International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Text-Dependent Speaker Identification and Verification Using Hindi Database in Adverse Acoustic Condition

Shrikant Upadhyay[1], Sudhir Kumar Sharma[2], Pawan Kumar[3], Aditi Upadhyay[4]

[1,3] Department of Electronics & Communication Engineering, Cambridge Institute of Technology, Ranchi, Jharkhand, India

[2,4] Department of Electronics & Communication Engineering, Jaipur National University, Jaipur, Rajasthan, India

Abstract: ***Human voice is one the medium through he/she communicate with the real world. Text-dependent voice or database can be used to protect your asset or privacy in many respects. Speaker identification and verification is one the issues that must be tested and verified so, that we can share information to the far location from the remote area and used for different application purposes like account access, password verification, pin access etc. Here, text-dependent Hindi database has been used from shunay to nau and we try to evaluate the efficiency and error rate in adverse acoustic condition. Original text might be corrupted or disturbed and person sitting at the access point may not identify the authenticate user in such situation. So, we put your effort with the help of this paper to identify and verify the speaker in adverse acoustic condition for real time applications. Here different combination of feature extraction method has been used to compute the performance of the database.***

Keywords: Hindi Database, Adverse Acoustic Condition, Feature Extraction, Text Dependent

1. Introduction

Speaker identification and verification is one the major challenges in the speech domain. This will help to solve many real time application issues. Text-dependent sample is quite easy to identify and create a database.

Text-dependent task involves some form of pre-determined or prompted password, in order to obtain the required text. It can be used for applications such as voice mode password or signature verification. In such applications, there is a need to change the password frequently and it can be done easily by changing the pre-determined text. In text-dependent speaker verification, during enrolment phase a limited number of utterances of the fixed text is collected. Therefore, approaches based on template matching are used for pattern comparison instead of approaches based on statistics or artificial neural networks, which needs a large amount of training data. Research in speech processing and communication, for the most part, was motivated by people‟s desire to build mechanicals models to emulate human verbal communication. Research interest in speech processing today has done well beyond the notion of mimicking human vocal apparatus [1]. People can reliably identify familiar voices and about 2-3 seconds is enough to identify a voice, although performance decreases for unfamiliar voices [2]. Even if duration of the utterance was increased, but played backward (which distorts timing and articulatory cues), the accuracy decreases drastically. Widely varying performance on this background task suggested that cues to voice recognition vary from voice to voice, and that voice patterns may consist of a set of acoustic cues from which listeners select a subset to use in identifying individual voices. Recognition often falls sharply when speakers attempt to distinguish their voices [3]. This is reflected in machines, where accuracy decreases when mimics act as impostors. Humans appear to handle mimics much better than machines do, easily perceiving when a voice is being mimicked [4]. If the target (intended) voice is familiar to the listener, he/she often associates the mimic voice with it. Certain voices are more easily mimicked than others, which lends evidence to the theory that different acoustic condition are used to distinguish different voices for real applications. Human performance in adverse conditions was also reviewed in [4], where it was reported clearly that human listeners are adept at using various cues to verify speakers in the presence of acoustic mismatch. Speaker recognition is one area of artificial intelligence where machine performance can exceed human performance- using short test utterances and N- number of speakers, machine accuracy often exceeds that of humans [4].

2. Speaker Identification and Verification

The objective of speaker identification is to classify an unlabeled utterance belonging to one of the N reference speakers [5]. It can be closed set identification or open set identification shown in Figure1. The objective of speaker identification is to decide the identity of speaker based on the speaker‟s voice, from set of N speakers i.e., one-to-many matching.

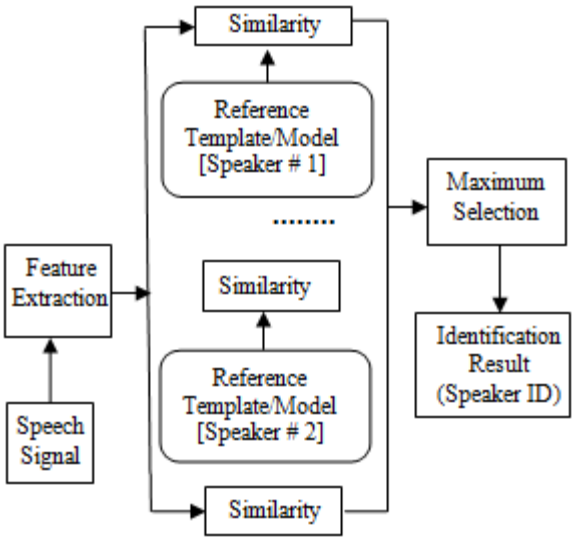

Figure 1: Speaker Identification System

The objective of speaker verification is to accept or reject the identity claim of speaker [6]. If the match between test and reference is above threshold level, the claim is accepted shown in Figure 2. Speaker verification is an open set problem. Speaker recognition systems can be further classified as text-dependent and text-independent systems. Speaker recognition by a machine involves three stages. They are: (1) Extraction of features to represent the speaker information present in the speech signal. (2) Modelling of speaker features. (3) Decision logic to implement the identification or verification task.

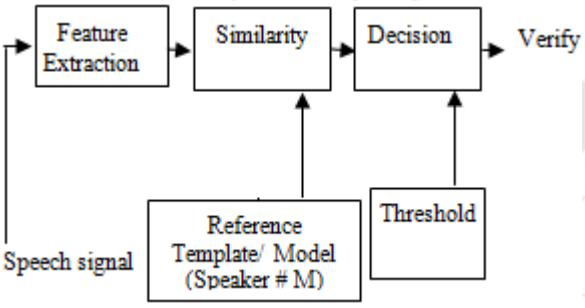

Figure 2: Speaker Verification System

3. Database Preparation

The base has been prepared for 50 speakers taking the voice of male and female of different age group using Goldwave 5.55 and cool software with the help of microphone. The database prepared is of isolated Hindi digits consisting of shunya, ek, do, teen, chaar, panch, cheh, saat, aath and nau. All speakers spoke 0-9 digits (in Hindi), each digit 10 times, which was recorded in a noise-free environment and saved as wave sound (.wav) file in some desired folder

Procedure for data collection

The voice of 50 different speakers has been recorded using monotype speaker. Each spectrum of Hindi digits from shunay to nau 10 times and each digit have been recorded in the form of spectrum by cutting each spectrum (frame) and by selecting "save option" from file menu each spectrum is saved using .wav file type in proper folder. Figure 3. shown below shows the spectrum of Hindi digits recorded.

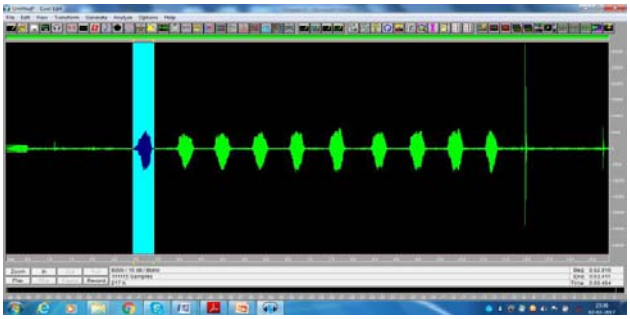

Figure 3: Recorded Spectrum

4. Feature Extraction

Feature extraction involved in signal modeling that performs temporal and spectral analysis. The need of feature extraction arises because the raw speech signal contains information to convey message to the observer or receiver and has a high dimensionality. Feature extraction algorithm derives a characteristics feature vector with lower physical or spatial properties.

A. Mel-Scale Cepstrum Co-efficient (MFCC)

MFCC technique is basically used to generate the fingerprints of the audio files. Let us consider each frame consist of „N" samples and let its adjacent frames be separated by „M" samples where M is less than N. Hamming window is used in which each frame is multiplied. Mathematically, Hamming window equation is given by:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

Now, Fourier Transform (FT) is used to convert the signal from time domain to its frequency domain. Mathematically, it is given by:

$$X_k = \sum_{i=0}^{N-1} x_i \; e^{\frac{2\pi i k}{N-1}} \quad (2)$$

$$M = 2595 log_{10}\left(1+\frac{f}{700}\right) \quad (3)$$

B. Linear Predictive Coding Analysis (LPC)

It is a frame based analysis of the speech signal performed to provide observation vectors [6]. The relation between speech sample S (n) and excitation X(n) for auto regressive model (system assume all pole mode) is explained mathematically as:

$$S(n) = \sum_{k=1}^{p} a_k \; s(n-k) + G.X(n) \quad (4)$$

The system function is defined as:

$$H(z) = \frac{S(Z)}{X(Z)} \quad (5)$$

A linear predictor of order „p" with prediction co-efficient (α_k) is defined as a system whose output is defined as:

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k \; S(n-k) \quad (6)$$

The system function is p^{th} order polynomial and it follows:

$$P(z) = \alpha_k \, z^{-k} \quad (7)$$

The prediction error e (n) is defined as:

$$e(n) = s(n) - \hat{s}(n)$$

$$= s(n) - \sum_{k=1}^{p} \alpha_k \; S(n-k) \quad (8)$$

The transfer function of prediction error sequence is:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k \; z^{-k} \quad (9)$$

Now, by comparing equation (5) and (10), if $\alpha_{k} = \alpha_{k}$ then A (z) will be inverse filter for the system H (z) of equation (6):

$$H(z) = G/A(z) \quad (10)$$

The purpose is to find out set of predictor coefficients that will minimize the mean squared error over a short segment of speech waveform.

C. Linear Prediction Cepestral Coefficients (LPCC)

The basic parameters for estimating a speech signal, LPCC play a very dominant role. This method is that where one speech sample at the current time can be predicated as a linear combination of past speech sequence or sample. LPCC algorithm in term of block diagram is shown in Figure 4. below

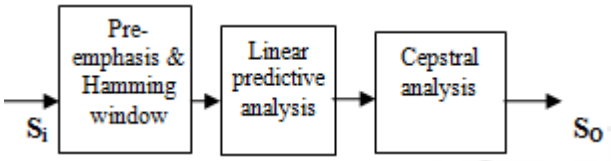

Figure 4: LPCC Processing

A digital all-pole filter is used to model the vocal tract and has a transfer function represented in z-domain as:

$$V(z) = \frac{G}{1-\sum_{k=1}^{p} a_{k}\ _{z^{k}}} \quad (11)$$

where, V (z) is the vocal tract transfer function, G is the gain of the filter, a_k is the set of auto regression coefficients known as linear prediction coefficients (LPC) and p is the order of all-pole filter. One of the efficient method for estimating the LPC coefficients and filter gain is autocorrelation [7]. The inverse FFT transform of the logarithm of the speech magnitude spectrum and it is defined as:

$$\hat{s}[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \ln[s(w)]\, e^{jwn}\ dw \quad (12)$$

5. Related Work

This paper will focused on the Hindi voice sample taken from „shunaya to nau" i.e. 0 to 9 and using various above mention feature techniques with proposed model are analyzed and the efficiency and error rate calculated. The analysis result proves to be useful for real time applications. The above mention feature extractions techniques has been analyzed in many papers but the combinations of these feature extraction has been less found in any research article. So, we try to combine these feature extraction and like LPC+LPCC, MFCC+LPCC and MFCC+LPC and try to judge the performance in adverse acoustic conditions using ROVER (Recognized Output Voting Error Reduction) and CNC (Combined Combination).

6. Simulation Results

The simulation result have been analyzed on the below mention Figure 5. using ROVER model.

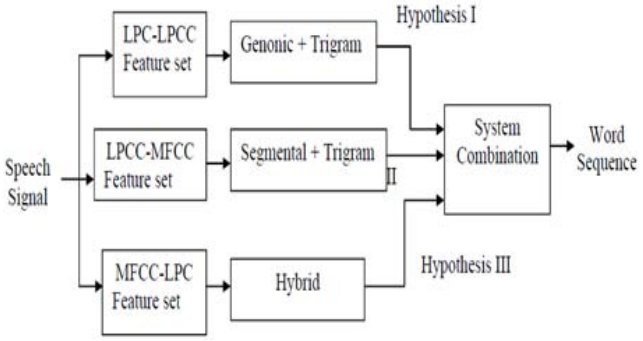

Figure 5: Proposed Model

Here the combinations of three feature extraction technique i.e. LPC + LPCC, LPC + MFCC and MFCC + LPCC have been analyzed for same database of 50 speakers in noisy condition. Table 1. shows the efficiency rate of Hindi dialects from „shunay to nau" and Figure below shows the in graph form the efficiency rate using proposed model in noisy environments.

Table 1: Efficiency Rate of Hindi Dialects

Hindi Dialects	LPC+ LPCC (%)	LPCC+ MFCC (%)	MFCC+ LPC (%)
SHUNYA	90.10	98.46	95.44
EK	91.78	97.65	95.89
DO	91.88	97.98	93.11
TEEN	90.12	97.09	91.08
CHAR	92.54	98.37	90.28
PANCH	87.32	99.55	90.11
CHEH	92.18	98.27	92.55
SAAT	93.91	96.78	91.07
AATH	90.55	97.99	91.28
NAU	93.10	96.98	94.14

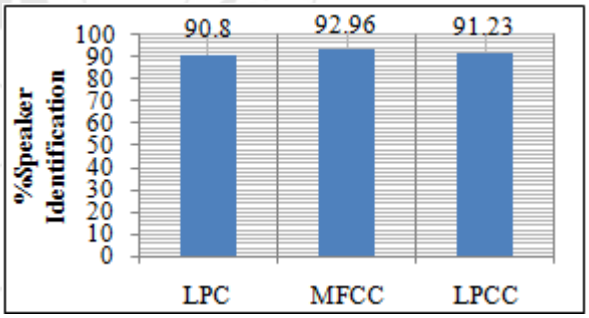

Figure 6: Efficiency rate for „SHUNYA"

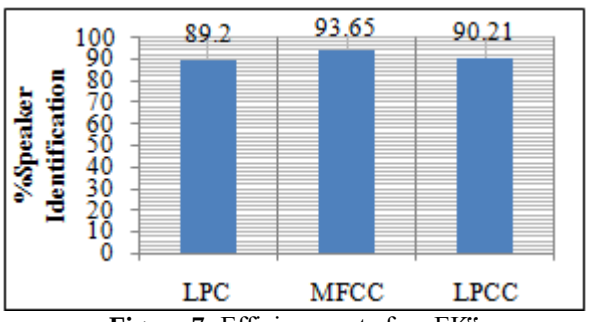

Figure 7: Efficiency rate for „EK"

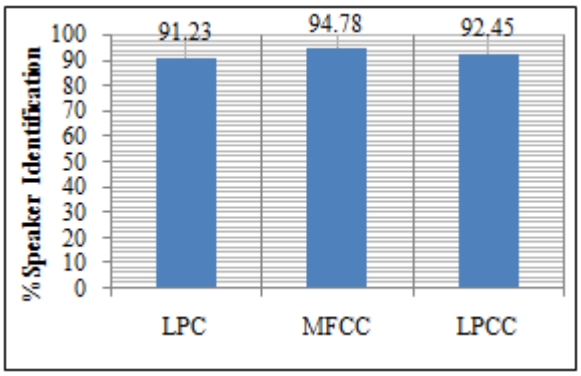

Figure 8: Efficiency rate for „DO"

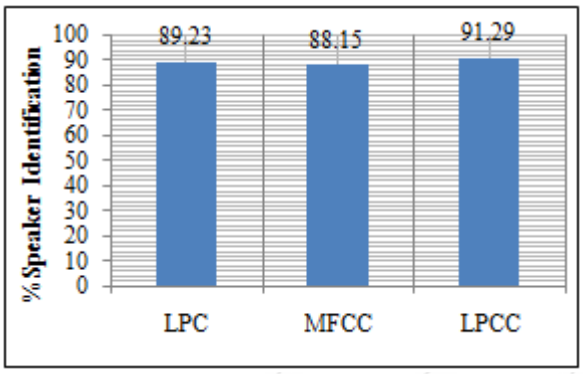

Figure 9: Efficiency rate for „TEEN"

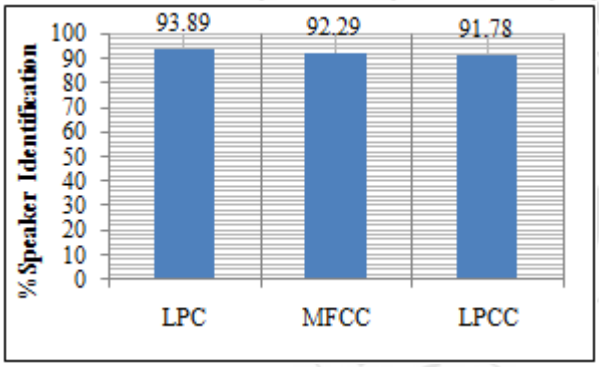

Figure 10: Efficiency rate for „CHAR"

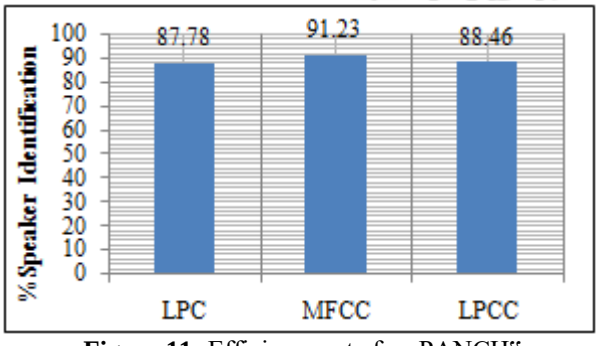

Figure 11: Efficiency rate for „PANCH"

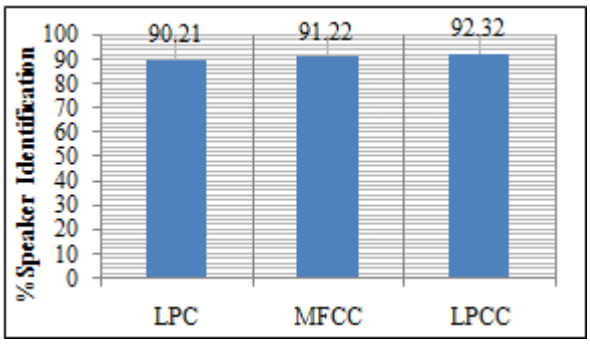

Figure 12: Efficiency rate for „CHEH"

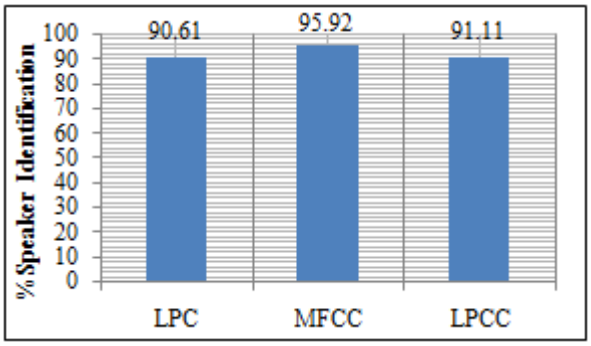

Figure 13: Efficiency rate for „SAAT"

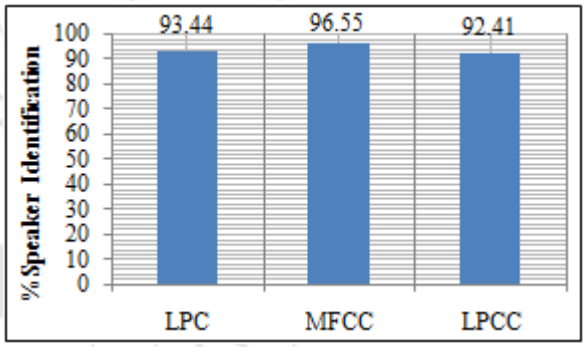

Figure 14: Efficiency rate for „AATH"

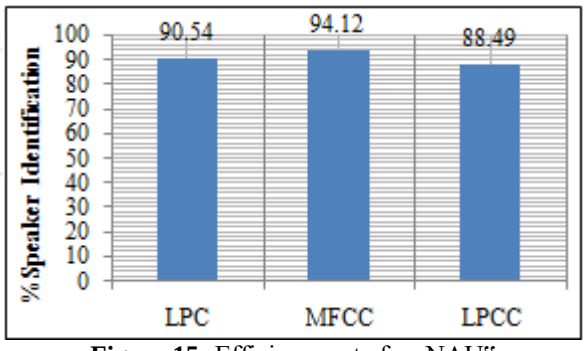

Figure 15: Efficiency rate for „NAU"

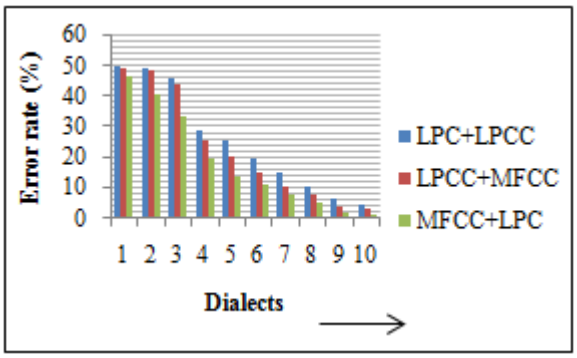

Figure 16: Error rate for combined feature extractions

The above Table 2 shows the error rate for the proposed model during training and testing.

Table 2: Error rate of proposed model in ideal condition (ROVER)

Dialects Hindi	Train (sec)	Test (sec)	Feature Extractions		
			LPC+LPCC (%ER)	LPCC+MFCC (%ER)	MFCC+LPC (%ER)
SHUNYA	5	5	50.22	49.20	46.54
EK	5	10	49.32	49.12	40.78
DO	5	15	46.11	44.43	33.56
TEEN	10	20	29.34	25.65	19.54
CHAR	10	5	25.65	20.56	14.11
PANCH	10	10	19.54	15.23	11.10
CHEH	20	15	14.98	10.54	7.76
SAAT	20	20	10.67	7.98	4.98
AATH	20	5	6.45	4.12	2.19
NAU	20	10	4.42	3.53	1.22

7. Conclusion

Efficiency rate and error rate of Hindi dialects from shunya to nau has been analyzed first in which MFCC will have less error rate compared to LPC and LPCCC in ideal condition. For noisy condition error rate is quite low for MFCC compared to LPC and LPCC. So, MFCC in noisy condition proves to be efficient. Combination of MFCC+LPC using the ROVER model will have higher efficiency compared to LPC+LPCC and LPC+MFCC in ideal condition. Error rate of MFCC+LPPC is found to be 1.22% compared to LPC+LPCC and LPCC+MFCC i.e. 4.22% and 3.53% in ideal condition. Also, in noisy environment same model for MFCC+LPC combination efficiency is quite high 98.12% compared to LPC+LPCC and LPCC+MFCC i.e. 90.35% and 96.89%. Error rate of MFCC+LPC is found to be 1.00% compared to LPC+LPCC and MFCC+LPCC. So, combination of MFCC+LPC is proves to be efficient for adverse acoustic condition and the proposed model work well in such scenario.

References

[1] D. G. Childers, R. V. Cox, R. Demori, B. H. Juang, J. J. Mariani, P. Price, S. Sagayama, M. M. Sondhi and R. Weischedel, "The past, present and future of speech processing," *IEEE Processing Magazine,* pp. 24-48, May 1998.

[2] D. Lancker, J. Kreiman and K. Emmorey, "Familiar voice recognition: Pattern and parameters- recognition of backward voices," *Phonetics*, vol. 13, no. 1, pp. 19-38, 1985.

[3] A. Reich and J. Duke, "Effect of selected vocal disguises upon speaker identification by listening," J.Acoust. Soc. Amer., vol.66, no. 4, pp. 1023-1028, 1979.

[4] M. Sigmund, "Speaker recognition identifying people by their voices," *Habilitation thesis, Brno University of Technology, Institute of radio electronics,* 2000.

[5] H. Gish and M. Schmidt, "Text-dependent speaker identification," *IEEE Signal Processing Magazine*, pp. 18-32, October 1994.

[6] S. Fururi, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 254-272, April 1984.

[7] Abdelnaiem, "LPC and MFCC performance evaluation with artificial neural network for spoken language identification," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, pp. 55-66, June 2013.