Speaker and Gender Identification using Multilingual Speech

Pankaj Kumar Mishra [Á], Prof. Anupam Shukla[B̀]

[Á] Research Scholor Dr. C.V. Raman University, Bilaspur, India.
[B̀]Professor ABV-IIITM, Gwalior, India.

Abstract

As the demand for multilingual speaker recognizers increases, the development of systems which combine automatic speaker and gender identification, models becomes increasingly important. In this work a speaker and gender identification system is developed using multilingual speech signal as input. MFCCs and delta-MFCCs, LPC, LPCC , Formants ,ZCR are used to build modal for classification and to reduce size of feature vector k-means clustering used. Radial basis function network and multi-layer perceptron are used for classification and their results are compared. Here resilient back propagation algorithm used to train MLP. Two separate modules are used for gender and speaker identification in each experiment. In this experiment accuracy of gender identification is 99% and speaker recognition is 91% using back propagation algorithm and 98% and 92% for gender and speaker identification using radial basis function.

Keywords: *ANN, Speech feature, Speaker Recognition, Gender identification, MFCC, LPC, LPCC, ZCR .*

I. INTRODUCTION

A multilingual speaker recognition system is becoming more popular in countries like India where more than one language are spoken but development of this kind of system is a challenge. Acoustic signal has many level of information like what is spoken, who is speaking, which language is speaking, emotions and gender information. Gender and speaker based differences in human speech are partly due to physiological differences such as vocal fold thickness or vocal tract length and partly due to differences in speaking style. Since these changes are reflected in the speech signal, we have to exploit these properties to automatically classify a speaker.

Now a days, to identify a person number of biometric properties is used because these are unique i.e. face, speech, finger prints etc. Identification using voice signal has been using since 1970 or earlier. In the past decade, some work have been done in the field of speech recognition and speaker identification. Speaker Networks Recognition Using Neural and Conventional Classifiers is implemented by Kevin R. Farrell et al. [2] in 1994. In this work evaluation of various classifiers for text-independent speaker recognition is presented. In addition, a new classifier is examined for this application. The new classifier is called the modified neural tree network (MNTN). The MNTN is a hierarchical classifier that combines the properties of decision trees and feed forward neural networks.. The MNTN is found to perform better than full-search VQ classifiers for both of these applications. In addition to matching or exceeding the performance of the VQ classifier for these applications, the MNTN also provides a logarithmic saving for retrieval.

The Jayant M. Naik et al [3] present the results of speaker-verification technology development for use over long-distance telephone lines. A description is given of two large speech databases that were collected to support the development of new speaker verification algorithms. Also discussed are the results of discriminant analysis techniques which improve the discrimination between true speakers and imposters. A comparison is made of the performance of two speaker-verification algorithms, one using template-based dynamic time warping, and the other, hidden Markov modelling. Gender identification technique also used for security purpose in multimedia, telephone communication and other area lot of studies proof it [4, 19]. Biometrics is seen by many researchers as a solution to a lot of user identification and security problems now days [1]. Speaker identification is one of the most important areas where biometric techniques can be used.

Speech recognition techniques using RBF network is implemented by phillips@tuns.ca et al [5]. In this paper they present a pattern recognition approach, based on whole word patterns, to speaker independent automatic speech recognition of isolated digits. They use the decomposition of the spoken word into sub acoustic words to ensure time alignment of the significant portions of the input's acoustic characteristics and those of the reference patterns. The ISO data clustering algorithm is used by the radial basis function (RBF) network to create reference templates and classification of the speech samples

Much research has been done for multilingual speaker recognition system using ANN, and there are using different model like statistical methods [9, 11]. Speaker recognition is defined as “the process of recognizing who is speaking on the basis of individual information included in speech waves” [1, 7].

In [12] presents a method for speaker identification, independent of language spoken. Pitch frequency and speaker specific vocal tract information are used for speaker identification. Many researchers 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].

Speaker identification along with gender identification as biometric system can strengthen identification process because gender identification reduces the search space of speaker identification by half. Speaker recognition system attempts to identify person on the basis of their speech. Speaker recognition can divided into two types one is speaker identification and other one is speaker verification. In speaker identification one to many comparisons are done. The goal of identification system is determine identity of an unknown user form the number of speaker whose speech features are saved. Speaker identification further classified as closed-set and open-set. As clear from name in closed set unknown speech belongs to registered users whereas in open-set it can belong to unregistered user, Whereas in speaker verification one to one comparison is made. The goal of system is to verify whether a person is one who the person claims to be. Speaker recognition further classified into text dependent and text independent. In text dependent system same speech is spoken in training and testing phase. In text independent system it is not necessary. Text independent speaker recognition is more difficult to develop. In proposed modal we implement text dependent multilingual speaker and gender identification system.

Three sources of signal variability, which exist in a typical ASI system, are speaker variations, channel variations, and content (as in words in a text description of the speech) variations. Sometimes a speaker attempts to do mimicry of other this is example of speaker variation. As mentioned earlier, the channel of communication is another element that is uncontrolled and causes the variability. Speech signals often need to be transmitted over some form of communication channel from the source to the recording devices.

Bandwidth limitations and other interference lead to a low signal-to noise ratio, especially when the transmission medium is the standard telephone wire, ultimately resulting in a poor recorded signal quality. Another important aspect that needs to be mentioned at this juncture is that usually there is no control over the content of the spoken speech, giving rise to the need for “text-independent” speaker identification systems.

Research and development on speech recognition and speaker recognition methods and techniques has been undertaken for well over four decades and it continuous to be an active area [1]. Gender identification technique also used for security purpose in multimedia, telephone communication and other area lot of studies proof it [2, 18]. Biometrics is seen by many researchers as a solution to a lot of user identification and security problems now days [1]. Speaker identification is one of the most important areas where biometric techniques can be used. There are various techniques to resolve the automatic speaker identification problem [4, 5, 6, 7, 8], gender identification problems [2, 3, 11] and both together [15].

Approaches have spanned from human aural and spectrogram comparisons, to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition approaches, such as artificial neural networks (ANNs) and Hidden Markov Models (HMMs) [4].

Much research has been done for multilingual speaker recognition system using ANN, and there are using different model like statistical methods [6, 7]. Speaker recognition is defined as “the process of recognizing who is speaking on the basis of individual information included in speech waves” [1, 4].

In [8] presents a method for speaker identification, independent of language spoken. Pitch frequency and speaker specific vocal tract information are used for speaker identification. Many researchers 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) [13, 14].

In [3] present a method for gender identification; two stages are used one for pitch and other for generating formants. A preprocessing modal is built using LABVIEW for filtering out the noise components. Mean of formants and pitch of all the samples of a speaker calculated. Using nearest neighbor method, calculating Euclidean distance from the Mean value of Males and Females of the generated mean values of Formant 1 and Pitch, the speaker was classified between Male and Female

II. Feature Extraction

A. Dataset

This database contain 18 utterances ISS, BAAR, TUM, JAO,NOW, TWAM, VELLU etc. Words in this database collected from 50 speakers (25 male and 25 female). Sentence “ISS BAAR TUM JAO” is spoken in four Indian languages Hindi, English, Sanskrit and Telugu. Speech Acquisition is done on 44.1 KHz sampling frequency and format of sound file is .wav.

B. Speech Feature Extraction

Feature extraction uses different steps like acquisition of speeches of different speakers in different language, preprocessing of speech and then different features are extracted.

C. Preprocessing

In pre-processing phase different steps are performed like re-sampling, filtering, noise removal, silence removal, framing and windowing. Mfccs and delta-mfccs are extracted for each frame. Since some signal is sample on how frequencies and some on low so re-sampling is done on 8khz frequency ,considering nyquist theorem state that sampling frequency is equal or more than twice of maximum frequency component. Digitized signal is passed through an all band pass filter, α=0.9 is used as filter parameter. After filtering normalization is performed so that variation in part of data that does not contain useful information. For normalization below equation is used-

Y=Y/max (abs(Y));

Speech is quasi stationary signal [9, 10] so for speech analysis, stationary signal is used therefore framing is performed on speech signal. In this modal we used 25ms duration frame with 10ms overlapping, overlapped frame is used to remove discontinuity between frames.

Each frame is multiplied by window; windowing is carried out to minimize the spectral distortion by using the window to taper the signal on both ends thus reducing the side effects caused by signal discontinuity at the beginning and at the end due to framing.

Extracted features should meet some criteria when use for speaker identification. It should not be mimicry prone, less complexity and memory requirement should be less.

Research on human auditory system shown that it does not follow a linear scale. Thus for a tone have actual frequency say f, is mapped on Mel-scale. The Mel-scale mapped linearly below 1000 Hz and logarithmically above 1000Hz.Below equation shows relation between Mel frequency and linear frequency-

$$M(f)=2595*\log_{10}(1+f/100)$$

Since speech signal is convolved combination of excitation signal and vocal tract impulse response. Each person has different structure of vocal tract Speech signal S(n) represented as

$$S\ (n) = e\ (n)*\Theta\ (n)$$

Here e (n) is excitation signal and Θ (n) is vocal tract impulse signal. To identify speaker we can use vocal tract impulse response which give better estimation of correct speaker. So cepstral analysis is performed to separate excitation signal and vocal tract impulse response .

Applying Fourier transform on it we will find

$$S\ (w) = e(w)\Theta(w)$$

After taking log these signal will be separated.

LPC and LPC derived feature

LPC is widely used in speech processing in which signal is predicted on basis of past p samples.LPC feature is widely used in speech coding and decoding and it is coefficients of filter that model Vocal tract. LPC is estimation of sample on basis of past p sample and formula is given by

$$S[n] \approx \sum_{k=1}^{p} a[k]s[n-k]$$

Error is square of difference between expected value and actual value and Error can be calculated by formula

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

Formant frequencies

Formants are defined as 'The spectral peaks of the sound spectrum ' of the voice. Formant is also used to mean an acoustic resonance and, in speech science and phonetic, a resonance of the human vocal tract. It is often measured by a peak in the frequency spectrum of the sound, using a spectrogram or a spectrum analyzer, though in vowels spoken with a high fundamental frequency, as in a female or child voice, the frequency of the resonance may lie between the widely-spread harmonics and hence no peak is visible. In sort, formant frequencies are the nominal center frequencies of the resonance.

Zero crossing rates

Zero-crossing signals have a direction attribute, which can have three values Rising, Falling and Either. In Rising, when a signal rises to zero to or through zero, or when a signal leaves zero and becomes positive zero crossing occurs. In case of Falling, a zero crossing occurs when signal falls to or through zero, or when a signal leaves zero and becomes negative. In case of Either, a zero crossing occurs if either a Rising or Falling occurs.

Clustering

After extracting features for each frame within word, to prepare feature vector for whole word feature vectors from each frame will be concatenated and passes to input layer of ANN. This will be convenient when each utterance has fixed length so that each utterance has same number of frames and each utterance has same number of features. But in real life each word has different length and even same word's length vary with time and speaker. In this we will get different number of frames for each word and can't use these feature vectors to train an ANN has fixed neuron in input layer. To overcome this problem we can calculate average number of frames for all utterances but in this way there is chance to loss data those have more discriminative capability. Other drawback is concatenation of feature vectors from each frame will gives large size of feature vector for whole utterance. So to overcome this problem in this work k-means clustering is used to find out the representing frames in utterance and then these frames are concatenated. Now feature vector has same size for each utterance because for each utterance same number of clusters is selected

III. RECOGNITION

A. Back Propagation Network

It is most popular multi-layer feed forward network trained using back propagation algorithm. It utilizes mean square and gradient decent to modify the connection weight of network. Here we trained using resilient back propagation algorithm that considers sign and magnitude of gradient whereas back propagation considers only magnitude for weight modification. Numbers of neurons in input layer are equal to number of element in feature vector and number of neurons in output layer depends on number classes.

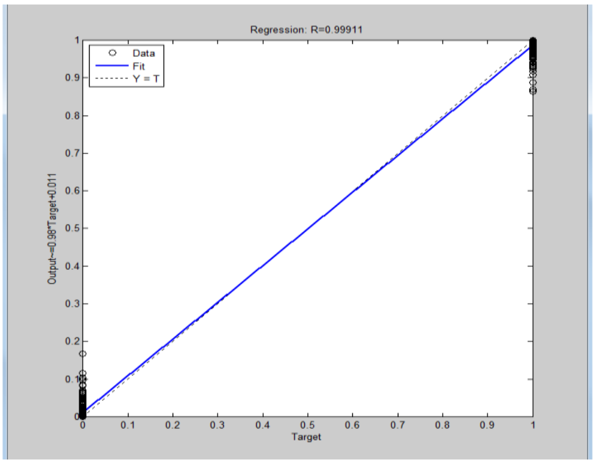

Fig 1

Overall speaker identification rate is 91% using BPA neural network. For gender identification recognition performance is 99%. Regression curve for Gender identification using BPA is shown in Fig 1.

Regression curve for Speaker identification using BPA is shown in Fig 2.

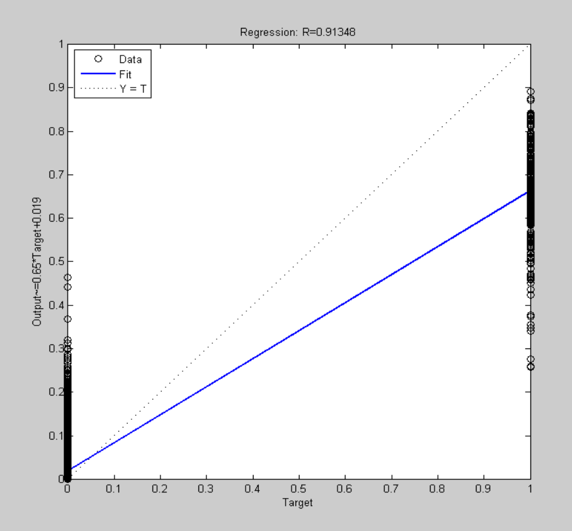

Fig2

B. Radial basis Function Neural Network

Radial Basis function network is a static type of feed forward network uses two layers one hidden layer and one output layer. In RBF network Gaussian or other basis kernel function is uses whereas in BPA network Sigmoid or S-shaped activation function uses. Since each hidden unit contains basis function, it has center and width. Let C_i is center for i^{th} hidden unit and V is feature vector. Euclidean distance Di is calculated at each hidden units.

$$Di= \| V – Ci \|$$

The output for each hidden unit computed by applying basis function G to this distance.

$$Oi = G(Di , \sigma i)$$

Here σi represent variance, in Gaussian function corresponding to variance. The σ value of function determines spread and by default spread constant is 1. Linear transfer function uses in output layer.

Speaker identification rate using RBF NN is 92%. Gender identification rate using RBF NN is 98%

In this experiment, we trained two separate network using same feature vector but for different target matrix. A data base of four languages (Hindi, English, Sanskrit and Telgu) has been prepared .We consider 50 speakers including 25 male and 25 female for each language and recorded sound *.wav file using microphone connected to personal computer. All speakers of respective languages uttered same paragraph for three minute duration in a noise free environment. Words taken for this research are in such formats that vowel and consonants both present one after other and the number of words recorded in Hindi and English are 5 and for Sanskrit and Telgu it is 4. We recorded 18 words of these three languages by 50 speakers so finally we have 900 speech utterances. The duration of speech utterances of all speakers are not fixed and they range from 22 sec to 70 sec.

Arrange these calculated features LPC, LPCC, MFCC and Δ-MFCC,ZCR for each of 900 speech utterances in a proper format (Matrix) that are further used for training and testing to BPA, RBF with one hidden layer and number of neurons in each hidden layer is in between number of inputs and target numbers. Regression curve for Gender identification using RBF is shown in Fig 3.

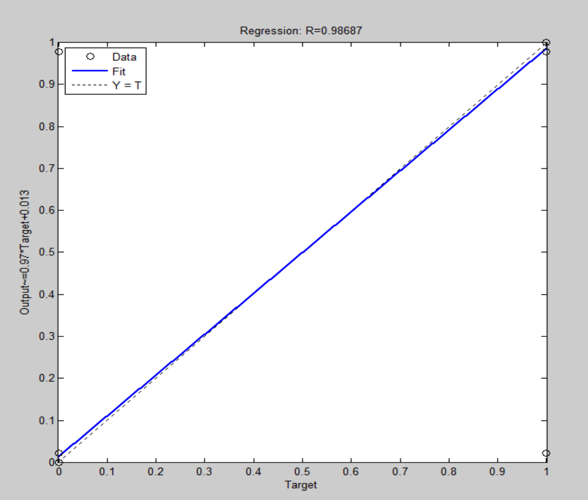

Fig 3

and Regression curve for Speaker identification using RBF is shown in Fig 4.

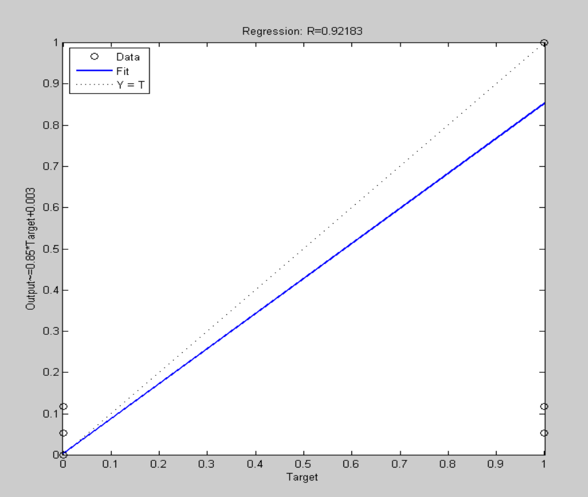

Fig 4

IV. CONCLUSION & FUTURE WORK

This work proposed the method to identify speaker cum gender identification on multilingual database. The experiment result shows BPA function performs better than RBF network. K-means clustering reduces the complexity of network by reducing number of hidden units.

In future, instead of using separate module for speaker and gender identification we can used single module for both in which no of neuron in output layer are equal to no. of neuron in gender identifier plus speaker identifier.

References

[1] R., Shukla A., Tiwari R., .,“ A Novel Approach to Classificatory Problem using Grammatical Evolution based Hybrid Algorithm”**,** 2010 International Journal of Computer Applications (0975 - 8887) Volume 1-No. 28.) Volume 1 – No. 28.

[2] Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh," Speaker Networks Recognition Using Neural and Conventional Classifiers" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 2, NO. 1,1994

[3] *Jayant* M. *Naik, Lorin P. Netsch and George R. Doddington,"* SPEAKER VERIFICATION OVER LONG DISTANCE TELEPHONE LINES", Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.

[4] C.S. Leung, M. Lee, and J.H. Chan (Eds.), “Gender Identification from Thai Speech Signal Using a Neural Network” ICONIP 2009, Part I, LNCS 5863, pp. 676–684, 2009

[5] Phillips, William J, Tosuner, C. ; Robertson, W. "SPEECH RECOGNITION TECHNIQUES USING RBF NETWORKS", WESCANEX 95. Communications, Power, and Computing. Conference Proceedings., IEEE (Volume:1), pp: 185 – 190.

[6] Kumar R.,Dutta S.,Kumara shama,”Gender Recognition using speech processing technique using LABVIEW” IJAET May 2011 .

[7] Md. Rabiul Islam1, Md. Fayzur Rahman,"Improvement of Text Dependent Speaker Identification System Using Neuro-Genetic Hybrid Algorithm in Office Environmental Conditions ",*IJCSI International Journal of Computer Science Issues*, Vol. 1, 2009.

[8] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet,” Front-End Factor Analysis for Speaker Verification” IEEE Transaction On Audio, Speech, And Language Processing, Vol. 19, No. 4, May 2011.

[9] Hui Lin, Li Deng, Jasha Droppo, Dong Yu, Alex Acero, “Learning Methods in Multilingual Speech Recognition”, *NIPS Workshop*, Whistler, BC, Canada ,2008.

[10] Hakkinen J., Jilei Tian"n-gram and decision tree based language identification for written words ", In Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE

[11] Stuker, S. Schultz, T. Metze, F. Waibel, A, “Multilingual articulatory features ”,*IEEE*, 2003.

[12] Tomi Kinnunen, ” Spectral Features for Automatic Text-Independent Speaker Recognition”, *Ph. Lic. Thesis, Department of Computer Science University of Joensuu* , 2004.

[13] Milan Sigmund. “Gender Distinction Using Short Segments Of Speech Signal”.

[14] J.CAMPBELL, J R. “Speaker Recognition : A Tutorial ”, *IEEE,*1997.

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

[16] Rehab F. M. F. Badran , Hany Selim , “Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes”, *Proceedings of ICSP*, 2000.

[17] K. Messer, J. Matas, J Kittler, J. Luettin, G. Maitre , “XM2VTSDB: The extended M2VTS data base ”, *2nd international conference on audio and video based biometric person authentication*,1999.

[18] W. Hess, Pitch determination of speech signals: algorithms and devices. Springer, 1983.

[19] J. M. Naik, L. P. Netsch, and G. R. Doddington, “Speaker verification over long distance telephone lines”, IEEE Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 1989, pages 524--527.