DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4912

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MULTIPLE SPEAKERS SPEECH RECOGNITION FOR SPOKEN DIGITS USING MFCC AND LPC BASED ON EUCLIDEAN DISTANCE

Ms. Munazah Gul
M. Tech Student
Department of Electronics and Communication,
Swami Devi Dyal Inst. of Engg. &Technology
Kurukshetra University, Kurukshetra

Sandeep Sangwan
Assistant Professor in Electronics and Communication,
Swami Devi Dyal Inst. Of Engg. & Technology,
Kurukshetra University, Kurukshetra

Muheet Ahmed Butt
Scientist, PG Department of Computer Sciences,
University of Kashmir, Srinagar

Majid Zaman
Scientist, Directorate of IT&SS.
University of Kashmir, Srinagar

Abstract: The fundamental part of human life that acts as one of the five senses of human body is speech recognition, for this reason the applications that come forth on the basis of speech recognition, have high level of favourable reception. This paper has endeavoured to examine multiple steps encompassed in artificial speech recognition by man-machine interface. In speech recognition multiple steps that we persued are distance calculation, feature extraction, dynamic time wrapping. Examining the similarity measuring algorithms in ASR systems was the most comprehensive aim of the proposed research. The feature vectors such as LPC, and MFCC were evaluated in the first place. Once the operations were carried we probed MFCC specifically. Moreover it was designated as the preferred mode of feature vector coding as they follow the human ear's reaction to the sound signals. Many new techniques of distance measurement were established and a comparison was drawn between them and thus the conclusion was made that Euclidean distance measure is a favoured one when the template database of sound is very poor. We carried out a rapid investigation of dynamic time wrapping algorithm and found the slightest path between two sounds. Then we devised a compact model by writing a simple code which was fit to identify small set of isolated words. The proposed speech recognition methodology has been implemented for multiple speakers also.

Keywords: The fundamental part of human life that acts as one of the five senses of human body is speech recognition

1. INTRODUCTION

In past years a considerable advancement in ASR has generated many practical applications viz., user-friendly speech interfaces in control consoles of cars, credit card number recognition and the verbal selection of menus over the telephone. Although after 50 years worth of the efforts utilized and significant advances in ASR notwithstanding, today also the robust speech recognition for human interface pursues to be a testing problem. In state of unfavourable conditions the performance of the modern speech recognizers perhaps prove to be poor, mostly when classifiers are directed under high signal-to-noise ratio (SNR) environments such as noise-free chambers (typically where SNR $\geq$ 30 dB) and performed in real-world surroundings of fairly lower SNR [1].On contrary under like training and operating situations, a robust human listener's performance is generally most stable on average. Unlikely great number of researchers concur that human-quality; adaptively-learning and noise-robust machines that elucidate and recognize human speech will not be attained in the near future [1,2].,Although ,gradual advancement heading towards this aim in ASR are of great significance.

In order to get into the domain of speech recognition, a brief initiation to how the speech signal is generated and recognized by the human system can be considered as a starting point. The figure 1 shows the process of production from human speech to human speech recognition, between the listener and speaker [3].

Fig: 1 Speech recognition of human speech

Translation of spoken words into text is speech recognition in electronics engineering which is also called as "computer speech recognition", "automatic speech recognition" (ASR), or sometimes it is just known as "speech to text" (STT). Similarity to the human speech communication system is established by speech recognition system. The main objective of human speech communication is in changing the ideas. They are first made within the speaker's brain and then, the source word sequence is performed to be delivered through her/his text generator. Speech generator component modeles the human vocal system which turns the source into the speech signal waveform which is transferred via air to the receiver (listener), being able to get affected by some external noise sources. Speech signal gets masked by noise interference and reduces its quality [4].When the acoustical signal is understood by the human auditory system, the

listener's brain starts processing this waveform to understand its content and then,this completes the communication.

Automatic speech recognition,another name for speech recognition is the method of changing a speech signal into sequence of words by using an algorithm executed as a computer program. In other words it can be defined as the capability of a computer to receive speech in audio format and then produce its content in text format.

Speech recognition in computer domain includes various steps with issues attached with them. In computer system voice detection has numerous steps with difficulties along with them. The amplitude time waveform is a basic model in which a user creates a voice signal.

The digitized voice signal is used to get many spectral and temporal features, like time energy, fundamental frequency, mfcc, zero crossing rate etc.

Some of them are used for silence detection word boundary detection etc., which takes place during preprocessing of the voice signal and the others are used for recognition in subsequent phases by making a feature vector.

These features vectors are compared with trained and stored data model to differentiate phonemes which are further

Linked to create the target words. Depending on the probabilistic confidence these words are either accepted or rejected. After so many years, voice recognition is still a challenging field, because of various features which creates issues in the performance of speech recognition recovery. Some of them are background noise effect, speaker variability (i.e.) same words spoken by different people. So researchers during their work also go through the literature survey.

Voice is the simple way of communication between different people. The aim of this voice recognition system is to develop communication between machine and humans. Voice quality also functions to signal the speaker's emotional or attitudinal [5].

However, it seems simple, but researchers from a longtime are trying to make it possible and it has been said that it is not very achieve as it seems. It faces multidimensional issues like non stationary nature of voice large size of vocabulary, and high processing time.

2. RELATED WORK

For most, the leading and natural way of communication is human voice. Automatic speech Recognition (ASR) is association of hardware and software that saves distinct features of speech with a source of input equipment, like a microphone and other processes these substitutes to match them to input speech to simulate to interact with computers, machines or human users. In 1950s the first primitive recognizer was developed at Bell Labs and in 1960s major break thoughts came in the field of ASR. Many of these achievements are not worthy to be mentioned, because they didn't create any useful tools for voice recognition but also developed the very simple concepts on which most of the research work is based. In 1965 , Turkey and Cooley led to the development of the Fast Fourier Transform (FFT) which minimized the load of Discrete Fourier Transform (DFT) with a faster algorithm, thereby endowing the hypothetical implementations of Digital Signal Processing (DSP) custom chips [6,2]. Oppenheim, Schafer, and Stockham introduced Cepstral Analysis which executes deconvolution of the speech signal to separate an excitation sequence from an impulse response convolved with it [7]. Cepstral coefficients and many derivatives have been widely used to potray the short-term spectral envelope of speech signals so far. In late 1960s and early 1970s another method for speech analysis, called as Linear Predictive Coding (LPC) was found. Atal and Schroeder [8,1] published one of the earliest and complete papers on the application of linear prediction to speech analysis .LPC uses a pole-only filter to design the speech signal. LPC coefficients and its derivatives are extensively used for transmitting speech spectral envelope information [2].Most notably, the foundations for the statistical technique of Hidden Markov Modeling, which models an observed sequence as produced by a sequence of hidden states, dates back to the 1960s as well. However, the first prosperous applications of Hidden Markov Modeling to speech recognition were accomplished in the 1970s [2].

Baum and his colleagues matured a popular expectation-maximization (EM) algorithm, called as the Baum-Welch Re-estimation Algorithm (or Forward-Backward Algorithm),to appraisal the parameters of a Hidden Markov Model (HMM) iteratively [9,10]. Hidden Markov Models (HMM) and the Baum-Welch Re-estimation Algorithm are commonly used today .

In 1970s Dynamic Time Warping (DTW), a deterministic two way approach to the statistical HMM was implemented.The various length articulations of the same word was normalized by DWT and appertains template based attributes to speech recognition. In 1970s many distinct approaches viz.,DTW,HMM and Artificial Neural Networks(ANN) were profound for recognition of speech. Among these studies, the task of the Advanced Research Projects Agency (ARPA), are obvious achievement in that it executed a 1000-word ASR system by making use of joined speech from a few speakers with a word error rate of less than 10% [2]. In the 1980s, the project of the Defense Advanced Research Projects Agency(DARPA Project) and the major other programs managed by Texas Instruments and the Massachusetts Institute of Technology (TIMIT Project) and the National Institute of Standards and Technology (NIST) firstly condenced on the collection of large corpora used for proclaimed ability and testing speech recognizers. These large corpora were accordingly used by the ASR research community at bulk for performance similarity of distinct approaches applied to speech recognition. The ASR community certified some other essential growths in the 1980sas well. Among those, the Mel-Cepstrum suggested research introduced by Davis [11],and the Dynamic Cepstral Coefficients proposed by Furui [12] can be extra ordinary techniques for speech feature-extraction due to important improval in recognition perfection. As for the speech recognizers of the 1980s, many researchers were investigating with frame-based HMM recognizers, ANN recognizers or hybrid schemes merging HMM and ANN in remote or endless contexts of speech [2]. Most essentially, the simultaneous speech recognition systems of today still use these subprograms basiclly. Lee and Waibel labeled the question of difficulty entangled in ASR in [13]as "dimensions of difficulty.

These factors ascertain the intricacy and the specifications of an ASR system.These resources are summarized by Deller et al. that interpret speech recognition techniques

difficult. Various researchers have lately taken into consideration the task to edifice more noise-robust recognizers that may be employed in noisy environments a noisy flight cabin or within a car) with higher precision [14, 15].

3. PROPOSED METHODOLOGY

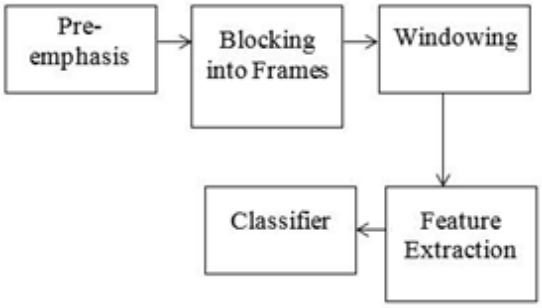


Fig: 2. Proposed method

A. Pre emphasis

The digitized speech signal is operated by a first order digital network so as to spectrally flatten the signal. This pre emphasis is carried out with ease in the time domain by considering difference.

$$\tilde{A}(n) = A(n) - a^* A(n-1)$$

a= scaling factor = 0.95, A(n)= Digitized Speech Sample, A(n-1) = Previous digitized Speech Sample, Ã(n) = Pre emphasised Speech Sample, n = No. of Samples in the whole frame.

B. Blocking into Frames

Section of N (e.g. 300) consecutive speech samples are utilizedas a single frame. Successive frames are spaced M (e.g. 100) samples independently.

$$X_l(n) = \tilde{A}(M^*l + n) \quad , \; 0 <= n <= N-1 \text{ and } 0 <= l <= L-1$$

N = Total No. of samples in a frame, M = Total No. of sample spacing between the frames. [Measure of overlap], L = Total number of frames.

C. Frame Windowing

Every frame is multiplied by an N sample window W (n). Here we make use of a hamming window. This hamming window is applied to reduce the adverse consequences of chopping an N sample section out of the running speech signal. While generating the frames the chopping of N sample from the running signal may have an untoward effect on the signal parameters. To reduce this effect windowing is performed.

$\hat{U}_l(n) = X_l(n) * W(n)$, $0 <= n <= N-1$

W(n) = Scale factor i.e. (0.54 - 0.46*Cos(2*pie*n/ N)) , $0 <= n <= N-1$

N = Total No. of samples in a frame.

The multiplicative scaling factor ensures appropriate overall signal amplitude

4. RESULT AND DISCUSSION

Here in the first place we introduced the exploration of multiple feature extraction processes. Then we made an attempt to propound the evaluation of MFCC as how it is an optimum approach of feature extraction. Subsequently we have made an effort to examine various procedures of distance measure used to evaluate the distance among the feature vectors obtained by us. Furthermore dynamic programming approach is utilized to do a slight examination of dynamic time warping. For the recognition of isolated words we attempt to introduce a small program for small speaker dependent recognition system.
Here we want to put forth that as we get attracted by the application of speech recognition in mobile phones we here make an attempt to recollect the English numerical digits from ‘zero’ to ‘nine’. As Matlab is the most methodical tool for mathematical and signal analysis so all the programming is done here.

The steps used in the Algorithm are as under

1. Create a Dataset by recording audio of spoken digits from 0 to 9 where we have taken 5 samples of each digit from multiple users where each user recorded voice has been kept in a separate .mat file in wave format. In the proposed research we have taken 3 users only but the algorithm can accommodate n users.
2. MPCC parameters are initialized as overlap size is kept at 0.5 and also MPCC matrix is initialized.
3. A training set is created for all digits by extracting various features using MPCC and the threshold speech samples are identified for faster comparison.
4. Gaussian Modeling is carried out for every digit from 0 to 9
5. Extraction of MFCC coefficient is carried out of Test Data of around voice samples per speaker.
6. Classification of MFCC is carried out using test data on Mahanalobis distance
7. LPC parameters are initialized
8. LPC Training is carried out to all voice samples from 0 to 9
9. Extraction of LPC coefficient is carried out of Test Data of around voice samples per speaker.
10. Classification of LPC is carried out using test data on Mahanalobis distance
11. Calculate the Euclidean Distance for MFCC and LPC matrices
12. The index provides us the appropriate Digit.

An important conclusion that we can make from the last set of experiments is that one of the main reasons for the need of large training databases for LPC based analysis (without filtering) ,is the large difference between the different telephone lines, which is reflected in a difference inspectral distortion.
Out of all the different options available for feature extraction we selected the MFC Coefficients as in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's

response more closely than the linearly spaced frequency bands obtained directly from the FFT (Fast Fourier Transform) or DCT (Discrete Cosine Transform). This can allow for better data processing. This feature of MFCC can be analyzed by a Matlab programme which takes in a speech waveform converts it into the MFCC coefficients and then reconstructs the waveform from the MFCC and thus compare the power spectra of the original sound and the reconstructed sound words and templates. The basic principle is to allow arrange of 'steps' in the space of (time frames in sample, time frames in template) and to find the path through that space that maximizes the local match between the aligned time frames, subject to the constraints implicit in the allowable steps. As the duration of speaking for different persons are different DTW is highly unavoidable. The most common algorithm used for this purpose is dynamic programming. Here we bring a Matlab program to calculate the DTW for two given signals, the input signal is two different versions of word 'one'.

The Experiment was carried out on 20 inputs and the results were collaborated.

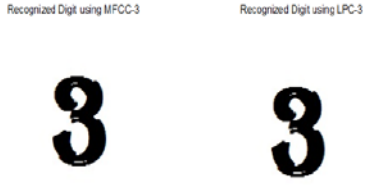
Iteration	Spoken Digit	User	Sample	MFCC	LPC	Accuracy LPC	Accuracy MFCC
1	5	Munazah	1	5	8	YES	NO
2	6	Munazah	2	6	6	YES	YES
3	7	Munazah	3	5	7	NO	YES
4	8	Munazah	4	8	6	YES	NO
5	3	Munazah	5	3	8	YES	NO
6	6	Basil	1	6	1	YES	YES
7	4	Basil	2	4	2	YES	NO
8	3	Basil	3	3	3	YES	YES
9	6	Basil	4	6	6	YES	YES
10	7	Basil	5	3	7	NO	YES
11	8	Mehak	1	8	8	YES	YES
12	9	Mehak	2	9	9	YES	YES
13	2	Mehak	3	2	6	YES	NO
14	5	Mehak	4	5	5	YES	YES
15	3	Mehak	5	3	3	YES	YES
16	8	Uzma	1	8	8	YES	YES
17	9	Uzma	2	9	2	YES	NO
18	1	Uzma	3	7	1	NO	YES
19	2	Uzma	4	2	3	YES	NO
20	5	Uzma	5	5	5	YES	YES

Accuracy %age using LPC 85%

Accuracy %age using MPCC 65%

In the above iterations of the recognition process we have seen that LPC method has more accuracy than MPCC. Thus it can be easily seen that even though Itakura-saito distance is a very good form of distance measure its performance for the case of isolated word recognition with very little database is very poor. Thus we have decided to use Euclidean distance for our purpose.

Experimental Results for Digit 3



In our Experiment LPC has shown more accuracy then MFCC

Dynamic Time Warping

Speech recognition faces difficulties one of which is that although different recordings of the same words may include more or less the similar sounds in the similar order, the exact timing – the interval of each sub word within the word - will not correspond. As a result, struggle to recognize words by ensembling them to templates will not give appropriate results if there is no secular alignment.

Although it has been mainly replaced by hidden Markov models, firstly speech recognizers used a dynamic-programming technique called Dynamic Time Warping (DTW) to reconcile divergence in timing between specimen. After analyzing the different parts of the speech recognition analysis here we try to present a small program which does two tasks. The first task is to produce a data base of templates for once spoken wards for example 'zero' to 'nine'. This is known as the training of the recognizer. The next task is to recognize. The MFCC feature coefficient is used here for reasons stated earlier Euclidean distance is used to measure the distance between the feature vectors. Here we first give the process of training. The Euclidean distance is given by:

$Dist(x,y)= \left| x-y \right| =[(x_1.y_1)^2+(x_2-y_2)^2 +\ldots\ldots\ldots+(x_n-y_n)^2]^{1/2}$

5. CONCLUSION

In this topic we at first calculated the different types of feature vectors such as LPC,RASTA and MFCC. After performing such operations we analyzed MFCC in particular, and selected it as the preferred mode of feature vector coding because they follow the human ear's response to the sound signals. We also found different methods of distance Measurement and compared them and concluded that euclidean distance measure is a preferred one when the template database of sound is very low. We also performed a quick analysis of dynamic time warping algorithm and found the least path between two sounds. Then we designed a small model by writing a simple code which was able to recognize small set of isolated words.

The performance of this model is limited by a single template generated by the training programmers, as it does not incorporate training algorithm of any sort. The performance factor can be optimized by using high quality audio devices in a noise free environment. There is a possibility that the speech can be recorded and can be used in place of the original speaker. This would not be a problem in our case because the MFCCs of the original speech signal and there corded signal re different. Finally I conclude that although the project has certain limitations, its performance and efficiency have outshined these limitations at large.

6. REFERENCES

[1] J. H. L. Hansen, J. R. Deller, and J. G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, 2000.

[2] N. Morgan andB. Gold , Speech and Audio Signal Processing, John Wiley & Sons, 2000.

[3] Meseguer, NoeliaAlcaraz, "Speech analysis for automatic speech recognition, "Norwegian University of Science and Technology, Master's Proposed research 109 (2009).

[4] Fujimoto, M. and Ariki, Y., 2000. Noisy speech recognition using noise reduction method based on Kalman filter. IEEE transactions on acoustic, speech and signal processing, vol. 3,727-1730.

[5] Yoon, T.J., Zhuang, X., Cole and Jhonsen, M.H., 2009. Voice Quality Dependent Speech Recognition. Linguistic patterns in spontaneous speech, Academia Sinica.

[6] J. W. Cooley and J. W. Tukey, "An algorithm for the machinecomputation of complex Fourier series," Mathematical Computations, Vol. 19, pp.297-301, 1965.

[7] A. V. Oppenheim, R. W. Schafer and T. G. Jr. Stockham,"Nonlinear filtering of multiplied and convolved signals," Proceedings of the IEEE,Vol. 56, No. 8, pp. 1264-1291, 1968.

[8] B. S. Atal and L. S. Hanauer, "Speech analysis and synproposed research bylinear prediction of the speech wave," Journal of the Acoustic Society of America,Vol. 50, pp. 637-655, 1971.

[9] L. E. Baum and T. Petrie, "Statistical inference for probabilisticfunctions of finite state Markov chains," Annals of Mathematical Statistics, Vol. 37,pp. 1554-1563, 1966.

[10] L. E. Baum, T. Petrie and G. Soules, "A maximization technique inthe statistical analysis of probabilistic functions of Markov chains," Annals ofMathematical Statistics, Vol. 41, pp. 164-171, 1970.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28,No. 4, pp. 357-366, 1980.

[12] S. Furui, "Speaker-independent isolated word recognition usingdynamic features of speech spectrum," IEEE Transactions on Acoustics, Speech andSignal Processing, Vol. 34, No. 1, pp. 52-59,1986.

[13] K. F. Lee and A. Waibel, Readings in Speech Recognition,Morgan-Kauffmann, Palo Alto, California, 1990.

[14] S. Moon and J. N. Hwang , "Robust speech recognition based on joint model and feature space optimization of hidden Markov models," IEEE Transactions on Neural Networks, Vol .8, No .2, 194-204, March 1997.

[15] M. Shozakai, S. Nakamura and K. Shikano, "Robust speechrecognition in carenvironments," Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP `98), Vol. 1, pp.269-272, May 1998.