International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 3, Issue 8, August 2014

COMPARATIVE STUDY OF MALE AND FEMALE VOICES USING MFCC AND DTW ALGORITHM IN SPEAKER RECOGNITION

Bhanu Priya, Sukhvinder Kaur
SWAMI DEVI DYAL INSTITUTE OF ENGINEERING AND TECHNOLOGY, BARWALA, PANCHKULA, INDIA

***Abstract*— In today's world there are a lot of applications in which voice and speaker recognition is required. Recognizing a person by her/his voice is known as speaker recognition .The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Feature extraction is the first step for speaker recognition. Many algorithms are suggested/ developed by the researchers for feature extraction. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. Applications like voice recognition in mobiles, locker's safety and conferences etc. where there is requirement for recognition of voice. In this paper, after extracting features using MFCC for speaker Recognition , we compare the voices of male and female speakers using DTW (Dynamic Time Warping) algorithm. From here the Euclidian distance is computed and compared and using warping technique in DTW we have tried to overcome the limitations of Euclidian distance..**

Index Terms— DTW, Feature Extraction, MFCC, Speaker Recognition.

I. INTRODUCTION

Speaker recognition is the task of recognizing people from their voices. Strictly speaking there is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). Speaker recognition system is categorized into speaker verification (to authenticate a claimed speaker identity from a voice signal based on speaker-specific characteristics reflected in spoken words) and speaker identification (to find the identity of a talker, in a known population of talkers, using the speech input).Speaker identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (and possibly a general world template) whereas speaker identification is a 1: N match where the voice is matched to N templates. The Speech is the most prominent and natural form of communication between humans. The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. The property of speech signal changes markedly as a function of time. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later.

The Euclidean distance between the MFCC's (Mel Frequency Cepstrum Coefficients) of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification satisfactorily.

Dynamic time warping is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and the warping which optimally deforms one of the two input series onto the other. A variety of algorithms and constraints have been discussed in the literature. The dtw package provides an unification of them; it allows R users to compute time series alignments mixing freely a variety of continuity constraints, restriction windows, endpoints, local distance definitions, and so on. The package also provides functions for visualizing alignments and constraints using several classic diagram types.

II. FEATURE EXTRACTION

Features extraction in speaker recognition is the computation of a sequence of feature vectors which provides a compact representation of the given speech signal. It is usually performed in three main stages. The first stage is called the speech analysis or the acoustic front-end, which performs spectra-temporal analysis of the speech signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.

2.1 Input Speech Signals with noise

Two input speech signals of male and female in fig. 1(a) and fig. 1(b) respectively.

ISSN: 2278 – 909X

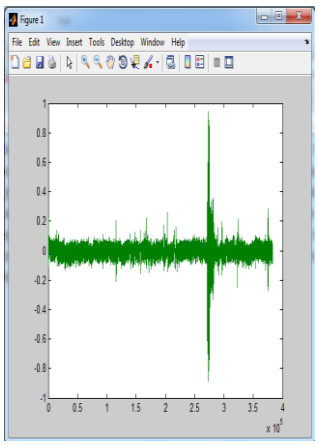

Fig. 1(a) Male voice 'm'

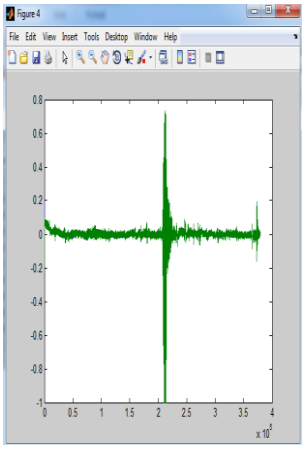

Fig. 1(b) Female voice 'm'

2.2 Input speech signals without noise

Two input speech signals after noise removal with the help of a filter is shown in fig. 2(a) and 2(b) respectively.

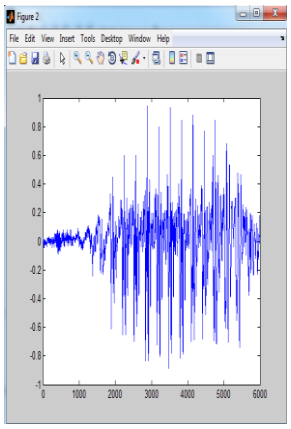

Fig. 2(a) Filtered male voice 'm'

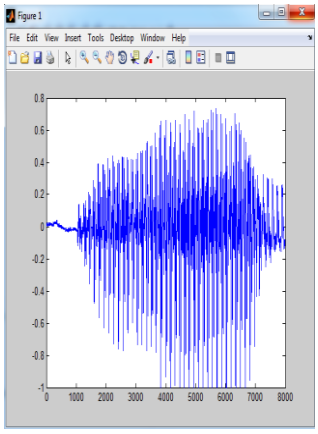

Fig. 2(b) Filtered female voice 'm'

2.3 MFCC Features Extracted using following steps:-

The following steps are followed while doing feature extraction using MFCC.

1) Pre-Emphasis
2) Framing
3) Windowing
4) Fast Fourier Transform
5) Mel filter bank processing
6) Discrete Cosine Transform
7) Delta Energy and Spectrum

The fig. 3(a) and fig. 3(b) shows the extracted features obtained from MFCC and all the features are labeled in the figure properly. After obtaining these features DTW is performed so that speaker recognition can be performed.

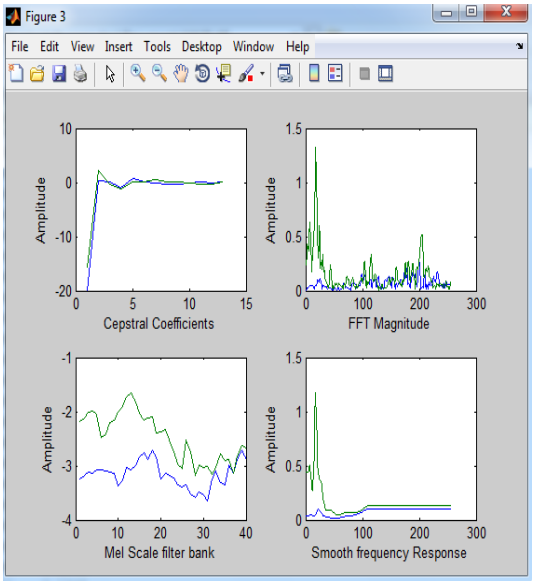

Fig. 3(a) MFCC features of male voice 'm'

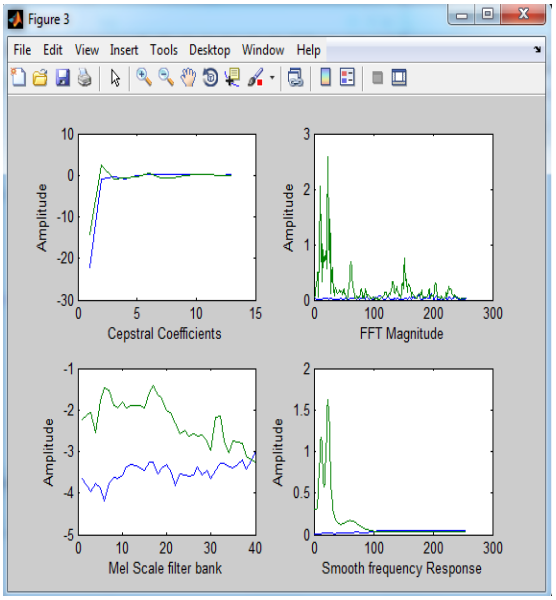

Fig. 3(b) MFCC features of female voice 'm'

After finding the features of the input voice signal using MFCC we use DTW for feature matching i.e to calculate the distance between different speakers and same speakers.

III. SPEAKER RECOGNITION

Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. Recognizing a person by her/his voice is known as speaker recognition. Since differences in the anatomical structure are an intrinsic property of the speaker, voice comes under the category of biometric identity. Using voice for identity has several advantages.

When two above voice signals i.e a male and female input is compared using their cepstral coefficients obtained from

ISSN: 2278 – 909X

MFCC then the results that were obtained are shown in Table I are:-

Input 1	Input 2	Euclidian Distance
Cepstral coefficients of male voice 'm'	Cepstral coefficients of female voice 'm'	20.1051
Cepstral coefficients of male voice 'm'	Cepstral coefficients of male voice 'm'	0
Cepstral coefficients of female voice 'm'	Cepstral coefficients of female voice 'm'	0

Table I. Comparison of Euclidian Distance between same and different speakers.

Why Dynamic Time Warping?

Any distance (Euclidean, Manhattan, …) which aligns the *i*-th point on one time series with the *i*-th point on the other will produce a poor similarity score.
A non-linear (elastic) alignment produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase in the time axis.

Let us understand with an example:
r=[1;2;3;4;5;6]
t=[0;0;1;2;3;4]
For these sequences the Euclidian distance is 20 but the DTW cost function is evaluated to be 0 by performing warping of the two signals.

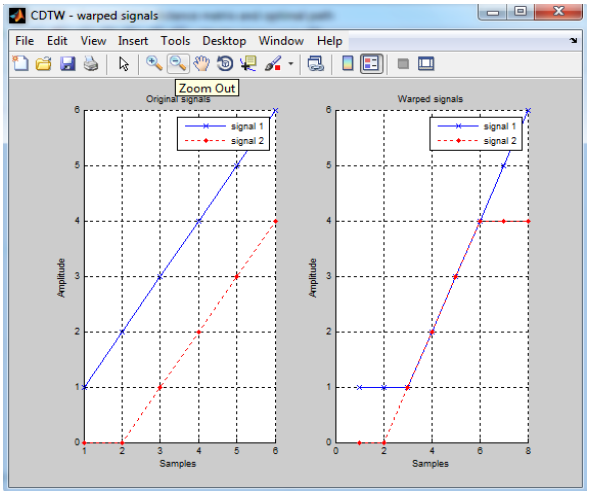

Fig.4(a) Original and warped signals of the input sequence given above as [r] and [t].

Similarly in the case of finding optimal path and warped signal in comparison of male and female voiced speech signals, we find results using their plots and study the differences.

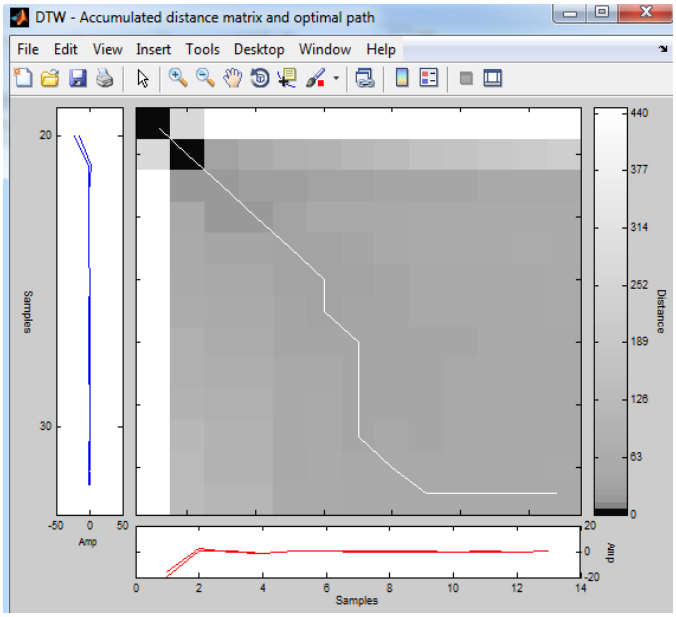

Fig. 4(b) Optimal path between above male and female speakers is calculated

The two input signals are then warped to find the common points and recognising who is speaking by comparing the distances.The white line indicates the region that is same or different. For a linear line the features are same and where the line is bnon-linear means that speakers are different.

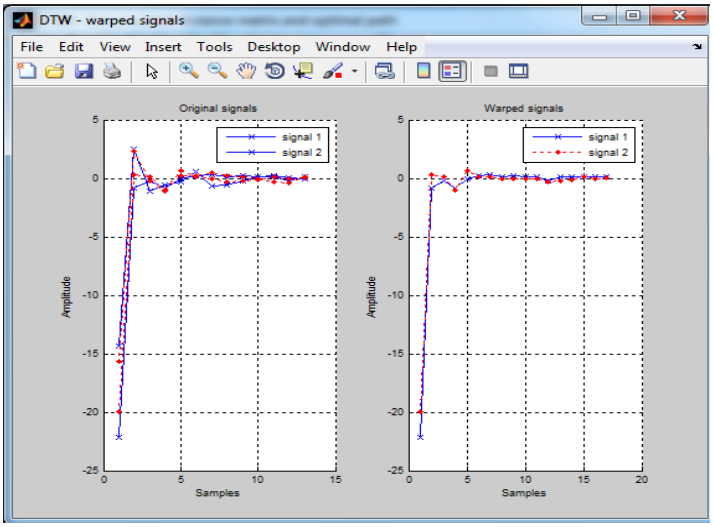

Fig. 4(c) Original and warped signals of male and female voice 'm'

The above red and blue lines show the two speech signals and their warping. The signal1 and signal2 are warped in such a way that each of them covers less distance.

IV. RESULTS

In this paper, I have tried to obtain results for speaker recognition by comparing their voices and using MFCC for feature extraction and DTW for feature matching. The results are studied and compared and shown in the graphs below. For same speakers the cost=0 and for different speakers have some value. It is studied using DTW.

ISSN: 2278 – 909X

IV. CONCLUSION AND FUTURE SCOPE

We use DTW generally for non-linear sequences where Euclidian distance do not work effectively. In future works, there is another technique called HMM which can also be applied on the extracted features and compared.

REFERENCES

[1] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah ,Wan Ahmad and Arpit Bansal, "Feature extraction using mfcc," *an international journal (sipij)* vol.4, no.4, august 2013

[2] Vibha Tiwari**,** Deptt. of Electronics Engg., Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA , **"**MFCC and its applications in speaker recognition," *International Journal On Emerging Technologies (Received 5 Nov., 2009, Accepted 10 Feb., 2010)* **,**ISSN : 0975-8364

[3] K.R. Aida–Zade, C. Ardil and S.S. Rustamov, " Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems," *Proceedings Of World Academy Of Science, Engineering And Technology* Volume 13 May 2006, ISSN 1307-6884

[4] K.R. Aida–Zade, C. Ardil and S.S. Rustamov, "Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems," *Proceedings Of World Academy Of Science, Engineering And Technology* Volume 13 May 2006, ISSN 1307-6884

[5] Anjali Bala, Kurukshetra University, Department of Instrumentation and Control engineering, H.E.C, Abhijeet Kumar, Mullana University, Department of Electronics and Communication Engineering, MMEC, "Voice Command Recognition System based on MFCC and DTW," *International Journal of Engineering Science and Technology* Vol.2(12), 2010, ISSN 7335-7342

[6] Ms Arundati, S. Mahandale and M.R Dixit, "Speaker Identification,**"** Signal & Image Processing *: An International Journal* (SIPIJ) Vol.2, No.2, June 2011

Bhanu Priya is pursuing her M.E in Communications from SDDIET, Barwala, Panchkula. She has done B.Tech in Electronics and Communications in 2012 from ACE, Mithapur, Ambala.

Er. Sukhvinder Kaur is persuing Ph.D in Communication. She is Assistant Professor in ECE at SDDIET

ISSN: 2278 – 909X