International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 09 | Sep-2016 www.irjet.net p-ISSN: 2395-0072

Text-Dependent Multilingual Speaker Identification using Learning Vector Quantization and PSO-GA Hybrid Model

PRIYATOSH MISHRA[1], Dr. PANKAJ KUMAR MISHRA[2]

[1]M.Tech, Electronics & Telecommunication Department, RCET, Bhilai, India
[2] Associate Professor & Head, Electronics & Telecommunication Department, RCET, Bhilai, India

---

Abstract - *In this work a multilingual speaker identification system is proposed. The feature extraction techniques employed in system extract Mel frequency cepstral coefficient (MFCC), delta mel frequency cepstral coefficient (DMFCC) and format frequency. The feature selection is done using hybrid model of particle swarm optimizatiom (PSO) and Genetic Algorithm (GA). We have used Learning Vector Quantization (LVQ) artificial Neural Network classifiers. The speech database consists of 40 speakers (20 males+ 20 females) speech utterance. The speech utterance is recorded for a specific sentence in three different languages viz. "Now this time you go" (in English), "Adhuna Asmin Twam Gachh "(in Sanskrit), "Ab Iss Baar Tum Jao" (in Hindi). Total word for this purpose is 14 including 4 for Sanskrit and 5 Hindi and English. The average identification rate 79.99% is achieved when the Network is trained by LVQ and it shows 80.52% when LVQ is trained using hybrid PSO-GA model.*

Key Words: **Mel frequency cepstral coefficient (MFCC), delta mel frequency cepstral coefficient (DMFCC), format frequency, particle swarm optimizatiom (PSO), Genetic Algorithm (GA), Learning Vector Quantization (LVQ)***)***

1. INTRODUCTION

Speaker recognition is divided into two types- speaker identification and speaker verification. The Speaker identification is task of finding identity of an individual based on his/her voice characteristics. Speech signal is basically meant to contain the information about the linguistic message. But, it also contains the speaker specific information and so it can be used to recognize (identify/verify) a person. Speaker identification determines who the speaker is among known voices enrolled in the system. Given an unknown speaker, the task of system is to compare his or her voice to a set of available models, thus makes this task a one-vs-all classification problem. On the other hand, Speaker verification a process of finding whether the speaker identity is who the person claims to be. It performs a one-to-one match between the features of an input voice A verification system is trained using not only the claimant's signal but also data from other speakers, called background speakers

The speaker identification system can again classified into two groups- Text dependent and text independent[1]. The Speaker identification is task of finding identity of an individual based on his/her voice characteristics. Speech signal is basically meant to contain the information about the linguistic message. But, it also contains the speaker specific information and so it can be used to recognize (identify/verify) a person. Speaker identification determines who the speaker is among known voices enrolled in the system. Given an unknown speaker, the task of system is to compare his or her voice to a set of available models, thus makes this task a one-vs-all classification problem. On the other hand, Speaker verification a process of finding whether the speaker identity is who the person claims to be. It performs a one-to-one match between the features of an input voice and those of the claimed voice that is enrolled in the system. The speaker who claims the identity is known as the test speaker; the signal is then compared against the model of the person who claims. A verification system is trained using not only the claimant's signal but also data from other speakers, called background speakers. Multilingual speaker recognition has thus been a field of active research in recent years [4] [12]. In this work a multilingual speaker identification system is designed using artificial neural network.

The Paper is organised in a way that section II will give complete process flow diagram of proposed system and the subsections viz. A to G will give complete detail of each steps used in proposed method. Section III contains the experiment and observations part. Section IV shows the result obtain after experiment and at last section V concludes the paper.

2. PROPOSED METHODOLOGY

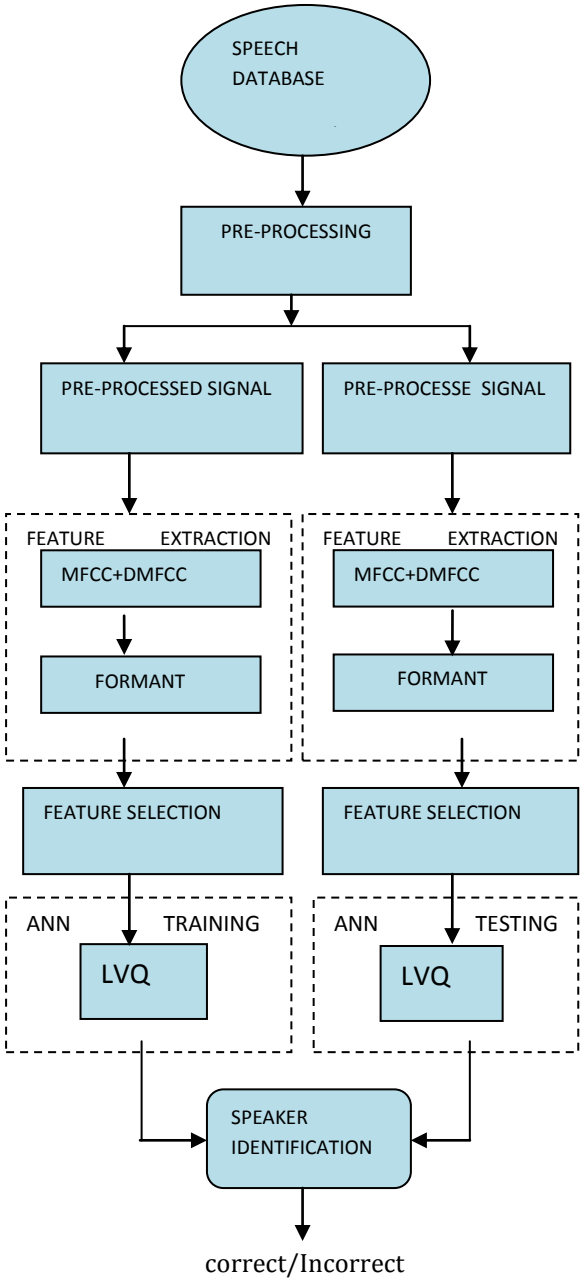

Fig -1: Proposed System for speaker Identification

The figure 1 shows complete flow diagram of proposed methodology. As it can be seen from figure, the speaker identification is divided into two phases- training phase and testing phase[2]. Training phase is also called as enrolment phase in which a speech model for every speaker is created using artificial neural network classifier. In testing phase, the model of test speech of a particular speaker is created and it is compared with earlier speech signal models stored in artificial neural network memory. The detail of each blocks of proposed speaker identification is explained in upcoming headings.

2.1 PRE-PROCESSING

In preprocessing step, separate different words of different languages from their respective sentences. Silence and noise are removed from speech signals. There are two steps in Preprocessing. One reduces noise and second is to separate word from sentences (by removal of silence between words) and new speech database is created using this preprocessed speech words and they are store in proper manner that they can use easily further. Feature extraction from these speech signals is most important task for this purpose.

2.3 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs)

MFCCs is considered to be low-level information which are based on spectral information derived from a short time-window segment of speech of about 20 millisecond [3][10].The feature MFCCs are represented by a real valued N-dimensional vector. The determination of the MFCC includes the following steps.

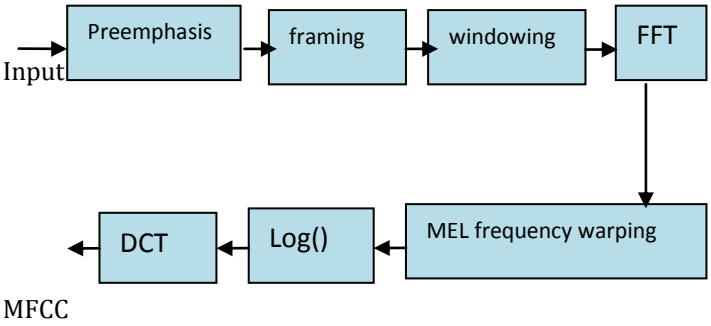

Fig -2: Extraction of MFCCs

1. Framing the Signal**:-** In this step, the input speech signal is segmented into frames of 15-20 ms with an overlap of 50% of the frame size. Overlapping is used to provide continuity within frames.

2. Windowing: -Each frame has to be multiplied with a hamming window so as to keep the continuity of first and the last points in the frame. If the input signal in a frame is denoted by X(n), n = 0,...N-1, then signal obtain after Hamming windowing[5] is,

$$Y(n) = X(n) * W(n) \quad \ldots\ldots (1)$$

$$W(n) = 0.54 - 0,46\cos\left(\frac{2n}{N-1}\right) \quad \ldots\ldots (2)$$

For $0 \leq n \leq N-1$

Where Y(n) is output, X(n) is input & W(n) is window function

3. FFT: -Spectral analysis shows that different tones in speech signals corresponds to the energy distribution over frequencies. Therefore FFT is performed to get the magnitude frequency response of each frame. When FFT is executed on a frame, the signal within a frame is considered to be periodic, and continuous when wrapping around. If it is not that case, FFT can still be used but the discontinuity at the frame's first and last points probably introduce undesirable consequences in the frequency response. To handle this problem, we multiply each frame by a hamming window in order to increase its continuity at the first and last points.[3]

4. Mel- Frequency Warping:- In order to realize model of human hearing 'mel' scale is used as human ear does not follow a linear scale. Thus for every tone with an actual frequency f which is measured in Hz, is measured on mel scale. Thus, we can use the following approximate expression to compute the mels for a given frequency *f* in Hz.

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

……(3)

5. Computing the Cepstral Coefficients: -In this last step, the conversion of the log mel spectrum back to time is done. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). As the mel spectrum coefficients (and so their logarithm) are real numbers, one can convert them to the time domain by using the discrete cosine transform (DCT). In this final step log mel spectrum is changed back to time. The result is called the Mel Frequency Cepstrum Coefficients .The DCT is done in order to transform the mel coefficients back to time domain.

$$C_n = \sum_{j=1}^{q} m_j \cos\left[\frac{\pi n}{q}(j - 0.5)\right]$$

.....(4)

Where, q is the number of filter and m_j is the coefficient of filter.As we have performed FFT, DCT transforms the frequency domain speactrum into quefrency domain which is like time domain. The determined features are similar to cepstrum, thus it is referred as the mel-scale cepstral coefficients, or MFCC.

2.4 DELTA MFCC

Delta Derivatives play an important role in identifying the speaking styles, pauses and duration of a particular speaker and is therefore an important feature set for speaker recognition. These dynamic temporal information are present between frames and are captured using [23].

$$\overrightarrow{d_t} = \frac{\sum_{p=1}^{P} p\left(\overrightarrow{c_{t+p}^{c}} - \overrightarrow{c_{t-p}^{c}}\right)}{2\sum_{p=1}^{P} p^2}$$

.....(5)

Where d_t = delta derivative. One can similarly derive the delta delta derivative by replacing c_t with d_t

2.5 FORMANT FREQUENCIES

Formants in Human speech signal are specified as the spectral peaks of the sound spectrum of the speech. In phonetics formant frequencies also mean an acoustic resonance of the human vocal tract which is generally measured as a peak amplitude in the frequency spectrum of the sound.

The vocal tract can be modeled as a linear filter with resonances and the resonance frequencies of the vocal tract are called formant frequencies [13]. Peaks of the vocal tract response correspond approximately to its formant frequencies. If vocal tract is modeled as a linear, time-invariant and all- pole system, then each conjugate pair of poles corresponds to a formant frequency or resonance frequency [13]. The vocal tract can be modeled by a stable allpole model [13].

Here is the brief review of the mechanics of computing a linear prediction model and the formant frequencies estimation. In fact, the speech signal in linear prediction model can be defined as:

$$X(n) = a(1)X(n-1) + a(2)X(n-2) + \cdots + a(L)X(n-L) + e(n)$$

.....(6)

Or we can write it as

$$X(n) = \sum_{i=1}^{L} a(i).x(n-i) + e(n)$$

……(7)

Where L is the number of coefficients in the model, a(i), where i=1,2,3.... represent the linear prediction coefficients and e(n) is the error in the model. LPC analysis typically produces predictor polynomial of degree 10 which, due to the stability requirement, has all its roots within unit circle

[13]. The equation (1) is written in Z-transform notation as a linear filtering operation:

$$X(z)\left[1-\sum_{i=1}^{L} a(i)z^{-i}\right]=E(z) \quad \text{.....(8)}$$

Or,

$$\frac{X(Z)}{E(Z)}=\frac{1}{1-\sum_{i=1}^{L} a(i)z^{-k}}=H(z) \quad \text{.....(9)}$$

E(z) and X(z) denote, respectively, the Z-transform of the error signal and the speech signal while H(z) is used for a linear prediction inverse filter

$$H(z)=\sum_{i=0}^{N} a(i).z^{-1} \quad \text{.....(10)}$$

There is another way to estimate formant frequencies which is based on the relation between formant and poles of the vocal tract filter [13]. Fig shows an all pole filter. It is an all pole filter as the numerator of its transfer function is a constant. The input to the system is the error function which can be taken as the unpredictable part of the signal or excitation that derives the all pole system. The frequencies of the complex poles of $H(z)$ are candidate frequencies for the formants since the poles in a system model the resonances in that system.

The denominator of the transfer function can be factored as,

$$1-\sum_{i=0}^{N} a(i).z^{-1}=\prod_{k=0}^{N}(1-C_k.z^{-1}) \quad \text{.....(11)}$$

Where C_k represents a set of complex values with each complex conjugate pair of poles represents a resonance at frequency F_k

$$F_k=\frac{F_s}{2\pi}.tan^{-1}\frac{im(C_k)}{Re(C_k)} \quad \text{.....(12)}$$

And the bandwidth is given by,

$$B_k=-\frac{F_s}{\pi}*In(C_k) \quad \text{.....(13)}$$

When the pole is closer to the unit circle then the root represents a formant.

$$R_k=\frac{(Im(C_k))^2+(Re(C_k))^2}{2} \quad \text{.....(14)}$$

Where, $R_k \geq 0.7$

2.6 FEATURE SELECTION

In this proposed system particle swarm optimization[15] with Genetic Algorithm[16] operator is used as optimization technique. It was figured out that PSO suffers from premature convergence, tending to get stuck in local optima, low solution precision and so on. In order to remove these shortcomings and get better results, numerous improvements to PSO have been proposed. One of the novel versions of PSO with crossover operator is proposed by adding a crossover step to the standard PSO. PSO uses iterative process to search the global optima in solution space. Crossover operator with PSO has a property of better exploration so by using crossover search area is explored in a relatively better manner. PSO has a higher convergence rate, by using crossover with PSO premature convergence is also reduced and PSO does not get trapped in local optima. Crossover can help the particles to jump out of the local optima by sharing the others' information. To improve the solution diversity in PSO, a crossover BPSO is introduced. In this paper BPSO with five different types of crossover applied to the five benchmark functions for testing which crossover gives the optimum result at what probability. Particles generated by BPSO are randomly selected for crossover operation and two new offspring's are formed. The best offspring (in terms of fitness) selected from the new offspring's. This new best offspring replaces the worst parent particle which is selected for crossover. The replacement is done if the new best offspring has the good fitness value than the parent particle. In this paper we analyze different types of crossover operators used with BPSO, all the crossover discussed in previous section applied to the BPSO one by one according to the algorithm shown in below algorithm.

2.7 CLASSIFICATION USING LVQ

Learning vector Quantization (LVQ)[17] is a neural net that combines competitive learning with supervision. It can be used for pattern classification. A neural network for learning vector quantization consists of two layers: an input layer and an output layer. It represents a set of *reference vectors*, the coordinates of which are the weights of the connections leading from the input neurons to an output neuron. Hence, one may also say that each output neuron corresponds to one reference vector.

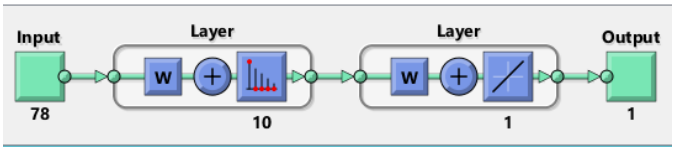

Fig -2: LVQ architecture in generated in MATLAB

A training set consisting of P training vector - target output pairs are assumed to be given

$\{S^{(p)} : d^{(p)}\}$, p= 1,2,3.....,P

Where $S^{(p)}$ are M dimensional training vectors, and $d^{(p)}$ are N dimensional target output vectors. N is the number of classes and so it must be smaller than P.

The target vectors are defined as

$$d_i^p = \begin{cases} 1, & if, S^{(p)} belongs\ to\ class\ i \\ 0, & otherwise \end{cases}$$

The LVQ is made up of a competitive layer, which includes a Competitive subnet, and a linear layer. In the first layer (not counting the input layer), each neuron is assigned to a class. Different neurons in the first layer can be assigned to the same class. Each of those classes is then assigned to one neuron in the second layer. The number of neurons in the first layer, P, will therefore always be at least as large as the number of neurons in the second layer, N. In the competitive layer, neurons in the first layer learn a prototype vector which allows it to classify a region of the input space. Closeness between the input vector and any of the weight vectors is measured by the smallness of the Euclidean distance between them.

Learning of LVQ:- The learning method of learning vector quantization is often called *competition learning*, because it works as follows: For each training pattern the reference vector that is closest to it is determined. The corresponding output neuron is also called the *winner neuron*. The weights of the connections to this neuron - and this neuron only: the winner takes all - are then adapted. The direction of the adaption depends on whether the class of the training pattern and the class assigned to the reference vector coincide or not. If they coincide, the reference vector is moved closer to the training pattern, otherwise it is moved farther away. This movement of the reference vector is controlled by a parameter called the *learning rate*. It states as a fraction of the distance to the training pattern how far the reference vector is moved. Usually the learning rate is decreased in the course of time, so that initial changes are larger than changes made in later epochs of the training process. Learning may be terminated when the positions of the reference vectors do hardly change anymore.

3. EXPERIMENT AND OBSERVATIONS

The speech database consist of 40 speakers(20 males+ 20 females) speech utterance. The speech utterance is recorded for a specific sentence in three different languages viz. "*Now this time you go"* (in English), "*Adhuna Asmin Twam Gachh* " (in sanskrit), "*Ab Iss Baar Tum Jao*" (in Hindi). Sentences of different languages Sound wave files (.wav) are recorded in acoustic rooms by using the handset microphone at sampling rate 44.100 KHz with mono channel. Total word for testing purpose is 14 including 4 for Sanskrit and 5 Hindi and English.

Table- 1: Speaker Identification accuracy(%) for LVQ

Speaker no.	Test input	Error LVQ	Error LVQ+ PSOGA	Efficiency(%) LVQ (X)	Efficiency(%) LVQ+ PSOGA (Y)
1.	14	3	3	78.57	78.57
2.	14	4	3	71.40	78.57
3.	14	2	3	78.57	78.57
4.	14	3	3	78.57	78.57
5.	14	2	2	85.71	85.71
6.	14	3	3	78.57	78.57
7.	14	2	2	78.57	78.57
8.	14	2	2	78.57	78.57
9.	14	3	3	71.40	78.57
10.	14	2	2	85.71	85.71
11.	14	3	3	71.40	71.40
12.	14	4	4	71.40	71.40
13.	14	2	2	85.71	85.71
14.	14	3	3	78.57	78.57
15.	14	3	3	78.57	78.57
16.	14	3	2	78.57	85.71
17.	14	2	2	85.71	85.71
18.	14	2	2	85.71	85.71

19.	14	2	2	85.71	85.71
20.	14	3	3	78.57	78.57
21.	14	4	4	71.40	71.40
22.	14	4	4	71.40	71.40
23.	14	3	2	78.57	85.71
24.	14	2	2	85.71	85.71
25.	14	4	3	71.40	78.57
26.	14	4	4	71.40	71.40
27.	14	3	3	78.57	78.57
28.	14	3	3	78.57	78.57
29.	14	3	3	78.57	78.57
30.	14	3	3	78.57	78.57
31.	14	2	2	78.57	78.57
32.	14	2	2	78.57	78.57
33.	14	3	3	78.57	78.57
34.	14	2	2	78.57	78.57
35.	14	3	3	78.57	78.57
36.	14	3	3	78.57	78.57
37.	14	4	4	71.40	71.40
38.	14	2	2	85.71	85.71
39.	14	2	2	85.71	85.71
40.	14	3	3	78.57	78.57
	Σ = 560	Σ= 120	Σ = 63	Σ = 79.99	Σ =80.52

4. RESULT

The experiment is done on MATLAB 15. There are 14 isolated words uttered by 40 speakers (20males + 20females) in English, Hindi and Sanskrit language. The LVQ is trained by2240sample i.e 56(14x4) samples for each speaker. All speakers with their input utterances, number of errors and efficiency is given in table 2. When using LVQ for this multilingual System the minimum performance is 78.57% while best performance reaches up to 92.85% for a speaker. Overall performance of the system is 79.99%.When using GA-PSO hybrid Algorithm for optimization this multilingual System the minimum performance is 71.40% and the peak performance is 85.71% and with the proposed optimization technique overall performance of the system increased up to 80.52%..

5. CONCLUSIONS

In this project, we implemented MFCC, DMFCC and Formant frequency for feature extraction and Radial Basis function feed neural network for training process and testing process of input vectors. We proposed an optimization technique based on GA-PSO hybrid model. It can be seen clearly from result the performance has increased to some value. Moreover, using this optimization scheme in conjunction with LVQ the benefit is that the number of feature sets gets reduced. Without using optimization the feature set of a speaker consist of 78 coefficients but after using optimization it get reduced to 35 coefficients. Clearly, the proposed optimization scheme succeeded in finding the most significant feature which can give sufficiently high identification rate.

REFERENCES

[1] Rabiner.L.,Juang B.H., "Fundamental of speech recognition, Prentice

[2] R.P.Lippmann,"Review of Neural Networks for Speech Recognition," Neural Computation, Vol. 1, No. 1, pp. 1-38, 1989.

[3] Suma Swamy, Shalini T.,Sindhu P. Nagabhushan, Sumaiah Nawaz, and K.V. Ramakrishnan,2012.Text Dependent Speaker Identification and Speech Recognition Using Artificial Neural Network, ObCom 2011, Springer-Verlag Berlin Heidelberg :160–168.

[4] U. Bhattacharjee and A. Sarmah, "A multilingual speech database for speaker recognition," in *Proc. IEEE International Conference on Signal Processing, Computing and Control (ISPCC)*, 2012, pp. 1–5.

[5] B. G. Nagaraja,H. S. Jayanna,2013.Efficient Window for Monolingual and Crosslingual Speaker Identification using MFCC,IEEE International Conference on Advanced Computing and Communication Systems (ICACCS):1-4.

[6] S G Bagul & R.K.Shastri,2013. Text independent speaker recognition system using gmm,IEEE International Conference on Human Computer Interactions (ICHCI): 1 – 5.

[7] Shahzadi Farah & Azra Shamim,2013.Speaker Recognition System Using Mel-Frequency Cepstrum Coefficients, Linear Prediction Coding and Vector Quantization,IEEE International

Conference on Computer,Control & Communication (IC4): 1 – 5.

[8] Sourjya Sarkar, K. Sreenivasa Rao, Dipanjan Nandi and Sunil Kumar,2013. Multilingual Speaker Recognition on Indian Languages, IEEE India Conference (INDICON): 1 – 5.

[9] Sumit Srivastava, Pratibha Nandi, G. Sahoo, Mahesh Chandra,2014.Formant Based Linear Prediction Coefficients for Speaker Identification, IEEE International Conference on Signal Processing and Integrated Networks (SPIN): 685 – 688.

[10] P. Suba &B. Bharathi,2014.Analysing the Performance of Speaker Identification task using different Short term and Long term Features, IEEE Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference: 1451 – 1456.

[11] Khan Suhail Ahmad , Anil S. Thosar Jagannath H. Nirmal and Vinay S. Pande,2015. A Unique Approach in Text Independent Speaker Recognition using MFCC Feature Sets and Probabilistic Neural Network,IEEE English International Conference Advances in Pattern Recognition (ICAPR): 1 – 6.

[12] L. Mary, K. S. Rao, and B.Yegnanarayana, "Neural network classifiers for language identification using syntactic and prosodic features," in *Proc. 2nd Int. Conf. Intelligent Sensing and Information Processing (ICISIP-2005)*, Chennai, India, January 2005.

[13] Dorra Gargouri, Med Ali Kammoun and Ahmed Ben Hamida, ENIS, University of SfaxTunisia, "A Comparative Study of Formant Frequencies Estimation Techniques", Proceedings of the 5th Wseas International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006 (pp15-19)

[14] John H.L. Hansen and Sanjay Patil, "Speech Under Stress: Analysis, Modeling and Recognition". M¨uller (Ed.): Speaker Classification I, LNAI 4343, pp. 108–137, 2007.

[15] Maider Zamalloa, Germacn Bordel, Luis Javier Rodriguez, Mikel Penagarikano "Feature Selection Based on Genetic Algorithms for Speaker Recognition" 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop 1-8.

[16] Nie ru.Yue Jianhua " A GA and Particle Swarm Optimization Based Hybrid Algorithm" 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence): 1047-1050.

[17] Rafik Djemili, Rocine Bourouba, Mohamed Cherif Amara Korba "A Speech Signal Based Gender Identification System Using Four Classifiers" Multimedia Computing and Systems (ICMCS), 2012 International Conference :184-187.