A Fuzzy Nearest Neighbor Classifier for Speaker Identification

Conference Paper · November 2012

DOI: 10.1109/CICN.2012.16

2 authors:

Seba Susan
Delhi Technological University
122 PUBLICATIONS **1,180** CITATIONS

SEE PROFILE

Srishti Sharma
Netaji Subhas Institute of Technology
9 PUBLICATIONS **51** CITATIONS

SEE PROFILE

A Fuzzy Nearest Neighbor classifier for Speaker Identification

Seba Susan and Srishti Sharma
Department of Information Technology,
Delhi Technological University,
New Delhi, India

Abstract— **Mel-frequency Cepstral coefficients (MFCC) are popular features extracted from speech data for speaker identification. The speech signal is fragmented into frames and the MFCC features extracted from each frame show some temporal redundancy which forms the basis of the fuzzy classifier proposed in this paper. We propose a fuzzy nearest neighbor classifier that defines a frame prototype for each training audio sample using a weighted mean technique with the weights being probability values, and the class label for each test sample is decided from fuzzy membership functions involving the frame prototypes. The classification results of the proposed classifier on audio samples from the VidTIMIT database show a superior performance to the Nearest Neighbor classifier, GMM, HMM and MLP neural networks. It is observed that the execution time of the fuzzy classifier is a very small fraction of the time taken by the HMM and neural network classifiers and the training database is significantly reduced due to the use of frame prototypes instead of actual frames.**

Keywords- *Mel-frequency cepstral coeffecients, Gaussian fuzzy membership functions, fuzzy classifer***.**

I. INTRODUCTION

Biometrics is a science that establishes the identity of an individual based on one or more intrinsic physiological or behavioral traits and uses modalities like voice, face, fingerprint, retina and iris [1] to uniquely identify or characterize an individual. Voice is one of the most reliable biometric trait used to identify a speaker from audio-visual clips and the Mel-frequency cepstral coefficients (MFCC) [2-4],[10] are the most preferred audio features in terms of the low computational cost involved and temporal consistency of features with higher inter-class variability [11-14]. Some of the other popular audio features are LPC [4]-[7],[10], PLP[3],[8],[10] and RASTA[6],[9]-[10] though the MFCC features are more popular among researchers due to its simple use and high reliability. Many classifiers have been proposed for audio classification over the years with the prominent ones being based on neural networks [3],[16-18], Gaussian Mixture models (GMM) [15] and Hidden Markov model (HMM) [18]. The multilayer perceptron neural network (MLP) [18] and radial basis function [3] neural networks are widely used for speaker classification due to their non-linearity property which is further enhanced by increasing the number of neurons in the hidden layer. The disadvantage of all the conventional non-linear classifiers based on neural networks, GMM or HMM is the total time involved for training and the large and varied training database required. Alternatively fuzzy [19] and neuro-fuzzy techniques [20] have been proposed for speaker identification with an aim to remove the training overhead and save memory space. In this paper a novel fuzzy classification of audio features is proposed that translates the Euclidean distance between training and test feature vectors into fuzzy domain using Gaussian membership functions. The proposed technique outperforms the non-linear neural network in terms of execution time and the size of the training database. The paper is organized as follows. Section 2 reviews the steps for the extraction of MFCC features from a audio file. In Section 3 the proposed fuzzy classification scheme is presented. Section 4 reports the experimental results and Section 5 outlines the conclusions from the results.

II. EXTRACTION OF MFCC FEATURES- A BASIC REVIEW

The Mel-frequency cepstral coefficients (MFCC) approximate the human audio perception since its frequency bands are positioned logarithmically quite similar to the human auditory response. The Mel-scale was developed in accordance with the pitch or frequency observed by the human ear with the basic unit being in mel. The speech signal is divided into frames to facilitate spectral analysis which is defined by short time segments in milliseconds called frames in which the signal is assumed to be stationary. A pre-emphasis is carried out initially to reduce the high frequency falloff and a hamming window is applied to each segment to minimize spectral leakage. The magnitude of the Discrete Fourier Transform of the speech signal sampled as per Nyquist criterion is warped into mel frequency using a filter bank. This is equivalent to a logarithmic mapping of the normal frequency scale into mel frequency in which the log total energy of the critical band around the center frequency is included. The Discrete Cosine Transform (DCT) of the mel frequency gives the DC coefficient and first 12 AC coefficients which when taken in order form the 13 MFCC features derived from a audio

frame. The MFCC Features along with their corresponding delta and accelerations (to capture changes between frames) form a 39 dimensional audio feature vector for each frame in the given sample.

III. PROPOSED FUZZY NEAREST NEIGHBOR CLASSSIFER

Any classification procedure is divided into two phases namely, the training and test phases. The term 'Fuzzy classifier' refers to the use of fuzzy membership functions to compute 'soft' class labels that are eventually defuzzified to classify the test sample to a known class. The training and test samples are initially segregated and 39 MFCC coefficients are extracted from a single frame of each speech sample in the manner explained in Section II. The first phase of fuzzy classification is the preparation of a compact training database by computing the expected values of the 39 MFCC features over all the temporal frames in the audio sample. The computational steps are detailed below:

Preparation of a compact training database using frame prototypes

The preparation of the compact training database exploits the temporal consistency of the MFCC coefficients derived from the different fragments or frames of the training audio sample. This is achieved by averaging the related MFCC coefficients over time. Let the histogram of h^{th} MFCC coefficient z_{ih} over all frames i=1,2,..,N, in an audio sample be denoted by the probability value $p(z_{ih})$ where h=1,2,...,39. Then the expected value of the h^{th} MFCC coefficient when averaged over all temporal frames is given by,

$$m_h = \sum_{i=1}^{N} z_{ih} p(z_{ih}) \tag{1}$$

Therefore the weighted mean m_h for each column h in the $Nx39$ feature matrix is calculated by (1) and a mean feature vector of dimension 1x39 is computed for each training audio sample. This mean feature vector constitutes the frame prototype used to represent the class or category of the training sample. Therefore similar to the conventional nearest neighbor classifier the computations are made between each test and training sample. However in this case while the test input is a 1x39 frame feature vector, the training sample is now a category or class representing a particular frame configuration to which the membership of the test frame is computed. The Categorical memberships are computed for all the training samples or categories. The speaker label is finally assigned to the test sample by assembling and comparing all the computed category memberships of its constituent frames in the manner as explained in the *testing phase* below.

Testing Phase:

A fuzzy membership function is computed for each frame of the test audio sample based on the Euclidean distance of its MFCC features from each frame prototype in the training database. The Gaussian function is used to compute the fuzzy membership due to its inherent advantages of being non-linear and a monotonically decreasing function. The fuzzy membership of the j^{th} test feature vector $\{x_{jh}\}$, h=1,2,…,39, to the k^{th} training class prototype vector $\{m_{kh}\}$, h=1,2,…,39, is given by,

$$\mu_{jk} = e^{-\|x_{jh} - m_{kh}\|^2} \tag{2}$$

where, ||.|| is the Euclidean distance norm. Assume there are M audio samples per speaker in the training database. This implies that the compact database would comprise of MxS feature vectors, where S is the number of speakers in the experiment. Therefore there are MxS classes and MxS fuzzy memberships. Then the speaker label p assigned to the j^{th} test feature vector is decided by averaging the fuzzy memberships of the M samples that belong to the same speaker and finding the maximum of the results for all speakers p=1,2,…,S.

$$\mu_{jp} = \frac{\left| \sum_{k=(p-1)M+1}^{(p-1)M+M} \mu_{jk} \right|}{M} \tag{3}$$

$$class_j = \arg\max_{p}(\mu_{jp}) \tag{4}$$

Each test frame j from the test audio sample, j=1,2,…,N, is thus labeled with a speaker label as in (4). Finally the entire test audio sample is classified as belonging to the speaker whose label is most frequent among the constituent frames of the test sample.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments are conducted on audio samples from the robust VidTIMIT database [22] that contains audio-visual recordings of 43 people reciting sentences from TIMIT corpus [23]. It has been recorded in 3 sessions with a gap of 7 days between sessions 1 and 2 and 6 days between sessions 2 and 3.The gap between sessions accounts for the possibility of mood and appearance changes that may occur in real life. There are a total of 10 sentences per person, 6 of them recorded in session 1 and two each in sessions 2 and 3. Two sentences are common to all speakers while the other eight sentences are different for each speaker.

We conducted the speaker identification experiments for a hierarchical subset of 5, 10, 15, 20, 25, 30, 35, 43 speakers and study the variability in results as the number of users is increased. Out of the ten audio samples per speaker, nine have been used for training and one has been used for testing taking into account the robust nature of the database.

This implies that S=5,10,15, 20, 25, 30, 35, 40, 43 and M=9, for the equations in Section III. Each audio sample in the ".wav" format is divided into 50 frames (N=50) using a 25 ms window with 50% overlap, and the 39-dimensional feature vector (comprising the MFCC coefficients and their delta and accelerations) is derived from each frame as explained in Section II. The training and test datasets are then segregated and the fuzzy classification procedure is applied for assigning a class to the test audio sample. The label of the speaker associated with nearest frame prototype is assigned to the test frame. The MATLAB 7.9 version software is used for implementing the code on a 2.16 GHz processor. For evaluating the efficiency of our classifier we use the Multi-layer Perceptron (MLP) neural network, Hidden Markov model (HMM), Nearest Neighbor, Gaussian Mixture Model (GMM) classifiers for comparison and the results are indicated in Table 1 and 2. It is observed from Table 1 that the fuzzy nearest neighbor classifier is highly successful in distinguishing between speakers as seen from the high classification rate of 80% for all three subsets S=5,10,15 while the MLP neural network (with 100 neurons in the hidden layer), GMM and HMM (for 3 states and 2 mixtures) are unable to perform in the absence of sufficient training. It is observed from Table 2 that the computational time is still much lower as compared to MLP neural network, GMM and HMM and comparable to that of the nearest neighbor classifier due to the simple computations involved. Another reason for its speedy execution is the reduced size of the training database since only a single prototype of the frame is stored per training sample as seen from the feature dimension D column in Table 2. The Nearest Neighbor classifier starts to become slower as the number of users is increased from 5 to 43 whereas its fuzzy counterpart is not much affected as seen from Table 2. The primary advantage of the proposed fuzzy nearest neighbor classifier for audio features is its high accuracy and a reduced or compact storage of training vectors and execution time of a few seconds.

V. CONCLUSION

A new fuzzy classification scheme for MFCC audio features is proposed in this paper that gives high classification accuracy for a compact database and execution time in a couple of seconds. The Gaussian membership function is used for fuzzifying the distance between training and test feature vectors. The performance is evaluated with respect to MLP neural network, Nearest Neighbor Classifier, HMM and GMM and is found to outperform all of them with a high accuracy at a reduced training database and extremely low computation time.

REFERENCES

[1] Ross, A. and Jain, A. K., "Multimodal biometrics:an overview," *Procc. EUSIPCO*, pp. 1221-1224, Sept.2004.

[2] Zhiping Dan,Sheng Zheng, Shuifa Sun, Ren Dong," Speaker Recognition based on LS-SVM**",** *3rd International Conference on Innovative Computing Information and Control* (ICICIC'08).

[3] Nima Yousefian, Azarakhsh Jalalvand, Pooyan Ahmadi, Morteza Analoui," Speech Recognition with a Competitive Probabilistic Radial Basis Neural Network", *IEEE Conference Intelligent Systems*, 2008

[4] Hemant A. Patil, Prakhar Kant Jain, Robin Jain," A Novel Approach To Identification Of Speakers From Their Hum"**,** *Seventh International Conference on Advances in Pattern Recognition*, 2009.

[5] Z. Uzdy, "Human Speaker Recognition Performance of LPC Voice Processors**",** *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. assp-33, no. 3**,** June 1985

[6] Rajparthiban Kumar, Aravind CV, Kanendra Naidu, Anis Fariza," Development of a Novel Voice Verification System using Wavelets", *Proceedings of the International Conference on Computer and Communication Engineering* 2008 May 13-15, 2008 Kuala Lumpur.

[7] Dr. Gwyn P. Edwards," A Speech/Speaker Recognition and Response System*, , IEEE Procc*. *ICASSP* 1980

[8] Wu Guo, Yanhua Long, Yijie Li, Lei Pan, Eryu Wang, Lirong Dai*,"* iFLY System For The NIST 2008 Speaker Recognition Evaluation", *IEEE Procc*. *ICASSP 2009*.

[9] Ramon F. Astudillo, Dorothea Kolossa, Reinhold Orglmeister," Uncertainty Propagation for Speech Recognition using RASTA Features in Highly Nonstationary Noisy Environments", *ITG-Fachtagung Sprachkommunikation* 8- 10 October 2008 in Aachen

[10] Tilo Schiirer," An Experimental Comparison Of Different Feature Extraction And Classification Methods For Telephone Speech*", 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* **(**IvTTA94),1994.

[11] Aleksic, P. S. and Katsaggelos, A. K., "Audio-visual biometrics," *Procc. IEEE*, vol. 94, no. 11, pp. 2025-2044, Nov. 2006.

[12] Sanderson, C., "Automatic person verification using speech and face information," *Ph.D. Thesis*.

[13] Chetty, G. and Wagner, M., "Speaking faces for facevoice speaker identity verification," *Proc. Interspeech*, pp. 513-516, Sept. 2006.

[14] Erzin, E., Yemez, Y., and Tekalp, A. M., "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability*," IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 840-852, Oct. 2005.

[15] D.Reynolds, T.Quatieri, and R.Dunn, "Speaker Verification Using Adaptive Mixture Models", *Digital Signal processing*, 2000, 10,pp.181-202.

[16] Zhiping Dan.,Sheng Zheng.,Shuifa Sun.,Ren Dong., "Speaker Recognition Based on LS-SVM" , *3rd International Conference On Innovative Computing And Information And Control,*2008.

[17]H.Hattori, "Text Independent speaker recognition using neural networks", *IEEE conf. proceeding ICASSP* 1992, pp. 153-156

[18] Y. Arriola, R A Carrasco," Integration Of Multilayer Perceptron And Markov Models For Automatic Speech Recognition", *UK IT* 1990 Conference

[19] Tongtao Zheng, Dat Tran and Michael Wagner "Fuzzy Nearest Prototype Classifier Applied to Speaker Identification", in *Proceedings of the European Symposium on Intelligent Techniques (ESIT'99),*1999

[20] Jyh-Shing Roger Jang and Jiuann Jye-Chen, " Neuro-Fuzzy and Soft Computing for Speaker Recognition**",** *IEEE conf. proceeding* 1997

[21] M. Hanmandlu, J. Grover, V. K. Madasu, S. Vasikarla, "Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals", *International Conference on Information Technology (ITNG'07).*

[22] Sanderson, C., Biometric person recognition : face, speech, and fusion. *VDM Verlag*, June 2008.

[23] Garofolo, J. S., Lamel, L.F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NIST order number PB91-100354, 1992.

Table 1: Percentage of correct classification for MLP-NN, Nearest Neighbor, HMM, GMM and Proposed Fuzzy Classifier (with highest values shown in bold)

No. of speakers	MLP Neural Network	Nearest Neighbor Classifier	HMM	GMM	Proposed Fuzzy Classifier
5	**80%**	**80%**	**80%**	60%	**80%**
10	70%	50%	50%	40%	**80%**
15	40%	46.67%	26.67%	40%	**80%**
20	45%	50%	40%	45%	**65%**
25	36%	40%	28%	40%	**56%**
30	30%	33.33%	30%	43.33%	**46.67%**
35	28.57%	31.43%	28.57%	40%	**51.43%**
40	30%	37.5%	25%	40%	**50%**
43	23.26%	48.17%	23.26%	41.86%	**48.84%**

Table 2: Time Complexity (T) in seconds and Feature Dimension of the Training data (D) for the MLP-Neural Network, Nearest Neighbor, HMM, GMM classifiers and the Proposed Fuzzy Classifier
(with most optimum values shown in bold)

No. of speakers	MLP Neural Network		Nearest Neighbor Classifier		HMM Classifier		GMM Classifier		Proposed Fuzzy Classifier	
	T(secs)	D	T(secs)	D	T(secs)	D	T(secs)	D	T(secs)	D
5	93.78	2250x39	**0.34**	2250x39	58.18	2250x39	1.08	2250x39	1.93	**45x39**
10	337.85	4500x39	**1.35**	4500x39	137.26	4500x39	8.68	4500x39	1.96	**90x39**
15	683.30	6750x39	3.08	6750x39	365.09	6750x39	30.83	6750x39	**2.02**	**135x39**
20	536.95	9000x39	5.62	9000x39	402.14	9000x39	30.13	9000x39	**2.45**	**180x39**
25	1122. 2	11250x39	7.96	11250x39	541.75	11250x39	33.41	11250x39	**3.06**	**225x39**
30	1021.68	13500x39	11.00	13500x39	995.43	13500x39	127.85	13500x39	**3.86**	**270X39**
35	1172.92	15750x39	15.11	15750x39	1215.76	15750x39	170.67	15750x39	**5.14**	**315x39**
40	2048.08	18000x39	19.55	18000x39	1386.35	18000x39	219.198	18000x39	**6.41**	**360x39**
43	1686.48	19350x39	23.12	19350x39	1420.71	19350x39	254.57	19350x39	**7.7**	**387x39**