as the cost decreases and performance improves, voice and speech recognition systems are entering the mainstream and are being used as an alternative to keyboards

LITERATURE SURVEY

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. The research in automatic voice and speech recognition by machine has attracted a great deal of attention over the past five decades.

- In the late 1960" s, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC), which greatly simplified the estimation of the vocal tract response from voice and speech waveforms.
- By the mid 1970" s, the basic ideas of applying fundamental pattern recognition system to voice and speech recognition, based on LPC methods, were proposed by Itakura, Rabiner and Levinson and others.
- Another system that was introduced in the late 1980" s was the idea of Artificial Neural Networks (ANN).

Digital processing of voice and speech signal and voice recognition technique is very important for fast and accurate automatic voice recognition system. The voice is a signal of Infinite information. A direct analysis and synthesizing the complex voice signal is tough due to too much information contained in the signal. Therefore the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal.

Linear Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN) and etc are evaluated with a view to identify a straight forward and effective method for voice signal. The Voice is a signal of infinite information. Nowadays it is being used for health care, telephony military and people with disabilities therefore the digital signal processes such as Feature Extraction and Feature Matching are the latest issues for study of voice signal.

The most common approaches to voice recognition can be divided into two phases: "template matching" and "feature analysis". Template matching is the simplest technique and has the highest accuracy when used properly, but it also suffers from the most limitations. As with any approach to voice recognition [1-7].

II. METHODOLOGY

The types of voice and speech differences that the speaker-independent method can deal with, but which pattern matching would fail to handle, include accents, and varying speed of delivery, pitch, volume, and inflection. Speaker-independent voice and speech recognition has proven to be very difficult, with some of the greatest hurdles being the variety of accents and inflections used by speakers of different nationalities. Recognition accuracy for speaker-independent systems is somewhat less than for speaker-dependent systems, usually between ninety and ninety five percent. Voice Recognition based on the speaker can be classified into two types namely: Speaker-dependent and Speaker-independent.

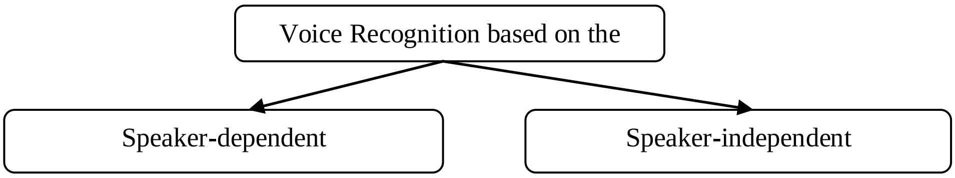


Fig1: Voice Recognition based on the speaker

Voice Recognition based on the speaker words can be classified into three types namely: Discrete, Connected and Continuous speech

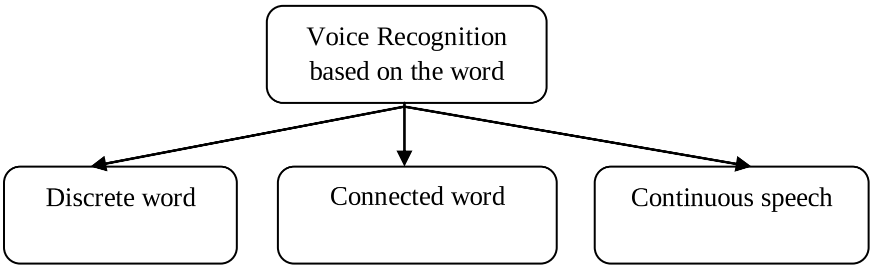


Fig 2: Voice Recognition based on the speaker words

III. DISCRETE WORD

Another way to differentiate between voice recognition systems is by determining if they can handle only discrete words, connected words, or continuous voice and speech. Most voice recognition systems are discrete word systems, and these are easiest to implement. For this type of system, the speaker must pause between words. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs.

IV. CONNECTED WORD

In a connected word voice recognition system, the user is allowed to speak in multiple word phrases, but he or she must still be careful to articulate each word and not slur the end of one word into the beginning of the next word. Totally natural, continuous voice and speech includes a great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words. A voice and speech recognition system that handles continuous voice and speech is the most difficult to implement.

Speech system is there are two types namely Discrete and Continuous

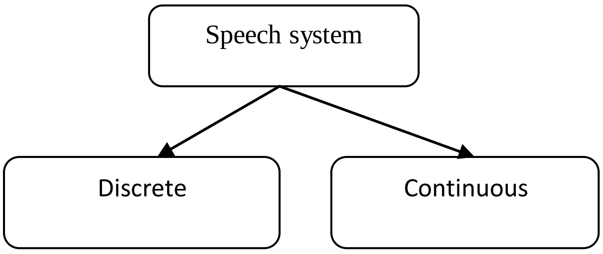


Fig3: Speech system

The template matching method of voice recognition is based on the general principles of digital electronics and basic computer programming. To fully understand the challenges of efficient speaker- independent voice recognition, the fields of phonetics, linguistics, and digital signal processing should also be explored.

The most powerful systems can recognize thousands of words. However, they generally require an extended training session during which the computer system becomes accustomed to a particular voice and accent. Such systems are said to be speaker dependent. Many systems also require that the speaker speak slowly and distinctly and separate each word with a short pause. These systems are called discrete voice and speech systems. Recently, great strides have been made in continuous voice and speech systems voice recognition systems that allow you to speak naturally.

- The voice recognition is available through feature analysis and this technique usually leads to speaker-independent voice recognition.
- Instead of trying to find an exact or near-exact match between the actual voice input and a previously stored voice template.
- This method first processes the voice input using Fourier transforms or Linear Predictive Coding (LPC), then attempts to find characteristic similarities between the expected inputs and the actual digitized voice input. These similarities will be present for a wide range of speakers, and so the system need not be trained by each new user.

OVERVIEW

- The first phase is for the user to speak a word or phrase into a microphone.
- The electrical signal from the microphone is digitized by an Analog-to-Digital (A/D) converter, and is stored in memory.
- To determine the "meaning" of this voice input, the computer attempts to match the input with a digitized voice sample, or template that has a known meaning. This technique is a close analogy to the traditional command inputs from a keyboard.
- The program contains the input template, and attempts to match this template with the actual input using a simple conditional statement.

NUMERAL RECOGNITION SYSTEM

There are several kinds of parametric representation of the acoustic signals. Among them the Mel-Frequency cepstral Coefficient (MFCC) is most widely used. We have developed the recognition system using MFCC and DTW

TRAINING

Since each person's voice is different, the program cannot possibly contain a template for each potential user, so the program must first be "trained" with a new user's voice input before that user's voice can be recognized by the program. During a training session, the program displays a printed word or phrase, and the user speaks that word or phrase several times into a microphone.

The program computes a statistical average of the multiple samples of the same word and stores the averaged sample as a template in a program data structure. With this approach to voice recognition, the program has a "vocabulary" that is limited to the words or phrases used in the training session, and its user base is also limited to those users who have trained the program. This type of system is known as speaker dependent. It can have vocabularies on the order of a few hundred words and short phrases, and recognition accuracy can be about ninety five percent.

FEATURE EXTRACTION AND MATCHING

Feature extraction techniques are emerging such as pith-synchronous signal analysis, phase spectrum based acoustic features, etc. The optimization of the existing features may draw attention

Feature matching in this phase, MFCC coefficients of both the voice and speech signals are compared using the concept of Dynamic Time Warping. This technique is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis.

One of the earliest approaches to isolated word voice and speech recognition was to store a prototypical version of each word (called a template) in the vocabulary and compare incoming voice and speech with each word, taking the closest match.

Comparing the template with incoming voice and speech might be achieved via a pair wise comparison of the feature vectors in each. The total distance between the sequences would be the sum or the mean of the individual distances between feature vectors. The problem with this approach is that if constant window spacing is used, the lengths of the input and stored sequences are unlikely to be the same.

The Dynamic Time Warping technique achieves this goal; it finds an optimal match between two sequences of feature vectors which allows for stretched and compressed sections of the sequence. A single recording of each word is used as the basis for the stored template in a DTW based recognizer. This approach will not be very robust since it takes no account of the variability of different utterances and does not ensure that the template is representative of the class as a whole.

A voice of Tamil word and numeral analysis is done after taking an input through microphone from a user. The design of the system, involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word.

The research work is to build a Tamil word and numeral recognition tool for Tamil language. This is an isolated word voice and speech recognition tool.

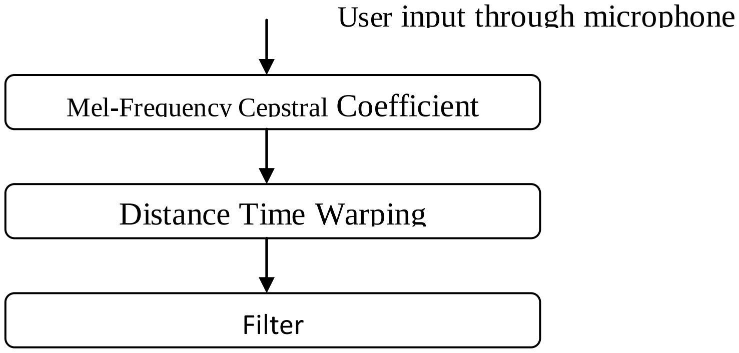

User input through microphone
Mel-Frequency Cepstral Coefficient
Distance Time Warping
Filter

Fig4: Feature Extraction

This work has discussed two phases used for voice recognition system which are important in improving its performance.

- First phase provides the information, to extract MFCC coefficients from the voice signal
- Second phase endow with the technique to compare or match them with the already fed user's voice features using DTW (dynamic time warping technique).

In voice processing, the Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and The Mel-Frequency Cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

MFCCs are derived as follows:

- Take the Fourier Transform of (a windowed excerpt of) a signal.
- Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
- Take the logs of the powers at each of the Mel frequencies.
- Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

MFCCs are used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone.

A cepstrum is the result of taking the Fourier transform (FT) of the logarithm of the estimated spectrum of a signal. There is a complex cepstrum, a real cepstrum, a power cepstrum, and phase cepstrum. The power cepstrum in particular finds applications in the analysis of human speech. The name "cepstrum" was derived by reversing the first four letters of "spectrum". Operations on cepstra are labeled quefrency analysis, cepstral analysis. The power cepstrum of a signal is defined as the squared magnitude of the Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal.

Power cepstrum of signal $\mathbf{=F\ \{Log([F\{f(t)\}]^2)\}]^2}$

The cepstrum is a representation used in homomorphic signal processing, to convert signals (such as a source and filter) combined by convolution into sums of their cepstra, for linear separation.

The result is called the Mel-Frequency Cepstrum or MFC (its coefficients are called Mel-Frequency Cepstral Coefficients, or MFCCs). It is used for voice identification, pitch detection and much more. The independent variable of a cepstral graph is called the quefrency. The quefrency is a measure of time, though not in the sense of a signal in the time domain.

DYNAMIC TIME WARPING (DTW)-BASED SPEECH RECOGNITION

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed.

A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g., time series) with certain restrictions. That is, the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

Dynamic time warping (DTW) is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds.

In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in t This example illustrates the implementation of dynamic time warping when the two sequences are strings of discrete symbols. d(x, y) is a distance between symbols, i.e. d(x, y) = | x - y |.

These both techniques have been worked out for same voice and speech signals as well as for different voice and speech signals and it have been found that if both voice and speech signals are same the cost will be 0 and if voice and speech signal are of different voices then cost will definitely have some value which shows the mismatching of the signals.

While voice and speech recognition is the process of converting voice and speech to digital data, voice recognition is aimed toward identifying the person who is speaking. Voice recognition works by analyzing the features of voice and speech that differ between individuals. Everyone has a unique pattern of voice and speech stemming from their anatomy (the size and shape of the mouth and throat) and behavioral patterns.

The applications of voice recognition are markedly different from those of voice and speech recognition. Most commonly, voice recognition system is used to verify a speaker's identity or determine an unknown speaker's identity. Speaker verification and speaker identification are both common types. Speaker verification is the process of using a person's voice to verify that they are who they say they are. Essentially, a person's voice is used like a fingerprint. [9-13]

V. CONCLUSION

The Voice and speech is the most prominent and natural form of communication between humans. There are various spoken the excitation signal is spectrally shaped by a vocal tract Equivalent filter. The outcome of this process is the sequence of exciting signal called voice and speech.

The digitized samples are then processed using MFCC to produce Tamil word and numeral features. After that, the coefficient of Tamil numeral features can go through DTW to select the pattern that matches the database and input frame in order to minimize the resulting error between them. The popularly used cepstrum based methods to compare the pattern to find their similarity are the MFCC and DTW. The MFCC and DTW features can be implemented using MATLAB.

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. Fast Fourier Transform to convert each frame of N samples from time domain into frequency domain FFT is being used. The Fourier Transform is used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain.

Discrete Cosine Transform (DCT) this is the process to convert the log Mel spectrum into time domain using DCT. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. Delta energy and delta spectrum the voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented.

Feature matching technique is based on Dynamic Programming and DTW. This technique is used for measuring similarity between two time series which may vary in time or speed. The theoretical and experimental comparisons of the methods are finding an obviously superior feature combination technique.

Experiments and results:

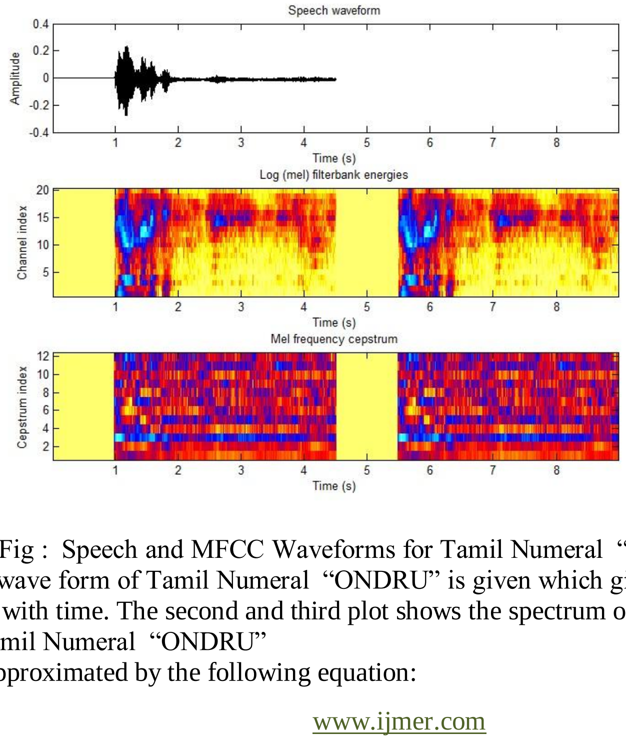


Fig : Speech and MFCC Waveforms for Tamil Numeral "ONDRU"

In the above figure first the wave form of Tamil Numeral "ONDRU" is given which gives the variation of amplitude of speech signal in accordance with time. The second and third plot shows the spectrum of Log (mel) filter bank energies Mel Frequency Cepstrum for Tamil Numeral "ONDRU"

The mel frequency can be approximated by the following equation:

$$mel(f)=2595\cdot\log_{10}(1+\frac{f}{700})$$

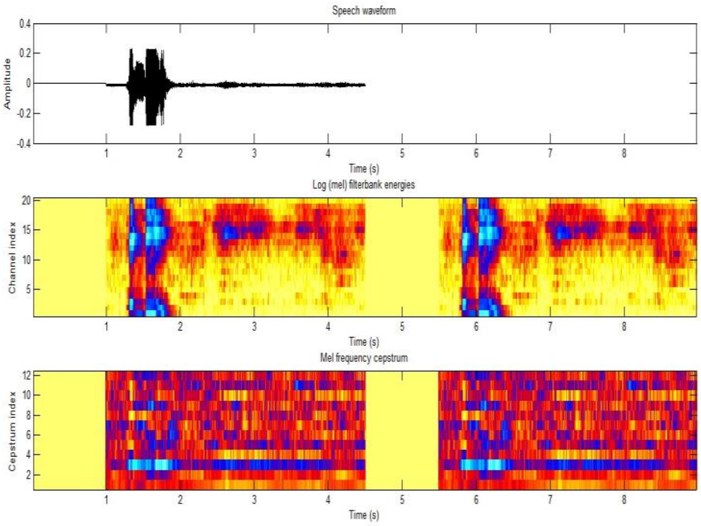
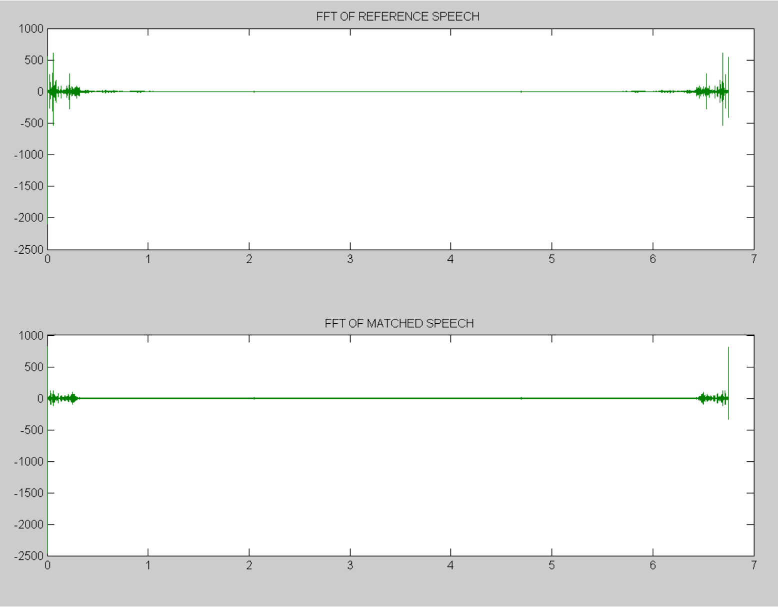


Fig: Speech and MFCC Waveforms for Tamil Numeral "IRANDU"

In the above figure first the wave form of Tamil Numeral "IRANDU"is given which gives the variation of amplitude of speech signal in accordance with time. The second and third plot shows the spectrum of Log (mel) filter bank energies and Mel Frequency Cepstrum for Tamil Numeral "IRANDU"

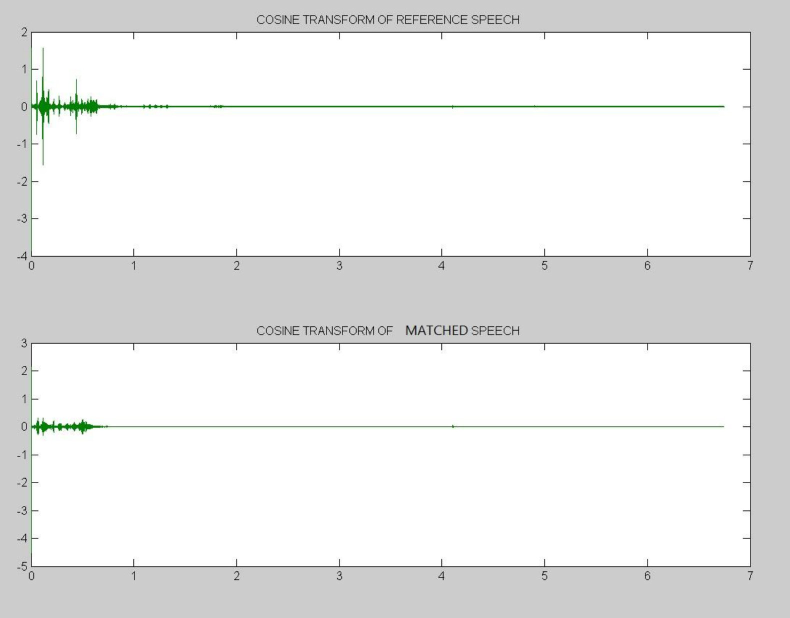
The mel frequency can be approximated by the following equation:

$$mel(f)=2595\cdot\log_{10}(1+\frac{f}{700})$$



The above figure shows the FFT of the Original speech and matched speech. Here the FFT is calculated for spectral estimation of the speech signal. FFT is calculated using the formula

$$X_k=\sum_{n=0}^{N-1}x_n\cdot e^{\frac{-2\pi ikn}{N}},k=0,\ldots,N-1$$



The above plot shows the discrete cosine transform of the original speech and the matched speech. The process of discrete cosine transform is carried out before calculating MFCC.DCT can be calculated using the formula

$$c[n]=\sum_{i=1}^{M}\log(Y(i))\cdot\cos(\frac{\pi n}{M}(i-\frac{1}{2}))$$

Distance coordinates for "IRANDU" matching speech

0	0	16	4777	9195	16028
16	16	0	5329	7378	14379
4777	4777	5329	0	484	486
9195	6986	7378	484	0	441
16028	13819	12170	486	441	0

The above distance coordinates give the distance between the spoken speech and the reference if the distance coordinate is zero it represents that the speech matches

Output window for "IRANDU" in matching speech

VOICE MATCHED

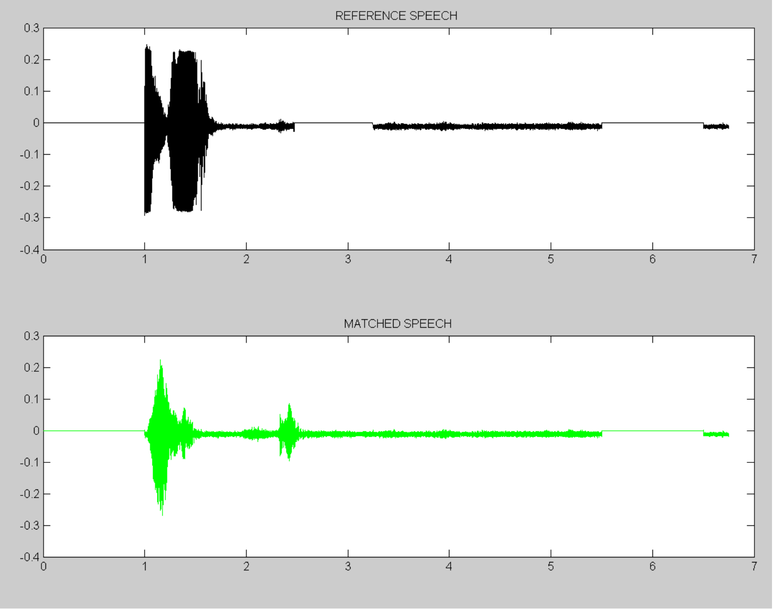
Distance coordinates for "IRANDU" matching speech

1	3	12	4912	9520	16585
18	17	3	5332	7513	14704
4779	4778	5332	3	487	489
9197	6988	7379	487	3	444
16030	13821	12172	489	444	3

Output window for "IRANDU" in matching speech

If the distance coordinate is a non-zero value which means that Spoken speech does not match with the template

VOICE NOT MATCHED



the above plot shows the matching speech spoken two different users where the first one is the reference speech and the second one is the word spoken by the speaker. Both the speech are matched by the concept of Dynamic time warping (DTW).

SPEECH BY USER	REFERENCE SPEECH	DISTANCE
ONDRU	PUJJIUM	94.968
	ONDRU	0
	IRANDU	116.780
	NANGU	56.743
	AYNTHU	76.873

COMPARISON OF SPEECH BY USER AND REFERENCE SPEECH

SPEECH BY USER	REFERENCE SPEECH	DISTANCE
AMMA	AMMA	0
	APPA	32.981
	AKKA	16.790
	ANNA	36.351
	THAMBHI	154.873

The above table compares the speech spoken by the user and the reference speech templates using DTW and obtains the distance betweeen the speech signals. The table shows that if there is a matching speech the distance is zero else a non zero value turns up as distance value.

REFERENCES

[1] Adams, Russ, Sourcebook of Automatic Identification and Data Collection, Van Nostrand Reinhold, New York, 1990.

[2] Cater, John P., Electronically Hearing: Computer Voice and speech Recognition, Howard W. Sams & Co., Indianapolis, IN, 1984.

[3] Fourcin, A., G. Harland, W. Barry, and V. Hazan, editors, Voice and speech Input and Output Assessment, Ellis Horwood Limited, Chichester, UK, 1989.

[4] Yannakoudakis, E. J., and P. J. Hutton, Voice and speech Synpaper and Recognition Systems, Ellis Horwood Limited, Chichester, UK, 1987.

[5] Christopher Hale, CamQuynh Nguyen,"Voice Command Recognition Using Fuzzy Logic" ,Motorola, Austin, Texas 78735, pp 608-613,ISBN no: 0-7803-2636-9.

[6] Hubert Wassner and Gerard Chollet, "New Time Frequency Derived Cepstral Coefficients For Automatic Voice and speech Recognition", 8thEuropean Signal Processing Conference (Eusipco'96).

[7] Marco Grimaldi and Fred Cummins,"Speaker Identification Using Instantaneous Frequencies" ,IEEE Transactions On Audio, Voice and speech, And Language Processing, VOL. 16, NO. 6, pp 1097-1111, ISBN: 1558-7916, August 2008.

[8] Mahdi Shaneh addnd Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ Techniques" ,World Academy of Science, Engineering and System 57 2009, pp 534-538.

[9] Muda Lindasalwa, Begam Mumtaj and Elamvazuthi I.," Voice Recognition Techniques using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal Of Computing, Volume 2, Issue 3, pp 138-143, ISSN 2151-9617, March 2010.

[10] Norhaslinda Kamaruddin and Abdul Wahab, "Voice and speech Emotion Verification System (Sevs) Based On MFCC For Real TimeApplications", School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.

[11] Paulraj M P1, Sazali Bin Yaacob1, Ahamad Nazri2 and Sathees Kumar1,"Classification of Vowel Sounds Using MFCC and FeedForward Neural Network" ,5th International Colloquium on Signal Processing & Its Applications (CSPA), pp 60 -63, ISBN: 978-1-4244-4152-5, March 2009.

[12] Rozeha A. Rashid, Nur Hija Mahalin, Mohd Adib Sarijari and Ahmad Aizuddin Abdul Aziz, "Security System Using Biometric System: Design and Implementation of Voice Recognition System (VRS)",Proceedings of the International Conference on Computer and Communication Engineering, pp 898-902, ISBN :978-1-4244-1692-9, May 2008.

[13] Suzuki H., Zen H., Nunkuku Y., Miyajima C., Tokuda K., and Kitumuru I:,"Voice and speech Recognition Using Voice-Characteristic dependent Acoustic Models" ,lCASSP 2003,pp 740-743,ISBN: 0-7803-7663-3103,2003.