SVM based Emotional Speaker Recognition using MFCC-SDC Features

Asma Mansour
University of Tunis El Manar
National School of Engineers of Tunis
Signal, Image and Information Technology laboratory
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

Zied Lachiri
University of Tunis El Manar
National School of Engineers of Tunis
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

Abstract—**Enhancing the performance of emotional speaker recognition process has witnessed an increasing interest in the last years. This paper highlights a methodology for speaker recognition under different emotional states based on the multiclass Support Vector Machine (SVM) classifier. We compare two feature extraction methods which are used to represent emotional speech utterances in order to obtain best accuracies. The first method known as traditional Mel-Frequency Cepstral Coefficients (MFCC) and the second one is MFCC combined with Shifted-Delta-Cepstra (MFCC-SDC). Experimentations are conducted on IEMOCAP database using two multiclass SVM approaches: One-Against-One (OAO) and One Against-All (OAA). Obtained results show that MFCC-SDC features outperform the conventional MFCC.**

Keywords—Emotion; Speaker recognition; Mel Frequency Cepstral Coefficients (MFCC); Shifted-Delta-Cepstral (SDC); SVM

I. Introduction

Emotional speaker recognition is one of research fields in Human-Computer Interaction (HCI) or affective computing [1]. The main motivation comes from the want to develop a human machine interface that's more intelligent, adaptive and credible. This may gives computers the ability to know person in such context for many real applications .Speaker recognition in emotional context can be used in criminal or forensic investigation to identify the suspected person who produces the emotional utterances. It can also be used in telecommunication to ameliorate the telephone based speech recognition performance,etc...

Emotional speaker recognition systems are composed of two mains components which are feature extraction and classification [2]. In littrature, different classifiers have been used to model speakers under emotional states. I.Shahin [3] has used Hidden Markov Model(HMM) and suprasegmental hidden Markov models (SPHMMs) to identify speaker using emotional cues. In the same context, Yingchun Yang et al. citeyang have used GMM-UBM classifier. Support Vector Machines (SVM) are used [5] to show the important influence of the emotional state upon text independent speaker identification.

In general, human emotions are complicated phenomenon. Thus, choosing a most suitable features that represent emotional utterances has been an important step in emotional speaker recognition process. Researches have demonstrated that features derived from the speech spectrum usually give best performances for the automatic recognition system. Indeed, the spectrum reflects the geometry of the system that generates the speech signal. Therefore, spectral features are widely developed for the speaker recognition such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients(LPCC) in addition to the other acoustic features [6].

Many features have been used to ameliorate the performance of speaker recognition system in emotional context [7]. MFCC features are the most common used features in speaker recognition in emotional context [8] [9]. Linear Predictive Cepstral Coefficients(LPCC) have also been used frequently in this context [10] .

MFCC coefficients are based on human auditory system [11]. However, these coefficients will be more efficient, if speech is of short duration. For long-term speech signals, Shifted Delta Coefficients (SDC) features are more appropriated, since they identify the dynamic behavior of the speaker along the prosodic features of speech signal. Kshirod Sarmah et al. [12] have employed MFCC-SDC features to identify language. N. Murali Krishna et al. [13] have used MFCC-SDC to recognize different human emotional states. Fred Richardso et al. [14] have introduced SDC features for speaker and language recognition. However, this method has not been used in emotional speaker recognition applications.

In this work, we propose to investigate MFCC-SDC features to improve the performances of the speaker recognition system in emotional talking environment. Hence in order to evaluate the proposed recognition system, it is advantageous to use two multiclass SVM approaches : One Against One (OAO) and One Against All (OAA) in classification step. We are also interested to compare obtained results from diffrent feature extraction methods.

This paper is organized as follows: Section II present the proposed emotional speaker recognition system . Section III presents the process of MFCC and MFCC-SDC features extraction and Section IV deals with multiclass Support Vector Machines approaches. Results and experiments are given in Section V. Finally, conclusion is given in Section VI.

II. System design

The proposed emotional speaker recognition system is displayed in figure 1. It can be divided into two main components: feature extraction and speaker classification. Firstly,

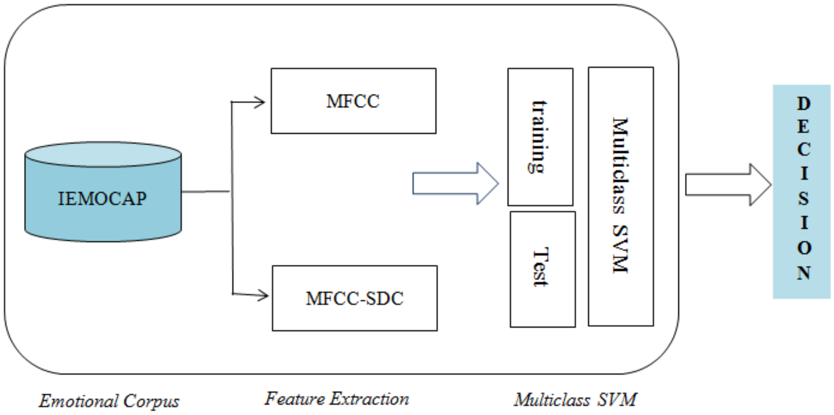


Fig. 1: System description

improvised emotional utterances of IEMOCAP database are considered to evaluate this system. We extract MFCCs and MFCC-SDC from the emotional speech signals. These obtained feature vectors are then divided into training and test sets. Then, classification is done using two well known multiclass SVM approaches which are One-Against-All(OAA), and One-Against-One(OAO). Finally, the decision of the recognition system is specified with accuracy rate using the test set.

III. Feature Extraction

Feature extraction is a critical step in emotional speaker recognition system. In fact, choosing a suitable features which represent useful information increase precision of recognition system. In this section, we describe the procedure of MFCC feature extraction and we introduce Shifted-Delta-Cepstra (SDC) technique in order to compare both the traditional MFCC coefficients and MFCC-SDC features.

A. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) is one of the commonly used technique of feature extraction in emotional speaker recognition. MFCC coefficients are based on human hearing perceptions which cannot perceive frequencies over 1Khz. After frame blocking and windowing step, the FFT is computed and the power coefficients are filtered by a triangular band pass filter bank also known as Mel-scale. They have been used to capture the phonetically important characteristics of speech signal. MFCC has two kind of filters which are filters spaced linearly at low frequency below $1000Hz$ and filters logarithmic spacing above $1000Hz$ [15] . Therefore, the following approximate formula can be used to compute the Mels for a given linear frequency f :

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}). \qquad (1)$$

The full extraction procedure of a Melfrequency cepstral coefficient is described in figure 2.

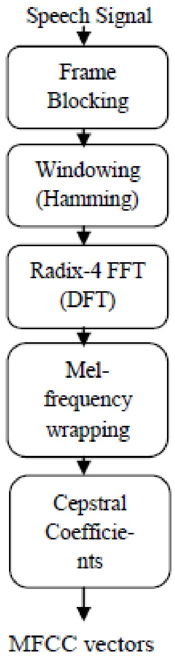


Fig. 2: Pipeline of MFCC extraction.

B. Shifted Delta Cepstra features

SDC features are widely used in language identification and speech recognition fields [12]. SDC feature vectors are an extension of delta-cepstra coefficients. Figure 3 describe the extraction procedure of SDC feature vectors. In fact, these vectors are obtained by stacking delta-cepstra computed across multiple speech frames. SDC coefficients depend on four parameters typically named as $N - d - P - k$. The parameter N represents the number of cepstral coefficients used to compute MFCC at each frame. So each frame is presented by a coefficient vector given as:

$$c(t) = [c_0 c_1 ... c_i ... c_{N-1}]. \qquad (2)$$

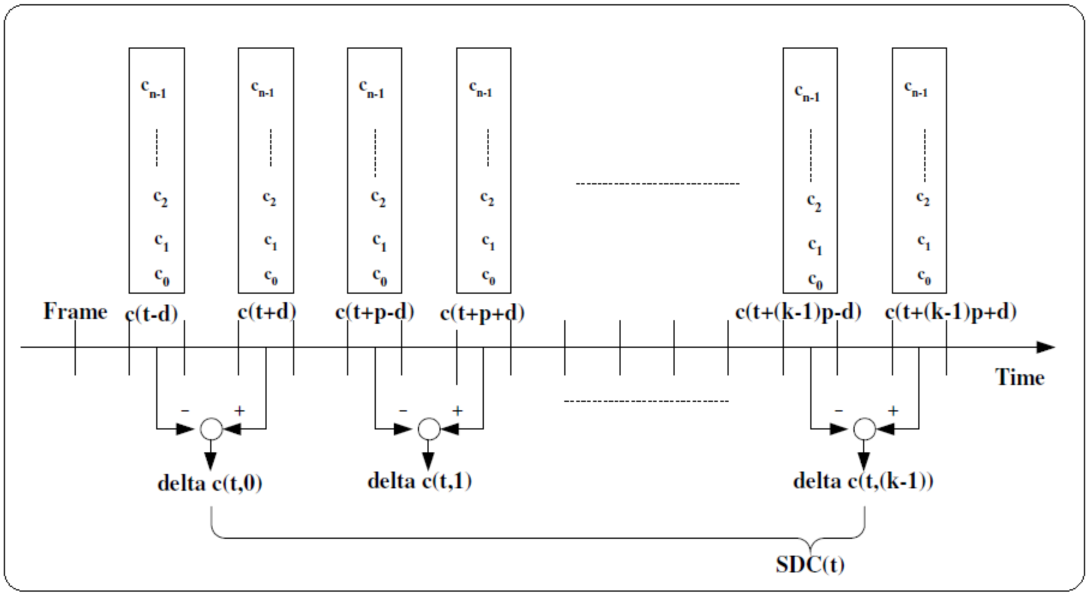


Fig. 3: Extraction procedure of SDC coefficients.

Where c_i are the MFCC coefficients and t is the coefficient index. The parameter d presents the spread over which delta are computed. The gaps between different delta computations is given by the parameter P. Parameter k determines the number of blocks whose delta coefficients are concatenated to obtain the final form of feature vector. For given time t, an intermediate calculation is done to obtain these k coefficients :

$$\Delta\mathbf{c}(t,i) = \mathbf{c}(t+i\times P+d) - \mathbf{c}(t+i\times P-d). \qquad (3)$$

Finally, the SDC coefficient vector of k dimension is obtained as:

$$\mathbf{SDC}(t) = [\Delta c(t,0)\Delta c(t,1)...\Delta c(t,k-1)]. \qquad (4)$$

Hence, SDC coefficients expressed in 4 are the stracked version of MFCC coefficents given in 1, and $k\times N$ parameters are then used for each SDC feature vector.

The SDC coefficients are able to interpret signal and capture features from the long duration speech samples or dynamically changing samples. Thus, it solves limitations of the traditional short time derivation of the cepstra features. This technique is widely successful in language identification system (LID) using GMM with high order (512-1024-2048) mixture models [16].

IV. Classification

In literature, various classification approches have been used to recognize speaker from his emotional speech such Gaussian Mixture Models [18], Hidden Markov Model [17], and Support Vector Machines [5].

Support vector machines (SVM) were well used for pattern recognition . It is a simple and efficient classifier which transforms the original input set into a higher dimensional space using kernel mapping functions. Its main goal is to find the optimal separating hyperplane using the maximized margin criteria to distinguish between classes. The most frequently employed kernel functions named standard kernels are linear, polynomial and Gaussian)kernels.

SVM are firstly introduced by Vapinik for binary classification [11]. Then, they are extended to solve multiclass problem using different approches. There are two main strategies for extending the binary SVM classifier to multiclass classifier: either by decomposing multiclass problem to a set of binary classifiers and combining it or by taking all the classes at once and considering the others instances in one optimization formulation [19]. In this work, we are interested to the first strategy and we are tested two most popular SVM methods which are One-Against-All (OAA) and One-Against-One (OAO) [20].

The OAA approach is the earliest and simplest one [21]. It builds kSVM binary classifiers in which k is the number of classes. The principle of OAA method is training the i^{th} binary SVM with all the positive labels examples in this class to separate it to the other classes. This j^{th} class is considered as positive class and all other examples are negative samples. Finally, the winner class corresponds to the highest output of SVM. This methods is considered fast, only k binary SVM classifiers are trained. However, OAA is an asymmetric method due to the small ratio of positive examples if compared with the negatives ones in every training hyperplane. Thus, many indecision regions may be created which degraded the performance of the classifier.

The OAO is symmetric method which involves $k(k-1)/2$ SVM binary classifiers [20]. Each SVM binary classifier is trained to distinguish between each pair of classes. Then, these classifiers will be combined using majority voting strategy. The decision of classification obtained by the maximal vote-number class. The training process of this method is long in terms of times.

V. Experiments

A. Emotional Database

The quality of the database plays an important role in performance of emotional speaker recognition. The emotional speech corpus selected for this study is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [22]. It consists of audio, video and motion-capture recordings of dyadic mixed-gender pairs of actors. It includes five sessions, in each session actors play improvisation of scripts or hypothetical scenarios. Each improvisation conveys a general emotional theme. The main goal is to have an expression that mostly resembles to natural emotion expression. Then, these expressions have been divided into utterances which were manually annotated in categorical labels: {angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted...} and in terms of three dimensional axes :valence, activation, and dominance.

In our study, we used only the speech utterances recorded under four emotional states which are:anger, happiness, sadness and neutrality. In our analysis, emotional speech of 10 speakers are considered that to say that 2218 utterances distributed over four emotions.

B. Experimental setup

In current experiments, IEMOCAP emotional database is used to evaluate the performance of SVM based speaker recognition in emotional context. The MFCCs and MFCC-SDC features are extracted from emotional utterances which represent four emotions: neutral, happy, angry and sad expressed by ten speakers (five male and five female): {SP1, SP2, SP3, SP4, SP5, SP6, SP7, SP8, SP9, SP10}.

All audio recording were sampled at a rate of 16 KHz. Speech samples were segmented into frames of $50ms$ length with 50% overlap between frames. The 70% of data set was used for training phase and 30% of this data formed the testing set.

The feature vector MFCC consists of 13 coefficients stacked with the set of shifted delta cepstra (SDC)features. We have tested some SDC schemes founded in literature [23], 12-1-3-3 SDC scheme was given the best results. The Shifted-Delta-Cepstra coefficients are extracted from MFCC coefficients.

Two multiclass SVM approaches including OAA and OAO were used in order to evaluate the proposed emotional speaker recognition. The multiclass SVMs methods were performed using the SVM-KM toolbox for Matlab [24]. Two kernel functions are used which are polynomial and gaussian kernels with both OAA and OAO strategies. Each kernel is characterized by paire of parameters typically written (C, σ) which C presents the regularization parameter parameter and σ is the gaussian width . To select suitable parameters (C, σ), a cross validation algorithm is employed by varying the C and σ in $[2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}]$ [25].

C. Results

In this study, a speaker recognition system in emotional context is implemented using MFCCs and MFCC-SDC features. Speakers are recognized under four different emotional states using two multiclass SVM approaches which are OAA and OAO tested with polynomial and gaussian kernels.

We evaluate MFCC and MFCC-SDC features using OAA SVM multiclass method. Different results are detailed in I.We remak firstly that the gaussian kernel gives best classification accuracies using both MFCC and MFCC-SDC features. Comparing the classification rates obtained with traditional MFCCs coefficients to those obtained with the MFCC-SDC features , we remark that the classification accuracies are improved ranging from 86.55% to 88.23% using polynomial kernel and 89.33% to 91.31% using gaussian kernel. With a 95% confidence interval in the range of $[78.33\%, 95.71\%]$, MFCC-SDC presents promising results with polynomial kernel (CI +/- 4%). A best speaker recognition rate obtained with the application of MFCC-SDC features and gaussian kernel by an average of 91.34% (CI +/- 2%).

The same experiment was conducted with the OAO multiclass SVM method. Results of classification accuracies are illustrated in table II. Similar to OAA approach, gaussian kernel gives best results with an average of 88% using MFCC coefficients and 90.90% using MFCC-SDC features. However, we can notice that the MFCC-SDC features have given an improvement with the gaussian kernel while they weren't successful with the polynomial kernel.In fact, with a 95% confidence interval, MFCC-SDC features give classification results with the polynomial kernel in the range of $[53.55\%, 70.80\%]$ (CI +/- 3%) and with the gaussian kernel in the range of $[83.33\%, 96\%]$(CI +/- 2%)

In general, classification rates of speaker recognition in emotional context have been improved using MFCC-SDC features except for polynomial kernel in OAO approach. The best results are obtained when we use the gaussian kernel associated to MFCC-SDC computed with OAA and OAO multiclass SVM startegies.

For a better presentation of results of the table I and the tzble II, rates were shown in figures .These graphs illustrate a comparative analysis between different used features and different multiclass SVM approaches. Indeed, MFCC-SDC features extracted from the traditional MFCC improve the performances of the emotional speaker recognition system using both OAA and OAO multiclass SVM methods. Moreover, the best classification accuracies are often obtained with the use of the gaussian kernel and OAA SVM multiclass method with an avreage of 91.31% (CI +/-2%).

VI. Conclusion

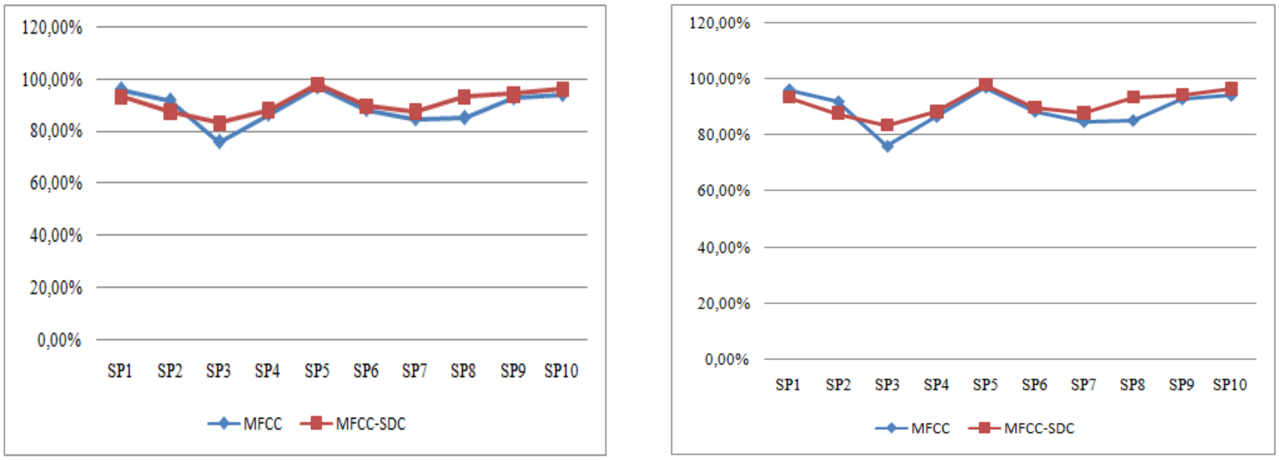
In this paper, it has been observed that the performance of speaker recognition system in emotional context has been improved applying MFCC-SDC as a feature extraction method and SVM as a classifier. The evaluation of emotional speaker recognition system is carried out on IEMOCAP emotional database using two multiclass SVM approaches which are OAA and OAO methods. The experimental results reveal that there is an improvement in the performances of the proposed emotional speaker recognition in improvised context using MFCC-SDC features extracted from the conventional MFCC coefficients and they outperform the baseline systems that using MFCC features.

TABLE I: Classification Results using OAA/SVM

Speakers	Features	Polynomial (%)	Gaussian(%)
SP1	MFCC	95.25	96.10
	MFCC-SDC	94.80	93.50
SP2	MFCC	80.72	91.89
	MFCC-SDC	87.83	87.83
SP3	MFCC	82	86.66
	MFCC-SDC	79.10	83.58
SP4	MFCC	83.11	97
	MFCC-SDC	78.33	88.33
SP5	MFCC	88	88.46
	MFCC-SDC	97	98
SP6	MFCC	84.34	84.84
	MFCC-SDC	82.05	90
SP7	MFCC	83.93	85.24
	MFCC-SDC	84.84	87.87
SP8	MFCC	94.28	92.85
	MFCC-SDC	88.52	93.44
SP9	MFCC	87	94.11
	MFCC-SDC	95.71	94.28
SP10	MFCC	86.85	98.79
	MFCC-SDC	94.11	96.50
AVERAGE	MFCC	86.55	89.33
	MFCC-SDC	88.23	91.34
Confidence Interval	MFCC	+/-0.03	+/-0.03
	MFCC-SDC	+/-0.04	+/-0.02

TABLE II: Classification Results using OAO/SVM

Speakers	Features	Polynomial (%)	Gaussian(%)
SP1	MFCC	96	96.10
	MFCC-SDC	70.80	92.20
SP2	MFCC	83.12	91.89
	MFCC-SDC	56.50	91.90
SP3	MFCC	82.40	82.08
	MFCC-SDC	53.55	89.55
SP4	MFCC	82	80
	MFCC-SDC	57.80	88.33
SP5	MFCC	88.45	91
	MFCC-SDC	55.10	96
SP6	MFCC	85.50	84.61
	MFCC-SDC	59.28	83.33
SP7	MFCC	84	86.36
	MFCC-SDC	61.11	90.91
SP8	MFCC	95	83.60
	MFCC-SDC	58.62	86.88
SP9	MFCC	87.35	91.42
	MFCC-SDC	65 .88	95.71
SP10	MFCC	88.36	92.94
	MFCC-SDC	68.50	94.11
AVERAGE AVERAGE	MFCC	87.22	88
	MFCC-SDC	60.74	90.90
Confidence Interval	MFCC	+/-0.03	+/-0.03
	MFCC-SDC	+/-0.03	+/-0.02



(a)

(b)

Fig. 4: Emotional speaker recognition using SVM classifier with MFCC and MFCC-SDC features:(a): OAA multiclass approach,(b): OAO multiclass approach.

For future work, MFCC-SDC features can be tested with other classifiers such as Deep Neural Network (DNN). Moreover, we can combined these features with other cepstral parameters to enhance performance of speaker recognition under different emotional states.

ACKNOWLEDGMENT

This work has been supported by the Research Laboratory LR-SITI-ENIT (Signal, Images and Information Technologies), ENIT.

REFERENCES

[1] Ghiurcau, M.V, Rusu C., Astola J., A study of the effect of emotional state upon text-independent speaker identification. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4944-4947. [doi:10.1109/ ICASSP.2011.5947465].

[2] B. Schuller, S. Steidl, and A. Batliner, The interspeech 2009 emotion challenge, Interspeech (2009), ISCA, Brighton, UK, 2009.

[3] Ismail Shahin ,Speaker Identification in Emotional Environments, IRANIAN JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING, VOL. 8, NO. 1, 2009.

[4] Yingchun Yang and Li Chen, Emotional Speaker Identification by Humans and Machines, Speech Commun: Springer, pp. 167173, 2011.

[5] Marius Vasile Ghiurcau, Corneliu Rusu and Jaakko Astola , SPEAKER RECOGNITION IN AN EMOTIONAL ENVIRONMENT, EURASIP, 2011.

[6] Sarika S. Admuthe and Shubhada Ghugardare, Survey Paper on Automatic Speaker Recognition Systems, International Journal Of Engineering And Computer Science ISSN:2319-7242, VOL. 4, Issue. 3, pp. 0895-10898, March 2015.

[7] Huang, T., Yang and Y.C., Applying pitch-dependent difference detection and modification to emotional speaker recognition. Proc. 9th Annual Conf. of the Int. Speech Communication Association, pp.2751-2754, 2008.

[8] Shashidhar G. Koolagudi, Shan e Fatima and K. Sreenivasa Rao, Speaker recognition in the case of emotional environment using transformation of speech features, CUBE 2012, September 35, 2012, Pune, Maharashtra, India.

[9] J. Sirisha Devi, Dr. Srinivas Yarramalle and Siva Prasad Nandyala, Speaker Emotion Recognition Based on Speech Features and Classification Techniques, I.J. Computer Network and Information Security, VOL. 7, pp.61-77, JUNE 2014.

[10] A. Mansour and Z.lachiri, Speaker Recognition in Emotional Context, Proceedings of Proceedings of Engineering Technology(PET), ACECS 2015.

[11] Li CHEN, Ying-chun YANG and Zhao-hui WU, Mismatched feature detection with finer granularity for emotional speaker recognition, Journal of Zhejiang University-SCIENCE (Computers and Electronics), pp. 903-916, 2014.

[12] Kshirod Sarmah and Utpal Bhattacharjee, GMM based Language Identification using MFCC and SDC Features, International Journal of Computer Applications,VOL. 85, NO. 5, January 2014.

[13] N. Murali Krishna, P.V. Lakshmi and Y. Srinivas, Inferring the Human Emotional State of Mind using Assymetric Distrubution,International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 4, No.1, 2013

[14] Fred Richardson, Doug Reynolds and Najim Dehak, A Unified Deep Neural Network for Speaker and Language Recognition, International Speech Communication Association, INTERSPEECH, September 6-10, 2015, Dresden, Germany.

[15] V.N.Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.

[16] Torres-Carrasquillo, P.A. 2002 Language identification using Gaussian mixture models, PhD, thesis, Michigan State University.

[17] Ismail Shahin, MEmploying both gender and emotion cues to enhance speaker identification performance in emotional talking environments, Springer, pp. 341351, 2013.

[18] Marius Vasile Ghiurcau, Corneliu Rusu and Jakko Astola, A STUDY OF THE EFFECT OF EMOTIONAL STATE UPON TEXT-INDEPENDENT SPEAKER IDENTIFICATION, ICASSP 2011.

[19] A. Hassan and R. I. Damper, Multi-class and hierarchical SVMs for emotion recognition, In Proc. Interspeech, 2010.

[20] C. Hsu and C. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, VOL. 13, NO. 2, pp. 415425, 2001.

[21] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge, UK: Cambridge University Press, 2000.

[22] C. Busso, M. Bulut, C. Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation 2008 vol. 42, no. 4, pp. 335-359.

[23] Muthusamy, Y.K., Barnard, and Cole R.A, Automatic Language Identification: A Review/Tutorials. Signal Processing Magazine, IEEE, Vol 11, Issue.4.pages.33-41, 1994.

[24] S. Canu, Y. Grandvalet, V. Guigue, A. Rakotomamonjy, SVM and

Kernel Methods Matlab Toolbox, Perception Systmes et Information, INSA de Rouen, France,2008.

[25] L.I.Kuncheva, Combining pattern classifiers methods and algorithms. New York: Wiley,2004.