A REVIEW ON PERFORMANCE OF VOICE FEATURE EXTRACTION TECHNIQUES

D.Prabakaran[1], Dr.R.Shyamala[2]
[1]Associate Professor, IFET College of Engineering, Villupuram, Tamil Nadu, India- 605108. erprabakarand@gmail.com
[2]Assistant Professor, University College of Engineering- Tindivanam, Tamil Nadu, India, 605001. vasuchaaru@gmail.com

Abstract— **In the digital era, the computing applications are to be secured from anonymous attacks by strengthening the authentication credentials. Numerous methodologies and algorithms have been proposed implementing human biometric as unique identity and one such identity is human voice print. The human voice print is a unique characteristic of the individual and has a wide variety of techniques in representing and extracting the features from the digital speech signals. The voice recognition techniques were executed on different platforms and exploit different mathematical tools in voice feature extraction, leading to dissimilarity in performance and results. In this paper, we investigate, analyze and present a review on performance of numerous voice recognition techniques.**

Keywords—Biometric identity, Voice print, Feature extraction techniques, Performance analysis.

I. INTRODUCTION

Biometric identity is one of the thrust areas to be concentrated on improving the security of computing applications and the sensitive data stored in it. The biometric is defined [59] as the biological and behavioral characteristic of an individual, the feature extraction can be employed for the purpose of a biometric authentication process. The biometric identity places a strong foundation for personal identification, verification and authentication of individual in granting access to any secured computing applications. The computing applications and transactions were highly vulnerable due to increase of fraudulent access and security policy violations, in turn, stimulates the organizations and researchers to create a strong secure network infrastructure by strengthening the authentication credentials. Strengthening of credentials paves to development of secured infrastructures of organizations, online banking and financial transactions which were treated as nerves of the global economy. The authentication credentials involves login ID and password created by an individual and is fortified involving individual's biometric identity. Integrating biometric identity with a password and encrypted key creates a strong, secure framework to perform online transactions. The key metric of biometric is cannot be copied and is unique to every individual. The biometric entities were classified under two categories, namely physiological biometric and behavioral biometric. The physiological biometrics is the unique statistical characteristics of any individual. The varieties of biometric entities [60] identified under physiological biometric are:

- Fingerprint
- Face recognition
- DNA matching
- Palm print recognition
- Hand geometry recognition
- Eyes-iris & retina recognition

The biological biometric is based on inner variants like individual mentality and health condition and is also named as psychological biometric identity. This biological or psychological biometric were not unique all time and is variable with respect to time. It is essential to design a database that to store the identity of an individual at different mentalities. The classifications of psychological biometric entities are:

- Voice/ Speaker recognition
- Typing rhythm
- Gait
- Signature recognition

This biometrics obtains common characteristics of uniqueness, universality, collectability, acceptability and performance measurability.

TABLE I. COMPARISON OF BIOMETRIC TYPES

Biometric Type	Feature extraction Algorithm	Computation latency	Device compatibility	Uniqueness	Duplication Endurance
Fingerprint	Simple	Medium	Yes	High	High
Face	Complex	High	No	High	Moderate
DNA Matching	Complex	High	No	High	High
Palm print	Simple	Medium	No	Moderate	Moderate
Hand Geometry	Simple	High	No	Low	Low
Eyes- Iris/ Retina	Complex	High	No	High	High
Voice/Speaker	Complex	Medium	Yes	High	High
Typing Rhythm	Simple	Low	Yes	No	Low
Gait	Complex	High	No	No	Low
Signature	Simple	Low	Yes	High	Low

The table 1 portrays the key parameters of any biometric that ease the selection to employ in securing the computer based sensitive informative system. It is clearly evident from the table 1 that, the biometric type fingerprint and voice recognition methods pocess device compatibility and the feature extraction algorithm are simpler that produces least computational latency. From this analysis, we pick fingerprint and voice print as an appropriate biometric for securing the information system with device independence. Furthermore, the voice recognition leads fingerprint in an authentication, race with special ability of remote access which is absent in fingerprint recognition system. Here in this paper, the voice print feature extraction techniques were analyzed and discussed about the salient qualities and pitfalls in the upcoming sections.

II. VOICE FEATURE EXTRACTION TECHNIQUES

Human voice is time variant digital signal when treated with mathematical tools of digital signal processing, employed in

978-1-5386-9371-1/19/$31.00 ©2019 IEEE

multiple authentication applications like language identification, voice analysis and synthesis, speech coding, voice enhancement applications and speaker recognition applications. Several methods have been designed to extract the identical features [61] of human voice and are listed as follows.

- Linear Predictive Coefficients (LPC)
- Linear Predictive Cepstral Coefficients (LPCC)
- Perceptual Linear Predictive Coefficients (PLP)
- Mel Frequency Cepstral Coefficients (MFCC)
- Relative Spectra filtering of log domain coefficients (RASTA)

The feature extraction is considered as a vibrant process involves intricate procedures in speech analysis. The core cepstral coefficients is a time invariant factor ease the speech recognition and analysis methods. Various methods of $|DFT(x(n))|^2 \to P(n)$ feature extraction have been discussed in the following section.

A. Linear Predictive Coefficients (LPC)

Linear Predictive Coefficients is an extensively used low bit rate coder, determines the power spectrum of digital speech signal. LPC technique [52] highly involved in formant analysis. In this technique, as an alternative of converting the entire digital speech signal, the difference of samples generated to predict the upcoming voice samples and is mathematically processed to extract the feature of digital speech signal. The digital speech signal can be rehabilitated by combining the "n" number of predictive samples. The prediction model is illustrated as follows.

Let x(n) and x(n-1) be the present and previous voice sample, the future sample can be predicted by

$$\hat{x}(n) = \sum_{i=1}^{n=\infty} p_i x(n-i) \pm e(n)$$

Where, p_i is the prediction factor. The prediction error e(n) can be computed by

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

The prediction errors were cleared in successive samples to obtain accurate predicted samples so as to avoid the noisy samples in voice print features. The processes involved in the LPC feature extraction are pre emphasis of voice frames, power estimation.

B. Linear Predictive Cepstral Coefficients (LPCC)

The Linear Predictive Cepstral Coefficients (LPCC) technique [53] is an improvised version of LPC to overcome the channel effects by executing Cepstral Mean Subtraction (CMS). The Cepstral is a sequence of power spectral density extracted from periodogram, is utilized in pitch tracking. The cepstrum is obtained by performing Inverse Fourier Transform for the obtained power spectrum of voice signals and may classify as real cepstrum, complex cepstrum, phase cepstrum and power cepstrum where the power cepstrum is employed in speech analysis.

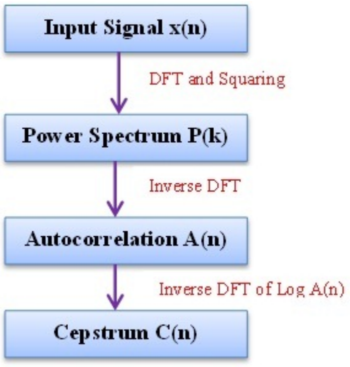

Figure 1: LPCC Computation process

From the Figure 1, let x(n) be the voice signal characterized in time domain on treating mathematically with the Discrete Fourier Transform (DFT) and squaring yields power spectrum P(k)as follows.

$$P(k) = |DFT(x(n))|^2$$

The inverse DFT of P (n) didn't produce x (n) but generates the Autocorrelation A(n) for the time domain speech signal x(n)

$$A(n) = DFT^{-1}(P(k))$$

The cepstrum C (n) is acquired by the logarithmic compressing of Power spectrum and treating with Inverse DFT as follows.

$$C(n) = DFT^{-1}(\log(P(k))$$

The cepstrum computed is a result of logarithmic of power spectrum {log(P(k)}, instead of obtaining from standard power spectrum {P(k)}of speech signal x(n).

C. Perceptual Linear Predictive Coefficients (PLP)

The Perceptual Linear Predictive (PLP) Coefficients is an improvised model of LPCC for noise reduction and to reduce mismatches of training and test voice samples.

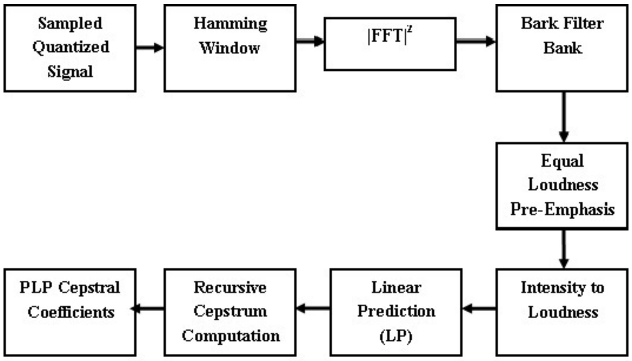

Figure 2: PLC Cepstral computation process

The steps implemented in PLP are illustrated in the Figure 2 that involves (i) Computation of the power spectrum. (ii) Frequency warping. (iii) Convolution of warped spectrum with power spectrum for critical bank integration. (iv)

Spectrum smoothening. (v) Pre-emphasis (vi) Linear prediction and Cepstral extraction.

The integration of frequency warping, smoothing and sampling yields the bark frequency bank. The quantized sampled signal is weighed employing Hamming window of window size 20ms which involves a 256 point Fast Fourier Transform (FFT). The FFT samples 200 samples of speech signal into 56 zero valued samples.

$$P(\omega) = \mathrm{Re}[S(\omega)]^2 + \mathrm{Im}[S(\omega)]^2$$

The power spectrum P(ω) is warped for barking frequency Ω by

$$\Omega(\omega) = 6\ln\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\}$$

The convolution of warped power spectrum and power spectrum is performed and the critical band is obtained. The linear prediction Cepstral coefficients were better on in contrast to linear prediction coefficients as it is sensitive to numerical errors.

D. *Mel Frequency Cepstral Coefficients (MFCC)*

Mel Frequency Cepstral Coefficients (MFCC) oriented on linear cosine transform of the log power spectrum. The frequency bands in MFCC [54] are equally spaced on Mel scale was the notified feature of MFCC in contrast with Cepstrum coefficient method.

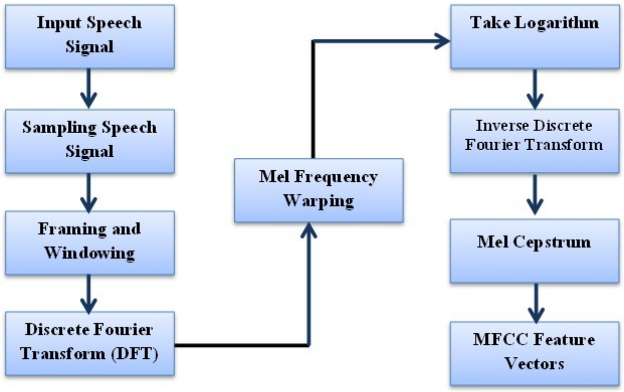

Figure 3: Steps involved in MFCC Extraction

The Figure 3 depicts the steps involved in extraction of MFCC from the input speech signal. The speech signal is sampled and framed for a limit of 20-40ms. The periodogram estimate of the power spectrum is determined for each frame and Mel filter bank [54] is applied to the power spectrum. The logarithmic result of all filter bank energies was treated with mathematical tools of Discrete Cosine Transform (DCT) and 2-13 out of 26 DCT samples or coefficients were stored while the residual were discarded.

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \quad ,1 \le k \le K$$

The frequency to Mel scale adaptation and vice versa is performed by

$$M(f) = 1125\ln(1 + f/700)$$
$$M^{-1}(m) = 700(\exp(m/1125) - 1)$$

The MFCC is a multiplatform oriented technique that might be operated in MATLAB and python. The performance comparison yields better in MATLAB in contrast with python programming.

E. *Relative Spectra filtering of log domain coefficients (RASTA)*

The RelAtive SpecTrA filtering of log domain coefficients (RASTA) method overcomes the shortcoming the suppression of zero frequency and slow varying components in speech signal when feature extraction is done by MFCC method. The RASTA filter [62] steps ahead of MFCC by dropping the impact of noise in speech signals and provides high rate of robustness.

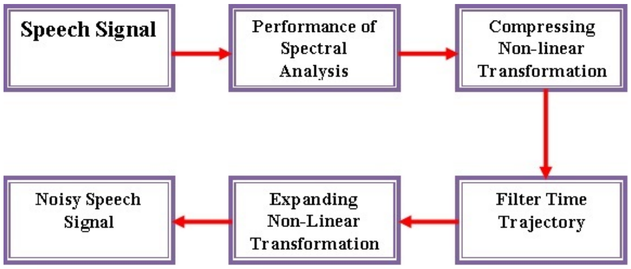
Speech Signal
Performance of Spectral Analysis
Compressing Non-linear Transformation
Filter Time Trajectory
Expanding Non-Linear Transformation
Noisy Speech Signal

Figure 4: RASTA filtering process

The Figure 4 portrays the steps performed in RASTA filtering of speech signal. The analysis library provides power to compensate the linear channel distortion results in passing each coefficient during filtering process. The estimate of clean speech is performed by choosing frequency "i" and frame index "k".

$$S_i(k) = \sum_{j}^{M} w_i(j)Y_i(k-j)$$

The RASTA method yields better performance ratio by removing the slow and fast varying variations. Whereas this method, grounds minor deprivation in performance to obtain clear information. The steps covered in RASTA are

- Computing the critical band algorithm.
- Taking Logarithm for critical band.
- Determination of temporal derivative of the log critical band spectrum
- Reintegration of log critical band temporal derivative.
- Addition of equal loudness curve.
- Performing inverse Logarithmic process.

The RASTA holds equally good merits with MFCC whereas the previous is better for speech signal recorded in noisy environment. The MFCC filter fails to record a good performance in filtering the speech signal recorded in noisy environment.

III. Feature Extraction Performance- A Review

The feature extraction from the speech signal plays vital role in multiple authentication applications. Many researchers had a greater contribution in deriving better results and performance in feature extraction from the speech signal. In this section, the significant research contribution of various dignified researchers has been highlighted to recognize the feature extraction techniques in a better way.

Zhen Tao Liu and et al. (2018) has proposed a model for speech recognition [1] based on improved brain emotion learning model. This model resolves underprivileged adaptation and performance problems utilizing MFCC and first order derivative algorithms. In addition, speaker's emotions were recognized employing CASIA, Surrey Audio-Visual Expressed Emotion (SAVEE) [63] emotion corpus and FAU Aibo data sets. The accuracy of this system is arrived upto 90.28% in an average.

Surekha Rathod and et al. (2015) proposed a security model based on speech recognition using MFCC method [2] with MATLAB approach. In this paper, the author executed the MFCC technique in MATLAB platform to provide security based on speech recognition.

Hynek Hermansky and et al. (1992) proposed and systemized a model based on RASTA – PLP [3] Speech recognition technique. This technique proves better performance when compared to LPC and PLP. The RASTA is highly immune to noise and is integrated with PLP to obtain better performance against noise in speech signal.

Hemanta Kumar Palo and et al. (2017) designed a wavelet based feature combination [4] to recognize the authenticated speaker's emotion during the recording of the input speech signal. The author executes analysis and achieves an accuracy of 90.55% on various databases, namely Berlin Database of Emotional Speech (EMO-DB) [64], SAVEE and self database.

Maged M.M.Fahmy (2010) designs and verified the performance of a model for palm print identification [5] using MFCC algorithm. This method was proven to be robust against noise in the input speech signal. This system fabricates good signal to noise ratio (SNR) and accuracy of 99.56%. In this system, the MFCC algorithm is integrated with Discrete Wavelet Transform (DWT) [65] for better performance in the feature extraction process.

Zulfiqar Ali and et al. (2016) proposed a model for detecting voice pathology [6] by utilizing Estimation of auditory spectrum and Cepstral coefficients based in All pole method. This system yields an average of 99.56% of accuracy during the feature extraction progression from the speech signal.

Dhanalakshmi.P and et al. (2011) proposed pattern classification model for indexing and classification of speech signal. The proposed model [7] utilizes Gaussian Mixture Model (GMM) [55] for extraction of feature vectors while the LPCC and k means clustering is for indexing audio signals. This model yields of about 93% accuracy in feature classification.

Shashidhar G. Koolagudi and et al. (2012) performs an experimentation process of determining the accuracy percentage of feature extraction in speech signal for multiple types of Indian languages. This system proves its performance [8] for about 88% covering fifteen Indian languages and is treated using MFCC algorithm. The notable feature of this algorithm is that it varies in sample count based on the language chosen by the speaker.

Claude Turner and et al. (2015) designed a wavelet based MFCC [9] feature extraction model for authenticating the privileged user and to provide a high level of security. The author involved Wavelet Packet Transform (WPT) is an extended version of Discrete Wavelet Transform (DWT) to harvest a result in lower order of Gaussian Mixture Model (GMM) [55]. The resulting performance lacks consistency in maintaining the best level of accuracy, but holds good error rate of 1/6.

Manal Abdel Wahed (2014) proposed a pathological voice feature identification model by computer aided recognition [10] method. In this model, the voice pathology recognition is performed on both time and frequency domain spectrum. This model has a common platform of MATLAB to get executed and to obtain better result of a maximum of 92% on employing MLP classifiers.

Guofu ZHai and et al. (2014) designed a pattern recognition approach [11] employing the modified MFCC technique and Hidden Markov Model (HMM) [56] data sets to categorize features from the speech signal. In this system, the modified MFCC which provides better accuracy in placing the frequency bands in Mel bank.

Jothilakshmi.S and et al. (2009) proposed an unsupervised method to improve the automatic speaker recognition [12] through Residual Phase (RP) and computing MFCC coefficients. This algorithm relies on Support Vector Machine (SVM) to generate an accuracy of 85.97% and high robust authentication procedure.

Xiao- Chen Yuan and et al. (2015) designed an audio watermarking scheme [13] centered on MFCC feature extraction technique. The dual tree complex wavelet transforms yield robust feature extraction from speech signal and produces tempo invariant pitch shifting of about 70% to resist any kind of audio signal attacks.

Eduardo Pavez and et al. (2012) proposed a Automatic Speech Recognition (ASR) model [14] based on Wavelet packet Cepstral coefficients (WPCC) to exploit the property of iconic coverage of time frequency property. On pinpointing the accuracy of proposed systems, relying on the number of bands the accuracy ranges from 61.19% to 63.37%.

Lalitha S and et al. (2015) designed a system for detection of emotions [15] of a speaker in Automatic Speaker Recognition (ASR) to provide a high robust security system. This system involves the technique of MFCC and Cepstrum features to identify emotions and match with the original feature of speech signal. This system produces an accuracy of 85.7% with an ability of identifying 7 various emotions of the authenticated speaker.

Xiaolan Zhao and et al. (2011) proposed a mechanism for voice pronunciation [16] with Gaussian Mixture Model (GMM) [66] and Entropy Coefficient for Decomposition (ECD) for de-noising effect. This method exhibits different resolution's subspace and renders better accuracy.

Shabnam Gholamdokht Firooz and et al. (2016) proposed nonlinear dynamical features [17] for the developed version of automatic speech recognition and feature extraction by evaluating the Recurrence Plot (RP) of the input speech signal. The author involves MFCC technique for feature extraction, before which executes Reconstructed Phase Space (RPS) for extraction of useful features. This system reduced the noisy content up to 3.62% in the original speech signal.

Johan de Veth and et al. (2003) employs phase corrected RASTA filtering technique [18] for the recognition and feature extraction from the continuous time speech signal. The corrected RASTA filtering highly robust towards the additive noise in speech signal.

P.L.Emiliani and et al. (1983) proposed an excitation function for Linear Predictive Coefficients (LPC) [19] for extraction of voice features. The proposed model is highly robust and improves the quality of synthesizing in incoming speech signal by reducing the buzz.

Tarek Mellahi and et al. (2015) proposed a speech enhancement model implementing LPC in Kalman filtering [20] method. This model is highly resistive towards colored noise and white Gaussian noise. In addition, the proposed model provides enhanced speech recognition to improve the accuracy of feature extraction.

Xuewen Luo and et al. (2009) designed an auditory model for speech recognition [21] and feature extraction process. The background noise is a vital factor that depreciates the accuracy of feature extraction. The author employs MFCC technique for feature extraction and AURORA2 database [73] for to obtain a high rate of accuracy, effectively improving the recognition rate in the noisy environment of speech acquisition.

Jose Novoa and et al. (2018) designed an Automatic Speech Recognition (ASR) system [22] employing Deep Neural Network (DNN) integrated with Hidden Markov Model (HMM). This model procures two different sources of acquiring speech signal and produces a word error rate of about 26% to 38% in contrast with the other speech recognition techniques.

Rama Hasan and et al. (2017) suggested ways to improve the performance of speech recognition techniques [23] through the integration of the notable features of feature extraction methods, namely MFCC, LPCC and PLP in Hidden Markov Model (HMM). The integrations provides escalation in the Speech recognition rate to 49.1%, which is already recorded as 44.80% in MFCC technique.

Anand R Mehta (2018) designed an authentication system exclusive for disabled user employing Optimization technique [24] is feature extraction from speech signal. The author utilizes Artificial Neural Network (ANN) and Genetic Algorithm (GA) in extracting unique features. The author executed the model in MATLAB 2016a simulator and the recognition rate of 97% is achieved.

Hynek Hermansky (1989) proposed a new technique named Perceptual Linear Predictive (PLP) [25] for analyzing the speech signal and for feature extraction. This technique has been proved to be speaker independent in Automatic Speech Recognition (ASR) and has implemented 5th order PLP which is proved to be consistent to frequency changes in human sensitivity.

Anthony Larcher and et al. (2014) designed RSR2015 data base [26] [74] to introduce text dependent speaker verification. The author tested the system with HiLAM system to produce state of the art performance, whereas the HiLAM depends on GMM and HMM data sets.

Mohsen Sadeghi and et al. (2017) designed an algorithm for speaker recognition based on Optimal MFCC feature extraction [27] by Differential Evolution algorithm. This system relies on differential Evolution Algorithm (EA) and Probabilistic Neural Network (PNN) to acquire an optimum number of Mel frequency coefficients. The system's recognition accuracy is of 87.5% and capitulates the needy 13MFCC frames for speaker recognition.

Diksha Sharma and et al. (2015) elucidate the effects of DC coefficient on mMFCC and mIMFCC [28] to obtain a robust automatic speaker recognition system. The author exploit YOHO [67] and POLYCOST [68] data set to obtain clean speech signal. This system proves that mMFCC is advanced in contrast with MFCC by eliminating the DC (noise) content in speech signal which deteriorates the performance of MFCC algorithm.

Ning Wang and et al. (2016) explores the effect of G.723.1 in automatic speaker recognition system. The author designs and implements Power Normalized Cepstral Coefficient (PNCC) [29] to improve the performance of the feature extraction as a part of automatic speaker recognition system. The system improvises the I vector in speaker recognition and on measuring the performance results in improving to 72% employing multi stream features.

Sandeep Rathor and et al. (2017) proposed a text independent speaker recognition model [30] orients on Wavelet Cepstral Coefficient and Butterworth filter. The proposed model's performance matches up to MFCC with a performance of 72% in noisy environment. The notable attributes of this system is that of text independent.

Wei-Hua Cao and et al. (2017) derived a random forest feature selection algorithm [31], a speaker independent model for speech emotion recognition. The algorithm concentrates to improve the performance on effective acoustic feature extraction. The performance of 78.6% is accomplished on utilization of Support Vector Machine (SVM) classifier [69].

Atik Charisma and et al. (2017) implements MFCC Vector Quantization (MFCC-VQ) [32] in the speaker recognition model. The Sum Square Error [70] matches unidentified speakers in the data base and achieves a performance of 83.3%. The threshold value of Sum Square

Error stays behind in achieving this performance and to recognize the unidentified speakers.

Anggun Winursito and et al. (2018) introduces a combined features of MFCC with Principle Component Analysis (PCA) [33] to overcome the pitfalls in achieving the accuracy in MFCC technique. The author has tested this model in Indonesian Language speech recognition and can succeed by creating an improvement in performance from 86.43% to 89.29% and deterioration of dimension of data from 26 to 10features.

Elvira Sukma Wahyuni (2017) proposed an Arabic language recognition model implementing MFCC feature extraction technique integrated with Artificial Neural Network (ANN) [34] classifier to improve the accuracy to 92.42% in an average.

K.N.R.K Raju Alluri and et al. (2016) analysed the system feature in recognizing the speaker emotions [35] during the Automatic Speaker Recognition process. The author implements Linear Prediction Residual Cepstral Coefficients (LPRCC) which is the integration result of MFCC and LPCC techniques. The author tests the model based on 3 emotions of speaker and achieved 16% of accuracy in excess when weigh against with conventional MFCC and LPCC techniques implemented identically.

Ashwini Rajasekar and et al. (2018) evaluates the performance of MFCC algorithm [36] in Automatic Emotion Recognition to implement in security based applications or Human Machine Interface (HMI). Targeting to reduce the error, the author utilized SVM classifier in this model for border maximization.

Hong Yu and et al. (2012) developed an automatic speech recognition model using DNN filter bank Cepstral coefficients [37] to detect spoofing. The author developed a new filter bank, Deep Neural Network-Filter Bank Cepstral Coefficients (DNN-FBCC) and Gaussian Markov Model-Maximum Likelihood classifier in detecting speech attacks.

Dipjyoti Paul and et al. (2016) proposed a feature extraction technique that positions on MFCC algorithm and Gaussian Markov Model (GMM) classifier [38] for the detection of synthetic speech.The performance graph evident that the average Equal Error Rate (EER) is 0.00% and the overall accuracy hikes by 7.12%.

Ahmed Kamil Hasan Al- Ali and et al. (2017) designed a model by the amalgamation of MFCC feature extraction [39] with Discrete Wavelet Transform (DWT) in a noisy environment for an enhanced verification of Forensic speaker. The performance is evaluated via Australian Forensic Voice Comparison (AFVC) [72] resulting in the diminishing of average Equal Error Rate from 21.33% to 13.28%.

Josue Fredes and et al. (2017) introduced modifications in locally normalized Filter banks applicable to Deep Neural Network (DNN) for a rigid automatic speech recognition [40]. The author adjoins Locally Normalized Cepstral Coefficients followed by the reduction of average Equal Error Rate from 11.4% to 9.4%.

Muhammad Amirul Azzim Zulkifly and et al. (2017) proposed a speech signal analysis model [41] employing RASTA-PLP integrated with Singular Value Decomposition (SVD) technique for noise decomposition in speech signal.

Detlef Hardt and et al. (1997) proposed a text dependent speaker recognition [42] model involving spectral subtraction and RASTA filtering technique. The author presents a post analysis result with Equal Error Rate (EER) [71] to 1.41% as the model concentrates on non-stationary noise in speech signal.

Ozlem Kalinli and et al. (2007) proposed a robust automatic speech recognition model [43] with early auditory processing. The author employs MFCC combined with RASTA filtering technique for feature extraction to depreciate the Equal Error Rate from 40% to 18% under noisy environment recording of speech signal.

Jia-Lin Shen (1997) designed a robust speech recognition model [44] using Discriminative Temporal Feature Extraction technique with MFCC and RASTA filtering model. The author intended to improve the recognition rate and achieved it to 84.98% with noisy content reduced to less than 7%.

Qi Li and et al. (2011) derived an auditory based feature extraction algorithm Cochlear Filter Cepstral Coefficients (CFCC) [45] which fits well in unmatched conditions. The model is tested to achieve 96% of Signal to Noise ratio.

Yu-Min Zeng and et al. (2006) proposed a gender classification model [46] based on Gaussian Markov Model (GMM) with RASTA-PLP featured parameters. The model proves the performance to 98% with the adaptability to multi language speaking gender detection.

Rekha Nair and et al. (2014) designed a speaker verification model [47] employing LPCC with Dynamic Time Warping (DTW) and the accuracy is verified to 97% during the testing phase.

George Frewat and et al. (2016) developed an Android application [48] for voice recognition using Multi speaker feature. The author involves MFCC technique for speech feature extraction to achieve the recognition rate to 90.3% so as to employ the model in security applications.

Hemant A.Patil and et al. (2009) designed a novel method for recognizing the speaker [49] accepting the “hum” sound from the authorized speaker. The author involves MFCC integrating LPCC algorithm with 2nd and 3rd order polynomial classifier [58] to achieve high recognition rate.

Drishya Vsudev and et al. (2014) proposed a text independent speaker identification model [50] based on Bessel function (FBCC) and Gaussian Mixture Model to obtain an recognition rate of 98%.

El Bachir Tazi and et al. (2017) proposed a hybrid front en speaker identification model [51] employing RASTA-PLP and MFCC technique to achieve an improvement in performance by 3.38% to reach 60.8% involving Gaussian Markov Model (GMM) classifier.

TABLE II. COMPARISON OF PERFORMANCE OF VOICE FEATURE EXTRACTION TECHNIQUES IN SECURITY APPLICATIONS

Ref	Feature Extraction Technique	Data set / Classifiers	Accuracy / Performance
[1]	Mel Frequency Cepstral Coefficient (MFCC)	CASIA, SAVEE emotion corpus	Accuracy: 90.28%
[2]	Mel Frequency Cepstral Coefficient (MFCC)	Self data set	-
[3]	RASTA-PLP	Self data set	-
[4]	MFCC & LPCC	SAVEE data base	Accuracy: 90.55%
[5]	MFCC with DWT	Self data set	-
[6]	PLP	Self data set	Accuracy: 99.56%
[7]	LPCC	Gaussian Mixture Model (GMM)	Accuracy: 93%
[8]	MFCC	Self data set	Accuracy: 88%
[9]	MFCC with DWT and WPT	Gaussian Mixture Model (GMM)	Error rate: 1/6
[10]	Computer Aided Recognition	MLP Classifiers	Accuracy: 92%
[11]	MFCC	Hidden Markov Model (HMM)	-
[12]	MFCC	Support Vector Machine (SVM)	Accuracy: 85.97%
[13]	MFCC with Wavelet transformation	Self data set	Accuracy: 70%
[14]	Wavelet Packet Cepstral Coefficient (WPCC)	Self data set	Accuracy: 63.37%
[15]	MFCC	Self data set	Accuracy: 85.7%
[16]	Wavelet transformation	Gaussian Mixture Model (GMM)	-
[17]	MFCC with Recurrence Plot	Self data set	Error: 3.62%
[18]	RASTA	Self data set	Accuracy: 95%
[19]	Linear Predictive Coefficient (LPC)	Self data set	Variation: ±10%
[20]	LPC with Kalman filtering	Self data set	-
[21]	MFCC	AURORA2 data set	Recognition rate: +10%
[22]	Deep Neural Network (DNN)	Hidden Markov Model (HMM)	Error rate: 26%
[23]	MFCC	Hidden Markov Model (HMM)	Recognition rate: 49.1%
[24]	Artificial Neural Network with GA	Self data set	Recognition rate: 97%
[25]	PLP (5th Order)	Self data set	-
[26]	HiLAM system	RSR2015 data base	-
[27]	MFCC with Evolution Algorithm (EA)	Self data set	Accuracy: 87.5%
[28]	mMFCC and mIMFCC	YOHO and POLYCOST dataset	-
[29]	Power Normalized Cepstral Coefficient	Self data set	Accuracy: 72%
[30]	MFCC with Wavelet Cepstral Coefficient	Self data set	Accuracy: 72%
[31]	Random forest feature extraction algorithm	Support Vector Machine classifier	Accuracy: 78.6%
[32]	MFCC Vector Quantization (MFCC-VQ)	Self data set	Accuracy: 83.3%
[33]	MFCC with Principle Component Analysis	Self data set	Accuracy: 89.29%
[34]	MFCC with Artificial Neural Network	Self data set	Accuracy: 92.42%
[35]	LPRCC	Self data set	Accuracy: +16%
[36]	MFCC	Support Vector Machine classifier	-
[37]	DNN- FBCC	Gaussian Mixture Model (GMM)	-
[38]	MFCC	Gaussian Mixture Model (GMM)	Accuracy: +7.12%
[39]	MFCC with Discrete Wavelet Transform	AVFC	Error rate: 13.28%
[40]	Locally Normalized Cepstral Coefficient	Deep Neural Network	Error rate: 9.4%
[41]	RASTA-PLP	Singular Value Decomposition	-
[42]	RASTA	Self data set	Error rate: 1.41%
[43]	MFCC- RASTA	Self data set	Error rate: 18%
[44]	MFCC- RASTA	Self data set	Accuracy: 84.98%
[45]	Cochlear Filter Cepstral Coefficient (CFCC)	Self data set	SNR: 96%
[46]	RASTA-PLP	Gaussian Mixture Model (GMM)	Accuracy: 98%
[47]	LPCC with Dynamic Time Warping (DTW)	Self data set	Accuracy: 97%
[48]	MFCC	Self data set	Accuracy: 90.3%
[49]	MFCC- LPCC	Polynomial classifier	-
[50]	Filter Bank Cestral Coefficient Bessel function	Self data set	Accuracy: 98%
[51]	MFCC- RASTA- PLP	Gaussian Mixture Model (GMM)	Accuracy: 60.8%

IV. Feature Extraction Performance- Discussion

The performance of feature extraction techniques from the Voiceprint was surveyed and the review results have been listed in Table II. The Table II portrays the various feature extraction algorithms, their utilization of data sets along with the key factors like accuracy and error rate. The accuracy of feature extraction is actually the recognition rate and Word Error Rate (WER). The evolution and development of Automatic Speaker Recognition (ASR) technique becomes vital, as it contributes high in security applications like cloud security, mobile security, banking security etc. The applications necessitate to inbuilt the voice print feature extraction technique with an anticipation of better recognition rate and reduced Word Error Rate and performance latency. The security applications like cloud computing [57] which involves banking transactions are time stamp oriented, that concentrates much in performance latency. The observation from Table II states that varieties of feature extraction algorithms, namely LPC, LPCC, PLP, MFCC, and RASTA evolve with varying panel of performances. The suggested solution defined from Table II for the selection of feature extraction algorithm is to combine the feature extraction algorithm like MFCC integrated with RASTA filtering with Gaussian Markov Model classifier and self data base will be suitable to achieve a greater accuracy. It is essential to note that the proposed feature extraction model must attain high recognition rate irrespective of speaker's emotion. Hence it is advisable to select the feature extraction technique that sounds good irrespective of emotions like angry, disgust, neutral, fear and happy. A strong suggestion for the design of emotion independent recognition model, the pronunciation approach and pitch rate may also be considered as vital parameters. By bringing these parameters into consideration, the speaker experiences high rate of recognition irrespective of emotion and during throat sickness.

V. Conclusion

Automatic Speaker Recognition (ASR) stays as the backbone in providing robust security against any attacks. The voice based authentication system experience a low rate of duplication in contrast with other biometric techniques. Countable feature extraction techniques are available with its unique features and pitfalls when taking performance and accuracy into account as discussed in previous sections. This paper reviews the performance traits of each feature extraction techniques and provides a strong suggestion that to combine the feature extraction techniques rather than employing unaccompanied. The integration of feature extraction techniques provides mutual cancellation of pitfalls and elevation of accuracy consuming reduced operation latency. In continuation of the above suggestion, the merging of MFCC and RASTA filtering awards better performance, less error rate and high rate of recognition.

References

[1] Zhen-Tao Liu and Qiao Xie, "Speech Emotion Recognition based on an improved brain emotion learning model", *Neurocomputing,* Vol. 309, Pp. 145-156, 2018.

[2] Surekha Rathod and Sangita Nikumbh, "Security based on speech recognition using MFCC method with MATLAB approach", *International Journal of Soft Computing and Artificial Intelligence,* ISSN:2321-404X, Vol.3, No. 2, Pp. 105-109, 2015.

[3] Hynek Hermansky and Nelson Morgan, "RASTA-PLP Speech Analysis", *International Conference in Acoustics, Speech and Signal Processing (IEEE),* 1992.

[4] Hemanta Kumar Palo and Mihir Narayan Mohanty, "Wavelet based feature combination for recognition of emotions", *Ain Shams Engineering Journal (Elsevier),*Vol. 9, Issue 4, Pp. 1799-1806, 2018.

[5] Maged M.M.Fahmy, "Palmprint recognition based on Mel Frequency Cepstral Coefficients feature extraction", *Ain Shams Engineering Journal (Elsevier),* Vol. 1, Issue 1, Pp. 39-47, 2010.

[6] Zulfiqar Ali and Irraivan Elamvazuthi, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and Cepstral coefficients based on the All Pole model", *Journal of Voice (Elsevier),* Vol. 30, Issue 6, Pp.1-13, 2016.

[7] P.Dhanalakshmi and S.Palanivel, "Pattern classification models for classifying and indexing audio signals", *Engineering Application of Artificial Intelligence, Elsevier,* Vol. 24, Issue 2, Pp.350-357, 2011.

[8] Shashidhar G.Koolagudi and Deepika Rastogi, "Identification of language using Mel-Frequency Cepstral Coefficients (MFCC)", *Procedia Engineering (Elsevier),* Vol. 38, Pp. 3391-3398, 2012.

[9] Claude Turner and Anthony Joseph, "A Wavelet packet and Mel-Frequency Cepstral Coefficients based feature extraction method for speaker identification", *Procedia Computer Science (Elsevier),* Vol. 61, Pp. 416-421, 2015.

[10] Manal Abdel Wahed, "Computer Aided Recognition of pathological voice", *IEEE 31st National Radio Science Conference,* 2014.

[11] Guofu Zhai and Jinbao Chen, "Pattern recognition approach to identify loose particle material based on modified MFCC and HMMs", *Neurocomputing (Elsevier)*, 2015.

[12] S.Jothilakshmi and V.Ramalingam, "Unsupervised speaker segmentation with residual phase and MFCC

features", *Expert Systems with Applications (Elsevier),* Vol. 36, Issue 6, Pp.9799-9804. 2009.

[13] Xiao- Chen Yuan and Chi-Man Pun, "Robust Mel Frequency Cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking", *Information Sciences (Elsevier),* Vol. 298, Pp. 159-179, 2015.

[14] Eduardo Pavez and Jorge F.Silva, "Analysis and design of Wavelet-Packet Cepstral coefficients for automatic speech recognition", *Speech Communication (Elsevier),* Vol. 54, Issue 6, Pp.814-835, 2012.

[15] S.Lalitha and D.Geyasruti, "Emotion Detection using MFCC and Cepstrum Features", *Procedia Computer Science (Elsevier),* Vol. 70, Pp. 29-35, 2015.

[16] Xiaolan Zhao and Zuguo Wu, "Speech Signal Feature Extraction Based on Wavelet Transform", *IEEE International Conference on Intelligent Computation and Bio-Medical Instrumentation,* 2011.

[17] Shabnam Gholamdokht Firooz and Farshad Almasganj, "Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals", *Computers and Electrical Engineering (Elsevier),* Vol. 58, Pp. 215-226, 2016.

[18] Johan de Veth and Louis Boves, "On the efficiency of classical RASTA filtering for continuous speech recognition: Keeping the balance between acoustic pre-processing and acoustic modeling", *Speech Communication (Elsevier),* Vol. 39, Issue 3, pp. 269-286, 2003.

[19] P.L.Emiliani and P.Graziani, "An excitation function for LPC synthesis of Voiced frames", *Signal Processing (Elsevier),* Vol. 5, Issue 6, Pp. 515-521, 1983.

[20] Tarek Mellahi and Rachid Hamdi, "LPC-based formant enhancement method in Kalman filtering for speech enhancement", *International Journal of Electronics and Communication (AEU) (Elsevier),* Vol. 69, Issue 2, Pp. 545-554, 2015.

[21] Xuewen Luo and Ing Yann Soon, "An Auditory Model for Robust Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing,* Pp. 1105-1109, 2009.

[22] Jose Novoa and Jorge Wuth, "DNN- HMM based Automatic Speech Recognition for HRI Scenarios", *IEEE International Conference on Human- Robot Interaction,* Pp. 150-159, 2018.

[23] Rama Hasan and Hussein Hussein, "Improvement of Speech Recognition Results by a Combination of Systems", *IEEE 23rd International Conference on Automation and Computing,* Pp. 1-4, 2017.

[24] Anand R Mehta, "Optimization Based Speech Authentication System to Web Content for Disabled Users", *International Journal of Enhanced Research in Science, Technology & Engineering*, ISSN: 2319-7463, Vol. 7 No. 3, 2018.

[25] Hynek Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech", *Journal of Acoustical society of America,* Vol. 87, No. 4, Pp. 1738-1752, 1990.

[26] Anthony Larcher and Kong Aik Lee, "Text-dependent speaker verification: Classifiers, databases and RSR2015", *Speech Communication (Elsevier),* Vol. 60, Pp. 56-77, 2014.

[27] Mohsen Sadeghi and Hossein Marvi, "Optimal MFCC Features Extraction by Differential Evolution Algorithm for Speaker Recognition", *IEEE 3rd Iranian Conference on Signal Processing and Intelligent systems"*, Pp. 169-173, 2017.

[28] Diksha Sharma and Israj Ali, "The Effect of DC Coefficient on mMFCC and mIMFCC for Robust Speaker Recognition", *IEEE International Conference on Advances in Computing, Communications and Informatics,* Pp. 313-317, 2015.

[29] Ning Wang and Lei Wang, "Robust Speaker Recognition Based on Multi-Stream Features", *IEEE International Conference on Consumer Electronics-China,* Pp. 1-4, 2016.

[30] Sandeep Rathor and R.S.Jadon, "Text Independent Speaker Recognition Using Wavelet Cepstral Coefficient and Butterworth Filter", *IEEE Conference ICCNT'17,* 2017.

[31] Wei-Hua Cao and Jian-Ping Xu, "Speaker-independent Speech Emotion Recognition Based on Random Forest Feature Selection Algorithm", *IEEE 36th Chinese Control Conference,* Pp. 10995-10998, 2017.

[32] Atik Charisma and M. Reza Hidayat, "Speaker Recognition Using Mel-Frequency Cepstrum Coefficients and Sum Square Error", *IEEE 3rd International Conference on Wireless and Telematics,* Pp. 160-163, 2017.

[33] Anggun Winursito and Risanuri Hidayat, "Improvement of MFCC Feature Extraction Accuracy Using PCA in Indonesian Speech Recognition", *IEEE International Conference on Information and Communications Technology,* Pp. 379-383, 2018.

[34] Elvira Sukma Wahyuni, “Arabic Speech Recognition Using MFCC Feature Extraction and ANN Classification”, *IEEE International Conference on Information Technology, Information Systems and Electrical Engineering,* Pp. 22-25, 2017.

[35] K N R K Raju Alluri and V.V.Vidyadhara Raju, “Analysis of Source and System Features for Speaker Recognition in Emotional Conditions”, *IEEE Region 10 Conference,* Pp. 2847-2850, *2016.*

[36] Ashwini Rajasekhar and Malaya Kumar Hota, “A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines”, *IEEE 4th International Conference on Advanced Technologies for Signal and Image Processing,* Pp. 114-118, 2018.

[37] Hong Yu and Zheng-Hua Tan, “DNN Filter Bank Cepstral Coefficients for Spoofing Detection”, *IEEE Access,* Vol. 5, Pp. 4779-4787, 2012.

[38] Dipjyoti Paul and Monisankha Pal, “Spectral Features for Synthetic Speech Detection”, *IEEE Journal on Selected Topics in Signal Processing”*, Vol. 11, Issue 4, Pp. 605-617, 2016.

[39] Ahmed Kamil Hasan AL-ALI and David Dean, “Enhanced Forensic Speaker Verification Using A Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions”, *IEEE Access,* Vol. 5, Pp. 15400-15413. 2017.

[40] Josué Fredes and José Novoa, “Locally-Normalized Filter Banks Applied to Deep Neural Network-based Robust Speech Recognition”, *IEEE Signal Processing Letters,* Vol. 24, Issue 4, Pp. 377-381, 2017.

[41] Muhammad Amirul Azzim Zulkifly and Norashikin Yahya, “Relative Spectral-Perceptual Linear Prediction (RASTA-PLP) Speech Signals Analysis Using Singular Value Decomposition (SVD)”, *IEEE 3rd International Symposium on Robotics and Manufacturing Automation,* Pp. 1-5, 2017.

[42] Detlef Hardt and Klaus Fellbaum, “Spectral Subtraction And RASTA-Filtering In Text-Dependent Hmm- based speaker Verification”, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. 2, Pp. 867-870, 1997.

[43] Ozlem Kalinli and Shrikanth Narayanan, “Early Auditory processing inspired features for Robust Automatic Speech Recognition”, *IEEE 15th European Signal Processing Conference,* Pp. 2385- 2389, 2007.

[44] Jia Lin Shen, “Discriminative temporal feature extraction for robust speech recognition”, *IEEE Electronic Letters,* Vol. 33, No. 19, Pp. 1598-1600, 1997.

[45] Qi Li and Yan Huang, “An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification under Mismatched Conditions”, *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 19, No.6, Pp. 1791-1801, 2011.

[46] Yu Min Zeng and Zhen Yang Wu, “Robust GMM based Gender classification using Pitch and RASTA-PLP parameters of Speech”, *IEEE 5th International Conference on Machine Learning and Cybernetics,* Pp. 3376-3379, 2006.

[47] Rekha Nair and Nirmala Salam, “A Reliable Speaker Verification System Based on LPCC and DTW”, *IEEE International Conference on Computational Intelligence and Computing Research,* Pp. 1-4, 2014.

[48] George Frewat and Charbel Baroud, “Android Voice Recognition Application with Multi Speaker Feature”, *IEEE 18th Mediterranean Electro technical Conference”,* Pp. 1-5, 2016.

[49] Hemant A Patil and Prakhar Kant Jain, “A Novel approach to Identification of Speakers from their Hum”, *IEEE 7th International Conference on Advances in Pattern Recognition”,* Pp. 167-170, 2009.

[50] Drisya Vasudev and Anish Babu K. K, “Speaker Identification using FBCC in Malayalam Language”, *IEEE International Conference on Advances in Computing, Communications and Informatics,* Pp. 1759-1763, 2014.

[51] El bachir Tazi and Noureddine El makhfi, “An Hybrid Front-End for Robust Speaker Identification under Noisy Conditions”, *IEEE Intelligent Systems Conference,* Pp. 764-768, 2017.

[52] Xuetao Xing and Jin Lin, “Model Predictive Control of LPC-Looped Active Distribution Network With High Penetration of Distributed Generation”, *IEEE Transactions on Sustainable Energy,* Vol. 8, Issue 3, Pp. 1051-1063, 2017.

[53] Harshita Gupta and Divya Gupta, “LPC and LPCC method of feature extraction in Speech Recognition system”, *IEEE 6th International Conference – Cloud system and Big Data Engineering,* Pp.498-502, 2016.

[54] Lindasalwa Muda and Mumtaz Begam, “Voice Recognition Algorithms using Mel Frequency Ceptral Coefficient (MFCC) and Dynamic Time Warping Techniques”, *Journal of Computing*, Vol.2, Issue 3, Pp.138-143, 2010.

[55] T.R.Jayanthi Kumari and H.S.Jayanna, “Comparison of LPCC and MFCC features and GMM and GMM-UBM modeling for limited data speaker verification”, *IEEE International Conference on Computational Intelligence and Computing Research”*, Pp. 1-6, 2014.

[56] Marzieh Razavi and Ramya Rasipuram, “On modeling context-dependent clustered states: Comparing HMM / GMM, hybrid HMM/ANN and KL-HMM approaches”, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Pp. 7659-7663, 2014.

[57] R.Shyamala and Prabakaran.D, “A Survey on Security Issues and Solutions in Virtual Private Network”, *International Journal of Pure and Applied Mathematics*” Vol.119, Issue.15, Pp. 3115-3122, 2018.

[58] William Campbell and Khaled Assaleh, “Polynomial Classifier techniques for speech verification”, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. 1, Pp. 321-324, 1999.

[59] ISO 9001:2015- Quality Management Systems. https://www.iso.org/standard/66693.html

[60] Biometric Types: https://www.elprocus.com/different-types-biometric-sensors/

[61] Feature Extraction Techniques: https://en.wikipedia.org/wiki/Feature_extraction

[62] RASTA: https://en.wikipedia.org/wiki/Rasta_filtering

[63] SAVEE: http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/Register.html

[64] EMO-DB: http://emodb.bilderbar.info/docu/

[65] DWT:https://en.wikipedia.org/wiki/Discrete_wavelet_transform

[66] GMM: https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem

[67] YOHO: https://catalog.ldc.upenn.edu/LDC94S16

[68] POLYCOST: http://www.speech.kth.se/cost250/polycost/be/v2.0/

[69] Support Vector machine (SVM): https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[70] Sum Square Error: https://hlab.stanford.edu/brian/error_sum_of_squares.html

[71] Sum Square Error Calculation: https://www.wikihow.com/Calculate-the-Sum-of-Squares-for-Error-(SSE)

[72] Australian Forensic Voice Comparison: http://databases.forensic-voice-comparison.net/

[73] Aurora 2: http://aurora.hsnr.de/aurora-2.html

[74] RSR 2015: https://www.accelerate.tech/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications