IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Speaker Identification with Whispered Speech

Nisha Beegum S

M.Tech Scholar, Department of ECE, MZCE, Kadammanitta, Pathanamthitta, Kerala

Abstract: For an access control system, which is a speaker identification system based on whispered speech. Speaker identification is a main function of an access control system. Hence, a novel speaker identification system using instantaneous frequencies is proposed. The input speech signals pass through both signal independent and signal dependent filters firstly. Then, we derive the signal's instantaneous frequencies by applying the Hilbert transform. The analysed instantaneous frequencies are proceeded to be modelled as probability density models. Using these probability density models as the feature in the proposed speaker identification system. Here, compare the use of parametric and nonparametric probability density estimation for instantaneous frequency modeling. Furthermore, propose an approximated probability product kernel support vector machine (APPKSVM). In the APPKSVM, Riemann sum is applied in approximating the probability product kernel. The whisper sounds from the chain speech corpus were used in the experiments. Results of the experiments show the superiority of the existing speaker identification system. For the proposed system using the empirical mode decomposition.

Keywords: MFCC, APPKSVM, FP model, LPCC, IF, EMD, RPS

I. INTRODUCTION

The number of house break-in cases increases nowadays. The personal safety at home has always been a concern in current daily life. An access control system serves as an essential role to prevent house break-in. Verification using password and PIN codes is inadequate since it is easy to be replaced. Therefore, biometric-based access control systems gradually become the mainstream. There are many commonly used biometric modalities, e.g., irises, faces, fingerprints, and speech . Among all possible biometric modalities, speech has the advantages of simple obtainment, non-invasiveness, and convenient usability. Rather than adopting natural speech, this work focuses on whispered speech because it is not easy to be copied and has high privacy. Speaker identification is a primary function of an access control system. As computer science technology has rapidly advanced in recent years, such system can be fooled, such as by using speaker voice conversion software. Normal speech-based speaker identification systems are thus unsafe since the normal speeches are easy to collect in the daily life, and these collected speeches can be utilized to help the speaker voice conversion software simulate well normal human speeches. However, people seldom whisper, so recording whispered speeches is much more difficult than recording normal speeches.

A whisper-based speaker identification system is accordingly safer than a normal speech-based speaker identification system. Thus to develop a novel speaker identification system that is based on whispered speech. Compared with neutral speech, whispered speech has different characteristics. It has much lower energy and signal-to-noise ratio (SNR). Moreover, the lower frequency formants of whispered speech are shifted to higher frequencies . To improve the performance of speaker identification based on whispered speech, Jin et al. suggested the use of feature warping and frame-based score competition for classification. The use of static linear frequency cepstral coefficient vectors, which is based on linear frequency scale, is proposed as features. In addition, Gu and Zhao proposed the use of a factor analysis and support vector machine for whispered speech speaker identification. A novel set of acoustical features for speaker identification based on whispered speech.

Feature extraction plays an important role when performing speaker recognition. The choice of features representing the speech signal influences the recognition performance. In the work of speaker recognition, there are two conventional acoustic feature extraction approaches, i.e., parametric methods and nonparametric methods. Parametric methods can match the resonant structure of the human vocal tract. An example of parametric methods is linear prediction cepstral coefficients (LPCCs) which are obtained using linear predictive analysis. The use of Mel frequency cepstral coefficients (MFCCs) is an example of nonparametric methods. In addition to the conventional LPCCs and MFCCs, AM-FM modelswhich have been applied to various speech processing that provide an alternative way to extract features for speakerrecognition.

The AM-FM model measures the amount of amplitude and frequency modulation that exists in speech resonances. Using this model to extract features can provide acoustic information that the conventional speech features lack. In modern speaker recognition systems, statistical Gaussian mixture models (GMMs) and support vector machines (SVMs) are commonly applied for classification. GMMs achieve a satisfied performance in speaker modeling by estimating efficiently parameters and maximizing the likelihood. Unfortunately, GMM may not fit a classification tack since it is a generative method. A discriminative approach thus becomes a key factor for performing effective speaker recognition. One of common discriminative approaches is support vector machine, which efficiently trains nonlinear decision boundaries.

DOI 10.17148/IJARCCE.2016.56122

Considering the advantages of an SVM, this paper developed an SVM-based speaker identification system. As mentioned above, the properties of whispered speech differ from those of normal speech. The instantaneous frequencies have been proved to capable of capturing the properties of a whispered speech. To match whispered speech and perform effective speaker identification, a Gabor filter and empirical mode decomposition are used to extract the instantaneous frequencies from the speech signal. The extracted instantaneous frequencies are further modeled as parametric or nonparametric probability densities, and then the probability densities can be used to train SVM with a probability product kernel. The experimental results demonstrate the effectiveness of the proposed system

II. EXISTING SYSTEM

A. System Overview

The overview of the speaker identification system is shown in Fig.1 illustrates the system flowchart.

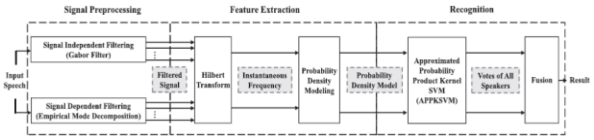

Signal Preprocessing
Feature Extraction
Recognition
Input Speech
Signal Independent Filtering (Gabor Filter)
Signal Dependent Filtering (Empirical Mode Decomposition)
Filtered Signal
Hilbert Transform
Instantaneous Frequency
Probability Density Modeling
Probability Density Model
Approximated Probability Product Kernel SVM (APPKSVM)
Votes of All Speakers
Fusion
Result

Fig. 1 Block diagram of Speaker identification system

The blocks with solid lines in the figure show the functionalities of the existing system, while the gray blocks with dotted lines show the outputs of the functional blocks. The existing system consists of three phases, i.e., signal pre-processing, feature extraction and recognition. First, the AM-FM model is utilized to model a speech resonance by means of an AM-FM signal.

$$u_i(t) = a_i(t)\cos(2\Pi \int_0^t fi(\tau)d\tau) \qquad (1)$$

where $f_i(\tau)$and $a_i(t)$represent the instantaneous frequency and amplitude, respectively.

Assume a speech signal with resonant components. The speech signal is a superposition of AM-FM signals, expressed by

$$S(t)= \sum_{i=1}^{k} ui(t)=\sum_{i=1}^{t} \cos(2\Pi \int_0^t fi(\tau)d\tau) \qquad (2)$$

A demodulation operation is required to calculate $f_i(\tau)$and $a_i(t)$ in a speech resonancecomponent . The Teager energy operator and Hilbert transform can be used for this purpose.Next, we want to further capture the property of the input speech. Boviket al. introduced the multiband demodulation analysis (MDA), a two-stage framework, to demodulate a signal. The MDA uses filtering to isolate each resonance signal from speech and then performs demodulation to each individual resonance. Gabor filter (signal independent filters) iscommonly used in the filtering stage for its optimal compactness and smoothness in the time and frequency domains. In addition, the empirical mode decomposition, regarded as a signal dependent filter, proposed by Huang searches for the expression of a signal using a mixture of few components, and each component is a purely oscillatory function. A such function is called intrinsic mode function (IMF). An IMF can be regarded as an AM-FM separated component. Integration of the EMD and the Hilbert transform are designated as the Hilbert–Huang transform (HHT) by NASA. The HHT received considerable attention in the last fault diagnosis, medical pathology, etc. In this work, signal independent multiband filters as well as signal dependent EMD filtering areutilized to analyze instantaneous frequencies (IF) and develop useful acoustical features for speaker identification. From this point, the signals resulting from the different filters are processed separately. Afterwards, the analysed instantaneous frequency set of the signals is modeled with a probability density estimator.

The probability density models are then fed as features to the recognition phase. Existing system adopts SVM in the recognition phase. Exploring the kernel operation in an SVM, as it is an effective approach to measure the similarity of two vectors in a high dimensional space. The use of different kernels may influence the classification performance of an SVM. Selecting a kernel to be used in an SVM is application dependent. Based on the instantaneous frequency features, we propose an approximated probability product kernel support vector machine (APPKSVM), which uses Riemann sum to approximate the integral of two probability density models. Finally, a fusion of the recognition results of bothapproaches is performed. The final output of the system is the recognition result of the input speech.

B. Instantaneous Frequency Extraction

The approach used in extracting the instantaneous frequencies of an input speech signal is described. The first two sections describe the two filtering approaches used, i.e., Gabor filtering and empirical mode decomposition. Then, how the instantaneous frequencies are derived from the filtered speech signal using the Hilbert transform is also shown.

1) Gabor Filtering:

The Gabor filters can be used to extract separated speech resonant components. The impulse response of a Gabor filter is defined by

$$g(t) = \exp(-\alpha^2t^2)\cos(2\Pi fct) \qquad (3)$$

The corresponding frequency response is expressed by

$$G(f)= (\frac{\sqrt{2\Pi)}}{2\alpha})[\exp\{-\Pi^2(f-fc)^2/\alpha^2\}+\exp\{-\Pi^2(f+ fc2/\alpha2\}] \qquad (4)$$

where f_c denotes the center frequency of the filter, and α represents the bandwidth control parameter which results in an effective bandwidth. $\alpha/(\sqrt{2\Pi})$Proper selection of the bandwidth of band pass filters is essential for signal

DOI 10.17148/IJARCCE.2016.56122

analysis and characterization. For formant tracking, Potamianos et al used a Gabor filter bank with constant bandwidth of 400 Hz to catch speech resonances and form a pyknogram. Besides a constant bandwidth of 400 Hz, Grimaldi and Cummins exploited another two bandwidths for speaker identification on the Mel scale: 266 Mel and 106 Mel. The chosen center frequencies of the Gabor filters were uniformly spaced on the Hertz scale.

2) Empirical Mode Decomposition :

Empirical mode decomposition decomposes a signal into several intrinsic mode functions. The decomposed signal's intrinsic property can be presented efficiently by these IMFs. IMFs center in the origin of a phase plane, and this meets the requirement to compute correct instantaneous frequencies using Hilbert transform. The process of EMD having the following steps .

1) For the original signal s(t) or intermediate signal h(t), the envelopes are wrapped to obtain upper envelope u(t)and lower envelope l(t) using the spline function.

2) Calculate the average of the envelopes obtained from step 1. The intermediate signal $h_1(t)$ is then obtained by subtracting the average of the envelopes $\mu_1(t)$ from the original source

$$\mu_1(t) = (u(t) + l(t))/2 \qquad (5)$$

$$h_1(t) = s(t) - \mu_1(t) \qquad (6)$$

3) Following the same procedure as steps 1 and 2, the average of the envelopes $\mu_{11}(t)$ for $h_1(t)$ is calculated. Then, the intermediate signal $h_{11}(t)$ for the first iteration is obtained by

$$h_{11}(t) = h_1(t) - \mu_{11}(t) \qquad (7)$$

4) Do the above three steps iteratively until the procedure reaches the stop criteria, where the threshold SD is usually set as 0.3. The first IMF $h_{1K}(t)$ is then generated

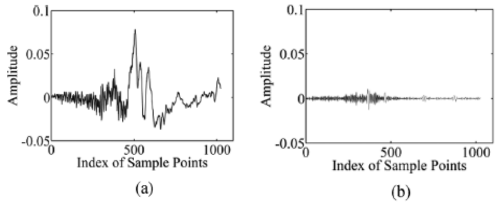
$$SD \geq \sum_{k=2}^{K} \frac{[h_{1k}(t) - h_{1k-1}(t)]^2}{h_{1k-1}^2(t)}. \qquad (8)$$

Similarly, the other IMFs can be generated by passing the remaining signal $r_1(t)$ back to the EMD algorithm above

$$r_1(t) = s(t) - h_{1K}(t) \qquad (9)$$

Finally, the original signal is decomposed into the combination of several IMFs .

$$s(t) \approx h_{1K}(t) + h_{2K}(t) + \dots\dots\dots + h_{NK}(t) \qquad (10)$$



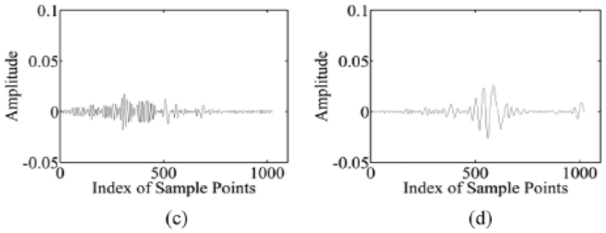


Fig. 2 Example of EMD

Fig. 2(a) shows an example of EMD. Figure displays the waveform of a frame, which has a size of 1024 sample points. We can obtain 13 IMFs for the frame using EMD. Fig. 2(b)–(d) shows three of the 13 IMFs.

3)Speech Demodulation :

The two fundamental approaches to AM–FM signal demodulation are the Hilbert transform and the energy operator. The former mainly adopts a linear integrator operator while the latter uses a nonlinear differential operator. This work adopts Hilbert transform to perform speech demodulation. For any signal s(t) , its Hilbert transform is H(t) defined as

$$H(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{s(\tau)}{t - \tau} d\tau \qquad (11)$$

where P is the Cauchy principle value of the singular integral.

With the signal s(t) and its Hilbert transform H(t) , we can obtain the analytic function Z(t)

$$Z(t)= s(t) +H(t) =\sqrt{S^2(t) + H^2(t)}e^{j\Theta}(12)$$

where the phase θ(t) = arctan (H(t) / s(t))

The instantaneous frequency can then be calculated from the first differential of the phase

C. Probability Density Modeling Of Instantaneous Frequency

Here applies both EMD and multiband Gabor filtering. Let $S_k(t)$ denotes the k^{th} IMF obtained by the EMD or the signal after going through the k^{th} band pass filtering by the Gabor filtering.

Through Hilbert transform, obtain the instantaneous frequency of $S_k(t)$, denoted as $w_k(t)$. For the k^{th} frame of filtered signal $S_k(t)$, its corresponding instantaneous frequencies set is described as :

$$W_{ik}=\{w_k[(i -1)L],w_k[(i -1)L+1],\dots,w_k[(i -1)L+(L-1)]\} \qquad (14)$$

Here, L is the frame size. Assume frame i has K sets of instantaneous frequencies, i.e.,

$W_i = \{w_{i1}, w_{i2}, \dots, w_{ik}\}$. Each is then modeled as the probability density function by performing the probability density estimation on it. Two approaches used in this work are parametric density modelling (PDM) and nonparametric density modeling (NPDM).

DOI 10.17148/IJARCCE.2016.56122

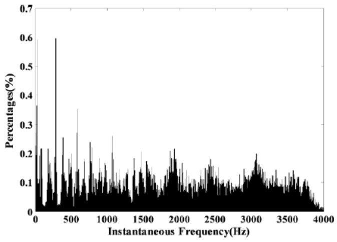


Fig. 3 Histogram of instantaneous frequencies

Fig. 3 shows the histogram of instantaneous frequencies. This histogram is computed from a frame data of a whispered speech. The frame size is 1024 points.

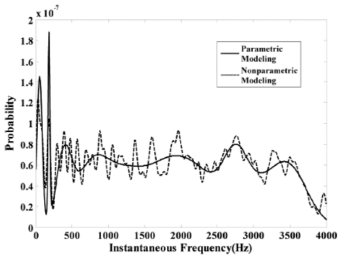


Fig. 4 Probability density functions of models

Fig. 4 shows the corresponding probability density functions (PDFs) obtained using parametric modeling and nonparametric modeling.

1)Parametric Density Modeling:
Gaussian mixture models are often used to model the probability density of complex signals such as images and audio, and yields excellent performance in many applications. GMMs are utilized to model the probability distribution of instantaneous frequencies. In a GMM, the data is assumed to belong to a mixture of Gaussian distributions. Given an instantaneous frequency, the GMM is expressed as follows:

$$f(w;\theta) = \sum_{c=1}^{C} v_c g(w;\mu_c,\sigma_c) \qquad (15)$$

wherev$_c$ is the mixing parameter which represents the weight of the cth Gaussian probability density function g(w;μ_c,σ_c). The mixing parameters should satisfy the following constraints:

$$v_c \geq 0 \quad \text{and} \; \sum_{c=1}^{C}(V_c) = 1 \qquad (16)$$

To obtain the best parameters of Gaussian probability density functions, the expectation-maximization (EM) algorithm is applied.

2)Nonparametric Density Modeling:
Kernel density estimation is a nonparametric density modelling approach that has been used to model important problems in various applications. Using kernel density estimation, an instantaneous frequency is modeled as follows:

$$f(w) = \sum_{l=1}^{K\times L} u_l \phi(w;w_l,\sigma_l) \qquad (17)$$

where ϕ(.)is the kernel function, and w$_l$is an IF instance which belongs to W$_i$; u$_l$ and σ_l are the weight and bandwidth of the lthkernel, respectively. A commonly used kernel function is the Gaussian function

$$f(w) = \sum_{l=1}^{K\times L} u_l \exp\left(-\frac{\|w-w_l\|^2}{2\sigma_l^2}\right). \qquad (18)$$

For the estimators, this work adopts a fixed bandwidth kernel which gives the following equation :

$$f(w) = \frac{1}{K \times L} \cdot \frac{1}{\sqrt{2\pi}\sigma} \sum_{l=1}^{K\times L} u_l \exp\left(-\frac{\|w-w_l\|^2}{2\sigma^2}\right) \qquad (19)$$

where σ denotes the bandwidth.

D. Approximated Probability Product Kernel SVM
The SVM theory is a model that can be used to perform various classification tasks. The key concept of SVM is finding an optimal hyper plane that separates the two classes withthe largest margin in the feature space. This idea comes from the work of structural risk minimization (SRM) induction principle, which considers the bound of generalization error instead of minimizing the mean square error. Besides the conventional SVM, many variations of SVM have also been proposed. Given a hyperplane ,w.x +b , and w ϵ R^N , and b ϵ R. The decision function of classifying a unknown point x is defined as

$$f(\mathbf{x}) = \text{sign}(\mathbf{wx}+b) = \text{sign}(\sum_{i=1}^{N_S} \alpha_i m_i \mathbf{x}_i \cdot \mathbf{x}) \qquad (20)$$

where N$_S$ denotes the number of support vectors, x$_i$ is a support vector with its corresponding Lagrange multiplier α_i; m$_i$ ϵ {-1, +1} refers to the binary class label. In fact, it is often hardto search an optimum hyperplane that can separate two classes of data in the input feature space. To solve the problem, SVM usually projects data to another feature space, and then searches the optimal hyperplane. This trick is implemented by kernel methods. Define φ (x) as the projection of the original feature vector x, and φ is a mapping function that projects a feature from the original space R^N to another feature space. Given a kernel K(. , .), which has the following form:

$$K(x_i,x_j) = \varphi(x_i).\varphi(x_j) \qquad (21)$$

Finally, the decision function becomes

$$f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{N_S} \alpha_i m_i K(\mathbf{x}_i,\mathbf{x}) + b). \qquad (22)$$

DOI 10.17148/IJARCCE.2016.56122

The probability product SVM, which considers a probability product kernel. A probability product kernel treats each IF data sets W_1,W_2,…..W_i as a probability distribution function p_1(w),p_2(w),…..p_i(w). The kernel is defined as

$$k(p_i,p_j)=\int_w p_i^{\rho}(w)p_j^{\rho}(w)dw=\left\langle p_i^{\rho},p_j^{\rho}\right\rangle_{L_2} \qquad (23)$$

where L_2 is a Hilbert space, and ρ is a positive constant. From the above equation, we see that the use of probability product kernel allows the introduction of prior knowledge of data. However, it is not always possible to obtain a closed form. Some approximation techniques for this problem are the Trapezoidal rule and Simpson's rule. We adopt the Trapezoidal rule to approximate.

The Trapezoidal rule supposes a given function f that is continuous and positive on the interval [a,b]. The integral of f on this interval is then the area under f along the interval [a,b] on the x axis.

Based on the rule, we use N_T trapezoids to approximate the definite integral. Suppose there are N_T partitions, (x_{k-1},x_k),x_{k-1}<x_k , in the interval [a,b], with each partition having the length of (b-a)/ N_T. The area of the kth trapezoid can then be written as

$$\gamma(k)=\frac{[f(x_k-1)+f(x_k)](b-a)}{2N_T} \qquad (24)$$

and the integral of f on the interval [a,b] can thus be approximated as the sum of all ϒ(k). Since p_i (w) is continuous on [0,fs/2] , where f_s is the sampling rate, the Trapezoidal rule is satisfied and can be approximated by

$$\begin{aligned}k(p_i,p_j)&=\int_w p_i^{\rho}(w)p_j^{\rho}(w)dw\\&\approx\frac{f_s}{4N_T}[p_i^{\rho}(x_0)p_j^{\rho}(x_0)+2p_i^{\rho}(x_1)p_j^{\rho}(x_1)+2p_i^{\rho}(x_2)p_j^{\rho}(\\&\cdots+2p_i^{\rho}(x_{N_T-1})p_j^{\rho}(x_{N_T-1})+p_i^{\rho}(x_{N_T})p_j^{\rho}(x_{N_T})]\end{aligned} \qquad (25)$$

where N_T denotes the number of subintervals [0,fs/2]. Equation (25) is the proposed approximated probability product kernel that can handle both PDM and NPDM of the instantaneous frequencies.

The advantage of this approach is that the error e is promised to be below the threshold, depending on the partition number N_T.

The threshold is defined as

$$e\leq\frac{(b-a)^3}{12{N_T}^2}\left[\max\left|f''(x)\right|\right],a\leq x\leq b \qquad (26)$$

Where f " (x)is the second derivative of f. The error decreases when the partition number N_T increases, resulting closer to the probability product kernel.

E. Frame-Based Speaker Identification By SVM And Fusion System

A multiclass support vector machine (multiclass SVM) is utilized as a classifier. A frame-level voting strategy is applied to the identification phase. For each frame of a speech utterance, there are C_2^M one-versus-one SVMs used to classify its speaker. Totally C_2^M classification results are obtained, and each result is regarded as a speaker vote for the frame. For a frame t, the obtained votes are merged to a voting vector $I_t \epsilon R^M$, where I_t (m) is the number of votes that frame t is classified to the mth speaker. After all frames of a speech utterance have been classified, the speaker model with the most votes is chosen as the identification result. Suppose we have N frames, then for each speech utterance, we have a voting vector

$$SC=\textstyle\sum_{t=1}^{N}(I_t), \qquad SC\epsilon R^M \qquad (27)$$

with SC (m) denoting the number of votes that the speech utterance is voted as belonging to the mth speaker model. The classification results is then given by

$$m'=\text{argmax}SC(m) \qquad (28)$$

In the existing speaker identificationsystem, two individual classification systems are used. To integrate the two systems, we adopt a linear fusion approach in performing the speaker identification task. Suppose the classification result of a speech utterance using the Gabor filter-based speaker identification system is denoted as SC_G and using the EMD filter-based speaker identification system is denoted as SC_E . Perform the fusion using

$$m' = \text{argmax}\,(\alpha * SC_G(m) + (1 - c) * SC_E(m)) \qquad (29)$$

where α is a weighting factor which value is 0 ≤. α ≤. 1

F. EXPERIMENTAL RESULTS

1)Experimental Setup :

Chain corpus is a conventional speech database for speaker recognition. For each sentence recorded from the same speakers, the chain corpus consists of various speaking styles from the same speakers. The speaking styles include normal, fast, whisper, etc. In this paper, we address the problem of speaker identification based on whispered speech.

Thirty-two speaker sets are chosen and each speaker set includes equal numbers of males and females. Each speaker uttered 33 sentences, and each sentence is about 1~2 seconds. In the training phase, we trained the SVM classifier using features extracted from the 24 sentences of speakers. The total length of these 24 sentences is about 25 seconds. The other three sentences were used to develop the fusion classifier. In the identification phase, the remaining 6 sentences were used to test the proposed identification system. The sampling rate is 8000 Hz and the frame size is 1024 samples, with a 50% overlap in two adjacent frames.

DOI 10.17148/IJARCCE.2016.56122

2)Optimization of the Subsystem Parameters :
The identification system can be decomposed into two subsystems: Gabor filter-based system and EMD-based system. Choose the applicative parameters of these two subsystems. The main parameters include GMM mixture number for the parametric modeling and the bandwidth used for nonparametric modeling. For EMD-based system, which IMFs should be chosen to extract IFs is also decided. For the EMD-based system, this work empirically chooses the first IMF to seventh IMF to extract IFs. The energy distribution of IMFs is like a Gaussian distribution, and the peak is at fifth IMF. The chosen IMFs contribute a large portion of the total energy.

After extracting the IF, both of the parametric modeling andnonparametric modeling are adopted. For the Gabor filter-based system, Gabor bandpass filter with 106 Hz bandwidth is exploited for speaker identification. The chosen center frequencies of the Gabor filters are uniformly spaced on the Hertz scale and the chosen numbers of center frequencies are 40. The experiments on both parametric density modeling (PDM) and nonparametric density modeling (NPDM) are conducted for evaluation of the proposed Gabor filter-based and EMD-based speaker identification system. The accuracy rates of the PDM and NPDM are given in Tables 1 and 2, respectively.

TABLE I
Accuracy rates of PDM based identification system

System Name	GMM Mixture Number			Average
	4	8	16	
Gabor-filter based system	36.46 %	66.32 %	70.14%	57.64 %
EMD-based system	15.97 %	27.08 %	23.96%	22.34%

Table II shows that the PDM for the Gabor filter-based system contributes to better identification performance with more GMM mixture number. The PDM for the EMD-based system gets the best performance when the GMM mixture number is equal to 8.

TABLE IIIIV
Accuracy rates of NPDM based identification system

System Name	Bandwidth				Average
	10 Hz	20 Hz	30 Hz	40 Hz	
Gabor-filter based system	81.60 %	74.65 %	76.39 %	68.40%	75.26 %
EMD-based system	39.58 %	40.28 %	39.58 %	31.94 %	37.85%

According to the comparison of Table V with Table VIVII, the accuracy rate of the NPDM for the Gabor filter-based system is 17.62% higher than the PDM on average, and the NPDM for the EMD-based system is 15.51% higher than the PDM on average. The NPDM for the Gabor-based system with best performance is 81.60% when the bandwidth is 10 Hz, and that for the EMD-based system gets the best performance when the bandwidth is 20 Hz.

3)Optimization of the Weighting Factor for the SVM:
The speaker identification system adopts a linear fusion approach with the weighting factor. The sensitivity of the weighting factor on the overall performance of the PDM and NPDM-based fusion system are considered. Fig. 5 shows the identification accuracy by varying values of on the development set. For the PDM-based fusion system, two subsystems getting the best performance in Table 1 are fused, which are the Gabor filter-based subsystem with 16 GMM mixtures and the EMD-based subsystem with 8 GMM mixtures. Similarly, for the NPDM-based system, the Gabor filter subsystem with the 10 Hz bandwidth and the EMD subsystem with the 20 Hz bandwidth are fused. According to the results in Fig. 5, the weighting factor is set 0.75 for the PDM fusion system and 0.7 for the NPDM fusion system.

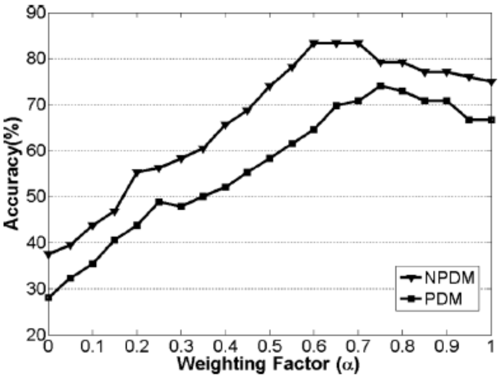


Fig. 5 Accuracy of PDM and NPDM based fusion system

4)Comparison of the Identification Systems:
Identification performance is evaluated by designing an experiment to test the following identification systems: I) The PDM-based fusion system: after the instantaneous frequencies are extracted from Gabor bandpass filtered signals and EMD filtered signals, the system then uses univariate GMMs with 16 mixture and 8 mixture to model the IFs for the Gabor-based and the EMD-based system, respectively. Using (25), a kernel SVM with approximated probability product kernel is adopted. The fusion weighting factor is set to 0.75 according to the results given in the above section. II) The NPDM-based fusion system: this system models the IFs with the 10 Hz bandwidth for the Gabor-based system and the 20 Hz bandwidth for the EMD-based system. The fusion weighting factor is set to 0.7. III) The baseline system for this work is the pyknogram feature set, which is referred to "Pyknogram-Based System." To obtain the pyknogram feature set, an input speech signal s (t) is passed through

Gabor filters, resulting in waveforms $S_k(t)$ for the k^{th}bandpass filter. The Hilbert transform is then applied to these waveforms, and derive the instantaneous frequency $w_k(t)$ and magnitude $a_k(t)$. The pyknogram feature is then obtained by calculating the spectral centroid of each sound frame as follows:

$$F_k = \frac{\sum_{t_0}^{t_0+L} \left[w_k(t) \cdot a_k^2(t)\right]}{\sum_{t_0}^{t_0+\tau} a_k^2(t)}. \qquad (30)$$

For the pyknogram feature set, we used a linear-SVM as the classifier, applying the frame-based voting approach. IV) This baseline (MFCC +GMM) is a major technique for speaker identification. The system firstly extracts MFCCs from the files, and then GMM is adopted to be the classifier.

TABLE VIIIIXX
Comparison between different systems

System	Accuracy Rate
PDM Based Fusion system (α= 0.75)	83.13 %
NPDM Based Fusion system (α = 0.70)	83.75 %
Pyknogram Based Fusion system	79.51 %
MFCC + GMM	76.04 %

Table.XIXIIXIII compares the results of the identification systems, where the first column of the table shows the test systems, and the second column shows accuracy of the corresponding test system. As shown in the Table. XIVXVXVI, the NPDM-based fusion system can achieve as high an accuracy rate as 83.75%. In comparison with the systems III and IV, the recognition rate of system II is increased by 4.24% and 7.71%, respectively. Additionally, the PDM-based fusion system achieves a great improvement in accuracy compared with two PDM-based subsystem in Table. XVII. Besides whisper speech-based speaker identification, use the NPDM-based fusion system (α =0.7) to identify speakers based on normal speech, or fast speech. Table.XVIIIV displays the experimental results.

Table.XIXV
Comparison between speaking styles

Speaking Style		Accuracy Rate
1.	Normal	98.12 %
2.	Fast	97.50 %

III.PROPOSED SYSTEM

Discusses the problem of robust parametric model estimation and classification in noisy acoustic environments. Characterization and modeling of the external noise sources in these environments is in itself an important issue in noise compensation. The techniques described provide a mechanism for integrating parametric models of the acoustic background with the signal model so that noise compensation is tightly coupled with signal model training and classification. Prior information about the acoustic background process is provided using a maximum likelihood parameter estimation procedure that integrates an a priori model of the acoustic background with the signal model. An experimental study is presented on the application of this approach to text-independent speaker identification in noisy acoustic environments. Considerable improvement in speaker classification performance was obtained for classifying unlabeled sections of conversational speech utterances from a 16-speaker population under cross-environment training and testing conditions

The problem of speaker identification and verification in noisy conditions, assuming that speech signals are corrupted by environmental noise, but knowledge about the noise characteristics is not available. This research is motivated in part by the potential application of speaker recognition technologies on handheld devices or the Internet. While the technologies promise an additional biometric layer of security to protect the user, the practical implementation of such systems faces many challenges. One of these is environmental noise. Due to the mobile nature of such systems, the noise sources can be highly time-varying and potentially unknown. This raises the requirement for noise robustness in the absence of information about the noise. This paper describes a method that combines multi condition model training and missing-feature theory to model noise with unknown temporal-spectral characteristics. Multi condition training is conducted using simulated noisy data with limited noise variation, providing a compensation for the noise, and missing-feature theory is applied to refine the compensation by ignoring noise variation outside the given training conditions, thereby reducing the training and testing mismatch. Focused on several issues relating to the implementation of the new model for real-world applications. These include the generation of multi condition training data to model noisy speech, the combination of different training data to optimize the recognition performance, and the reduction of the model's complexity. The new algorithm was tested using two databases with simulated and realistic noisy speech data. The first database is a redevelopment of the database by rerecording the data in the presence of various noise types, used to test the model for speaker identification with a focus on the varieties of noise. The second database is a handheld-device database collected in realistic noisy conditions, used to further validate the model for real-world speaker verification. The new model is compared to baseline systems and is found to achieve lower error rates. Speech separation of the signal from noisy from noise environment using adaptive filters. Then the noise free signal give as input to whispered speech using SVM classifier. Finally find out attack speech with whispered speech.

IV. CONCLUSION

The work presents an access control system, which is a speaker identification system based on whispered speech. The experiments are conducted using whispered speeches from the chain speech corpus. Signals filtered with Gabor filters and empirical mode decomposition are then transformed using the Hilbert transform. We then model the probability distribution of the instantaneous frequencies, comparing the performance of using parametric and nonparametric probability density modeling the work. An approximated probability product kernel support vector machine is also presented, which applies Riemann sum to approximate the probability product kernel for the feature set. An interesting issue with the work is the choice of appropriate probability density modeling for IF.

The experimental results show that nonparametric modelling outperforms parametric modeling by 17.62% for the Gabor filter-based system and by 15.51% for the EMD-based system. As a baseline system, we choose the pyknogram feature set, since the pyknogram is also based on the instantaneous frequency of a signal. In comparison with the pyknogram-based system, the recognition rate of the proposed NPDM-based fusion system was better by 4.24%. In this work, we already have satisfying performance of speaker identification with whispered speech. A possible future work is to increase the robustness for identifying different speaking style data with a pre trained classifier and to extend this framework to online learning for speaker adaptation, so that we can achieve a robust and safety access control system. For proposed system speech separation of the signal from noisy from noise environment using adaptive filters. Then the noise free signal give as input to whispered speech using SVM classifier. Finally find out attack speech with whispered speech.

ACKNOWLEDGMENT

I sincerely thank to all those who helped me in completing this task.

REFERENCES

[1] JiaChing Wang, Yu Hao Chin, “Speaker identificationwith whispered speech for the access control system,” IEEE Trans. On Automation science and Engineering , vol. 12, Oct. 2015.

[2] J. Poignant, L. Besacier, and G. Quenot, “Unsupervised speaker identificationin TV broadcast based on written names,” IEEE Trans. Audio,Speech, Language Proc., vol. 23, pp. 57–68, Nov. 2014.

[3] J. Xu and H. Zhao, “Speaker identification with whispered speech using unvoiced-consonant phonemes,” in Proc. Int. Conf. Image Anal.Signal Process., Nov. 9–11, 2012.

[4] M. Grimaldi and F. Cummins, “Speaker identification using instantaneousfrequencies,” IEEE Trans. Audio, Speech, Language Process.,vol. 16, no. 6, pp. 1097–1111, 2008

[5] J. C. Wang, C. H. Yang, J. F. Wang, and H. P. Lee, “Robust speaker identification and verification,” IEEE Comput. Intell. Mag., vol. 2, no.2, pp. 52–59, May 2007

BIOGRAPHY

Nisha Beegum S received the B.Tech degree in Electronics and Communication Engineering from M.G. University, Kerala at Amal Jyothi College of Engineering. Now pursuing her M.Tech degree in Communication Engineering under the same university in Mount Zion College of Engineering.

DOI 10.17148/IJARCCE.2016.56122