Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

A Review on Extraction of Speaker Related Variability Features Using Short-Term Feature Extraction Techniques

Dr. Sujiya Sreedharan[1*], Dr. A. Devi[2]

[1]*Assistant Professor, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore – 641049, Tamil Nadu, India*

[2]*Associate Professor, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore – 641049, Tamil Nadu, India*

Abstract:-In the rapidly developing digital environment, speaker verification is becoming more and more popular with prominent applications in security, automation, and authentication. Techniques for verifying speakers reject the brief fluctuations in the feature extraction stage that contain significant speaker-related characteristics. This page discusses a variety of feature extraction methods, ranging from short-term to long-term characteristics. Also the article discusses about speaker related variability features extraction with various fusion schemes such as Mel-Frequency Cepstral Coefficients (MFCC), Frequency Domain Linear Prediction (FDLP), Mean Hilbert Envelope Coefficients (MHEC) and Power-Normalized Cepstral Coefficients (PNCC) which acts as a complimentary for extracting common short-term speaker-related features. To combat accurate speaker verification, the review article with give a base idea for extracting short-term features for accurate recognition of speaker by compacting classifier efficiency.

Keywords*: Speaker verification, MFCC, FDLP, MHES, PNCC.*

1. Introduction

Voice is a highly information-rich signal that carries pitch modulation, harmonics, noise, power, interval, resonance activity, and more. It is modulated by frequency, amplitude, and time. The voice signals carry several levels of data. In the primary level, it carries a message through words [1]. In the remaining levels, it carries data about the language being spoken and the sensation, gender, age and so on i.e., the individuality and state of the speaker [2]. Normally, the Speaker Recognition intends to identify the word spoken in the voice signal. In contrast, the aim of automated Speaker Recognition system is to extract, distinguish and identify the data in the voice signal carrying speaker individuality. Speaker Recognition is the task of automatically identifying who is speaking via the speaker specific data involved in the speech waves [3].

Speech signal is a technology innovated in 1970s which allows speaker's individuality to be verified by using unique features of their voiceprints. Voice is the most usual, smallest part of spoken content that represents noticeable discrete sound in a continuous speech which is crucial and effective means of interaction between people [4]. It is a complex signal generated as a result of many conversions taking place at different levels such as articulatory, acoustic, linguistic and semantic. Speaker Recognition (SR) system utilize vocal sound as a distinctive feature that can recognize an individual based on their voice [5-6]. The process of extracting and modeling acoustic properties from voice data such that a person can be identified from others is known as speaker recognition. A plentiful understanding of the human speech procedure is needed before understanding automated Speaker Recognition system [7].

1.1 Feature Extraction

Every individual has distinct qualities derived from their voiceprint. Speech consists of several features but all of these features contained in the speech signal are not required for discriminating the speakers. The desired characteristics of an ideal feature are:

- Its variability should be limited within speakers and great between speakers.
- It was simple to extract from the speech signal.
- Variability in age and session shouldn't have an impact.
- In speech, it ought to come readily and often.
- It should be challenging to replicate or imitate.
- It must be resistant to distortion and noise.

Since no single feature can have all of these qualities, multiple features must be used to identify speakers. However, since methods like the Gaussian mixture model cannot handle high-dimensional data, the number of features that are taken into consideration for processing and recognition must also be small. For accurate density estimation, the number of training samples needed grows exponentially with features [8]. The "curse of dimensionality" refers to this. The specific application, the size of the voice database, the computational resources available, and the sort of speaker to be recognized—cooperative or not—all influence the choice of features. Prosodic and high-level characteristics are resistant against noise, but short-term spectral features are straightforward to compute and perform well. These characteristics can have disadvantages, such as being more easily imitated and having less discrimination. High-level features also demand a more complicated system. Therefore, it may be said that features are preferable for recognition, and that choosing a feature involves balancing robustness, discriminative property, and system implementation feasibility [9-11]. The following features can be used to identify a speaker:

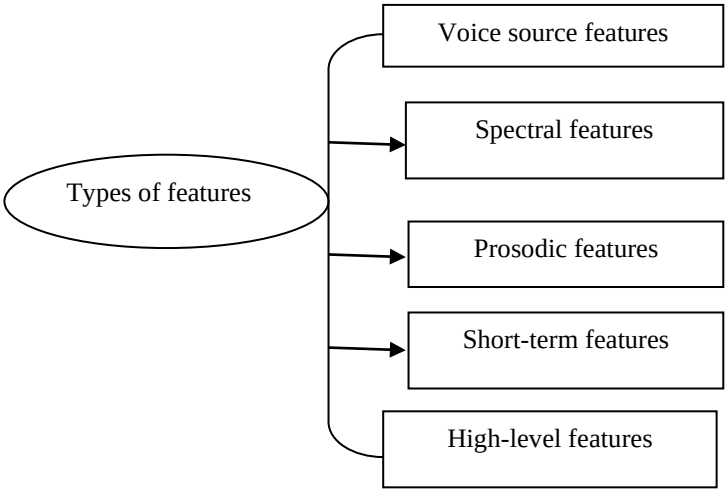

Figure 1. Different Types of Features

There are several speaker-specific properties in speech signals that are necessary for user detection and recognition. In order to obtain speaker-specific data in the speech and minimize statistical redundancies, the feature extraction unit transforms the pre-processed speech signal into feature vectors [12]. An ideal feature would be:

- Possess a large range of inter- and intra-speaker variability, or strong discriminative energy.
- Don't react negatively to noise.
- Remain resilient to voice impersonation and mimicry.
- Take place in speaking clearly and frequently.
- Remain unaffected by the speaker's health or chronic speech abnormalities.

- Have good computational efficiency. Based on individual physical analysis, feature representation is broadly divided into five categories: high-level characteristics, prosodic features, spectro-temporal features, voice source features, and short-term spectral features [13].

Short-term features: The most popular and often utilized feature representation for the Speaker Recognition process is this type of feature. These characteristics provide information about the vocal band resonance properties that are useful for speaker discrimination. Because of the articulator's movements, the vocal band's form changes every 20 to 30 milliseconds. Due to these vibrational variations in the speech generating mechanism, the resulting speech utterance is inherently highly non-stationary. Nonetheless, characteristics can be retrieved from the brief speech frames—roughly 20 to 30 ms—where the speech signal is thought to be fixed. Requiring framing and windowing using an appropriate window function is necessary to extract the short-term properties [14].

Initially, each short-term speech segment was represented by the spectral envelope obtained from the global profile of the Discrete Fourier Transform (DFT)-magnitude spectrum. The simplest spectral envelope framework achieves power mixing across neighboring frequency bands by using a collection of band-pass filters. Additionally, as supported by psycho-acoustic studies, the number of narrow band filters assigned allowed for a higher level of resolution in indicating the lower frequency value.An alternative spectrum estimation method to DFT with more obvious interpolation was created: coding analysis. Instead of being used directly as features, the prediction coefficients obtained from LPC analysis [15] are first transformed into less correlated features known as Linear Predictive Cepstral Coefficients (LPCCs), Line Spectral Frequencies (LSFs), and Perceptual Linear Prediction (PLP) coefficients before being used for Speaker Recognition processes [9]. Furthermore, these characteristics have been taken out of the speech signal's Fourier transform step rather than the DFT magnitude. Numerous methods of processing the phase spectrum have been proposed, such as those based on instantaneous frequency, inter-frame phase variance, and the negative derivative of the phase, also known as the group delay factor [16].However, the reliable extraction of phase characteristics is still difficult, which prevents its widespread use in Speaker Verification applications. MFCCs [17], which are obtained by using a psycho-acoustically stimulated mel filter bank, logarithmic compression, and the Discrete Cosine Transform (DCT), have been taken for granted as the standard characteristic for speaker verification in recent years. Heuristic representations of acoustic properties that approximate human hearing are called MFCCs [18].

Voice Source Features: These characteristics convey data including the degree of vocal fold opening and the length of the closure phase, as well as speaker-specific information like the shape of the glottal pulse and fundamental frequency, among others. The speech quality is determined by these factors and can be classified as creaky, breathy, modal, or pressured. The vocal tract filtering effect makes it impossible to measure the glottal properties directly. The vocal tract parameter can be estimated using the LP model, and the speech signal can then be estimated from the source by inversely filtering the original waveform. Close-phase covariance analysis is an additional option if the vocal folds are closed. This improves the estimation of the vocal tract, but it also necessitates that the closed phase be detected accurately, which is very difficult in noisy conditions [19].

After inverse filtering, the signal features can be extracted using an autoassociative neural network. Other methods have made use of parameters such as cepstral coefficient, glottal flow model, residual phase, higher order statistics, and many others. Voice source features are less dependent on phonetic content than vocal tract features, which require a high phonetic coverage due to their greater reliance on phonetic elements. This results in a massive demand for training and testing data. Given that vocal tract features are more discriminative than voice source features, the requirement for a large volume of data can be justified. However, it's also important to note that combining these two traits can increase accuracy [20].

Spectro-temporal Features: Spectrophotoral features, like as energy modulation and formant transitions, yield a wealth of speaker-specific information. Temporal information about the features can be included using the double-delta ($\Delta 2$) and delta (Δ) coefficients, which are estimates of the first and second order derivatives, respectively. Time differences between subsequent feature vector coefficients are utilized to calculate these

coefficients, which are then mixed with the original coefficient component. Additionally, "data-driven temporal filters" may be employed. Modulation frequency can also be utilized as a characteristic in speech recognition. The rate at which a speaker speaks is disclosed in modulation frequency, along with some other stylistic characteristics. For speech intelligibility, modulation frequencies lower than 20 Hz are taken into account.To get the most efficiency out of this feature, a 300 ms temporal window and modulation frequencies lower than 20 Hz were employed in. The feature vector's dimension is determined by the number of frames and FFT points. DCT can be used on the temporal trajectories rather than the spectrogram magnitudes to decrease the dimensionality of the spectro-temporal characteristics. For instance, with Δ & Δ2 coefficients, the total number of coefficients will be 3n if n is the number of original coefficients. Every frame goes through this procedure again. An alternative approach that is more reliable fits a regression line to the temporal curves. However, studies indicate that straightforward distinction might potentially yield comparable or superior results.

Prosodic Features Utilizing prosodic qualities to improve speech processing is crucial since prosody plays a significant role in the process of hearing speech. Prosodic features are essentially mixed with other acoustic features to be used in speaker recognition systems; however, there are a few drawbacks. Firstly, prosodic features have a much wider range than phonemes, which means that the framework designed to handle segmental features cannot handle these features. For this reason, these characteristics—which include pause length, pitch, syllable stress, speaking rate or tempo, intonation patterns, and energy distribution—are often referred to as suprasegmental characteristics. Determining speaker differences through the processing of prosodic information—which can be rapid or long-term—is another challenge. Furthermore, the attributes could rely on the aspects which can be changed intentionally by the speaker.

High-level Features Speakers can also be discriminated on the basis of the type of words a speaker generally uses during conversation. Initial research in this area was started by Doddington in 2001. An idiolect (Specific vocabulary used by a speaker) was used to discriminate the speakers. Higher-level cepstral system outperforms standard systems, (2) a prosodic system shows excellent performance individually and in combination, (3) other higher-level systems provide further gains, and (4) higher-level systems provide increasing relative gains as training data increases.

1.2 Feature Extraction based Speaker Verification System

In speaker verification, speaker individuality is composed of spectral, high-level and prosodic features. Short-term spectral features which are acoustical correlation of voice timbre extracted from short frames of 20-30 ms duration. Popular short-term features namely MFCC, LPCC, PLP are spectral features reflects the spectral attributes of the original intend speaker. According to the author, most speaker verification system utilise short-term features [21]. Voice Production is highly non-stationary in nature. The acoustic characteristics of speech signal varies continuously over a period of time. Features are extracted commonly from short-term frames of 20-30 ms duration which are time derivative and generally normalized for effective modelling of speaker verification system [22].

MFCC has greater impact in the process of speech processing and speaker verification system. From the literature point many researchers have made many attempts to improve the robustness of MFCC feature vector. Cepstral mean and variance normalization (CMVN) [23], Temporal structure normalization, RASTA filtering feature warping, MHEC [24], FDLP [25], PNCC [26] are commonly used feature extraction techniques for improving the performance of MFCC against additive noises and distortions towards channel variations. In article [23] a new MFCC features is proposed which reduces the MFCC estimation variations without depending on the statistical means beyond a speech frame to work under noisy conditions.

In speaker verification [24], uncertainty in feature space is modelled using variances in the GMM model and recently and subspace modelling of speaker and variability of session in a supervector are proposed [25]. MFCC has experienced with smaller variance in subsequent speaker and session variability models. The small variances are adopted based on multitapers with extension to DFT using multiple window functions. Additionally, [26] to form a spectrum estimate conventional windowing DFT is used to form an uncorrelated spectral estimate to reduce the variance which provided an encouraging result in speaker verification results. Finally, short-term

signal spectrum is often represented using MFCCs computed from a windowed Discrete Fourier transform (DFT). Windowing reduces spectral leakage but variance of the spectrum estimate remains high [27].

From the literature point, MFCC features are still utilized in most of the conventional speaker verification system for their better efficiency. Most of the speech parameterizations utilized in speaker verification systems relies on the cepstral representation of speech. This following section reviews feature extraction methods used in speaker verification system.

2. Literature Review

A novel method [28] was proposed for designing a feature extractor in Speaker Identification system based on the discriminative feature extraction method. In this method, a mel-cepstral estimation method based on the second-order all-pass system to the feature extractors in GMM-based and VQ text-independent Speaker identification system was proposed for finding a frequency scale relevant to recognition. Also, the frequency warping parameters of mel-cepstral estimation and the speaker model parameters were jointly optimized with the reduced classification error. The experiment was conducted with small-scale datasets.

Wang et al. [29] discussed about the robust speaker recognition to capture the vocal source and vocal tract features under noisy environment. The key idea was to reduce additive noise and convolutive reverberation. To address this limitation, authors proposed two phases. The first phase was to remove background noise through binary masking using deep neural network classifier. The second phase was to perform robust speaker identification with speaker models trained in selected reverberant conditions, on the basis of bounded marginalization and direct masking. The performance analysis results improved speaker identification over the related system in wide range of reverberation time and signal-tonoise ratios.

Nakagawa et al. [30] discussed about speaker characterization and recognition. The key role was to perform MFCC and phase information. The phase information method normalizes the change variation in the phase according to the frame position of the input speech and combines the information with MFCCs in text-independent speaker identification and verification methods. The experimental results shown that the combination of phase information with MFCC as provided better results under noisy speech data. The limitation occurred when comparing two phase values with the original phase information method under channel and background conditions.

Effectiveness of MHEC [31] was analyzed in i-vector Speaker Verification using Heavy-Tailed Probabilistic Linear Discriminant Analysis (HT-PLDA) as the compensation or back-end model. In this system, i-vectors were computed for MHECs and LDA was applied to reduce the dimensionality of feature space. Then, Within Class Covariance Normalization (WCCN) method was used to minimize the intra-speaker variability. Finally, scoring i-vectors was achieved via the Cosine Distance (CD) measure and HT-PLDA model. Moreover, the effect of i-vector dimension on system efficiency was explored. The experimental results confirmed that the system was trained with MHEC and traditional MFCC. The traditional MFCC provides better results under noiseless data. As conclusion additional techniques are required to develop a verification system under noisy and reverberant condition.

An improved algorithm [32] was proposed for abnormal audio recognition on the basis of improved MFCC. In this algorithm, different functions were included such as signal pre-processing, features extraction, features modeling, pattern matching. Initially, the audio signals were pre-processed and the features were extracted using improved MFCC algorithm that reduces the windowing process. Then, the extracted features were classified by using GMM classifier. But, the higher order GMM increases the number of parameters and loss in the convergence during the training process. As a conclusion, appropriate mixed order GMM is required to overcome the limitation caused by convergence rate during the training phase.

A new approach called MFCC, GMM and UBM [33] was developed to perform speaker identification under low-noise condition. The key idea of the approach was to use a decision tree to hierarchically partition the whole population into groups of small size and determine which speaker group at the leaf node a speaker under test belongs to and additionally MFCC+GMM were applied to the selected speaker group for speaker identification.

The performance analysis demonstrated that the approach outperforms MFCC+GMM and MFCC+GMM+UBM with higher accuracy and lower complexity for large population identification under Additive White Gaussian Noise (AWGN) conditions. However, the rise of population size can cause performance degradation of these schemes under noisy conditions.

Speaker identification was performed [34] by using features extracted from steady vowel regions. In this approach, first the combined temporal and spectral processing was performed to enhance the noisy speech signals. Secondly, the steady vowel regions were computed according to the knowledge of accurate vowel onset points and epochs. Moreover, GMM-based modeling was used to develop the speaker models. Finally, the improvements in the speaker identification were observed by the features extracted from steady vowel region in the presence of noisy atmosphere. Speaker Identification sometimes vary in high SNR value. As a conclusion, combination of spectral features with prosodic features can enhance the performance of Speaker Identification system.

An alternative noise-robust acoustic feature front-end [35] was proposed for obtaining the speaker identity including language structure or details conveyed in the speech signal. Particularly, a feature extraction process motivated by the human auditory processing was proposed. The proposed feature was based on the Hilbert envelope of Gammatone filter bank outputs that denote the envelope of the auditory nerve response. The sub-band amplitude modulations captured via smoothed Hilbert envelopes were used for carrying useful acoustic information. The proposed system was implemented to overcome the limitation of the conventional MFCC technique under noisy data.

A simple method [36] was proposed for capturing and characterizing the spectral variance through the Eigen structure of the sample covariance matrix. This covariance was estimated by sliding window over spectral features. This newly formulated feature vectors representing local spectral variances were used with standard speaker verification system. Also, the local variability features were extracted by using MFCCs including three different features such as FDLP, MHEC and PNCC. Then, the extracted features were classified by the GMM-UBM classifier for verifying the speakers. However, the length of sliding (temporal) window to derive the feature was not optimal and the sliding window length must be decided properly to achieve higher efficiency.

A simple HMM-based extension of i-vector method [37] was proposed which facilitates i-vectors to be successfully applied to the text-dependent Speaker Verification. In this method, the UBM was used for training phrase-independent i-vector extractor based on the set of monophone HMMs. Also, precondition i-vectors were proposed by using a regularized variant of WCNN for compensating the channel variability. The verification scores were cosine similarities between i-vectors normalized using phrase-dependent snorm. However, additional techniques are required for channel compensation and score normalization in the text-independent cases.

A text independent speaker authentication method [38] was proposed for cellular phone tools by means of amplifiers which are text-independent. The main objective was enhancing the Speaker Verification method for increasing the confidence rate. In this method, three basic processes were performed. The primary process was extracting a chosen set of person’s accent features called LPCC from an acoustic signal for constructing the dataset. In the second process, the dataset was taken as input and training was performed by using a Naive Bayes classifier. In the last process, a verification choice was computed that verifies the speaker. On the other hand, additional feature extraction techniques are required to improve the reliability and efficiency.

3. Conclusion

Short-term features play a crucial role in speaker verification systems by capturing the temporal characteristics of an individual's speech signal over short durations. These features are essential for distinguishing between speakers and creating robust speaker verification models. t's important to note that the choice of short-term features depends on the specific requirements of the speaker verification system, the available data, and the computational resources. Combining multiple types of features or using feature fusion techniques can enhance

the robustness and generalization capabilities of the speaker verification model. Additionally, techniques like data augmentation and normalization may be applied to further improve the system's performance.

Refrences

[1] Abd El-Samie, F. E. (2011), "Information Security for Automatic Speaker Identification", Information Security for Automatic Speaker Identification, pp. 1-122.

[2] Kiktova, E., and Juhar, J. (2015), "Speaker Recognition for Surveillance Application", Journal of Electrical and Electronics Engineering, Vol. 8, No. 2, pp. 19-22.

[3] Abualadas, F. E., Zeki, A. M., Al-Ani, M. S., and Messikh, A. E. (2019), "Speaker Identification based on Hybrid Feature Extraction Techniques", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 3, pp. 322-327.

[4] A. (2017), "Automatic Speaker Recognition for Mobile Forensic Applications", Mobile Information Systems, Vol. 2017, pp 1-6.

[5] Alsaadi, I. M. (2015), "Physiological Biometric Authentication Systems, Advantages, Disadvantages and Future Development: A Review", International Journal of Scientific & Technology Research, Vol. 4. No. 12, pp. 285-289.

[6] Backes, M., Doychev, G., Dürmuth, M., and Köpf, B. (2010), "Speaker Recognition in Encrypted Voice Streams", In: European Symposium on Research in Computer Security, pp. 508-523.

[7] Banumathi, A. C., and Chandra, E. (2015), "An Overview of Speech Recognition and its Challenges", Journal of Management and Science, Vol. 5, No. 1, pp. 91-100.

[8] Bansod, N. S., Kawathekar, S., and Dabhade, S. B. (2012), "Review of Different Techniques for Speaker Recognition System", Advances in Computational Research, Vol. 4, No. 1, pp. 57-60.

[9] Ciota, Z. (2004), "Speaker Verification for Multimedia Application", In: IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, pp. 2752-2756.

[10] Dabike, G. R., and Barker, J. (2021), "The use of Voice Source Features for Sung Speech Recognition", In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6513-6517.

[11] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010), "Front-End Factor Analysis for Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788-798.

[12] Devi, J. S., Yarramalle, S., Nandyala, S. P., and Reddy, P. V. B. (2017), "Optimization of Feature Subset Using HABC for Automatic Speaker Verification", In: Second IEEE International Conference on Electrical, Computer and Communication Technologies, pp. 1-6.

[13] Fazel, A., and Chakrabartty, S. (2011), "An overview of Statistical Pattern Recognition techniques for Speaker Verification", IEEE Circuits and Systems Magazine, Vol. 11, No. 2, pp. 62-81.

[14] Ganchev, T., Zervas, P., Fakotakis, N., and Kokkinakis, G. (2006), "Benchmarking Feature Selection techniques on the Speaker Verification task", In: Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing, pp. 314-318.

[15] Hasan, M. R., Jamil, M., and Rahman, M. G. R. M. S. (2004), "Speaker Identification using Mel Frequency Cepstral Coefficients", Variations, Vol. 1, No. 4, pp. 565-568.

[16] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N. (2016), "End-to-end Text- dependent Speaker Verification", In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5115-5119.

[17] Hermansky, H. (2011), "Speech Recognition from Spectral Dynamics", Sadhana, Vol. 36, No. 5, pp. 729-744.

[18] Jothilakshmi, S., Sangeetha, J., and Brindha, R. (2017), "Speech based Automatic Personality Perception using Spectral Features", International Journal of Speech Technology, Vol. 20, No. 1, pp. 43-50.

[19] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008), "A Study of Interspeaker Variability in Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 5, pp. 980-988.

[20] Kim, C., and Stern, R. M. (2016), "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 24, No. 7, pp. 1315-1329.
[21] Kinnunen, T., and Li, H. (2010), "An Overview of Text-independent Speaker Recognition: From features to supervectors", Speech Communication, Vol. 52, No. 1, pp. 12-40.
[22] Kumar, S., Bhattacharya, S., and Patel, P. (2014), "A New Pitch Detection Scheme based on ACF and AMDF", In: IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp. 1235-1240.
[23] Liu, X., Sahidullah, M., and Kinnunen, T. (2021), "Learnable MFCCs for Speaker Verification", In: IEEE International Symposium on Circuits and Systems, pp. 1-5.
[24] Liu, Y., He, L., Liu, J., and Johnson, M. T. (2019), "Introducing Phonetic Information to Speaker Embedding for Speaker Verification", EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2019, No. 1, pp. 1-17.
[25] Mary, L. (2011), "Extraction and Representation of Prosody for Speaker, Speech and Language Recognition", Springer Science & Business Media, pp. 19-33.
[26] Mittal, A., and Dua, M. (2021), "Automatic Speaker Verification Systems and Spoof Detection Techniques: Review and Analysis", International Journal of Speech Technology, pp. 1-30.
[27] Nakagawa, S., Wang, L., and Ohtsuka, S. (2011), "Speaker Identification and Verification by combining MFCC and Phase Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1085-1095.
[28] Bhimani, N. V. (2014), "Speaker Recognition System based on MFCC and VQ Algorithms", International Journal of Engineering Research and Technology, Vol. 3, No. 2, pp. 772-774.
[29] Wang, N., Ching, P. C., Zheng, N., and Lee, T. (2011), "Robust Speaker Recognition using Denoised Vocal Source and Vocal Tract Features", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 1, pp. 196-205.
[30] Nakagawa, S., Wang, L., and Ohtsuka, S. (2011), "Speaker Identification and Verification by combining MFCC and Phase Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1085-1095.
[31] Suh, J. W., Sadjadi, S. O., Liu, G., Hasan, T., Godin, K. W., and Hansen, J. H. (2011), "Exploring Hilbert Envelope based Acoustic Features in i-vector Speaker Verification using HT-PLDA", In: Proceedings of NIST 2011 Speaker Recognition Evaluation Workshop, pp. 1-4.
[32] Xie, C., Cao, X., and He, L. (2012), "Algorithm of Abnormal Audio Recognition based on Improved MFCC", Procedia Engineering, Vol. 29, pp. 731-737.
[33] Hu, Y., Wu, D., and Nucci, A. (2012), "Fuzzy-Clustering-based Decision tree approach for large population Speaker Identification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 4, pp. 762-774.
[34] Vuppala, A. K., and Rao, K. S. (2013), "Speaker Identification under Background Noise using Features Extracted from Steady Vowel Regions", International Journal of Adaptive Control and Signal Processing, Vol. 27, No. 9, pp. 781-792.
[35] Sadjadi, S. O., and Hansen, J. H. (2015), "Mean Hilbert Envelope Coefficients (MHEC) for robust Speaker and Language Identification", Speech Communication, Vol. 72, pp. 138-148.
[36] Sahidullah, M., and Kinnunen, T. (2016), "Local Spectral Variability Features for Speaker Verification", Digital Signal Processing, Vol. 50, pp. 1-11.
[37] Zeinali, H., Sameti, H., and Burget, L. (2017), "HMM-based Phrase-Independent i-vector Extractor for Text-Dependent Speaker Verification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 7, pp. 1421-1435.
[38] Thullier, F., Bouchard, B., and Menelas, B. A. (2017), "A Text-Independent Speaker Authentication System for Mobile Devices", Cryptography, Vol. 1, No. 3, pp. 1-22.