Comparative Study on the Performance of Mel-Frequency Cepstral Coefficients and Linear Prediction Cepstral Coefficients under different Speaker's Conditions

Kamil Ismaila Adeniyi
Department of Electrical& Electronic Engineering
University of Ibadan
Ibadan, Nigeria

Oyeyiola Abdulhamid K.
Department of Electrical& Electronic Engineering
University of Ibadan
Ibadan, Nigeria

ABSTRACT

This paper compares Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs) features under three speaker conditions: waking up, being fully awake and being tired, to determine which is better at handling the effect of these variations. A Gaussian Mixture Model (GMM) Classifier was used for both features. Experimental results show an identification rate of 83.3% in the MFCC based system when the speakers were just waking up, while the LPCC based system had a lower identification rate of 75%. Also, when the speakers were either fully awake or tired, the MFCC based system achieved an identification rate of 100%, while the LPCC based system had an Identification rate of 91.7%. In speaker verification, under the first condition (Waking Up), there is a significant difference between the equal error rates (EER), 7.9% for MFCC and 22.0% for LPCC. Also, there is a significant difference between the total success rates (TSR) under this condition. 82.5% for MFCC and 65.0% for LPCC. Overall, MFCC achieved a better total success rate under the three conditions studied.

General Terms

Speaker Recognition, intra-speaker variability, session variability.

Keywords

Mel-frequency cepstral coefficients, linear prediction cepstral coefficients, speaker recognition, speaker's conditions.

1. INTRODUCTION

Speaker Recognition is a multi-disciplinary technology which uses the vocal characteristics of speakers to deduce information about their identities [1]. In speaker recognition, the interest lies in answering the question who said it? Rather than understanding what is being said. The later is the focus of speech recognition technology. It is a branch of biometrics (behavioral biometrics) because a Speaker's voice is unique. As a result, it may be used in Forensics and Access Control Applications. When there is something valuable that needs protection e.g. information in a bank's database, there is a need for gate-keeping or access control. An automatic speaker recognition system can be used for such purpose. When used in gate-keeping applications, a speaker recognition system can be operated in two modes, speaker identification and speaker verification. During speaker identification, an unknown speaker only records his/her voice without supplying a user ID. The system performs a one-to-many comparison through its database and returns the ID of the closest match. In a situation where the test subject is not an "enrolled" Speaker, the returned ID would indicate a system failure. Speaker identification should therefore be performed on a closed-set or in applications where misidentification has a mild consequence. On the other hand, for verification a claimed Speaker Supplies an ID and also records his/her voice. The system performs a one-to-one comparison with the "voice print" of the speaker in its database and makes a decision based on the outcome. It is also called binary detection, because the decision is to either accept or reject the claimed speaker [2-5]. Speaker recognition is very important as a means of verification because unlike many other biometrics, it requires no special infrastructure. It can make use of existing ones such as the telephone network, and it is easily tested remotely over a network [1].

Mel-frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs) are two spectral features widely used in implementing practical speaker recognition systems. They are extracted by mapping the frequency domain of the speech signal to that of either the speech production model or speech perception model, and are related to the spectrum of the log of spectrum of a speech signal. MFCC is inspired by human speech perception model. Results of psycho-acoustic experiments reveal that human perception of pitch is linear up to 1000 Hz and then becomes non-linear (logarithmic) for higher frequencies. MFCCs are computed by warping the frequency domain of the speech signal to the Melody (Mel) scale [1, 6, 7, 8], with the aid of a psycho-acoustically motivated filter bank, followed by logarithmic compression and discrete cosine transform (DCT) [6, 8]. On the other hand, LPCC has its origin in speech production model. In order to reflect the resonance properties of the supralaryngeal vocal tract, speech production is intuitively modeled as an all-pole filter, with complex poles capable of producing the resonant frequencies (Formant frequencies). By exploring the correlation property found in adjacent speech samples, the filter coefficients are estimated using linear prediction. These coefficients are rarely used as features but transformed into more robust and less correlated features [6], one of which are the LPCCs.

An ideal voice feature must have small intra-speaker variability and an inter-speaker variability large enough to prevent overlap between speakers in the decision space. If intra-speaker variability becomes too large, significant overlap will occur, leading to identification errors. Intra-speaker variabilities are unavoidable since it is virtually impossible for a speaker to repeat the same set of phrase(s) exactly the same way on different occasions. This is further complicated by the ever changing channel properties. Some factors responsible for changes due to the speaker him/herself are state of health, mood, aging, alertness or awakeness etc. Changes in a speaker's condition may be substance induced (foods or drugs) or physiological in nature. In this work we studied how three speaker states/conditions which are: waking up, being fully awake and being tired affect the performances of MFCCs and LPCCs when used in speaker recognition systems.

2. TYPES OF SPEAKER RECOGNITION

Speaker recognition systems fall into two categories: text-dependent and text-independent [4, 5, 6, 8].

2.1 Text-dependent Speaker Recognition

If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers or unique [4]. Text is fixed during training and the same is used during testing. Therefore both instances of the utterance can be regarded as an imperfect replica of each other, and be aligned temporally so as to measure their degree of similarity.

2.2 Text-independent Speaker Recognition

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker [4]. Text during training may bear no correlation to text during testing. Therefore any form of alignment between the two may be meaningless or devoid of any useful geometric interpretation. Text-independent speaker recognition uses more of acoustic features than speech dependent features.

3. THE SYSTEM

For this work we developed a text-dependent speaker recognition system based on MFCC and LPCC voice features using prompted digits (0-9). Figure 1, shows the flowchart of the system.

3.1 Speech Database

The database contains 15 speakers (5 males and 10 females), with 6 utterances from each speaker, each of spoken digits 0 to 9. Since the purpose is to study how voice features MFCC and LPCC perform under three speaker conditions, waking up, being fully awake and being tired, a set of utterances were recorded under each of these conditions. The utterances were recorded in a low noise environment. The utterances were recorded in wav format with 16-bit per speech sample and at 8 kHz sampling rate.

3.2 Preprocessing

This refers to all transformations performed on a speech signal before feature extraction. These include speech pre-emphasis, voice activity detection, frame blocking and windowing.

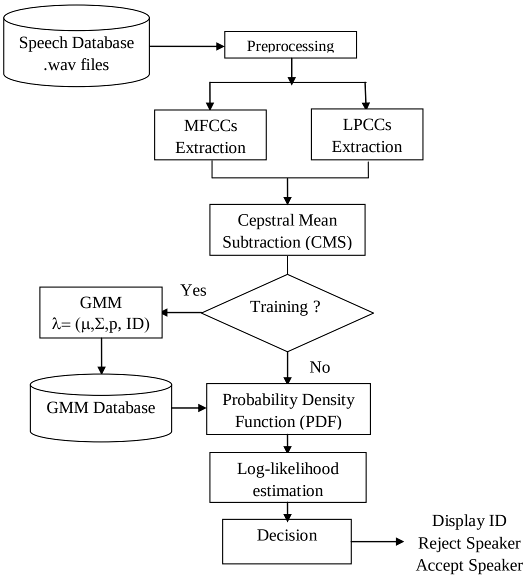

Fig 1: Flowchart of Speaker Recognition System

Table 1. Spoken Digit Database

Speakers	15 (5 males and 10 females)
Sessions/Speaker	6
Type of Speech	Prompted digits(0 to 9)
Microphone	Standard Microphone
Acoustic Environment	Recording room (±55dB)
Sample width	16 bit
Sampling rate	8 kHz
File format	wav

In our system we converted input signals to 20 millisecond frames with 50% overlap and used a Hamming window, before extracting feature vectors from each frame. On the average each utterance is 5 sec. in length, giving an average of 996 frames per speaker per condition.

3.3 Extraction of MFCCs

MFCCs were extracted using a triangular filter-bank with 26 filters and 257 points Discrete Fourier transform. The pre-computed magnitude (frequency response) of our filters $H_m[k]$, is given by [8, 9]:

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{(k - f[m-1])}{(f[m] - f[m-1])}, & f[m-1] \le k \le f[m] \\ \frac{(f[m+1] - k)}{(f[m+1] - f[m])}, & f[m] \le k \le f[m+1] \\ 0, & k > f[m+1] \end{cases} \quad (1)$$

where m is the mth filter and k is the kth point of its DFT. The procedure for obtaining the various values of f can be found in [8].

Given a frame of preprocessed speech $x[n]$, we used the following algorithm [8, 9] in calculating our MFCCs

- Compute the frame DFT

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{j2\pi nk}{N}}, \quad 0 \le k \le N \qquad (2)$$

- Weight the power spectrum $X[k]^2$ by the triangular filters

$$S[m] = \sum_{k=1}^{N} X[k]^2 H_m[k], \quad 0 \le m \le M \qquad (3)$$

- Compute the discrete cosine transform (DCT) of the logarithm of $S[m]$ to form the MFCCs as

$$mfcc[i] = \sum_{m=1}^{M} log[S[m]] cos\left[i\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right], \quad (4)$$

$$i = 1,2, \ldots \ldots L$$

where L is the number of cepstral coefficients.

3.4 Extraction of LPCCs

We extracted our LPCCs using a 10th order linear predictor. The predictor coefficients were transformed to LPCCs using the equations stated below [10]:

$$c_m = -a_m - \sum_{k=1}^{m-1}\left[\left(1 - \frac{k}{m}\right) . a_k . c_{(m-k)}\right], 1 \le m \le p \qquad (5)$$

$$c_m = -\sum_{k=1}^{p}\left[\left(1 - \frac{k}{m}\right) . a_k . c_{(m-k)}\right], \qquad m > p \qquad (6)$$

where p is the order and a_k is the kth coefficient of the linear predictor. c_m is mth cepstral coefficient.

For each frame of speech with N samples, a set of predictor coefficients ($a_k s$) are computed using Linear Predictive Coding (LPC) [11]. The basic idea behind LPC is that the nth speech sample can be approximated as a weighted sum of p past samples.

$$\hat{s}_n = \sum_{k=1}^{p} a_k \, s_{n-k} \qquad (7\,)$$

The difference between the approximated or predicted sample given by (7) and the actual sample is called error signal and it is defined as:

$$e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k \, s_{n-k} \qquad (8)$$

This error signal, or more importantly its energy E must be minimized to obtain the desired predictor coefficients.

$$E = e_n^2 = \left(s_n - \sum_{k=1}^{p} a_k \, s_{n-k}\right)^2 \qquad (9)$$

E is minimized by taking its partial derivatives with respect to a_i as given in the equation below:

$$\frac{\partial E}{\partial a_i} = 2\sum_{n=1}^{N} s_{n-i} \, s_n - 2\sum_{k=1}^{p} a_k \sum_{n=1}^{N} s_{n-i} s_{n-k} = 0 \qquad (10)$$

Rearranging the terms (10) becomes

$$\sum_{k=1}^{p} a_k \sum_{n=1}^{N} s_{n-i} s_{n-k} = \sum_{n=1}^{N} s_n \, s_{n-i} \quad i = 1,..,p \qquad (11)$$

A linear system with p equations is derived from (11) in terms of autocorrelation function as:

$$A_{i,k} \, a_k = b_i \qquad (12)$$

$$A_{i,k} = \sum_{n=1}^{N} s_n s_{n+i-k} \qquad (13)$$

$$b_i = \sum_{n=1}^{N} s_n \, s_{n-i} \qquad (14)$$

Matrix A has two special properties, it is symmetric and the diagonal elements have the same value. It is thus called a toeplitz matrix. An efficient recursive algorithm known as Levinson-Durbin [12] is used to solve the system.

3.5 Cepstral Mean Subtraction

In order to make our feature extraction robust, we used a simple form of feature normalization known as cepstral mean subtraction (CMS) [6]. The mean vector is subtracted from each feature to ensure that two feature sets obtain from different channels will have zero-mean [6].

4. TRAINING

It involves developing a parametric model for each speaker using features extracted from training speech samples. The condition whereby the speakers are awake was taken as their normal condition and the models were trained under this condition.

4.1 Speaker Modeling

For each speaker we trained a Gaussian mixture model (GMM) λ. A GMM is composed of a finite mixture of multivariate Gaussian components [6, 7]. Training involves estimating the parameters $\lambda = \{P_k, ⍰_k, \Sigma_k\}_{k=1}^{K}$ using the extracted training features. P_k is the prior probability (mixing weight) of the kth Gaussian component. $⍰_k$ is the mean vector and Σ_k is the covariance matrix. K is the number of Gaussian components. We used K-mean clustering [7, 13] algorithm to populate our Gaussian components. Our K was chosen to be 2 due to limited number of available training samples.

5. TESTING

5.1 Probability Density Function

During testing, test feature vectors were used to build probability density function(s) using the parameters of the model(s) obtained during training stage. Given a test vector $\mathrm{x} = \mathrm{x}_1, .. \mathrm{x}_\mathrm{T}$ and a modelλ, the probability density function is defined as (7) [6]

$$p(\mathrm{x}|\lambda) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{P_k} N(\mathrm{x}|\mu_\mathrm{k}, \Sigma_\mathrm{k}) \qquad (15)$$

where

$$N(\mathrm{x}|\mu_k, \Sigma_k) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathrm{x} - \mu_k)^T \Sigma_k^{-1} (\mathrm{x} - \mu_k)\right\} \qquad (16)$$

5.2 Average Log-likelihood

It is a measure of the likelihood that an unknown vector X originates from a known model λ and it is calculated from the probability density function as (9) [6, 7]

$$LL_{avg}\,(X,\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log \sum_{k=1}^{K} P_k N(\mathrm{x}|\mu_k, \Sigma_k) \qquad (17)$$

The higher the value of this function, the higher the indication that the unknown vector originates from the model [6]. It is used as a modality for scoring models in speaker identification and verification and forms the basis for decision making.

5.3 Speaker Identification

Speaker identification is a 1:N matching between an unknown vector X and N known models. Average log-likelihood is computed for all N models and the model with the maximum log-likelihood is picked as the target model.

5.4 Speaker Verification

Speaker verification is a 1:1 matching between an unknown vector X and a target model. The average log-likelihood is computed for the model and compared against a threshold. Verification is successful if and only if the value is greater than the threshold.

6. EXPERIMENTAL RESULTS

6.1 Performance Criteria

The performances of the Identification and Verification parts of the systems were measured separately. Identification rate is the criteria used for speaker identification. In speaker verification, there are two types of errors that can occur, false acceptance and false rejection. Therefore, the basic error measures of a verification system are false acceptance rate (FAR) and false rejection rate (FRR). Speaker verification systems are tuned to minimize both of these error rates by finding the "break-even" point where the two are equal and this point is referred to as equal error rate (EER). The Overall performance can be obtained by combining these two errors (FAR and FRR) into total success rate (TSR) [14].

$$FAR = \frac{Number\ of\ accepted\ imposter\ claims}{total\ number\ of\ imposter\ accesses} x100 \qquad (18)$$

$$FRR = \frac{Number\ of\ rejected\ genuine\ claims}{total\ number\ of\ genuine\ claims} x100 \qquad (19)$$

$$TSR = 100\% - \left(\frac{FAR + FRR}{Total\ number\ of\ accesses}\right) x100 \qquad (20)$$

6.2 Results and Discussion

Table 2 shows the results of the Speaker Identification experiments performed. It shows the identification rate under the three conditions. The identification rate in the MFCC based system when the speakers are waking up is 83.3%, while that of the LPCC based system is 75%. The MFCC based system achieved a 100% Identification rate when speakers were either fully awake or tired, while the LPCC based system had a lower Identification rate of 91.7% under the same conditions. Table 3 shows the False Acceptance and False Rejection Rate for the MFCC and the LPCC systems under the three Conditions. The highest values of FAR and FRR were recorded in the LPCC based system, 12% and 30% respectively. Figure 2 and Figure 3 show the plot of Detection Error Tradeoff (DET) curves for MFCC and LPCC respectively. Under the first condition (Waking Up), there is a significant difference between the equal error rates, 7.9% for MFCC and 22.0% for LPCC. This is a difference of 14.1%. Also, there is significant difference of 17.5% between the total success rates under this condition as can be seen in Table 4.

Table 2. Identification Rates

Conditions	Identification rate	
	MFCC	***LPCC***
Waking up	83.3%	75.0%
Awake	100%	91.7%
Tired	100%	91.7%

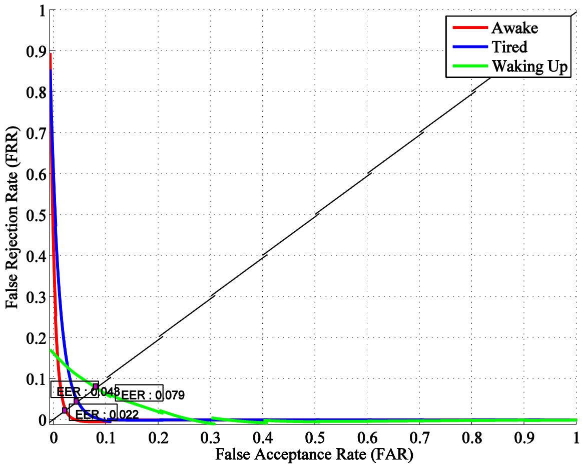

Fig 1: DET curves for the conditions using MFCC

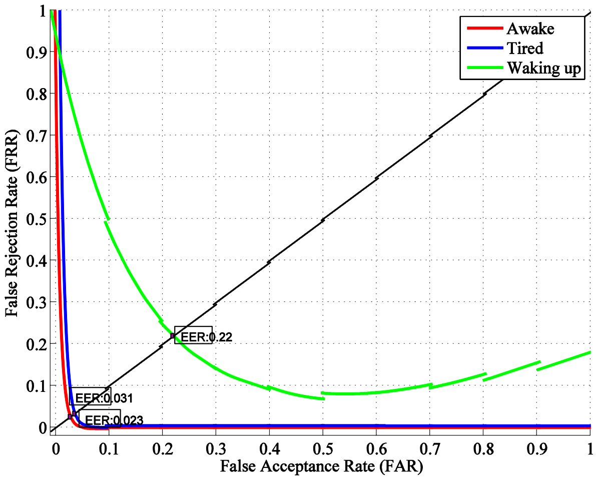

Fig 2: DET curves for the conditions using LPCC

Table 3. Speaker Verification Results

Conditions	MFCC		LPCC	
	FAR	***FRR***	***FAR***	***FRR***
Waking up	1.0%	20.0%	12.0%	30.0%
Awake	3.0%	0.0%	7.0%	0.0%
Tired	5.0%	5.0%	10.0%	10.0%

Table 4. Equal Error Rates and Total Success Rates

Conditions	MFCC		LPCC	
	EER	***TSR***	***EER***	***TSR***
Waking up	7.9%	82.5%	22.0%	65.0%
Awake	2.2%	97.5%	2.3%	94.2%
Tired	4.3%	91.7%	3.1%	83.3%

7. CONCLUSION

This paper has shown that the effect of variations in speaker's conditions can be significant on both MFCC and LPCC features used in speaker recognition systems, leading to significant variations in system performance. The lowest total success rates were recorded when the speakers were waking up, 82.5% for MFCC and 65.0% for LPCC. The highest total success rates were recorded when the speakers were awake, 97.5% for MFCC and 94.2% for LPCC. Thus, allowing a significant range for the total success rate variation, 15% for MFCC and 29.2% for LPCC. Based on the recorded values for false rejection rates, it can be concluded that a speaker's condition does affect his/her recognition by the system. This was most severe when the speakers were just waking up. These values suggest that in a system with a population of 100 speakers, 20 and 30 speakers are likely to be falsely rejected due to the fact that they are just waking up using MFCC and LPCC respectively. But, based on the recorded values for TSR and EER, it can be concluded that MFCC will always produce a better performance than LPCC under the three conditions studied.

8. REFERENCES

[1] Beigi, H. 2011, Fundamentals of Speaker Recognition, New York: Springer Science and Business Media, Inc.

[2] Rashid, R.A., Mahalin, N.H., Sarijari, M.A., Abdul Aziz, A.A., 2008."Security system using biometric technology: Design and implementation of Voice Recognition System (VRS)," Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on, vol., no., pp.898,902, 13-15.

[3] Currie, D., 2003. Shedding some light on Voice Authentication, SANS Institute, GSEC- V1.4b.

[4] http://en.wikipedia.org/wiki/Speaker_recognition

[5] http://www.nou.edu.ng/NOUN_OCL/pdf/SMS/MBF%20841%20EMERGING%20TECHNOLOGIES%20IN%20INFORMATION%20TECHNOLOGY.pdf

[6] Kinnunen, T. and Li, H. 2009. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors, pp. 1-14.

[7] Reynolds, D.A.; Rose, R.C.; , 1995. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech, Audio Processing. vol. 3, no. 1, pp. 72-83.

[8] Abdallah, S. J., Osman, I. M. and Mustafa, M. E. 2012. "Text-Independent Speaker Identification Using Hidden Markov Model", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 6, pp 203-208.

[9] Ibiyemi T.S., Aliyu S.A., Akintola A.G. 2012. "Face and Speech Recognition Fusion in Personal Identification", International Journal of Computer Applications, vol.47, no. 23, pp 36-41.

[10] Ouzounov, A. 2010. "Cepstral Features and Text-Dependent Speaker Identification – A Comparative Study", Cybernetics and Information Technologies, Vol. 10, No 1, pp 3-12.

[11] https://tv.unsw.edu.au/files//unswPDF/Chapter3.pdf

[12] www.ee.ucla.edu/~ingrid/ee213a/speech/vlad_present.pdf

[13] Teknomo, Kardi. K-Means Clustering Tutorials. http:\\people.revoledu.com\kardi\ tutorial\kMean\

[14] Ilyas, M. Z., Samad, S. A., Hussain, A. and Ishak, K. A. 2007. "Speaker Verification using Vector Quantization and Hidden Markov Model", Proc. The 5th Student Conference on Research and Development, Malaysia.