Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis

Songhita Misra
ECE Dept.
NIT Silchar
Assam, India
msonghita@gmail.com

TusharKanti Das
ECE Dept.
NIT Silchar
Assam, India
tusharkdnits@gmail.com

ParthaSaha
ECE Dept.
NIT Silchar
Assam, India
psaha089@gmail.com

UjwalaBaruah
CSE Dept.
NIT Silchar
Assam, India
b.ujwala@gmail.com

Rabul H.Laskar
ECE Dept.
NIT Silchar
Assam, India
rabul18@yahoo.com

***Abstract*—This paper reports some of the observations perceived on uncontrolled environment database for comparison of the Mel Frequency Cepstral Coefficients(MFCC) and Linear Predictive Cepstral Coefficients (LPCC) for development of a robust fixed phrase speaker verification system. MFCC are Cepstral coefficients computed on a warped frequency scale based on known human auditory perception whereas LPCC are Cepstral coefficients that represents the human articulatory system based on linear prediction. This paper compares the accuracy level of both the systems based on MFCC and LPCC, also it compares the systems from an equal error rate point of view (EER). The result suggests that LPCC performs more accurately as compared to that of MFCC by 2.59% for authenticating a speaker. The study also suggests that on basis average time required for giving a decision, MFCC outperforms LPCC significantly by 3.73 sec. Furthermore, the paper includes an analysis for the failure of the system on some of the tests which are not correctly detected as genuine or imposter.**

***Keywords*—MFCC, LPCC, Fixed Phrase Speaker Verification, Time Complexity, Failure Analysis, EER**

I. Introduction

Fixed phrase Speaker Verification (FDSVS) [1] [2] [3] [5] finds its application in many voice biometric based applications. In any vocal conversation, along with the message conveyed, the speech contains certain aspects of the speaker identity, emotion, gender etc. [2] [3]. Speaker authentication finds its application in the speaker identity associated with the speech conveyed. In the present scenario, FPSVS is believed to be one of the emerging person authentication technique mostly because of its remote accessibility and also because of its lesser time and space complexity as compared to that text independent speaker identification [4]. Fixed phrase speaker verification gives a binary decision of accepting or rejecting speaker on basis of the same phrase/text prompted at training and the testing phase [3].

In order to develop and compare the systems, a database of 30speakers each of 12 utterances were collected in an uncontrolled environment. The prompt selected was of duration 2-3 seconds using the standard TIMIT database phoneme rich prompts. The database collected was sampled at 8 KHz using a mono channel.

The baseline system uses an energy based Voice Activity Detection (VAD) [8] [9] technique for removal of the silence and the non-speech/ low-voice regions, completely extracting the speech for speaker specific feature extraction. This paper addresses two of the most popular Cepstral based feature extraction technique MFCC [11], frequency mapped into mel scale then converted to Cepstral domain and LPCC [6], linearly predicted frequency mapping converted to the Cepstral domain. The time series alignment of the speech is done using Dynamic Time Warping (DTW) [13] [14] for mapping one speech sample with another. The decision used for verifying a speaker is cohort based decision which believes that intra-speaker variability should be less as compared to that of the inter-speaker variability.

This paper also addresses the time required by a particular system for verifying a single test utterance and also does a failure analysis of the wrongly rejected true and wrongly accepted imposter speaker.

The rest of the paper is organised as follows. Section II deals with the different phases of the fixed phrase speaker verification system. Section III illustrates the system modelling approach. Time complexity analysis is discussed in Section IV. Section V highlights the experimental observations. Section VI analyses the reasons for failure of the system on some tests. Section VII summarizes and concludes the study along with future scope of work.

II. Fixed Phrase Speaker Verification

A fixed phrase speaker verification system consists of four phase [12]: pre-processing, speaker specific feature extraction, pattern mapping and decision making of the system.

978-1-4799-7075-9/15/$31.00 ©2015 IEEE

A. *Pre-processing*

Pre-processing deals with the correct detection of the speech part of the sample. The speech sample is being extracted using Voice Activity Detection (VAD) [9] algorithm whose outline is as:

- Noise reduction
- Calculation of threshold of energy for distinguishing the speech part on the basis of minimum and maximum energies
- Extracting the above speech region part using a threshold energy, which is the speech part extracted.

B. *Feature Extraction*

This paper addresses two feature extraction techniques, one is based on the human perception, MFCC[12][13] and the other is on human vocal tract, LPCC. The speaker specific feature extraction techniques are discussed as:

a) Mel Frequency Cepstral Coefficients (MFCC):

MFCC is most well-known feature extraction technique in speech recognition developed by Mermelstein [11].In this paper 13 dimensional MFCC is used to define the speaker specific feature. The speech frequencies are mapped into non-linear Mel filter bank and transforming it into the Cepstral domain. The steps involved in calculation of the MFCC [10] are shown as:

- Pre-emphasis the speech sample to boost the high frequency components of speech
- Framing (20 ms with 50% overlap) & Windowing (Hamming window) of the speech sample
- Calculate Fourier Transform
- Map the frequencies to mel scale using:

  $m = 2595 \log_{10}\left(\frac{f}{700} + 1\right)$... (1)

  where f is in Hz and m is in Mel
- Calculate Discrete Cosine Transform
- Amplitude of the resultant spectrum are the MFCC

b) Linear Predictive Cepstral Coefficients (LPCC):

In the context of speaker verification, LPCC [6] are used to capture speaker specific information manifested through vocal tract characteristics of the speaker. In this paper, the 13th order LPCC has been calculated per speech frame of 20 ms using a frame shift of 10ms (20±50%)

The key idea behind LP analysis [7] is that m^{th} speech sample can be predicted by linear combination of its previous *n* samples i.e.

$$S(m) \approx a_1 S(m-1) + a_2 S(m-2) + \cdots + a_n S(m-n) \ldots (2)$$

where $a_1, a_2, a_3, \ldots$are the constants (known as LP Coefficients) over a speech analysis frame. These coefficients are employed to predict the speech samples. Further the error in prediction$\big(e(m)\big)$is calculated by

$$\big(e(m)\big) = S(m) - \hat{S}(m) = S(m) - \sum_{k=1}^{n} a_k\, S(m-k) \ldots (3)$$

where $S(m)$ and $\hat{S}(m)$ are the original and predicted speech signals respectively. Now in order to compute a unique set of predictor coefficients, sum of squared

$$E_m = \sum_p [S_m(m) - \sum_{k=1}^{n} a_k S_m(p-k)]^2 \ldots (4)$$

where p is the number of samples in a particular analysis frame. To solve above equation for LPCC, $E_m$has to be differentiated with respect to each a_k and equating the result to zero as

$$\frac{\partial E_m}{\partial a_k} = 0 \qquad for\ k\ =\ 1, 2, 3 \ldots n \qquad \ldots (5)$$

After evaluating all a_k, we may find Cepstral coefficients using the following recursion:

$$C_0 = \log_e n \ldots (6)$$

$$C_p = a_p + \sum_{k=1}^{p-1} \frac{k}{p} C_k\, a_{p-k} \quad for\ 1 < p < n \ldots (7)$$

$$C_p = \sum_{k=p-n}^{p-1} \frac{k}{p} C_k\, a_{p-k} \qquad for\ p > n \ldots (8)$$

d) Pattern Mapping

Alignment of two speech samples is performed using DTW developed by Dan Ellis [13]. DTW finds an optimal warping path based on the distance matrix of two speech samples, the warping path used as an index modifies the two speech samples to map them together of same dimension.

e) Decision Making

The system developed for the study uses a cohort based decision. The Euclidean distance of the claimant utterance with that of the claimed utterance should be less than the non-claimed utterance for the claimant to be a true speaker.

III. System Modelling

The baseline system developed which serves as the FDSVS consists of two phases training and testing as:

A. *Training Phase*

This phase consists of the training the system to define a speaker's identity. Out of the 12 utterances of a speaker, 3 utterances were used for training of the system for a particular speaker. The three training utterances pass

through VAD [9], feature extraction using MFCC [12] or LPCC [6] and thus three reference templates for testing an individual is created using DTW [13] mapping using 1st, 2nd and 3rd utterance as reference.

B. Testing Phase

This phase of the system consists of testing the identity through true and imposter trials on the speaker definition templates created during the training phase. 9 utterances of each speaker were used true trials giving a total of $9 \times 30 = 270$ true trials. For each speaker, the remaining 9 utterance of other 29 speakers, i.e. $9 \times 29 = 261$utterances are imposter, out of which 9 random imposters were selected for each speaker to give the same number of imposter trials.

For making a decision, the distance of the extracted feature of the test utterance with that of the three templates saved from training session of that claimed speaker is being noted. Also the distance of the test utterance was evaluated with 9 of the random templates from non- claimed speakers. The test utterance was accepted to be a true user if the distance with the claimed speaker from that of the non-claimant speaker.

IV. Time Complexity Analysis

Time complexity is the time taken by a test trial to either get accepted or rejected. It is a very significant parameter to be analysed in order to design a robust real time system. The times taken by each of the true as well as the imposter trails are recorded. The mean of the time gives the average time taken by a true or an imposter trial and the mean of the true average time taken and imposter average time taken gives the overall average time taken for a trial. The overall average time parameter provides us with a parameter to compare the time complexity of the two approach discussed in this study.

V. Experimental Observations

The experimental observations incurred from both the feature extraction techniques, MFCC [12] and LPCC [6], for true speakers are shown in Table I and Table II for imposter speakers. Table III cumulatively shows the total success rate for both true and imposter speakers and parameterize the systems on the basis of Equal Error Rate (EER). Table IV shows the failure analysis.

It has been observed in Table I for true trials, 234 trials were accepted of 270 for MFCC based system for a true success rate of 86.67% whereas 230 trials were accepted for LPCC based system for a true success rate of 85.19%. The result shows are significantly better average time taken per trial in case of MFCC (0.95 sec) than that for LPCC (3.77).

Table I: Genuine Trials for both MFCC and LPCC based systems

	No. of observations	Accepted Trials	Success Rate %	Avg. time(sec)
MFCC	270	234	86.67	0.95
LPCC	270	230	85.19	3.77

In Table II, out 270 imposter trials, 222 were rejected by system based on MFCC for imposter success rate of 82.22% and 240 trials were rejected for LPCC based system providing am improved imposter success rate of 88.89%. The average time taken per trial for MFCC is 0.91 sec and that for LPCC is 5.16 sec.

Table II: Imposter Trials for both MFCC and LPCC based systems

	No. of observations	Rejected Trials	Success Rate %	Avg. time(sec)
MFCC	270	222	82.22	0.91
LPCC	270	240	88.89	5.16

Table III shows the overall / Total Success Rate (TSR) of genuine and imposter speaker, the system based on LPCC gave a TSR of 87.04% which shows a significant improvement for person authentication as compared to that of the system based on MFCC which gave a TSR of 84.45%. In terms of error, Equal Error Rate (EER) of MFCC is 15.55% and that of LPCC is 12.96% which are complementary to that of TSR. The total average time taken per trial is 0.93 sec for MFCC and 4.66 sec for LPCC.

Table III: Results for both MFCC and LPCC systems based on TSR, EER & Total Average Time

	Total Success Rate / Accuracy %	EER %	Total avg. time(sec)
MFCC	84.45	15.55	0.93
LPCC	87.04	12.96	4.66

VI. Failure Analysis

As the database was collected in an uncontrolled and unmonitored environment, this may inflict some of the probable failures in the correct detection of the test utterances. Our analysis is this section is to detect the probable reasons of failure of correct recognition. Out of total 540 genuine trials (270 for MFCC and 270 for LPCC) and 540 imposter trial, 154 of the trials were wrongly detected as imposter and genuine respectively. A detail probable failure analysis of the wrongly recognised trials is shown in Table IV. The perceived reasons of failure of these trials are:

- *Utterance Defect:* This occurs due to the incorrect phrase prompted by the speaker during the recording or when only half of the sentence is been

recorded by the system because the speaker speaks the sentence before the recording is been started.

- *Inaudible:* These are due to inaudibility of the text, prompted by the speaker or also the speaker's energy varies significantly during every next trial.
- *Noise Dominant:* Failure may also occur due to the noise dominance as compared to the speech.
- *Prompt Pace:* Significant change in the pace of the prompt spoken during each trial may be one of the reasons. Also, the limitations of the DTW to map two speech samples correctly, if their duration varies in the range of $0.7 - 1.4$ times, stands as one of the reason of failure.
- *Miscellaneous:* Some reasons for failure includes some reasons such stammering of the speaker, sudden pause of the speaker between the prompts, including external exclamations in to the prompts etc.

Table IV: Failure Analysis for both the systems

	No. of utterances	% of total failed utterances
Utterance defect	37	24.01
Inaudible	27	17.53
Noise dominant	29	18.83
Prompt Pace	28	18.18
Misc.	33	21.43

VII. Summary & Conclusion

The MFCC and LPCC are two of the widely used acoustic features used for speech recognition. This paper compares the relative performance of the two acoustic features for fixed phrase speaker verification performed on uncontrolled and unmonitored microphonic database. It exploits the time complexity of the two systems for both genuine and imposter trials and evaluates their relative performance for implementation on real time systems. It has been observed that LPCC provide a comparatively better performance as compared to MFCC for a robust FDSVS and on basis time taken for each trial, MFCC provide a better real based FDSVS. It may also be concluded that LPCC serve as a better acoustic feature as compared to MFCC for higher accuracy in designing an FDSVS. Through failure-analysis, some of the reasons are observed for failure of correct detection of trails by the systems, which can be avoided for better designing a FDSVS in future. Rigorous analysis can be made on this initial study to analyse the significance of acoustic features with respect time-complexity and failure analysis. A robust real time system can be designed, by erasing the faults in the test trails, and enhancing the database features and parameters to achieve higher success rate.

Acknowledgement

The authors highly acknowledge the Department of Electronics & Information Technology (DeitY), Ministry of Communications & IT, Government of India for the resources provided and also their never ending support and motivation for the research work.

References

[1] H. Matthieu, "*Text-Dependent Speaker Recognition*," Springer Handbook of Speech Processing, pp. 743-762, 2008

[2] "*A Tutorial on Text-Independent Speaker Verification*," EURASIP Journal on Applied Signal Processing, pp. 430–451, 2004

[3] K. Tomi, L. Haizhou, "*An overview of text-independent speaker recognition: From features to supervectors*," Speech Communication, Vol. 52, Issue 1, 2010

[4] S. Shukla, S. R. M. Prasanna and S.Dandapat, "*Speech Recognition under Stress Condition*," in 15thNational Conference on Communications, IIT Guwahati, pp. 299-302, Jan 2009.

[5] M. Campbell William, T. A. Khaled, and C. B. Charles, "*Speaker recognition with polynomial classifiers,*" Speech and Audio Processing, IEEE Trans., pp. 205-212, 2002.

[6] B. S. Atal, "*Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,*" the Journal of the Acoustical Society of America, pp. 1304-1312, 1974.

[7] L. Rabiner, B. H. Juang and B. Yegnanarayana, A Text Book on "*Fundamentals of Speech Recognition*," Pearson, 2009.

[8] Boyd, Ivan, and Daniel K. Freeman. "*Voice activity detection*," U.S. Patent No. 5,276,765, Jan 1994

[9] Kondoz, A. M. "*Voice Activity Detection,*" Digital Speech: Coding for Low Bit Rate Communication Systems, Second Edition, pp. 357-377, 2004

[10] Hasan, MdRashidul, et al. "*Speaker identification using mel frequency cepstral coefficients*," 3rd International Conference on Electrical & Computer Engineering ICECE, Bangladesh, 2004

[11] Davis, Steven, and Paul Mermelstein, "*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,"* Acoustics, Speech and Signal Processing, IEEE Trans, pp. 357-366, 1980.

[12] K.S.R. Murty, B. Yegnanarayana, "*Combining evidence from residual phase and MFCC features for speaker recognition*," IEEE Signal Processing Letters, pp. 52-55, 2006

[13] http://www.ee.columbia.edu/ln/rosa/matlab/dtw/

[14] R.R. Lawrence, E.R. Aaron and E.L. Stephen, "*Considerations in dynamic time warping algorithms for discrete word recognition*," Acoustical Society of America, 2005