Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 143 (2018) 250–257

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Computing and Communication (ICACC-2018)

Speaker identification based on combination of MFCC and UMRT based features

Anett Antony, R. Gopikakumari

Mtech studunt, Division of Electronics Engg., School of Engineering CUSAT, Cochin Pin:682022 , India
Professor, Division of Electronics Engg.,School of Engineering CUSAT,Cochin, Pin:682022 , India

Abstract

This paper introduces an isolated word speaker identification system based on a new feature extractor and using Artificial Neural Network. The system is designed for both text independent and text dependent speaker identification system for English words. The speech is recorded using audio wave recorder. Then the preprocessing is applied for the given speech signals. UMRT is a transform which has been used for image compression. Combinations of MFCC and UMRT are taken and are used as a feature extractor. The classification of the features is done using Multi-layer perceptron with back propagation algorithm. The accuracy is taken using confusion matrix. The accuracy achieved is around 97.91% for speech dependent systems while for speech independent system the accuracy is around 94.44%.

Keywords: Speaker identification; MFCC; UMRT; ANN

1. Introduction

When somebody says something, humans have the ability to guess who the speaker is even though we might not have seen their face. This is speaker identification. Voice of almost all human beings is different. So it acts as a

Tel.: +91 9947595849.
E-mail address: anettantony94@gmail.com

10.1016/j.procs.2018.10.393

biometric measure. Nowadays Speaker identification is used in many services like voice dialing, security services, telephone shopping etc.

Speaker recognition involves speaker verification as well as speaker identification. Speaker verification is a procedure of verifying the claimed identity of a speaker based on speech signal from speaker (voiceprint) [1]. While speaker identification is the task to identify the speaker [1]. Speaker identification systems can also be classified into speech dependent systems and speech independent systems. A speech dependent system is a system in which the identification takes place on the basis of a particular text, where the people are required to read a particular text and on the basis of which the identification of the person is done. In speech independent systems the speaker can say anything but the system still identifies the particular user. Developing a speech independent system is much difficult than a speech dependent system.

Two main steps are involved in speaker identification and they are feature extraction and feature classification. In feature extraction, certain features necessary for identification of the person is extracted from the speech data. While in feature classification, the features of an unknown person is taken and is compared with the features of different speakers and thus identifies the particular speaker. An ideal speaker recognition system should have high inter speaker and low intra speaker variation, easily measurable features, less prone to noise, gives correct responses to mimicry and not dependent on other features.

In this paper a new feature extraction method has been introduced which is a combination of both Mel frequency cepstrum coefficients(MFCC), a widely used method in speaker identification and speech recognition, and Unique mapped real transform (UMRT), an extension of Mapped real transform (MRT)[4]. Multi-layer perceptron (MLP) model using Back propagation algorithm is used for feature classification. The accuracy is tested using confusion matrix.

2. Related works

In the past few years lots of research work has been done in this area. There exists different techniques for feature extraction like LPC, MFCC, LFCC and various classifiers are used to classify these speech features like DTW, GMM, VQ, SVM etc [2]. A text independent speaker identification system was developed based on wavelet analysis and neural networks. The wavelet analysis comprises discrete wavelet transform, wavelet packet transform, wavelet sub-band coding and MFCC, while for learning combinations of neural networks were used. The system improved the identification rate by 15% compared to MFCC and the identification time also reduced by 40% [1]. For text independent speaker recognition systems, a comparison of various classifiers based on statistical pattern recognition and neural networks was performed. A modified neural tree network (MNTN) was developed. The error rates achieved by both the MNTN and full search VQ were comparable, but the MNTN demonstrates a logarithmic saving in retrieval time [2].

A Phoneme recognition speaker dependent system was developed using time delay neural network. The performance was compared with several discrete hidden markov models and it was found that the TDNN achieves better accuracy as compared to the hidden markov model [5]. A convolutional recurrent neural network was developed and was applied in polyphonic sound event detection task. It was a combination of both Convolutional neural network and recurrent neural network. The new hybrid system exhibits considerable improvement in performance as compared to the individual methods [6]. A speaker identification system was developed by cascading three MFCC frames together and was named as super mel-frequency cepstrum coefficients. The probability density function of these coefficients was estimated by the Histogram transform method [7]. The performance of MFCC degrades drastically when the interference due to noise is high. The logarithmic transformation in MFCC is replaced by a combination function. Speech enhancement methods like spectral subtraction and median filter were also used to supress the noise. In noisy environment, the proposed system performs far much better than MFCC [8].

3. Methodology

Sound data is required for speaker identification. The sound component in a speaker identification system includes sound equipment and an audio wave recorder. The sound is recorded using a microphone. Speech data usually contains a lot of silent regions, they are removed during preprocessing. The speech data before and after silence removal is shown in fig 1 and 2 respectively.

3.1. Feature extraction

Feature extraction techniques are applied onto the preprocessed data. There are different feature extraction techniques out of which MFCC is one of the mostly used feature extractor.

3.1.1. MFCC

This feature extraction technique was suggested by Bridle and Brown in 1947 and developed by Mermelstein in 1976. MFCC mimics the human speech production and reception system. The human ear receives frequencies less than 1 KHz at linear scale while frequencies higher than 1 KHz are being received at logarithmic scale. This property of human ear is used in MFCC.

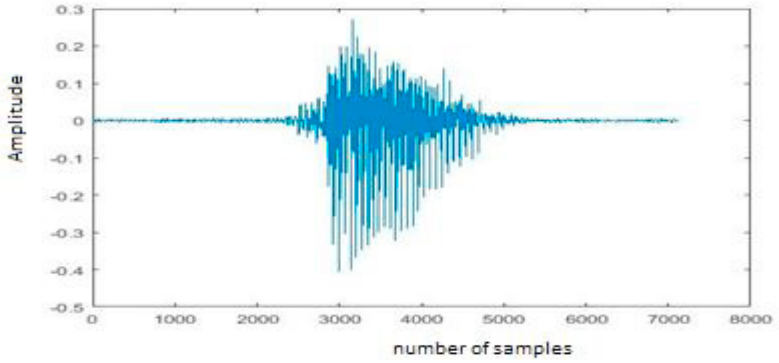

Fig. 1. Speaker saying the word DOWN without removing silent regions

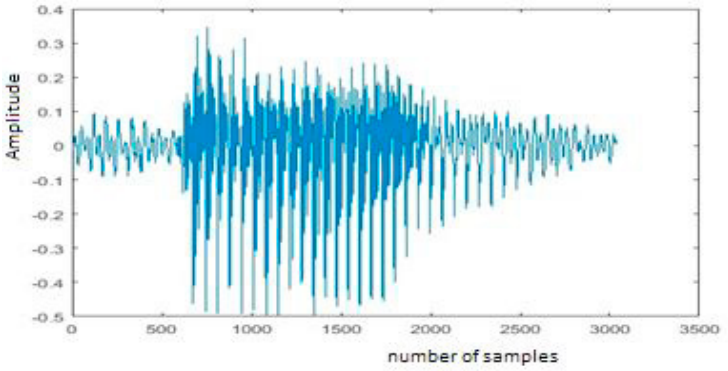

Fig. 2. Speaker saying the word DOWN after removing silent regions

Triangular Mel filters are used during the calculation of MFCC.
The MFCC computation consists of five main steps and they are:

- Pre-emphasis is done in order to achieve a higher signal quality. Here the amplitude of the higher frequency signals is increased as higher frequency signals are more important for signal disambiguation than lower frequencies.
- The audio signals are quasi stationary signals. That is, they remain stationary for a given time frame which is approximately 15-20 ms. So framing of audio signals is done. But the frame size should not be too small or too large or else the correct spectral properties will not be obtained. Framing is accompanied by overlapping which helps in providing time locality of information. The frame size used is 512 with an overlapping of 200.
- Each frame is then multiplied with a hamming window. It helps in minimizing the signal discontinuity.
- In order to convert a time domain signal to frequency domain signal, FFT is applied. The power spectrum of each frame is multiplied by Mel filter banks. The first filter is narrow and the size of the filters becomes wider and wider as frequency increases.
- Then the logarithm of the particular Mel filters is taken and after that DCT is applied. The coefficients obtained are the MFCC coefficients. Here 20 filters are used.

3.1.2. UMRT

DFT is a widely used transform in signal processing. FFT is a fast algorithm to implement DFT, exploiting periodicity and symmetry properties of the exponential kernel of the DFT in the complex domain, converting real data into complex form. the DFT computation was modified in terms of real addition rather than complex multiplication exploiting the property that 2x2 DFT does not require any complex multiplication and mapping to the complex domain by multiplying with N/2 twiddle factor[3] for an N×N data. MRT [4] is derived from the modified DFT [3] by eliminating the N/2 complex multiplication. In MRT, the arrangement and grouping of data based on the corresponding phase of the exponential kernel is utilized. The total number of MRT coefficients for a data sequence of length N is N× (N/2) and it also contains redundancies like complete redundancy and derived redundancy [4]. The complete redundancy was removed and a more compact version of MRT, named as UMRT was derived. The UMRT for a sequence of length N where N is a power of 2 and $x_1, x_2 .. x_n$ are the input sequence is given by

$$Y_0^{(0)} = \sum_{n=0}^{N-1} x_n \quad (1)$$

Which is the first UMRT coefficient and it gives the DC value of a particular signal. Rest of the UMRT coefficients are calculated by the equation

$$Y_k^{(p)} = \sum_{j=0}^{k-1} \left(x_{\frac{jN+p}{k}} - x_{\frac{(2j+1)N+2p}{2k}} \right) \quad (2)$$

Where M=N/2, $k=2^t$, $0 \le t \le \log_2 M$ and $p=rk$ where $0 \le r \le \frac{M}{k} - 1$

Here p is the phase index while k is the frequency index.

UMRT consists of both phase as well as frequency terms. In a frame size of length 512, there exists 10 frequency terms while the rest are phase terms. The length of phase terms changes at the decreasing order of power of 2.

After the silent regions are removed, framing of the speech data is done with a frame size of 512. After that, the UMRT for N a power of 2 is applied, then UMRT coefficients of length N is obtained. Now, grouping of UMRT coefficients is done in order to be used as a feature extractor. The zeroth UMRT coefficient is the DC value and is used as it is. Sum of the next 256 coefficients, which contains the same frequency term and different phase terms, is used as the second feature, then the sum of the next 128 coefficients, which contains the same frequency term and different phase terms, is used as the third coefficient and then, the next 64 coefficients are summed and used as the fourth coefficient and so on, till the length of the phase terms reaches the value 1. In this way, UMRT based features are obtained. For a frame of size 512, 10 features representing 10 frequencies are obtained.

3.2. Feature classification using Artificial neural network(ANN)

The human brain is made up of billions of neurons. One end of a neuron is a long slender projection known as the axon and it helps in transferring information to other neurons. While the other end of a neuron is made up of a branched like structure which is known as dendrites. They help in receiving information from other organs. According to the stimulus, the weights changes. This concept of neuron is taken in the development of ANN. ANNs are computing systems which are inspired by the animal brain. Such systems learn tasks by considering examples.

3.2.1 MLP

MLP is a class of feed-forward ANN. It helps in distinguishing data that is not linearly separable. The MLP consists of an input layer, an output layer and a hidden layer and all of these layers are interconnected. The learning used here is a batch based learning because here the system weights are kept constant while computing error associated with each sample in the input.

ANN has been used a lot for speaker identification [2], [5], [6]. In this project MATLAB Neural network toolbox has been used to train and simulate the networks. Here the speaker is identified based on the word DOWN spoken by 15 users. 70% of the data is used for training while the remaining 30 % is used for both validation and testing, that is 15% for both. In this project, the number of input neurons required changes based on the percentage combinations of the features used of both MFCC and UMRT. The best result is obtained when the number of hidden neurons is 88. The output layer contains 15 neurons one for each of the 15 speakers.

3.2.2 Confusion Matrix

It is a summary of prediction results of a classification problem. The number of incorrect and correct predictions is summarized with count values and is broken down by each class. This is the principle behind confusion matrix. The confusion matrix shows the ways in which the classification model gets confused when it makes predictions. It gives insight into not only the errors made by the classifier but also the types of errors made.

There are four types of

- True positive: observation is positive and is predicted to be positive
- False negative: observation is positive but is predicted negative
- True negative: observation is negative and is predicted to be negative
- False positive: observation is negative but is predicted to be positive

4. Results and discussions

Ten samples of words like DOWN, UP, LEFT, RIGHT, START, STOP and PAUSE, each spoken by 15 persons consisting of 8 female and 7 male speakers of different age groups is taken using the microphone at a sampling frequency of 8 KHz. The speaker identification is done for isolated words. For speech independent systems, the word DOWN is used for training while the words like UP, RIGHT, LEFT, START, STOP, DOWN and PAUSE, which were not used for training, are used for testing. For speech dependent systems, the word DOWN is used for both training and testing. The words like LEFT, RIGHT, UP, START and STOP are not used for speech dependent systems. Accuracy for different combinations of MFCC for the word DOWN is shown in table 1.

Table 1. Combinations of MFCC for the word DOWN

MFCC (%)	Speech independent (%)	Speech dependent (%)	Input neurons	Remarks
30% MFCC	84.5	90.3	96	[1:6] of each MFCC frame
40% MFCC	88.5	91.2	128	[1:8] of each MFCC frame
50% MFCC	87.6	93.6	170	[1:10] of each MFCC frame
60% MFCC	86.3	89.6	204	[1:12] of each MFCC frame
70% MFCC	89.5	91.3	138	[1:14] of each MFCC frame
All MFCC	91.6	92.5	266	Used all MFCC coefficients

Non overlapping UMRT based features are calculated for the first two frames of data. Now combination of both MFCC and UMRT based features are used. Classification is done using Multi-layer perceptron and average accuracy is measured using Confusion matrix. The accuracy is shown in table 2.

Table 2. Combinations of MFCC and UMRT for the word DOWN

Features	Speech independent (%)	Speech dependent (%)	Input neurons	Remarks
30%MFCC+2x10 UMRT	83.38	87	104	[1:6]of each MFCC frame and [1:2] frames of UMRT
40% MFCC +2x10 UMRT	88.94	95.6	132	[1:8] of each MFCC frame and [1:2] frames of UMRT
50%MFCC+2x10 UMRT	94.44	95.65	160	[1:10] of each MFCC frame and [1:2] frames of UMRT
60%MFCC+2x10 UMRT	91.66	97.91	188	[1:12] of each MFCC frame and [1:2] frames of UMRT
70%MFCC+2x10 UMRT	94.3	95.65	216	[1:14] of each MFCC frame and [1:2] frames of UMRT
Full MFCC+2x10 UMRT	83.3	97.91	286	All MFCC and [1:2] frames of UMRT

The comparison between the performance of both speech dependent and speech independent systems using the transforms MFCC and combination of both MFCC and UMRT based features is shown in the figure below.

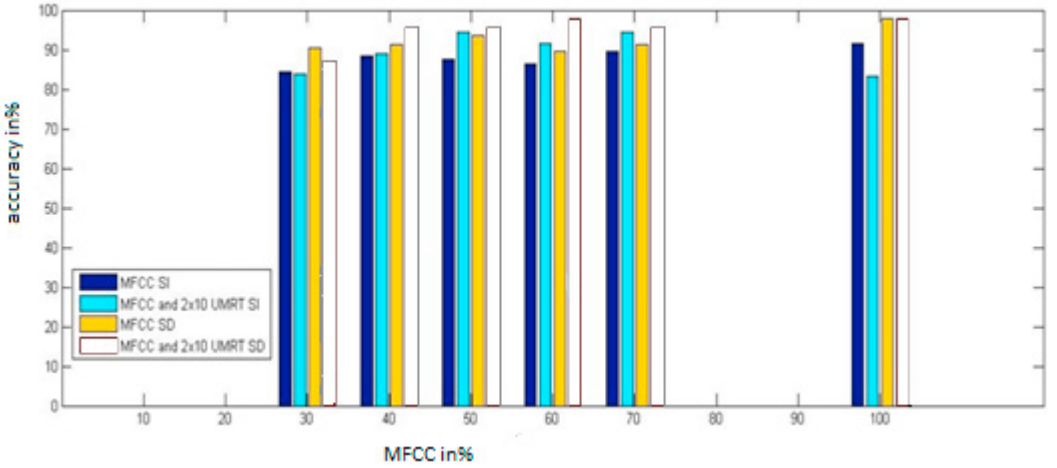

Fig. 3. Comparison of MFCC and combination of MFCC and UMRT based features for the word DOWN

The best average accuracy achieved for speech independent systems is around 94.4%, for the combination of 50% MFCC with first two frames of UMRT based features and with 60.15% of decrease in the total number of input neurons, The best average accuracy for speech dependent systems is 97.91 for the combination of 60% MFCC with two frames of UMRT and with 30% of decrease in the total number of input neurons. The additive white Gaussian noise with an SNR of 10 dB is added to the system and the following results are obtained.

Table 3. Combinations of MFCC and UMRT based features in presence of AWGN

Features	Speech independent	Speech dependent	Input neuron	Remark
MFCC	87.5	86.8	266	All MFCC coefficients
50%MFCC+2x10 UMRT	90.5	91.6	150	[1:10] of MFCC and [1:2] frames of UMRT

When the combinations of both MFCC and UMRT based features are used, the average accuracy increases by 3% for both speech dependent and independent cases as compared to MFCC coefficients alone and the number of input neurons also decreases by 43.6%.

5. Conclusion

A speaker identification system is developed which uses a new feature extractor, which is a combination of both MFCC and UMRT based features. This new feature extractor provides better results as compared to MFCC and that to with lesser number of input neurons. So the system complexity also reduces.

References

[1] Noor Almaadeed, Amar Aggoun, and Abbes Amira. (2015) "Speaker identification using multimodal neural networks and wavelet analysis." *IET Biometrics* **4(1)**: 18–28.

[2] Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh. (1994) "Speaker Networks Recognition Using Neural and Conventional Classifiers." *IEEE Transactions on Speech and Audio processing* **2(1)**: 194–204.

[3] R Gopikakumari.(1998) "Investigations on the development of an ANN model and visual manipulation approach for 2-D DFT in image processing." *PhD dissertation CUSAT*.

[4] Rajesh Cherian Roy. (2009) "Development of a new transform: MRT.", *PhD dissertation CUSAT*.

[5] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. (1989) "Phoneme Recognition Using Time-Delay Neural Networks." *IEEE Transactions on Acoustics, Speech and Signal processing* **31(3)** : 328–339.

[6] Emre C¸ akır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. (2017) "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection." *IEEE/ACM Transactions on Audio, Speech, and Language processing* **25(6):** 1291–1303.

[7] Zhanyu Ma, Hong Yu, Zheng-Hua Tan, and Jun Guo. (2016) " Text-Independent Speaker Identification Using the Histogram Transform Model." *IEEE Access* **4**: 9733–9739**.**

[8] WU Zunjing, and CAO Zhigang. (2005) "Improved MFCC-based feature for robust speaker identification." *Tsinghua Science and Technology* **10(2)**: 158–161.