Research Journal of Applied Sciences, Engineering and Technology 11(9): 983-987, 2015
DOI: 10.19026/rjaset.11.2138
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2015 Maxwell Scientific Publication Corp.
Submitted: May 16, 2015 Accepted: July 2, 2015 Published: November 25, 2015

Research Article

Speaker Recognition Using Cepstral Coefficient and Machine Learning Technique

[1]C. Sunitha and [2]E. Chandra
[1]Department of CA and SS, Sri Krishna Arts and Science College,
[2]Department of Computer Science in Bharathiar University, Coimbatore, India

Abstract: Speaker recognition is one of the important tasks in the signal processing. In this study we perform speaker recognition using MFCC with ELM. First noise is removed in the speech through low pass filter; the purpose of the filter is to remove the noise below 4 kHz. After enhancement of individual speech, feature vector is formed through Mel-frequency Cepstral Coefficient (MFCC). It is one of the nonlinear cepstral coefficient function, features are extracted using DCT, Mel scale and DCT. The feature set is given to Extreme Learning Machine (ELM) for training and testing the individual speech for speaker recognition. Compared to other machine learning technique, ELM provides faster speed and good performance. Experimental result shows the effectiveness of the proposed method.

Keywords: Discrete cosine transform, filter bank, mel-frequency cepstral coefficient, mel scale

INTRODUCTION

One of the most secured features of biomedical recognition is speaker recognition. To generate the speaker identity, we have to extract the features from the individual voice, which is the process of speaker recognition. In this speaker recognition, two types of tasks are available such as verification and identification (Ai *et al*., 2012; Jain *et al*., 2004). Speaker verification determines if a person is the claimed identity based on a piece of voice sample; speaker identification describes which one matches the input sample voice from the group of training voices.

Speech recognition is called as a sister technology to speaker verification. The function of the speech recognition is to correctly identify what the person says. This speaker recognition is has sbecome most popular and it is pathway for the speaker authentication. Total voice solution is a method used to interact the individual person with the system; this method is formed by the combination of speaker recognition and speaker authentication.

Operation of speaker recognition is carried out in three ways (Judith and Markowitz, 1998), first operation is called as speaker identification or speaker recognition. Second operation had many names such as speaker verification, speaker authentication, voice recognition and voice verification. Speaker separation and speaker classification falls under the ~~on~~ third operation.

Already, we know speaker recognition is a biometric authentication process and here human voice is one of the characteristics and it is used as an attribute (Kinnunen and Li, 2010; Campbell *et al*., 2009). Speaker recognition system have three fundamental section such as; to describe~~d~~ the speech signal in a compact manner using noise removal and feature extraction. Then to characterize those features by some statistical approach, finally speaker classification is used to find out the unknown utterance. The literature about several speaker identification algorithms is given in Clarkson *et al.* (2001) and Hui-Ling and Fang-Lin (2007).

In this study, proposed work is focused on designing the techniques by effectively preserving the information related to speaker and it is used to improve the speaker recognition system. In this study, the system is split into three models, they are:

- Speech enhancement is carried to improve the voice signal or remove the unwanted voice signal through low pass filter.
- Speech signal features are extracted through combination of subband based cepstral parameters and Mel-Frequency Cepstral Coefficients (MFCC) as feature vectors.
- Find out the unwanted utterance in the classification stage, here Extreme Learning Machine is used as a classification.

LITERATURE SURVEY

Linear transformation technique is implemented in Sahidullah and Saha (2012), it preserves the speech information effectively for speaker recognition improvement. Here, block based transformation approach is used to all Mel filter bank log energy at a

Corresponding Author: C. Sunitha, Department of CA and SS, Sri Krishna Arts and Science College, Coimbatore, India

time. Multi-block DCT is used for the formation of Cepstral coefficient. Better performance of speaker recognition is obtained by using combination of both systems. Performance is evaluated between NIST SRE 2001 and NIST SRE 2004.

Feature selection is one of the important tasks in speaker recognition and identification. Because large numbers of features are extracted from the same from of speech, so redundancy is presented in the extracted speech. So remove the redundancy and select the possible feature vector is most important. In Sandipan and Gowtam (2010), proposed technique for feature selection using Singular Value Decomposition (SVD) followed by QR Decomposition with Column Pivoting (QRcp). This feature technique is baseline to MFCC and LFCC.

Two types of approach are in speaker identification that is text-dependent and text-independent. In this, text-independent speaker identification is proposed in Kumari *et al*. (2012); here they identified the speaker for individual person through two different types of feature set. Two feature sets are Mel Frequency Cepstral Coefficient (MFCC) and Inverted Mel Frequency Cepstral Coefficient (MFCC) features. Finally this individual speaker features are trained using Expectation Maximization algorithm. Testing the data using GMM for two feature sets.

For the improvement of recognition rate of speaker combination of two features with traditional one (MFCC and LPCC) are proposed in Chetouani *et al*. (2009), here features are depend on LP-residual signal.

Probabilistic Neural Network (PNN) is used to recognize the speaker, but problem this technique is recognized is based on the smoothing factor. So to overcome this problem, combination of smoothing factor with PNN is implemented in Fan-Zi and Hui (2013). Hybrid algorithm is proposed for the improvement of speaker recognition using (DFOA-SOM-PNN), first Self-Organization Map (SOM) is to cluster the speaker features it is extracted through MFCC, then double fly fruit optimization algorithm is used to smoothing factor of PNN.

METHODOLOGY

This section describes about the proposed work behind the speaker recognition. Block diagram for the proposed work is illustrated in Fig. 1.

Speech enhancement: In the time of recording, speech signal is affected by noise or unwanted signal. Usually information of the input signal is present in the higher frequency. Noise may be occurred due to the channel fading, loss of speech segment, echo or reverberation.

So, in this study low pass is filter is used to remove the noise. This filter passes, if it below cut off frequency and stops above the cutoff frequency.

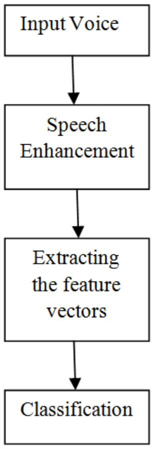

Fig. 1: Block diagram for proposed methodology

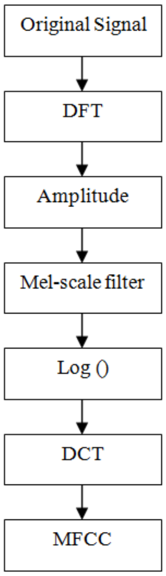

Fig. 2: Block diagram for MFCC

Feature extraction: This section is used to extracts the original signal into number of features for dimensionally reduction and probabilistic modeling. In speech recognition, there are many methods to extract features such as Mel frequency Cepstral coefficient (MFCC), Linear Prediction Coefficients (LPC), Perceptual Linear Prediction coefficients (PLP), etc. In this study, features are extracted through MFCC and this method is one of the popular methods for extraction of speech signal. Sounds are represented in two ways, such as linear Cepstral and nonlinear Cepstral. MFCC is derived from nonlinear Cepstral representation of sound. Mel scale is used in the MFCC and it is more responsible for human auditory system than linear Cepstral representation of sound (Bahoura, 2009). Block diagram for MFCC is given in Fig. 2.

In this process, first transform the original signal from time domain to frequency domain by using Discrete Fourier Transform (DFT), here power spectrum is used. Before DFT, hamming window is used for the reduction of frequency distortion due to segmentation.

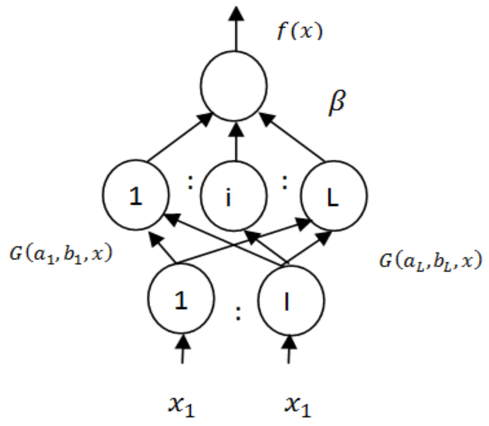

Fig. 3: ELM with I input neurons and L hidden neurons

After this process, filter bank is used wrapping the frequency from hertz scale to Mel scale. Finally Discrete Cosine Transformation (DCT) is used for the extraction of feature vectors on the logarithm of Mel scale power spectrum.

Step 1: In the first stage, original signal is multiplied by using Hamming window and then the window speech frames are processed under DFT. This is obtained from Fourier transform:

$$x(k) = \sum_{n=0}^{N_D-1} x(n)e^{-j2\pi nk/N_D}$$

In the above equation, $N_D$defines the number of points in the DFT.

Step 2: Filter bank is created:

$$e_s(i) = In\left[\sum_{k=0}^{N_f-1} |x(k)|^2 T_i(k)\right]$$

The above equation defines the energy spectrum $e_s(i)$, where number of filter is indicated by N_f and $i = 1,2,..,N_f$.

$$T_i(k) = \begin{cases} 0 & for \quad k < k_{b_{i-1}} \\ \frac{k-k_{b_{i-1}}}{k_{b_i}-k_{b_{i-1}}} & for \quad k_{b_{i-1}} \leq k \leq k_{b_i} \\ \frac{k_{b_{i+1}}-k}{k_{b_{i-1}}-k_{b_i}} & for \quad k_{b_i} \leq k \leq k_{b_{i+1}} \\ 0 & for \quad k > k_{b_{i+1}} \end{cases}$$

The above equation describes the band pass filter $x(k)$ by triangular filter bank $T_i(k)$.

Filter boundary points are indicated by $\{k_{b_i}\}_{i=0}^{N_f+1}$, where k denotes the index of the N_D point DFT.

Step 3: Mel-scale calculation using O'shaughnessy (Ai *et al.*, 2012). It is given by below equation:

$$f_{mel} = 2595 \times log_{10}\left(1 + \frac{f}{700}\right)$$

In the above equation, f_{mel} denotes the sampling frequency:

$$k_{b_i} = \left(\frac{N_D}{F_S}\right) f_{me1}^{-1}\left[\frac{f_{mel}(f_{min})}{+\frac{i\{f_{mel}(f_{max})-f_{mel}(f_{min})\}}{N_f+1}}\right]$$

In the above equation f_{min} and f_{max} denotes the low and high frequency boundary of the filter banks. Inverse transform f_{me1}^{-1} is given by below equation:

$$f_{me1}^{-1}(f_{mel}) = 700\left[10^{f_{mel}/2595} - 1\right]$$

Step 4: MFCC coefficient is calculated, that the output of logarithmic filter bank is given to the DCT:

$$MFCC(n) = \sum_{i=0}^{N_f-1} e_s \, cos\left(\frac{(\pi n(i-0.5))}{N_f}\right) 0 \leq n \leq N_f - 1$$

where, n defines the number of MFCC coeffiecints.

Extreme Learning Machine (ELM) for speaker recognition: Extreme Learning Machine is one of the useful statistical tools for machine learning techniques and it has been successfully applied in the pattern recognition tasks. ELM is proposed by Huang et al and it is developed for Single Hidden Layer feed forward Networks (SLFNs) with a wide variety of hidden nodes. This system can be represented as linear system; this system obtains the smallest training error and good performance. ELM has several methods such as optimization method based ELM (Sahidullah and Saha, 2012) regularized ELM and kernelized ELM (Sandipan and Gowtam, 2010).

Consider a number of N training samples $\{(x_1, t_1), \ldots, (x_N, t_N)\}$, here $x_i \in \mathbb{R}^I$ and $t_i \in \{-1,1\}$, usually SFLN have I input neuron and L hidden neuron and it is illustrated in Fig. 3.

The below equation give the function for binary classification:

$$f_L(x) = sign(\sum_{i=1}^{L} \beta_i h_i(x)) = sign(h(x)\beta)$$

In the above equation weights are present in the vector $\beta = [\beta_1, .., \beta_L]^T$ this weight connecting the hidden neurons and output neurons. The output of the hidden layer is given by $h(x) = [h_1(x), \ldots, h_L(x)]$ with respect to input x. nonlinear piecewise continuous function is defined by $G(a, b, x)$ it is derived from the following equation:

$$h(x) = [G(a_1, b_1, x), \ldots, G(a_L, b_L, x)]$$

Then the universal approximation capability theorems are satisfied by above nonlinear piecewise continuous function.

H is the hidden-layer output matrix is defined by below equation:

$$H = \begin{bmatrix} h_1(x_1) & : & h_L(x_1) \\ : & : & : \\ h_1(x_N) & : & h_L(x_N) \end{bmatrix}$$

To minimize the $\|H\beta - T\|$ and $\|\beta\|$ for training the ELM, in which $T = [t_1, t_2, \dots, t_N]^T$. The solution to the problem can be calculated as the minimum norm least-square solution of the linear system:

$$\hat{\beta} = H^{+}T$$

In the above equation, H^{+} defines the Moore-Penrose generalized inverse of matrix H. ELM have speed training phase as well as good performance for computing the output weights analytically.

ELM algorithm:
Input: Training Set, hidden node activation function, number of hidden nodes
Output: Weight vector

```
Step 1: Hidden node parameters are randomly generated.
Step 2: for i-1: L do
        $a_i, b_i$ Randomly assigned
Step 3: End
Step 4: Hidden layer output matrix H is examined
Step 5: $for\ i = 1: L\ do$
        $for\ j = 1: N\ do$
        $H(i,j) = G(a_i, b_i, x_j)$
        end
        end
Step 6: Finally output weight vector β is calculated
Step 7: $\hat{\beta} = H^{+}T$
```

EXPERIMENTAL RESULTS

NTT database (Matusi and Furui, 1993) and a large-scale Japanese Newspaper Article Sentences (JNAS) database Itou *et al.* (1999) were used to evaluate proposed method. The proposed work is compared with the previous work is classified by RVM and some other existing methods through performance metrics such as accuracy. Speaker verification performance will be reported using the True Positive (TP) samples and True Negative (TN) samples:

TP : Abnormal class classifies as abnormal
TN : Normal class classifies as normal

Table 1 gives true positive and true negative rate for proposed method. Table 2 provides the speaker recognition rate, compared with existing approaches of MFCC with GMM (Rajalakshmi and Revathy, 2013) is the combination of Perceptual Linear Predictive cepstrum with Gaussian Mixture Model and provide the identification accuracy of 77.36% IMFCC (Inverted MFCC) (Chakroborty and Saha, 2009) for polycost database using triangular filter gives 77%, Kullback-Leibler divergence (Saeidi *et al*., 2009) for different gender provide the accuracy of 93%, the proposed method of DT-CWT with RVM gives the recognition rate of 93.5%, compared with this RVM proposed, MFCC-ELM proposed gives the better result of 98.4%. From the result clearly observed that the proposed method of MFCC-ELM gives better speaker recognition rate than previous proposed method as well as existing method.

Table 1: Performance metrics

Performance metrics	Accuracy (%)
True Positive (TP)	98
True Negative (TN)	99

Table 2: Speaker identification rate

Techniques	Speaker identification rate
MFCC with GLM	77.36%
IMFCC (Inverted MFCC)	77%
Kullback-Leibler divergence	93%
Proposed DT-CWT with RVM	95%
Proposed MFCC-ELM	98.4

CONCLUSION

In this study, we have investigated an MFCC-ELM approach for speaker recognition. In this study, three steps are carried over. First is removed through low pass filter if the signal has below 4 kHz, this is used to improve the speaker recognition accuracy. Second, standard MFCC features are extracted using linearly spaced filters in Mel scale. Third, classification for speaker recognition based on ELM, it is more suitable for particle acoustic signals, leading to high material recognition accuracy than that of other system. The comparison study with existing methods also demonstrated the performance of proposed method. Compared to the first proposed method of DT-CWT with RVM and existing method proposed method provides better results. The results have demonstrated the effectiveness of the proposed method for speaker recognition through accuracy.

REFERENCES

Ai, O.C., M. Hariharan, S. Yaacob and L.S. Chee, 2012. Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst. Appl., 39(2): 2157-2165.

Bahoura, M., 2009. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. Comput. Biol. Med., 39(9): 824-843.

Campbell, J., W. Shen, W. Campbell, R. Schwartz, J.F. Bonastre and D. Matrouf, 2009. Forensic speaker recognition. IEEE Signal Proc. Mag., 26(2): 95-103.

Chakroborty, S. and G. Saha, 2009. Improved text-independ-ent speaker identification using fused MFCC and IMFCC feature sets based on Gaussian filter. Int. J. Signal Proc., 5(1): 11-19.

Chetouani, M., M. Faundez-Zanuy, B. Gas and J.L. Zarader, 2009. Investigation on LP-residual representations for speaker identification. Pattern Recogn., 42: 487-494.

Clarkson, T.G., C.C. Christodoulou, Y. Guan, D. Gorse, D.A. Romano-Critchley and J.G. Taylor, 2001. Speaker identification for security systems using reinforcement-trained pRAM neural network architectures. IEEE T. Syst. Man Cy. C, 31(1): 65-76.

Fan-Zi, Z. and Z. Hui, 2013. Speaker recognition based on a novel hybrid algorithm. Proc. Eng., 61: 220-226.

Hui-Ling, H. and C. Fang-Lin, 2007. ESVM: Evolutionary support vector machine for automatic feature selection and classification of micro array data. Biosytems, 90: 516-528.

Itou, K., M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, 1999. Japanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Jpn., 20(3): 199-206.

Jain, A.K., A. Ross and S. Prabhakar, 2004. An introduction to biometric recognition. IEEE T. Circ. Syst. Vid., 14(1): 4-20.

Judith, A. and J. Markowitz, 1998. Speaker recognition. Inform. Secur. Tech. Rep., 3(1): 14-20.

Kinnunen, T. and H. Li, 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun., 52(1): 12-40.

Kumari, R.S.S., S. Selva Nidhyananthan and G. Anand, 2012. Fused Mel feature sets based text-independent speaker identification using Gaussian mixture model. Proc. Eng., 30: 319-326.

Matusi, T. and S. Furui, 1993. Concatenated phoneme models for text-variable speaker recognition. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93). Minneapolis, MN, USA, 2: 391-394.

Rajalakshmi, R. and A. Revathy, 2013. Comparison of MFCC and PLP in speaker identification using GMM. Proceeding of the International Conference on Computing and Control Engineering (ICCCE, 2012), pp: 12-13.

Saeidi, R., P. Mowlaee, T. Kinnunen, Z.H. Tan, M.G. Christensen, S.H. Jensen and P. Franti, 2009. Signal-to-signal ratio independent speaker identification for co-channel speech signals. Proceeding of the 20th IEEE International Conference on Pattern Recognition (ICPR, 2009), pp: 4565-4568.

Sahidullah, M.D. and G. Saha, 2012. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun., 54: 543-565.

Sandipan, C. and S. Goutam, 2010. Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification. Speech Commun., 52: 693-709.