IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 4, June 2014.
www.ijiset.com
ISSN 2348 – 7968

Speaker and Gender Identification on Indian Languages using Multilingual Speech

Samiksha Sharma[1], Anupam Shukla[2] and Pankaj Mishra [3]

[1] Department of Information Technology, IIITM
Gwalior, MP,India

[2] Department of Information Technology, IIITM
Gwalior, MP,India

[3] Department of Electronics And Telecommunication,SSIPMT
Raipur, CG,India

Abstract

In this paper an attempt is made to develop speaker & gender identification system using continuous speech signal spoken in different languages as input. MFCCs and delta-MFCCs are used to build modal for classification .

Radial basis function network is used for classification. Here resilient back propagation algorithm used to train Multilingual Speech signal . Two separate modules are used for gender and speaker identification in each experiment. In this experiment accuracy of gender identification is 98.89% and speaker recognition is 87.22% using back propagation algorithm and 99.44% and 96.11% for gender and speaker identification using radial basis function. Radial basis function network perform much better than BPA network.

Keywords: *ANN, Speech feature, Speaker Recognition, Gender identification, Classification.*

1. 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. 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 attempt 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, 25]. Biometrics is seen by many researchers as a solution to a lot of user identification and

ISSN 2348 – 7968

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 [18].

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) [15, 16].

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.

2. Feature Extraction

2.1 Dataset

This database contain 18 utterances ISS, BAAR, TUM, JAO, NOW, TWAM, VELLU etc. Words in this database collected from 10 speakers (5 male and 5 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.

2.2 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 [9].

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 high 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-

$$Y=Y/\max(\text{abs}(Y));$$

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-

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

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.

3. Recognition

3.1 Back Propagation Network

It is most popular multi layer feed forward network trained using back propagation algorithm. It utilizes mean square

ISSN 2348 – 7968

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.

In this experiment, we trained two separate network using same feature vector but for different target matrix. There are 10 units in output layer because speech collected from 10 speakers and for gender identification two neurons are in output layer one for male another for female. Overall speaker identification rate is 87.22% using BPA neural network. For gender identification recognition performance is 98.89%. rate using BPA network

3.2 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 ith 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 96.11%. Gender identification rate using RBF NN is 99.44

4. CONCLUSION & FUTURE WORK

This work proposed the method to identify speaker cum gender on multilingual database. The experiment result shows radial basis function perform better than BPA network . In future, instead of using separate module for speaker and gender identification we can used single module for both.

REFERENCES

[1] Kala 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] 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

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

[4] 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.

[5] 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.

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

[7] Stuker, S. Schultz, T. Metze, F. Waibel, A, "Multilingual articulatory features ",*IEEE*, 2003.

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

[9] Rabiner, L.R., Schafer, R.W.: Introduction to Digital Speech Processing. Foundations and trends in Signal Processing 1, 1–194 (2007).

[10] Ting, H., Yingchun, Y., Zhaohui, W.: Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition. In: ICSP (2006)

[11] Milan Sigmund. "*Gender Distinction Using Short Segments Of Speech Signal*".

[12] J.CAMPBELL,JR. ," Speaker Recognition : A Tutorial ",*IEEE*,1997.

[13] Kevin R. Farrell, Richard J. Mammone, Khaled T. Assaleh, "Speaker Networks Recognition Using Neural and Conventional Classifiers ", *IEEE Transcations on speech and audio*, 1994.

[14] 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.

ISSN 2348 – 7968

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

[16] 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.

[17] Yakun Hu , Dapeng Wu, and Antonio Nucci " Pitch-based Gender Identification with Two-stage Classification".

[18] A. Acero and X. Huang, "Speaker and gender normalization for continuous-density hidden Markov models," in IEEE International conference, vol. 1. Citeseer, 1996.

[19] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of Intelligent Information Systems*, vol. 24,no. 2, pp. 179–198, 2005.

[20] M. Gelfer and V. Mikos, "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels," *Journal of Voice*, vol. 19, no. 4, pp. 544–554, 2005.

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

[22] H. Ting, Y. Yingchun, and W. Zhaohui, "Combining MFCC and pitch to enhance the performance of the gender recognition," in *SignalProcessing, 2006 8th International Conference on*, vol. 1, 2006.

[23] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on acoustics, speech and signal processing*, vol. 22, no. 2, pp. 135–141, 1974.

[24] R. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1,no. 2, pp. 129–134, 1993.

[25] 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.