Effectiveness of LP Based Features for Identification of Professional Mimics in Indian Languages

Hemant A. Patil[1], P. K. Dutta[2] and T. K. Basu[2]
[1]*Department of Electronics and Instrumentation Engineering,*
Dr. B.C. Roy Engineering College, Durgapur West Bengal, India.
[2]*Department of Electrical Engineering,*
Indian Institute of Technology, Kharagpur, PIN 721302, India.
E-mail: hemant_patil1977@yahoo.com, {hemant, pkd,tkb}@ee.iitkgp.ernet.in

Abstract

Automatic Speaker Recognition (ASR) is an economic tool for voice biometrics because of availability of low cost and powerful processors. For an ASR system to be successful in practical environments, it must have high mimic resistance, i.e., the system should not be defeated by determined mimics which may be either identical twins or professional mimics. In this paper, we demonstrate the effectiveness of Linear Prediction (LP) based features viz. Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC) over filterbank based features such as Mel-Frequency Cepstral Coefficients (MFCC) and newly proposed Teager energy based MFCC (T-MFCC) for the identification of professional mimics in Marathi and Hindi languages.

1. Introduction

Automatic Speaker Recognition (ASR) has been an active area of research in speech processing. The use of standard speech corpora for evaluation of ASR is the most crucial task in speech and speaker recognition systems. In addition to this, the ASR system should have high mimic resistance, i.e., the system should not be defeated by determined mimics. Mimics can be of two types viz. one based on physiological characteristics such as identical twins or triplets and the one based on behavior or learned characteristics such as professional mimic (or Bahurupee in India) [14].

Four mimic experiments have been reported in references [14]. The first, being done by Lumins and Rosenberg at Bell Labs, reported the significance of

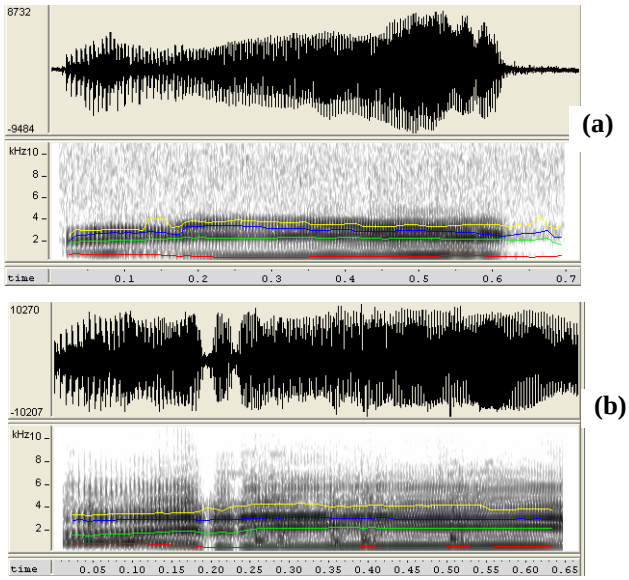

Figure 1. Speech signal and corresponding spectrogram of Hindi word 'Arrye' spoken by (a) Professional mimic and (b) target speaker.

formant information in mimic recognition task [8] whereas the second experiment, was done by Doddington at TI [5]. Both of these experiments revealed that mimic acceptance into the system were significantly greater than acceptance of casual impostors. Third experiment reported by Hair and Rekieta found that mimic acceptance was successful for individual features used in their verification system but it was unsuccessful when all the features were combined [6]. Finally, the fourth experiment reported by Luck found that only the worst casual impostor got some success for acceptance but not enough to be accepted into the system [9]. All of these experiments were performed in speaker verification mode. The

present study is concerned with *identification* of professional mimics in the presence of target speakers.

The evaluation of a system's resistance to mimics depends on the definition of a skilled mimic. Since the skills required of a mimic vary from system to system and since subjective impressions can be quite faulty, the most reliable definition of skilled mimic is not based on a priori appraisal. It is simply one who can significantly increase his acceptance as an impostor by deliberate imitation. This makes *mimic evaluation difficult* since one cannot be certain in advance who is a skilled mimic [14]. Figure 1 shows speech signal and corresponding spectrograms along with *formant contour* (F1-F4) of the Hindi word, *"Arrye"* spoken by professional mimic and target speaker. It is evident that the spectrograms are quite similar.

In this paper, a methodology and a typical experimental setup used for evaluation of mimic resistance (in terms of success rates) of ASR system against professional mimics in Indian languages viz. Marathi and Hindi has been described. And the results with different LP based and filterbank based feature sets have been reported. To the best of the authors' knowledge, there is no publicly available corpus in Indian languages for ASR of professional mimics in real life settings; so it was decided to design and develop a suitable corpus for this purpose.

Next section describes the experimental setup, data collection and corpus design procedure used in this study. Different speech features used in this study are discussed followed by discussion on polynomial classifier techniques for speaker modeling. Finally, an assessment of different feature sets for the proposed problem is presented.

2. Experimental Setup

A typical experimental setup consists of a close talking microphone, voice activated tape recorder and Pentium-III machine having speech processing software. Other studies in corpus design for ASR can be found in [2]. In this paper, two major experiments have been performed viz. real and fictitious. For real experiment, the mimic is imitating actual target speakers' voices in Hindi whereas in fictitious experiments, mimic is imitating imaginary target speakers (selected with perceptual judgments from different dialectal zones of Maharashtra) in Marathi. Pre-recorded cassettes of professional mimic in Marathi (who has performed more than 2000 mimic experiments till date) and Hindi (famous for his comic role) have been played back. 22 mimic voices of people in Marathi from different places and age groups have been produced. Another set of 21 voices of different speakers in Marathi have been used as the training data set. Mimic's original voice was also retained during the training process. The recording was done with the help of voice activated (VAS) tape recorders (Sanyo model no. M-1110C Aiwa model no. JS299) with microphone input and close talking microphones (viz. *Frontech and Intex*). The data is recorded on the Sony high fidelity voice and music recording cassettes (C-90HFB). A list consisting of five questions, isolated words, digits, combination-lock phrases, read sentences and a contextual speech of considerable duration was prepared in Marathi. The contextual speech consisted of description of nature or memorable events etc. of community or family life of the speaker. The topics were generally easy and simple for the speaker to think instantaneously and interact and the speech was usually conversational and quite varied. The interview was started with some questions to know about the personal information of the speaker such as his\her name, age, education, profession, etc. The data was recorded with 10 repetitions except for the contextual speech. The target speakers for real experiment in Hindi are collected with the help of different videos. Such video files are separated into audio files and image frames of video. The audio files had a sampling frequency 44.1 KHz which was downsampled to 22050Hz and audio files are normalized to peak amplitude value in the audio data. The silence periods are removed. All these operations are done in software. These files are saved as *. wav files for further processing. Corpus is designed into training segments of 30s, 60s, 90s and 120s durations and testing segments of 1s, 3s, 5s, 7s, 10s, 12s and 15s in order to find the performance of the system for various training and testing durations [2].

3. Speech Features

In this paper, LP based features such as LPC and LPCC and filterbank based features such as MFCC and T-MFCC have been considered for mimic recognition task. Computational details of MFCC are given in [4]. In the next subsection, we will briefly review the computational details of LPCC and T-MFCC.

3.1. LPC Based Features

Given that all the poles $z=z_i$ are inside the unit circle and the gain is 1, the causal LP cepstral coefficients (LPCC) of H (z) is given by [1],[10]:

$LPCC(n) = \frac{1}{n}\sum_{i=1}^{p}|r_i|^n \cos(\theta_i n), n > 0$, for complex $z_i = r_i \exp(j\theta_i)$. where z_i's are the poles of LP transfer function.

3.2. Filterbank Based Features- MFCC and T-MFCC

Recently, Tegaer energy based MFCC has been proposed by Patil and Basu for twin identification problem [12]. And it was felt desirable that it should be also employed for the present problem. Traditional methods of extraction of MFCC based features involve Mel-spectrum of pre-processed speech, followed by log-compression of subband energies and finally DCT computation [4]. For the computation of T-MFCC, we employ Teager Energy Operator (TEO) for calculating the energy of speech signal. Speech signal $x(n)$ is first passed through pre-processing stage (which includes frame blocking, Hamming windowing and pre-emphasis) to give pre-processed speech signal $x_p(n)$. TEO of a signal $x_p(n)$ is defined as [7], [15]:

$$\Psi[x_p(n)] = x_p{}^2(n) - x_p(n+1)x_p(n-1) = \psi_1(n)(say) \quad (1)$$

Now, one may apply TEO in frequency domain, i.e., TEO of each subband at the output of Mel-filterbank, but there is difficulty from implementation point of view. As discussed below in brief. In frequency-domain, eq (1) for pre-processed speech $x_p(n)$ implies

$$F\{\psi_1(n)\} = F\left\{x_p{}^2(n) - x_p(n+1)x_p(n-1)\right\} \Rightarrow F\left\{x_p{}^2(n)\right\} - F\left\{x_p(n+1)x_p(n-1)\right\} \quad (2)$$

Using shifting and multiplication property of Fourier transform, we have

$$F\left\{x_p{}^2(n)\right\} = \frac{1}{2\pi}\int_{2\pi} X_p(\theta)X_p(\omega-\theta)d\theta$$

$$F\left\{x_p(n+1)x_p(n-1)\right\} = \frac{1}{2\pi}\int_{2\pi} X_{1p}(\theta)X_{2p}(\omega-\theta)d\theta$$

where $X_{1p}(\omega) = e^{-j\omega}X_p(\omega)$ and $X_{2p}(\omega) = e^{j\omega}X_p(\omega)$

Hence, eq (2) becomes

$$F\{\psi_1(n)\} = \frac{1}{2\pi}\int_{2\pi}\left(1 - e^{j\omega}e^{-2\theta}\right)X_p(\theta)X_p(\omega-\theta)d\theta \quad (3)$$

It is evident that eq (3) is difficult to implement in discrete-frequency domain and is also time-consuming. So we have applied TEO in the time-domain. Let us now see the computational details of T-MFCC.

For T-MFCC, the magnitude spectrum of the TEO output is computed and warped to Mel frequency scale followed by usual log and DCT computation (of MFCC) to obtain T-MFCC. The Mel-filterbank is shown in figure 2.

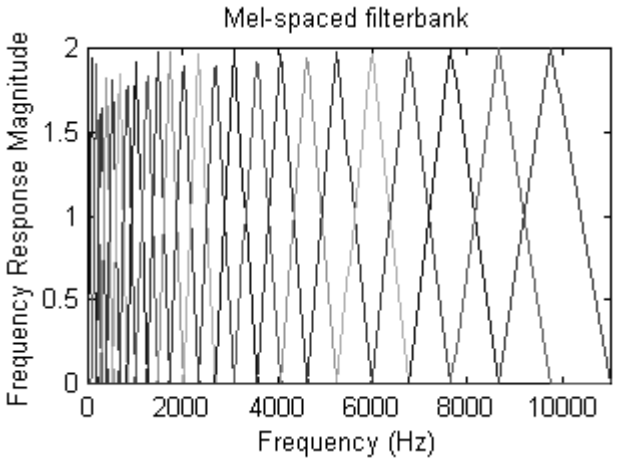

Figure 2. Mel-spaced filterbank

$$T - MFCC = \sum_{l=1}^{L}\log\left[\Psi_1(l)\right]\cos\left(\frac{k(l-0.5)}{L}\pi\right), k = 1, 2, ..., Nc.$$

where $\Psi_1(l)$ is the filterbank output of $F\{\psi_1(n)\}$ and $\log\left[\Psi_1(l)\right]$ is the log of filterbank output and $T - MFCC(k)$ is the k[th] T-MFCC. T-MFCC differs from the traditional MFCC in the definition of energy measure, i.e., MFCC employs L^2 energy in frequency domain (due to Parseval's equivalence) at each subband whereas T-MFCC employs Teager energy in time domain [12].

4. Polynomial Classifier

Campbell *et al.* first proposed polynomial classifier for speaker verification application [3] and later Mitra *et al.* applied this to speaker identification in Indian languages viz. Marathi and Hindi [11]. The basic structure of the classifier is shown in figure 3. The feature vectors are processed by the polynomial discriminant function. Every speaker i has a speaker specific vector $\mathbf{w}_i$, to be identified during training and the output of a discriminant function is averaged over time resulting in a score for every $\mathbf{w}_i$.

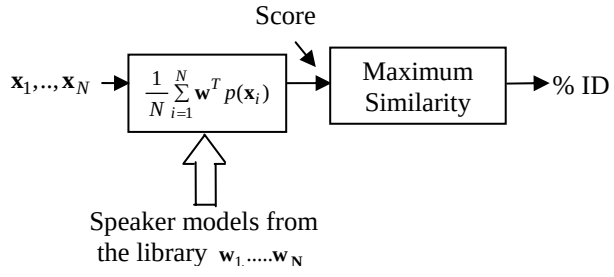

Figure 3. The Classifier Structure

The score is then given by, $s_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}^T p(\mathbf{x}_i)$

where $\mathbf{x}_i$ = i[th] input test feature vector,

$\mathbf{w}$ = Speaker model/voiceprint vector and

$p(\mathbf{x})$ = Vector of polynomial basis terms of the input test feature vector.

Training polynomial classifier is accomplished by obtaining the optimum speaker model for each speaker using discriminatively trained classifier with mean-squared error criterion, i.e., for a speaker's feature vector, an output of one is desired, whereas for an impostor data an output of zero is desired.

For the two-class problem, let $\mathbf{w}_{spk}$ be the optimum speaker model, ω the class label, and $y(\omega)$ the ideal output, i.e., $y(spk) = 1$ and $y(imp) = 0$ .The resulting problem using MSE is

$$\mathbf{w}_{spk} = \arg\min_{w} E\left\{\left(\mathbf{w}^T p(\mathbf{x}) - y(\omega)\right)^2\right\},$$

where $E\{.\}$ means expectation over X and ω .This can be approximated using the training feature set as

$$\mathbf{w}_{spk} = \arg\min_{\mathbf{w}}\left[\sum_{i=1}^{N_{spk}}\left|\mathbf{w}^T p(\mathbf{x}_i) - 1\right|^2 + \sum_{i=1}^{N_{imp}}\left|\mathbf{w}^T p(y_i)\right|^2\right] \quad (4)$$

where $\mathbf{x}_{1,.....,}\mathbf{x}_{N_{spk}}$ are speaker's training data and $y_{1,.....,}y_{N_{imp}}$ is the impostor data. This training algorithm can be expressed in matrix form.

Let $\mathbf{M}_{spk} = \left[\, p(\mathbf{x}_1) \quad p(\mathbf{x}_2) \quad ..... \quad p(\mathbf{x}_{N_{spk}}) \,\right]^T$ and a similar matrix for $\mathbf{M}_{imp}$. Also let $\mathbf{M} = \left[\mathbf{M}_{spk} \quad \mathbf{M}_{imp}\right]^T$ and thus the training problem in eq (4) is reduced to the well-known linear approximation problem in normed space as

$$\mathbf{w}_{spk} = \arg\min_{w} \left\|\mathbf{M}\mathbf{w} - \mathbf{o}\right\|_2,$$

where $\mathbf{o}$ consists of N_{spk} ones followed by N_{imp} zeros. We define $\mathbf{R}_{spk} = \mathbf{M}_{spk}^T\mathbf{M}_{spk}$ and define $\mathbf{R}_{imp}$ similarly; and then the problem can be solved using the method of normal equations,

$$(\mathbf{R}_{spk} + \mathbf{R}_{imp})\mathbf{w}_{spk} = \mathbf{M}_{spk}^T\mathbf{1} \quad (5)$$

where $\mathbf{R}_{spk} = \mathbf{M}_{spk}^T\mathbf{M}_{spk}$ and $\mathbf{1}$ is the vector of all ones. Also define $\mathbf{R} = \mathbf{R}_{spk} + \mathbf{R}_{imp}$. Thus eq (5) reduces to

$$\mathbf{w}_{spk} = \mathbf{R}^{-1}\mathbf{M}_{spk}^T\mathbf{1} \quad (6)$$

One of the advantages of training algorithm (6) is that, optimum speaker model does not depend upon the duration of the training speech but it is the length of the feature vector which predominantly determines the computational load on the machine. Since elements of polynomial basis vector form a semigroup of monomials, we can use a mapping algorithm based on semigroup isomorphism property of monomial which allows one to transform a symbol manipulation (polynomial basis terms) onto a number manipulation (set of primes). The details of training algorithm for multi-class problem, polynomial basis determination and mapping algorithm based semi-group isomorphism property of monomials for computing unique terms in $\mathbf{R}_{spk}$ and hence $\mathbf{R}$ are given in [3], [11].

5. Experimental Results

In this paper, polynomial classier of 2[nd] order approximation is used as the basis for all the experiments. The results are shown in two parts viz. for real and fictitious experiments. A 12[th] order LPC were extracted for frame of 23.22ms (512 samples) duration after pre-processing. LPCC was calculated from roots of LPC polynomial. The standard MFCC computations were performed as per method suggested in [4]. The pre-processing for MFCC and T-MFCC are similar except for mean removal. For calculating T-MFCC, we have taken 514 samples for each frame for TEO processing.

5.1. Results on Real Experiment

Database organization for this experiment is shown in figure 4. The training template contains the normal voice of professional mimic and 7 real target speakers whereas testing template contains the mimic's imitations for 7 target speakers, his normal voice and normal testing voices of the 7 target speakers. For this experiment, the success rates are found by counting the number of mimic's testing voices correctly identified as mimic's normal voice plus number of correctly identified target speakers' normal testing voices and mimic's normal testing voice. Results are shown as average success rates (average computed over testing segments of 1s, 3s, 5s, 7s, 10s, 12s, and 15s) for different training (TR) durations in table 1. It is clear from the results that the LP based features (i.e., LPC and LPCC) performed slightly better when compared with the filterbank based ones (i.e., MFCC and T-MFCC).

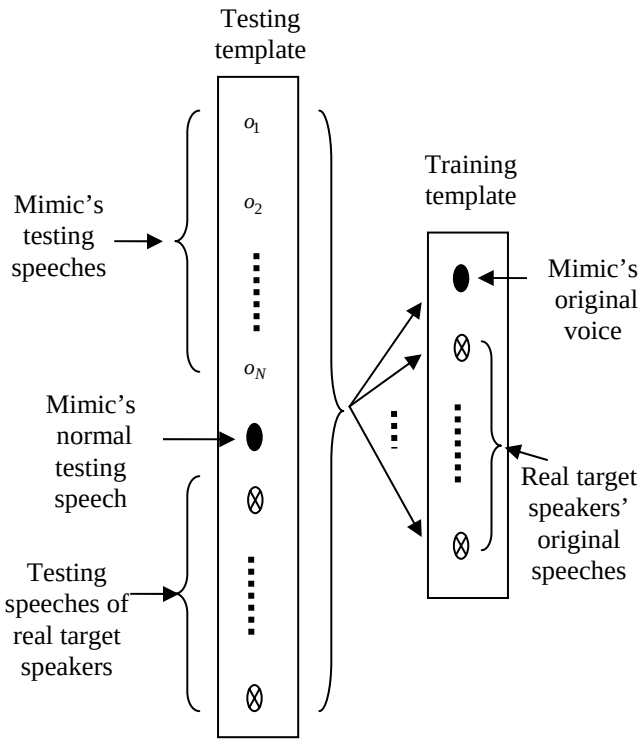

Figure 4. Database organization for real experiment

TABLE 1
AVERAGE SUCCESS RATES (%) FOR REAL EXPERIMENT WITH 2ND ORDER APPROXIMATION (HINDI MIMIC)

FEATURE \ TR	30s	60s
LPC	**98.09**	**99.04**
LPCC	**100**	**99.04**
MFCC	99.04	99.04
TMFCC	94.28	97.14

5.2. Results on fictitious experiments

Database organization for this experiment is shown in figure 5. The training template contains the normal voice of professional mimic and 22 imaginary target speakers whereas testing template contains the mimic's imitations for 22 imaginary target speakers only. For this experiment, the success rates are found by counting the number of mimic's testing voices correctly identified as mimic's normal voice. The results are shown in Tables 2. Some of the observations from the results are as follows:

1) LPC model gives 50-60% average success rates for almost all the cases of training durations.
2) The performance of MFCC is also very high but it proves to be less effective than LPC and LPCC. model whereas T-MFCC does not perform so well. Its success rate is in the range of 30%.

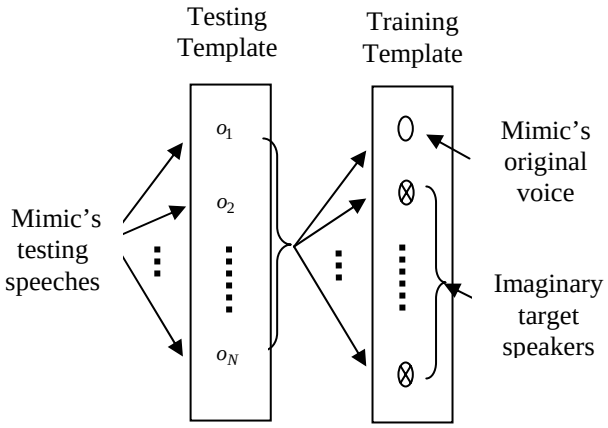

Figure 5. Database organization for fictitious experiment

TABLE 2
AVERAGE SUCCESS RATES (%) FOR FICTITIOUS EXPERIMENT WITH 2ND ORDER APPROXIMATION (MARATHI MIMIC)

FS \ TR	30s	60s	90s	120s
LPC	**57.14**	**58.43**	**59.08**	**61.03**
LPCC	**62.98**	**64.28**	**66.23**	**65.58**
MFCC	50.64	49.34	49.34	50.66
T-MFCC	27.26	26.61	27.26	27.91

3) On the whole, filterbank based features such as MFCC and T-MFCC do not perform well as compared to LP based features for this problem. This is contradictory to the result in normal ASR where filterbank based features perform better than LP based features [13]. This is probably due to the fact that filterbank based features are based on the human perception process and also the concept of energy (of the speech frame) involved in these models. So, when the mimic is performing, human perception process (in turn MFCC and T-MFCC features) will perceive these as the voice of a person whose voice the mimic is imitating. Hence the chances of misclassification will go up with these features.
4) LP based features perform well in this problem, because LPC model represents the combined effect of vocal tract (formant frequencies and their bandwidths and thus in turn emphasizes the *formant structure* more dominantly), glottal pulse

and radiation model and in turn the physiological characteristics of mimic's vocal tract. So even if mimic is imitating other person's voice to fool human perception process (so to the features based on it viz. MFCC and T-MFCC), he cannot change his\her physiological characteristics of the vocal tract which are known to be nicely tracked by LP based features. So, in the testing phase, LP based features track these properties dominantly as compared to filterbank based features and hence outperform MFCC and T-MFCC feature sets in the identification process.

5.3. Analysis of Results through MSE

Results reported in table 1 and 2 are justified by following experiment. In this experiment, Mean Square Error (MSE) is calculated between testing and training feature vectors for two cases viz. case 1 represents MSE between mimic'c imitations for target speakers and his normal voice whereas case 2 represents MSE between mimic'c imitations for target speakers and normal voice of the target speakers.

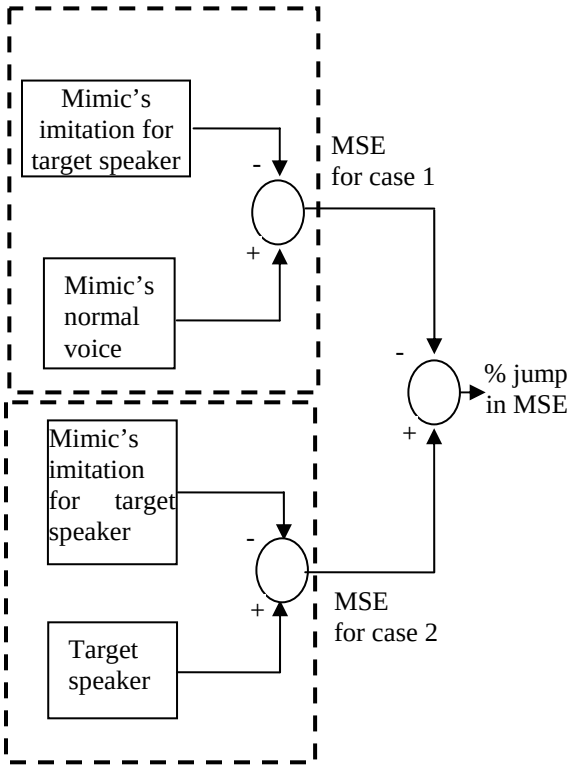

Figure 6. Schematic for calculation of % jump in MSE

Percentage jump in MSE from case 1 to case 2 is also calculated (as shown in figure 6). The main objective of this experiment is to investigate the effectiveness of LP based features for the present problem as compared filterbank based features. Figure 7 and 8 shows MSE for LPC, LPCC, MFCC and T-MFCC for first 429 frames corresponding to an utterance of 5s durations

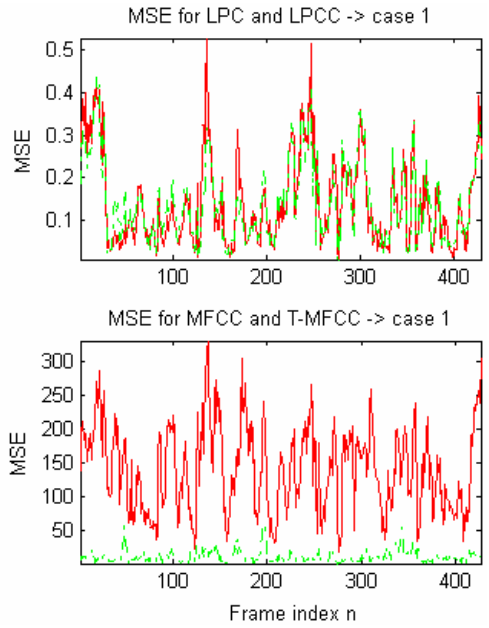

Figure 7. MSE for case 1

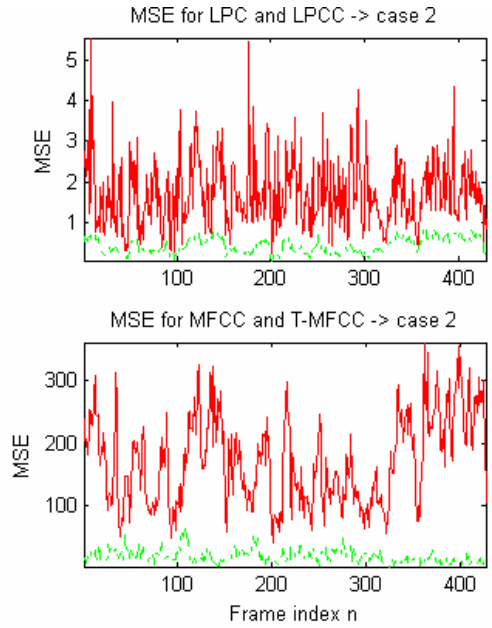

Figure 8. MSE for case 2

for case 1 and case 2 whereas Table 3 shows overall MSE (average calculated over 429 frames) for case 1 and case 2 and % jump of average MSE from case 1 to case 2 for LPC, LPCC, MFCC and T-MFCC. For a D-dimensional feature vector of nth frame, the equation for MSE is given by

$$MSE(n) = \frac{1}{N_D} \sum_{i=1}^{N_D} \left| \mathbf{x}_{tr_i}^n - \mathbf{x}_{te_i}^n \right|^2$$

where

$MSE(n)$ =Mean Square Error for n[th] frame.

$\mathbf{x}_{tr_i}^n$ =i[th] feature value in n[th] training feature vector for normal voice of the professional mimic (case 1) or normal voice of the target speaker (case 2).

$\mathbf{x}_{te_i}^n$ = i[th] feature value in n[th] testing feature vector for normal voice of mimic's imitations for the target speaker.

N_D =dimension of the feature vector.

TABLE 3
ANALYSIS OF RESULTS SHOWN IN TABLES 1-2 THEROGH OVERALL (OVER 429 FRAMES) MSE

FS / Av. MSE	LPC	LPCC	MFCC	T-MFCC
Case1	0.1433	0.1405	140.05	11.03
Case 2	1.7161	0.4211	172.05	19.28
% jump	**91.65**	**66.62**	18.16	42.79

It is evident from figures 7-8 and table 3 that for case 1 LP based features show very less error between testing and training feature vectors compared to filterbank based features whereas LP based features show relatively very large error (% jump of 91 % for LPC and 66.62% for LPCC) for case 2 as compared to filterbank based features (18.16% for MFCC and 42.79% for T-MFCC). Thus, LP based features show *close match* between mimic's imitations for target speaker and his normal voice and *strong discrimination* between his imitations and target speaker's voice. This may be due to the fact that LP based features emphasize *formant structure* dominantly whereas in case of filterbank based features *formant peaks are blunted/distorted due to the averaging process in Mel frequency warping.*

6. Summary and Conclusions

ASR is the use of machines to identify a person's voice. An ASR system for identification of extremely skillful professional mimic in Indian languages is presented to demonstrate the effectiveness of different LP based features over filterbank based features. The major contributions of the paper are as follows

1. Specialties of ASR for mimic recognition in the sense that the results obtained in this paper are exactly contradictory to that of normal ASR where filterbank based features (such as MFCC) perform normally better than that of LP based features [13].
2. A new feature set T-MFCC which proved effective in some earlier ASR studies has been attempted for mimic recognition problem [12].

Acknowledgments

The authors would like to thank the authorities of IIT Kharagpur and EU-India Culture Tech Project for their support to carry out this research work.

7. References

[1] B. S. Atal, "Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.* vol. 55, no.6, pp.1304-1312, 1974.

[2] J. P. Campbell Jr. and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," *Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP'99, no.* 2, pp. 829-832, 1999.

[3] W. M. Campbell, K. T. Assaleh and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Trans. on Speech and Audio Proc., vol.* 10, no. 4, pp.205-212, 2002.

[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech and Signal Proc., vol.* 28, no.4, pp. 357-366, 1980.

[5] G. R. Doddington, "Speaker verification-Final report," *Rome Air Development Center,* Griffiss AFB, NY, Tech. Rep. RADC 74-179, Apr. 1974.

[6] G. D. Hair and T. W. Rekieta, "Mimic resistance of speaker verification using phoneme spectra," *J. Acoust. Soc. Amer.,* vol. 51, p. 131(A), 1972.

[7] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. Int. Conf. on Acoustic, Speech and Signal Process.* vol. 1, pp.381-384, 1990.

[8] R. C. Lummis and A. E. Rosenberg, "Test of an automatic speaker verification method with intensively trained mimics," *J. Acoust. Soc. Amer.,* vol. 51, p.131 (A), 1972.

[9] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.,* vol.46, pp. 1026-1031, 1969.

[10] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition-A feature based approach," *IEEE Signal Proc. Mag,* vol. 13, pp.58-71, 1996.

[11] S. Mitra, Patil Hemant A. and T. K. Basu, "Polynomial classifier techniques for speaker recognition in Indian languages," *National System Conference, NSC'03*, IIT Kharagpur, India, 304-308, 2003.

[12] Hemant A. Patil and T. K. Basu, "The Teager energy based features for identification of identical twins in multilingual environment," N.R. Pal et al. (Eds.): ICONIP 2004, *Lecture Notes in Computer Science, LNCS, Springer-Verlag*, Berlin Heidelberg, Germany, vol. 3316, pp.333-337, 2004.

[13] D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," *IEEE Trans. on Speech and Audio Process., vol.*2, no.4, pp.639-643, 1994.

[14] A. E. Rosenberg, "Automatic Speaker Verification: A review," *Proc. IEEE, vol.* 64, pp. 475-487, 1976.

[15] H.M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Process, vol.*28, pp. 599-601, 1980.