MULTILINGUAL SPEAKER RECOGNITION USING NEURAL NETWORK

Rajeev Kumar rajeevkumariiitm@gmail.com

Rajesh Ranjan iiitm.rajesh@gmail.com

Sanjay Kumar Singh sksinghiiitm@gamil.com

Rahul kala rahulkalaiiitm@yahoo.co.in

Dr. Anupam Shukla dranupamshukla@gmail.com

Dr. Ritu Tiwari rt_twr@yahoo.co.in

Department of Information and Communication Technology,
Indian Institute of Information Technology and Management Gwalior, India

Abstract: In the present paper an attempt is made to recognize the speaker based on speech feature and independent of the language. Speaker recognition systems attempt to recognize speaker's identity from his/her speech utterances. Every speaker has different individual characteristics embedded in his /her speech utterances. These characteristics can be extracted from utterances and neural network model is used to get the desired results. The simulated model of neural network based multilingual Speaker recognition system for Indian languages Hindi, Punjabi, Telugu, Sanskrit and English has been developed.

The utterances of the speaker are stored in digitized form to evaluate the speech data bank. The sampling frequency and speech feature namely LPC, LPCC, RC, LAR, LSF, ARSCIN are extracted from speech signal and formed feature vectors. These features are fed into Neural Network back propagation learning algorithm for training and identification processes of different speakers and languages. The database used for this system consists of 25 speaker including both male and female from different regions. Five different speaking texts of different languages having same meaning are used to get the best speaker identification accuracy.

The ANN model consist of 575 neuron in inputs layer, two hidden layers of 52 and 38 neurons respectively and 25 neurons in output layer. The average recognition score is 85.74%.

Keywords: Multilingual Speaker Recognition, Back propagation Algorithm (BPA), Liner Prediction Coefficients (LPC) , Reflection coefficients (RC), Linear prediction Cepstral Coefficients (LPCC), Log Area Ratio (LAR), Arcus Sin Coefficients (ARCSIN), Line Spectral Frequencies (LSF).

1. Introduction:

Multilingual speaker recognition and language identification are key to the development of spoken dialogue systems that can function in multilingual environments [4]. In India, which officially recognizes more than twenty five languages and whose citizens almost without exception speaks more than one of these languages fluently, the development of such multilingual system is an especially relevant Challenge. This multilingual system should then be adapted to the target language with the help of a language identification system. In this paper, we have developed Speaker Recognition systems based on continuous speech recognition, and evaluate these for five Indian regional languages i.e. Hindi, English, Telugu, Punjabi and Sanskrit spoken by 25 speakers in each language. Speaker recognition can be divided into speaker verification and speaker identification.

1.1 Speaker Verification

Speaker verification is mainly concerned with the verification whether a speaker is the person he/she claims to be or not, and involves a binary decision whether the test utterance matches the features of the claimed speaker [2]. In a speaker verification trial an identity claim is provided or asserted along with the voice sample. In this case, the unknown voice sample is compared only with the speaker model whose label corresponds to the identity claim [6]. If the quality of the comparison is satisfactory, the identity claim is accepted; otherwise the claim is rejected. Speaker verification is a special case of open-set speaker identification with a one-speaker target set. The speaker verification decision mode is intrinsic to most access control applications. In a speaker verification trial only one comparison is required, so speaker verification performance is independent of the size of the speaker population. Speaker verification systems are mainly used in security access control. In addition to security applications, speaker verification may be used to offer personalized services to users. For example, once a speaker verification phrase is authenticated, the user may be given access to a personalized phone book for voice repertory dialing.

1.2 Speaker identification

The purpose of a speaker identification system is to determine the identity of an unknown speaker among several speakers of known speech characteristics, from a sample of his or her voice. In speaker identification a voice sample from an unknown speaker is compared with a set of labeled speaker models. When it is known that the set of speaker models includes all speakers of interest the task is referred to as closed-set identification [2]. The label of the best matching speaker is taken to be the identified speaker. In speaker identification, the number of decision alternatives is equal to the size of the sample. Most speaker identification applications are open-set, meaning that it is possible that the unknown speaker is not included in the set of speaker models [14]. It can readily be seen that in the speaker identification task performance degrades as the number of speaker models and the number of comparisons increases. Speaker identification systems are mainly used in criminal investigation [3].

Moreover Speaker recognition systems can be either text-dependent (constraint on what is spoken) or text-independent (no constraint on what is spoken). Text-independent recognition systems are more versatile but their accuracy is considerably lower than that of comparable text-dependent systems. To achieve acceptable results in this case more speech data is usually necessary for both training and testing. Three stages are generally involved in building a speaker-recognition system: Training, testing, and implementation [5].

1.3 ARTIFICIAL NEURAL NETWORK (BACK PROPAGATION)

Speech recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of sub word units (e.g., phonemes), words, phrases, and sentences [2]. Artificial neural networks have emerged as a promising approach to the problem of speech recognition [7, 8]. ANN can be most adequately characterized as computational models' with particular properties such as the ability to adapt or learn, to generalize, or to cluster or organize data, and which operation is based on parallel processing which are the requirements for speech and speaker recognition. ANN can learn complex features from the data, due to the non-linear structure of artificial neuron [7, 10]. Various ANN training algorithms such as BPA, Radial Basis Function, Recurrent networks etc., are being used for training purpose.

In this paper we have used **Back propagation** Neural Network [5] for the recognition system. It has been successfully applied to many pattern classification problems including speaker recognition [10].

Back propagation is the generalization of the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions [9]. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way as required.

Properly trained back propagation networks tend to give reasonable answers when presented with inputs that they have never seen. Typically, a new input leads to an output similar to the correct output for input vectors used in training that are similar to the new input being presented [19]. This generalization property makes it possible to train a network on a representative set of input/target pairs and get good results without training the network on all possible input/output pairs [10].

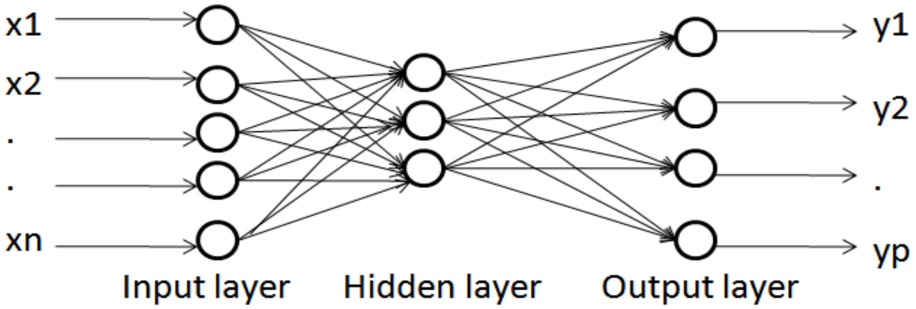

Fig. 1 Architecture of Feed Forward Neural Network

2. Literature survey

Speaker recognition is defined as “the process of recognizing who is speaking on the basis of individual information included in speech waves” [2, 4]. There are two separate problems occurs in speaker recognition, identification and verification [4].

Recognition methods are of two types. One is Text-dependent which require that the speaker say a specific text for both training and testing, and another is Text-independent methods can be used with varying text [5].

Automatic Speaker Recognition (ASR) is decoding of information in speech signal and its transcription into set of characters. It is used for practical application are of the small, medium and large vocabulary recognition systems [15].

Many researches have been done for multilingual speaker recognition system using ANN, and there are using different model like statistical methods Hidden Markov Model (HMMs), Harmonic Product Spectrum (HPS). [16, 17].

In identification of the speakers in single/different language ANNs have been used. In pattern classification or recognition phase, there are various methods used as vector quantization technique from signal processing to store features in codebooks [16, 14].

3. Feature Extraction

Feature extraction is most important part of speaker recognition. Features of speech play important role to separate one speaker from other. LPC is widely used in digital speech processing systems. It is one of the most powerful speech analysis techniques, and useful methods for encoding good quality speech at a low bit rate [11]. They vary from speaker to speaker and depending up on various factors like gender, emotions, filings, moods etc. Features extracted for this purpose are Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), Linear Predication Cepstral Coefficients (LPCC), Log Area Ratio (LAR), Arcus Sine Coefficients (ARCSIN) and Line Spectral Frequencies (LSF).

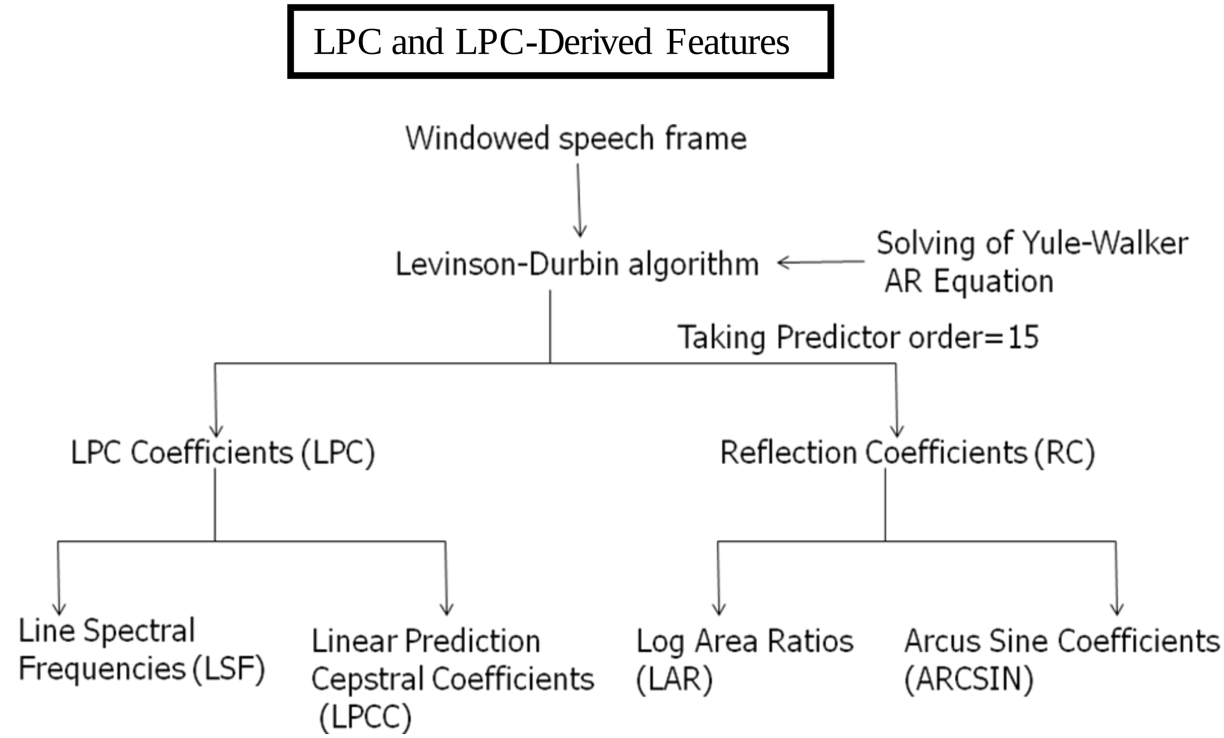
LPC and LPC-Derived Features
Windowed speech frame
Levinson-Durbin algorithm
Solving of Yule-Walker AR Equation
Taking Predictor order=15
LPC Coefficients (LPC)
Reflection Coefficients (RC)
Line Spectral Frequencies (LSF)
Linear Prediction Cepstral Coefficients (LPCC)
Log Area Ratios (LAR)
Arcus Sine Coefficients (ARCSIN)

Fig. 2 Features of speech calculated in this approach

LPC and LPC-Derived Features

The Linear Prediction Code (LPC) is approach to form feature or spectral vector [11]. Linear prediction coefficients are a highly effective representation of the speech signal. In this analysis, each speech sample is represented by a weighted sum of *p* past speech samples plus an appropriate excitation [12]. The main aim is to determine LPC coefficients minimizing the prediction error in the least squares sense.

The linear predictive model of speech production [12, 20] is given in the time domain:

$$s[n] \approx \sum_{k=1}^{p} a[k]s[n-k]\,, \qquad (1)$$

Where s[n] denotes the speech signal samples, a[k] are the predictor coefficients and p is the order of the predictor.
The total squared prediction error is:

$$E = \sum_{n} (s[n] - \sum_{k=1}^{p} a[k]s[n-k])^2 \qquad (2)$$

The objective of linear predictive analysis is to determine the coefficients a[k] for each speech frame so that error (2) is minimized. The problem can be solved by setting the partial derivatives of (2) with respect to a[k] to zero. This leads to so called Yule-Walker equations that can be efficiently solved using so-called Levinson-Durbin recursion [21].

The Levinson-Durbin recursion generates as its side product so-called Reflection Coefficients (RC), denoted here as k[i], i = 1, . . . , p. The name comes from the multitube model, each reflection coefficient characterizing the transmission/reflection of the acoustic wave at each tube junction.

Arcus Sin Coefficients (ARCSIN) derived from reflection coefficients k. This is more stable than reflection coefficient [2].

In the frequency domain, the linear predictive coefficients specify an IIR filter with the transfer function:

$$H(z) = \frac{1}{1-\sum_{k=1}^{p} a[k]z^{-k}} \qquad (3)$$

The poles of the filter (3) are the zeroes of the denominator. They are denoted here as z1, z2, . . . , zp, and they can be found by numerical root-finding techniques. The coefficients a[k] are real, which restricts the poles to be either real or occur in complex conjugate pairs.

Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) filter is introduced by Itakura [22]. LSFs are closely related to formant frequencies and they have some desirable properties which make them attractive to represent the Linear Predictive Coding (LPC) filter. The quantization properties of the LSF representation are recently investigated in [23].
Let the m-th order inverse filter $A_m(z)$,

$$A_{m\,(z)} = 1 + a_{1\,z}{}^{-1} + \cdots .. + a_{m\,z}{}^{-m} \qquad (4)$$

Given the LPC coefficients a[k], k = 1, . . . , p, the LPCC coefficients are computed using the recursion [24]:

$$C[n] = \begin{cases} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k]a[n-k], & 1 \leq n \leq p \\ \sum_{k=n-p}^{n-1} \frac{k}{n}\ C[k]a[n-k], & n > p \end{cases} \qquad (5)$$

The log area ratio (LAR) coefficients are derived from the linear prediction (LPC) coefficients. LPC can be transformed into other coefficients called Log area ratio coefficients (LAR). In LAR analysis, the vocal tract of a person is modeled as a non-uniform acoustic tube formed by cascading *p* uniform cylindrical tubes with different cross-section areas having equal lengths [2].The relationship between the LAR coefficients and the LPC is:

$$LAR_i = \log\left(\frac{A_i}{A_{i+1}}\right) = \log\left(\frac{1+\propto_i}{1-\propto_i}\right), A_{p+1=1} \qquad (6)$$

Where αi can be found by:

$$\propto = a_i^{(i)}, 1 \leq i \leq p \qquad (7)$$

Where $a_i^{(i)}$ is the ith LPC calculated by the ith order LPC model [2].

4. Approach

Proposed approach is done in various steps like Collection of speech utterances of different speakers of different languages, Preprocessing, Feature extraction, Neural Network training and Testing as shown in figure bellow.

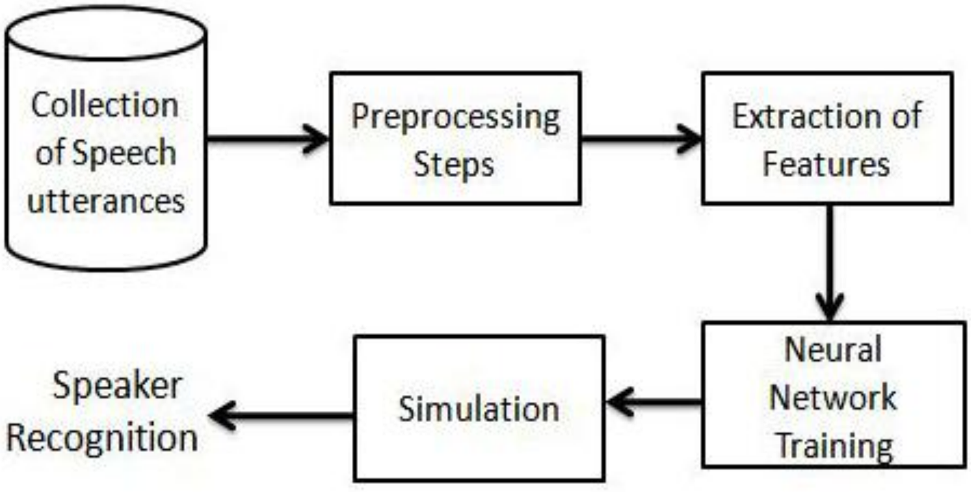

Fig. 3 various steps in this approach

Collection of speech utterances consists of sentences of different languages taken from a public domain. Sound wave files (.wav) are created by using microphone connected with Personal computer at sampling rate 44.100 KHz and 16 bit per sample with mono channel. For recording of sentences and Preprocessing we use software Cool Edit Pro 2.0 software. For this purpose one sentence ("*AB ISS BAAR TUM JAO* ") is recorded from 25speakers (15 male+10 female) in 5 different languages Hindi, Punjabi, Telugu, English and Sanskrit. The sentences in such format that in each word every consonant succeed a vowel and vice versa. Main reason behind this is that vowel sound is always generated with pronunciation of any letter.

In preprocessing step, separate different words of different languages from their respective sentences. Silence and noise are removed from speech signal. There are two steps in Preprocessing. One reduces noise (using Cool Edit Pro 2.0) and second is separate word from sentences (by removal of silence between words).

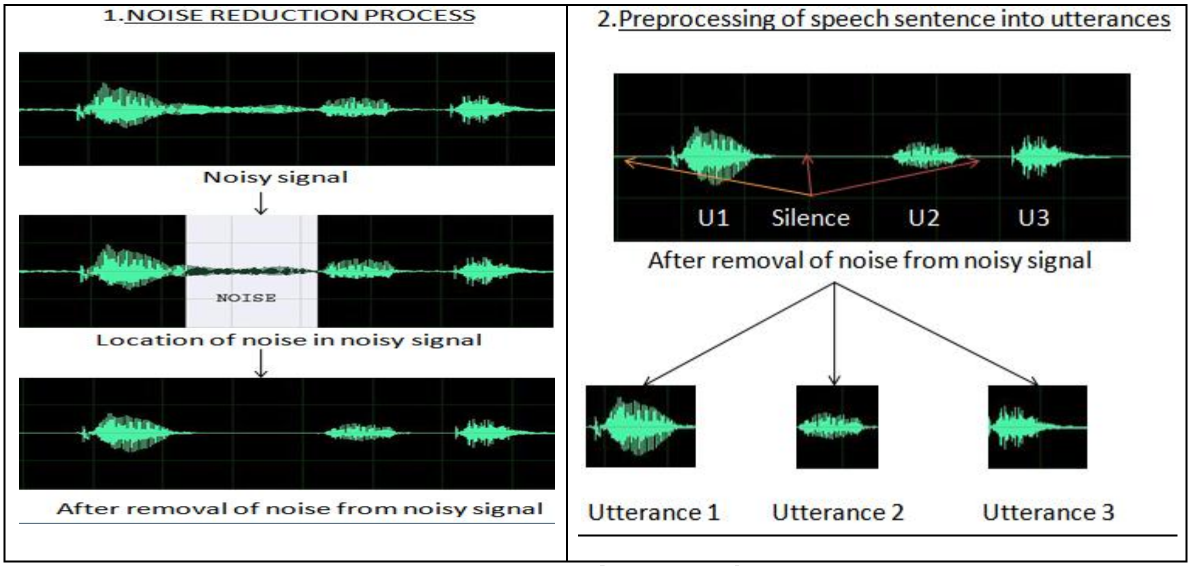
1.NOISE REDUCTION PROCESS
Noisy signal
NOISE
Location of noise in noisy signal
After removal of noise from noisy signal
2.Preprocessing of speech sentence into utterances
U1
Silence
U2
U3
After removal of noise from noisy signal
Utterance 1
Utterance 2
Utterance 3

Fig. 4 Preprocessing Steps in this approach

New speech database is created using this preprocessed speech (i.e. utterances) words and they are store in proper manner that they can use easily further. Next most important step is Feature extraction from these speech signals (A Computational Language). Mainly six different type of features are extracted they are Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), Linear Prediction Cepstral Coefficients (LPCC), Log area ratio (LAR), Arcus Sin Coefficients (ARCSIN) and Line Spectral Frequencies (LSF). Arrange this calculated data in a proper format (Matrix) and trained using BPA with two hidden layer and number of neurons in each hidden layer is in between number of inputs and target numbers.

After completion of training, the main step is to simulate trained Network properly and check weather Target output and Actual output is same or not. For this purpose a sample data known as test data is created which is totally different from input data used in training of network.

5. Result

When proposed system is trained using ANN as used by another researchers for such type of problems with limited number of neurons and layers, the average performance was 85.74% with 82 errors of 575 input data size. The various training parameter are illustrated in table 1.

Table 1: Various Parameter values for training of Network

Serial no.	Name of parameter	Corresponding values
1.	Error goal (δ)	0.011
2.	Momentum (µ)	0.9
3.	Training parameter (α)	0.26
4.	Maximum epochs	10,000
5.	Non linear function	Tan-sigmoid
6.	No. of hidden layer	2
7.	No. of neurons in hidden layers	52 and 38
8.	No. of Target	25

Number of input utterances by 25 speakers is 23 with maximum number of error is up to 6 and minimum is 2. All speakers with their input utterances, number of errors and efficiency is given in table 2.

Table 2: Training Result Data of Different Speakers

Speaker no.	**No. of input utterances**	**No. of Errors**	**Efficiency (%)**
1	23	3	86.95
2	23	2	91.30
3	23	4	82.60
4	23	2	91.30
5	23	2	91.30
6	23	3	86.95
7	23	6	73.91
8	23	2	91.30
9	23	2	91.30
10	23	3	86.95
11	23	2	91.30
12	23	4	82.60
13	23	3	86.95
14	23	2	91.30
15	23	4	82.60
16	23	3	86.95
17	23	5	78.26
18	23	4	82.60
19	23	3	86.95
20	23	2	91.30
21	23	6	73.91
22	23	5	78.26
23	23	4	82.60
24	23	4	82.60
25	23	2	91.30
Total	**∑=575**	**∑=82**	**∑=85.74**

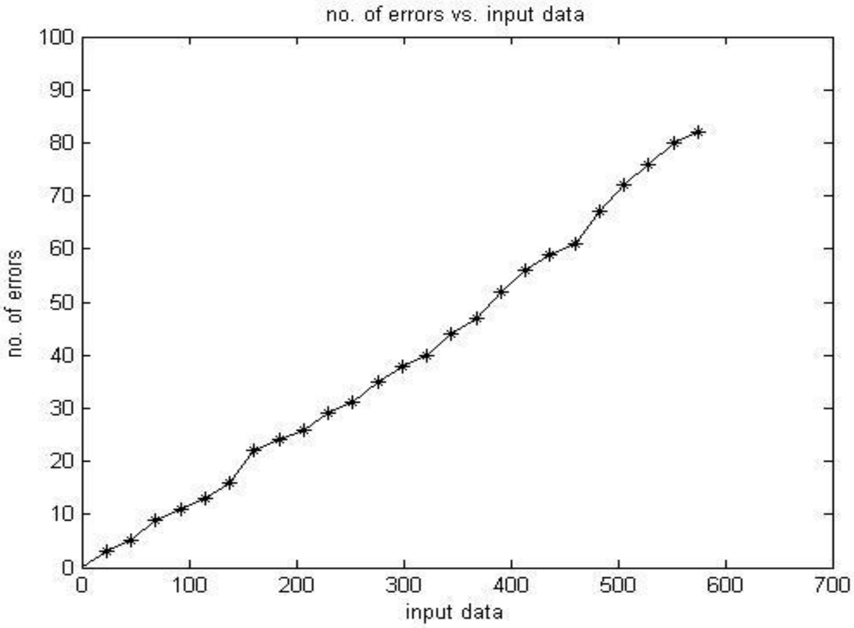

Fig. 5 Input data Vs. Error Graph

6. Conclusion

LPC Speech features and a neural network with back propagation training algorithm are appropriate to use for Text-Independent Multilingual Speaker Recognition.

The above result shows that BPA can be used for multilingual System .The minimum performance of system is 73.91% while best performance reach up to 91.30%.Overall performance of the system is 85.74%.The goal of the system which can recognize a text independent expression uttered by different speakers using various languages in different environment with different parameter may be the further enhancement of this research.

To improve system capability, mixed tone speech in appropriate length of speech duration should be selected for speaking sentence since it can cover more personal characteristics than using each tone in all utterances.

7. References

[1] Ing. Milan Sigmund, CSc. “Speaker Recognition, Identifying People by their Voices”, Brno University of Technology, Czech Republic, Habilitation Thesis, 2000.

[2] Campell J.P. and Jr. ,” Speaker recognition: a tutorial”, Proceeding of the IEEE, 1997, Vol 85, pp. 1437-1462.

[3] T. Kinnunen: *Spect*ral Features for Automatic Text-Independent Speaker Recognition, Ph.Lic. , Department of Computer Science, University of Joensuu, 2004.

[4] Li Deng, Jasha Droppo, Dong Yu, and Alex Acero “Learning Methods in Multilingual Speech Recognition”Speech Research Group Microsoft Research Redmond, WA 98052.

[5] Ehab F., M. F. Badran , Hany Selim “Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes” Electrical Engineeling Department, Assiut University.

[6] Stuker, S. Schultz, T. Metze, F. Waibel, A. “Multilingual articulatory features” Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference 7.

[7] Y.-Yan, M. Fanty, and R. Cole, "Speech Recognition Using Neural Networks with Forward-backward Probability Generated Targets", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Munich, April 1997.

[8] Shukla, Anupam;Tiwari, Ritu, “A novel approach of speaker authentication by fusion of speech and image features using Artificial Neural Networks”, Int. J. of Information and Communication Technology 2008-Vol.1,No.2 pp . 159 – 170.

[9] Zahorian, S. A. (1999), “Reusable Binary-Paired Partitioned Neural Networks for Text-Independent Speaker Identification, Proc. ICASSP-99, pp. II: 849- 852.

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

[11] Zebulum S. Ricardo, P. Guy, “A comparison of different spectral analysis models for speech recognition using neural networks”, IEEE 1997,pp. 1428-1431.

[12] J. Makhoul,” Linear prediction: a tutorial review”, Proceedings of the IEEE 1975, pp. 64(4):561–580.

[13] B.H. Juang, C.H. Lee and Wu Chou, “Minimum classification error rate methods for speech Recognition”, IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997.

[14] Chougule,S. and Rege,P.,“Language independent speaker identification,” ieeexplore pp 364-368, May 2006.

[15] Z. Bin, W. Xihong, C. Huisheng, “On the Importance of Components of the MFCC in Speech and Speaker Recognition”, Center for Information Science, Peking University, China, 2001.

[16] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," Proc. IEEE Int. Confi Acoust. Speech Signal Processing, S6.3, pp. 377-380 (1991).

[17] Faltlhauser,R.; Ruske G. “Robust speaker clustering in eigenspace”, Inst. for Human-Machine-Communication, Technische Universitat Munchen, Munich,Germany,2002 IEEE.

[18] XM2VTSDB: The Extended M2VTS Database, Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), Washington D.C, 1999.

[19] A comparison of different spectral analysis models for speechrecognition using neural networks. Zebulum, R.S. Vellasco, M. Perelmuter, G. Pacheco, M.A. Departamento de Engenharia Eletrica, PUC, Rio de Janeiro, Brazil; 1996 IEEE.

[20] J.R. Jr. Deller, J.H.L. Hansen, and J.G. Proakis,”.Discrete-Time Processing of Speech Signals”, IEEEPress, NewYork, second edition, 2000.

[21] J. Harrington and S. Cassidy,”Techniques in Speech Acoustics”, Kluwer Academic Publishers, ordrecht, 1999.

[22] E Itakura, “Line spectrum representation of linear predictive coefficients of speech signals”, J. Acoust. Sot. Amer**.** 1975, p. 535a.

[23] E. Erzin and A.E. Getin, “Interframe differential vector coding of Line Spectrum Frequencies”, Proc. Internat. ConjI Acoust. Speech Signal rocess. 1993 (ICASSP ‘93), Vol. II, April 1993, pp. 25-28**.**

[24] B. Atal, ”Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”, Journal of the Acoustic Society of America 1974, pp. 55(6):1304–1312.