Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 55–62

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017,

22-24 August 2017, Cochin, India

Performance Evaluation of Different Modeling Methods and Classifiers with MFCC and IHC Features for Speaker Recognition

Suma Paulose [a], Dominic Mathew [a,*], Abraham Thomas [a,*]

[a]Rajagiri School of Engineering and Technology, Rajagiri Valley, Kochi 682039,India

Abstract

Automatic speaker recognition system identifies a person from the information contained in the speech signal. These systems are the most user-friendly means of biometric recognition and are being used in applications like teleconferencing, banking, forensics etc. The accuracy of these depends on the methods used to extract features from the speech signal, modeling methods, classifiers used to identify the speaker and amount of data available for training and testing. In this paper, recognition systems are implemented using both spectro-temporal features and voice-source features. Classification is done with two different classifiers for i-vector method and the accuracy rates are compared.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications.

Keywords: MFCC; IHC; i-vector; GMM; CDS; PLDA

1. Introduction

Speaker recognition is the process of identifying a person from his voice signal. Each individual's sound will be different because of the differences in the shape of the vocal tract, size of the larynx and other parts of the voice production organs [1]. Apart from this, there will be other features that make each person different as manner of

* Corresponding author.
E-mail address: dominicmathew@rajagiritech.ac.in (Dominic Mathew), abrahamt@rajagiritech.ac.in (Abraham Thomas)

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications
10.1016/j.procs.2017.09.076

speaking, accent, pronunciation pattern, rhythm etc. Basically speaker recognition systems falls into two categories: speaker verification and speaker identification. Speaker identification is the task of determining who is talking from a set of known voices of speakers. Such a system is difficult to implement since the test speaker makes no claim on his identity and the system must perform a 1: N classification where N is the number of speakers enrolled. Commonly it is assumed the test speaker's voice come from a known set of speakers, and is often referred to as closed-set identification. Speaker verification is the task of determining whether a speaker is who he/she claims to be. Since it is assumed that false claims are not known to the system, it is referred to as an open-set task [2].

There exists two modalities for any speaker recognition systems: text dependent and text independent. In text dependent the test speaker is restricted to utter certain particular words or phrases whereas in text independent the speaker has the freedom to utter any word or phrase of his own. Hence implementation of a text independent system is much more difficult when compared to text dependent system. As far as the speaker is concerned, a text independent system offers much more flexibility to the speaker. For both modalities, there are two distinct operational phases: training and testing.

In training also called enrolment, models of all known speakers that need to be identified are built using the speech signals of those speakers. In testing, speech from an unknown utterance is compared against each of the trained speaker models [3]. In this work, a text independent speaker recognition system is implemented using closed set identification. Accuracy rates are compared using both spectral and voice source features and identification is performed using different modeling methods and classifiers.

The rest of this paper is organized as follows: Section 2 gives a description about feature extraction techniques, section 3 describes the speaker recognition methods used, section 4 gives the implementation details and section 5 give results and discussion. Finally section 6 summarizes the conclusion.

2. Feature Extraction

Fig.1 shows the block diagram of a speaker recognition system. The main blocks include feature extraction and speaker modelling. The feature extraction process aims to extract a compact, efficient set of parameters that represent the acoustic properties observed from input speech signal, for subsequent utilization [4]. Actually feature extraction is the method of reducing the dimension of data of the speech signal while retaining the required information. Speech signal contains several information of which not all are necessary for the identification of the speaker.

Ideal features must be robust against noise and distortion, occur frequently and naturally in speech, be easy to measure from speech signal, and be difficult to mimic etc. [1]. The features extracted may be categorized into short term spectral features, voice source features, spectro-temporal features, prosodic features etc. Short term spectral features are extracted from speech signals by dividing them into short frames of 20-30 ms duration. Voice source features make use of the features of the vocal tract. Here we make use of MFCC and IHC which are short term features and pitch and formants which are voice source features.

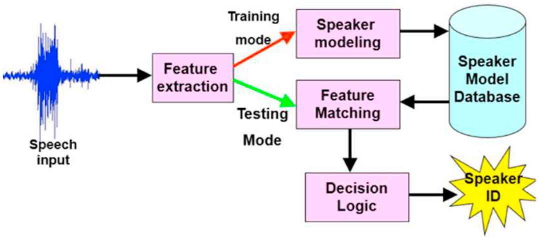

Fig. 1. Block Diagram of Speaker Recognition System

2.1. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most popular and commonly used technique in feature extraction for speaker recognition since it uses a MEL scale which is mimics the human ear scale. MFCC is considered as a frequency domain feature and is more accurate than time domain features [5]. These coefficients are found to be robust and reliable even if there are variations in the speakers and recording conditions [5]. The block diagram of MFCC is as given in Fig.2 [5]. The input speech signal is divided into frames of 25ms each with an overlap of 15ms. Usually overlapping of frames is done to make the transition from frame to frame more smoother [6]. In order to avoid the discontinuities occurring at the edges of the frames, each of these frames is multiplied with a Hamming window.

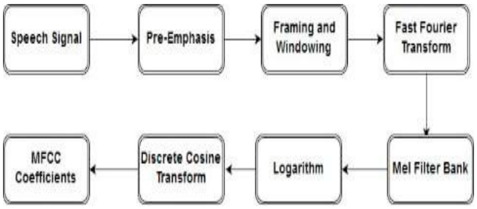

Fig. 2. Block Diagram of MFCC

The window function for hamming window of length n is given as in equation (1).

$$w(n) =; 0 \leq n \leq L - 1(1)$$

where *w(n)* is the Hamming window and *N* is the total number of samples and *L* is the window length [4]. Next DFT is done to extract the spectral information from the windowed signal. Since acoustic perceptions does not follow the linear frequency scale, MFCC make use of a perceptual pitch scale called Mel scale for feature extraction. The equation used to convert linear scale frequency *f* to Mel scale frequency is given in equation (2) [4].

$$Mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

During MFCC computation, this conversion is implemented by using a set of triangular filters which collects energy from each of the frequency bands. The number of filters used here is 26, with 10 filters linearly spaced below 1000 Hz and the remaining filters spaced logarithmically above 1000 Hz. Accordingly 26 MFCC coefficients are extracted. Since the filter banks are overlapped, the filter energies will be correlated [5]. To make them de-correlated, Discrete Cosine Transformation is done and of the 26 coefficients the first 13 coefficients leaving out the zeroth are being taken. An important fact about the speech signal is that it is not stationary from frame to frame. For this reason we add features related to the change in cepstral features over time. We do this by adding to each of the 13 features, 13 delta features, and 13 double delta features, thus making the total dimension of the MFCC feature vector to 39.

2.2. Inner Hair Cell Coefficients (IHC)

The speech signal is often corrupted by noises such as other speech signals, background noises, etc. Humans can communicate with each other by paying attention only to the necessary information under those situations. Therefore, in order to improve the conventional speech signal processing system a system which mimics some aspects of the human auditory system may be used. The Meddis Auditory Periphery (MAP) model is one such model. Fig.3 shows the block diagram of such a psycho-acoustic filtering model proposed by Roy Patterson [7].

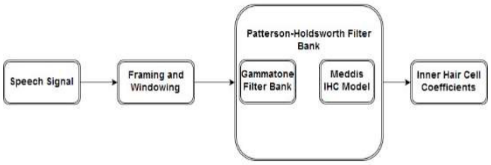

Fig. 3. Block Diagram of IHC

This model given in Fig.3 combines a Gammatone filter bank with an inner hair cell model proposed by Ray Meddis [8]. The speech signal is split up into short frames of 25ms with an overlap of 15ms. Before picking up features, we multiply the framed signal by a Hamming window to reduce spectral leakage caused by the framing of the signal. The most common model that fits the response of the auditory filter is a bank of gammatone filters. The Gammatone filter is called so because its impulse response is the product of a carrier tone at the center frequency and a gamma distribution function which gives the impulse response its shape. The gammatone filter is defined in the time domain with the impulse response is given by equation (3) [17].

$$h(t) = at^{n-1}e^{-2\pi\beta t}co\,s(2\pi f_c t + \psi) \tag{3}$$

where, *h(t)* is the impulse response of the filter, *a* is the filter gain, n is the filter order, *b* is the filter bandwidth in Hz, f_c is the center frequency in Hz and ψ is the phase in radians. Firing impulses are sent to the neurons in the brain depending on the vibrations of the inner hair cells in the ear. This firing activity is implemented using the Meddis IHC model [8]. Here 64 gammatone filters are being used and accordingly 64 IHC coefficients were extracted.

2.3. Pitch and Formants

The main characteristic of the voice source features is that, unlike the traditional short-term spectral features, it lasts over long segments like words and utterances and reflects differences in style of speaking, language background, type of sentences etc. Combining fundamental frequency related features with spectral features has shown to increase the speaker recognition rates, especially in noisy conditions [1]. Pitch frequency is the fundamental frequency of vibration of the vocal folds, which are present at the top of the trachea. Pitch period can be estimated by quantifying the period by using autocorrelation, or measuring the harmonics. Formant frequencies can be found by linear prediction analysis from the poles [18]. Pitch and frequency are also influenced by the affective state of the user [15]. Formant frequencies mainly vary due to the phoneme being articulated and not so much by factors such as age, weight, etc.

3. Speaker Recognition Methods

After extracting the features from input speech signals, a speaker model is required to be created for the purpose of storing in the database for later comparisons. Speech production is not deterministic, in that, a particular sound is not produced by a speaker with exactly the same vocal tract shape and glottal flow due to coarticulation and anatomical variations. Hence statistical models are used. The modeling methods used are: Gaussian Mixture Modeling (GMM) and i-vector method. Two classifiers are used for i-vector method: Cosine Distance Scoring (CDS) and Probabilistic Linear Discriminant Analysis (PLDA).

3.1. GMM

A Gaussian mixture model (GMM) represents feature distribution as the weighted sum of multiple Gaussian distributions. A C-component Gaussian mixture density is represented by the equation (4) as,

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i\, b_i \overset{\vec{x}}{)} \tag{4}$$

where $\vec{x}$ is a D dimensional random vector, $\underset{b_i}{\overset{\vec{x}}{}}$) with $i = 1,2,..,M$ are component densities and p_i with

$i = 1,2,..,M$are the mixture weights [9]. Each component density is a D -variate Gaussian function of the form as represented in equation (5).

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_i|^{\frac{1}{2}}} \exp(\frac{-1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \qquad (5)$$

with parameters, mean vector $\vec{\mu_i}$ and covariance vector Σ_i . The mixture weights satisfy the constraints, $\sum_{i=1}^{M} p_i = 1$ and $p_i \geq 0$ [9]. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and the mixture weights from all component densities. These parameters are collectively represented by the notation given in equation (6).

$$\lambda = \{p_i, \vec{\mu_i}, \Sigma_i\}, i = 1, \dots\dots\dots., C \qquad (6)$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her modelλ. GMM is trained using Expectation-Maximization (EM) algorithm [1]. A Universal Background Model (UBM) is generated using speech samples from all the different speakers and then MAP adapted to obtain models for each of the individual speakers. For testing purpose, the feature vectors extracted from test signal are compared against the available speaker models in the database and the model with the highest likelihood score is identified to be the speaker.

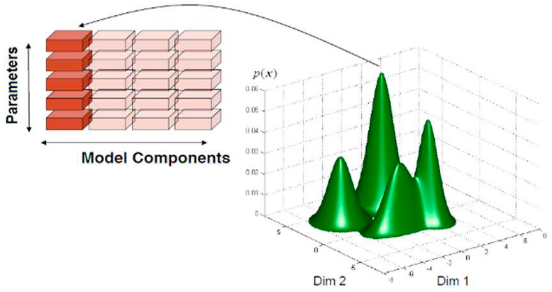

Fig. 4. Two Dimensional GMM

3.2. i-Vector method

The speaker and channel variability in GMM can be modelled by using the concept of GMM supervector, which consists of the stacked means of all mixture components. The traditional MAP adaptation performed for GMM systems does not take into consideration the effects of channel distortion, especially when the train and test sessions are from different channels. Hence we go for a different method known as the i-vector method, where every utterance is represented by a low dimensional feature vector. The i-vector based speaker recognition system makes use of a combination of channel and speaker spaces. There is no separate modeling done for the channel and speaker variabilities. Instead, every utterance is projected onto a low dimensional space known as the Total Variability space or TV space and the low dimensional feature vector used to represent each utterance is called the identity-vector or i-vector [10]. Once the MFCC or IHC feature vectors are obtained from the speech samples of different speakers, the next step is to represent them in the form of i-vectors.

Each speaker and channel-dependent GMM super vector ***m*** can be modelled using equation (7).

$$m = m_u + Tw \qquad (7)$$

where $m_u$is a speaker and channel independent super vector, whose value is often taken from UBM super vector,***T*** is the total variability matrix with low rank, which expands a subspace containing speaker and channel-dependent information and ***w*** is a standard normal distributed vector. The i-vector for a given utterance can be extracted using equation (8) [16] as,

$$\hat{w} = (I + T^T\Sigma^{-1}N(u)T)^{-1}T^T\Sigma^{-1}F(u) \quad (8)$$

Where, ***I*** is a *RXR* identity matrix *C* is the number of GMM components and *F* is the dimension of features), ***N***(u) is a diagonal matrix of dimension *CFXCF*, Σ is a diagonal covariance matrix of dimension *CFXCF*, ***F***(u) is a super vector of dimension *CFX1* . The covariance matrix Σ represents the residual variability not captured by ***T*** [10]. The i-vector, denoted as ***w***, can be obtained as the posterior distribution conditioned on the Baum-Welch statistics of each utterance. **T** matrix is trained using the EM algorithm using the whole training data.

3.2.1. Channel compensation methods

The extraction of i-vectors is done in such a way that there is no distinction made between the speaker and channel variability. So the separation or removal of channel variability is taken care of before creating the classifiers for the recognition of speakers, which take i-vectors as the input features. Channel compensation methods are estimated based on the within-class and between-class variances [10]. Here, a combination of Linear Discriminant Analysis (LDA) and the Within Class Covariance Normalization (WCCN). The main purpose of using LDA is to achieve dimensionality reduction while retaining as much of speaker discriminatory information as possible and it is a supervised method. LDA maximizes the between-class variability and at the same time, minimizes the within-class variability. Within class covariance normalization is done as a pre-processing step for performing the PLDA classification [10]. This helps to attenuate high variance within the same speaker class and hence helps in achieving better recognition during testing phase.

3.2.2. Classifiers used for i-vector method

In this work two classifiers are used for classification: CDS and PLDA. Cosine distance scoring (CDS) is a popular classification algorithm. It uses a cosine kernel to directly compare two input feature vectors and give out the degree of similarity between the two feature vectors. Here they are the i-vectors that represent reference speakers speech sample and test speaker's sample. The similarity score is given in equation (9),

$$S_{CDS}(w_t, w_r) = \frac{<(w_t, w_r)>}{(||w_t||)\,(||w_r||)} \quad (9)$$

where $w_t$is the test i-vector and w_r is the reference i-vector [11]. The use of the cosine kernel as a decision score for speaker verification makes the process faster and less complex.

In PLDA the within-speaker variability is being modelled by a residual term and we omit the channel subspace. The residual term ϵ is assumed to be having Gaussian distribution with the covariance matrix ofΣ. Given two i-vectors for test and reference, the PLDA method computes the identification score using the likelihood ratio as shown below in equation (10).

$$\mathbf{S}(\mathbf{w_t}, \boldsymbol{w_r}) = \frac{\boldsymbol{p(w_t, w_r|H_1)}}{\boldsymbol{p(W_r|H_0)\,p(w_t|H_0)}} \quad (10)$$

where w_t is the test i-vector, , w_r is the reference i-vector, H_1 is the hypothesis that both i-vectors belong to the same speaker, and H_0 is the hypothesis that both i-vectors belong to different speaker [12].

4. Implementation

The implementation was done in MATLAB with speech signals taken from TIMIT database, which contains 10 speech signals each for 630 speakers. All signals were sampled at a rate of 16 KHz and each of these signal duration is 2s to 3s. Out of these, 100 speakers with 10 speech signals each were used for the implementation. Of the 10 signals, 7 were used for training and 3 for testing each speaker. Each of these signals were framed into 25ms frames with 15ms overlap. Hamming window was used and the simulation was performed for both full speech and voiced signals. Using the parameters zero crossing rate (ZCR) and energy, the voiced and unvoiced parts of the speech signal, were separated and we used normalized speech lying in the range -1 to +1. To obtain the voiced speech signals, all frames with a power greater than 0.5 or ZCR less than 100 is taken [13]. From each frame of the speech signal 39 MFCC and 64 IHC coefficients were extracted. Pitch and formants were appended to MFCC and IHC coefficients to see the changes in the accuracy rates. The number of Gaussian components selected is 32 and for i-vector modelling, the size of i-vector is chosen to be 100. CDS and PLDA were used as classifiers for i-vector.

5. Results and Discussion

The accuracy rates of speaker recognition has been obtained for a text-independent speaker recognition system using GMM and i-vector methods. The i-vector method was classified using CDS and PLDA and it was found that the recognition system works well for PLDA. Two feature extraction methods were used: MFCC and IHC. The accuracy rates obtained for different modeling methods and classifiers using voiced speech is shown in Table.1. From that, it can be understood that PLDA performs better than CDS for i-vector method.

The performance comparison of the system for full speech and voiced speech was tested and it shows that MFCC works better than IHC for both full speech and voiced signals. Also, full speech recognition is better than voiced speech recognition with both features. The results are given in Table.2.Here we used short utterances obtained from TIMIT database of duration 2-3s for testing.

Table 1. Accuracy Rate Obtained for Voiced Short Utterance (Test data).

Feature	GMM	i-vector using PLDA	i-vector using CDS
MFCC	89.33	73.66	56.33
IHC	68.33	63.33	44.66

Table 2. Accuracy Rates Obtained for Full Speech and Voiced Signals

Training	Testing	Feature	GMM	i-vector(PLDA)
Full speech	Full speech	MFCC	94.33	79.66
Voiced	Voiced	MFCC	89.33	73.66
Full speech	Full speech	IHC	69.66	70.66
Voiced	Voiced	IHC	68.33	63.33

From the results shown in Table 1and Table 2, it is understood that GMM outperforms i-vector because i-vector based speaker recognition systems require large amount of data for estimating its parameters. To get a better performance, we concatenated the three test speech signals and verified the accuracy rates [14]. By concatenating, the duration of test signals is increased from 6s to 9s. The results obtained are tabulated in Table 3. Also the recognition accuracy was tested by appending the speaker specific information like pitch and formants onto the two features. One pitch and six formants were taken from each frame thereby increasing the dimensions of MFCC to 46 and IHC to 71. The results are as shown in Table.3.The accuracy rates was found to increase when the features were appended with pitch and formants. When training data was used as test data, the accuracy was always almost 100 %.

Table 3. Accuracy Rate obtained after combining with pitch and formants

Feature	GMM(short utterance	GMM (long utterance	i-vector(PLDA) (short utterance)	i-vector(PLDA) (long utterance)
MFCC	89.33	96	73.66	94
MFCC+Pitch+Formants	96.66	99	91.33	98
IHC	68.33	81	63.33	79
IHC+Pitch+Formants	82.33	89	80.33	88

6. Conclusion

The performance comparison of two different speaker recognition systems was implemented. MFCC is considered as a human peripheral auditory system and human perception of sound does not follow a linear scale. In MFCC we are using a Mel filter bank which maps the linear frequencies onto a Mel scale. IHC takes into account the physiological changes of the mammalian peripheral auditory system. These neural response are modelled using the Patterson-Holdsworth Meddis hair cell model. Even though the IHC is modelling the physiological variations of the speech signal, it is found to be less accurate than MFCC. From the analysis, it is also clear that in case of short utterances GMM performs better than i-vectors and there was a significant increase in the accuracy rates when concatenated test signals were used. Also, it can be found that PLDA outperforms CDS for i-vector method. There was a considerable increase in the accuracy rates when pitch and formants were appended.

References

1. Tomi Kinnunen, Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 2010; 52, 12-40.
2. Douglas A. Reynolds. An Overview of Automatic Speaker Recognition Technology.IEEE 2002; 4072-4075.
3. Roberto Togneri, Daniel Pullella, An Overview of Speaker Identification: Accuracy and Robustness, IEEE Circuits and Magazine, 2011, 23-61
4. Siddhant C. Joshi, A.N.Cheeran. MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition. International Journal of Science, Engineering and Technology Research, 3 2014; 1820-1823
5. Namratha Dave. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. International Journal for Advanced Research in Engineering and Technology, 2013.
6. Diksha Sharma, Israj Ali, A Modified MFCC Feature Extraction Technique for Robust Speaker Recognition, International Conference on Advances in Computing, Communications and Informatics 2015, 1052-1057.
7. R. D. Patterson et al.,Complex sounds and auditory images, in Acoustical Signal Processing in the Auditory System, Advances in the Biosciences, 83, 1992.
8. R. Meddis, Simulation of mechanical to neural transduction in the auditory receptor, Journal of the Acoustical Society of America, 79, 1986
9. Douglas A Reynolds, Speaker Identification using Gaussian Mixture Speaker Models,IEEE Transactions on Speech and Audio Processing, 3, 72-83, 1995
10. Najim Dehak, Patrick J. Kenny, Dehak, Pierre Dumouchel, and Pierre Ouellet ,Front-End Factor Analysis for Speaker Verification,, Transactions on Audio, Speech and language Processing,, 19, 2011
11. Lei Lei, She Kun, Speaker Recognition Using Wavelet Cepstral Coefficient, I-Vector, and Cosine Distance Scoring and Its Application for Forensics, Journal of Electrical and Computer Engineering, 2016
12. Padmanabhan Rajan, Anton Afanasyev, Ville Hautamaki,Tomi Kinnunan, From Single to Multiple Enrollment i-vectors : Practical PLDA Scoring Variants for Speaker Verification, Digital Signal Processing, 2014
13. Dominic Mathew, V.D. Devassia and Tessamma Thomas, A K-means Clustering Algorithm for Frequency Estimation and Classification of Speech Signals, IEEE Conference/ICSIP, 2006
14. A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, I-vector based speaker recognition on short utterances, Proceedings of the 12th Annual Conference of the International Speech Communication Association, .2341.
15. D. Sharma and P. A. Naylor, Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment, Proc European Signal Processing Conf. Cite seer, 2009, 2514-2518.
16. Wei LI, Tianfan Fu and Jie Zhu, An Improved i-vector Extraction for Speaker Verification, EURASIP journal on Audio, Speech and Music Processing, 2015; 345-354
17. Masahiro Abuku, Tadahiro Azetsu, Eiji Uchino and Noriaki Suetake, Application of peripheral auditory model to speaker identification, Second World Congress on Nature and Biologically Inspired Computing, 2010, 666-671.
18. Bageshree V. Sathe-Pathak and Ashish R. Panat, Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person, International Journal of Computer Science Issues, 9, 2012, 296-299.