Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2017/16125

Feature Extraction for Speaker Recognition: A Systematic Study

Pardeep Sangwan*

Department of ECE, Maharaja Surajmal Institute of Technology, Delhi, India; sangwanpardeep@gmail.com

Abstract

The objective of a speaker recognition system is to identify a specific speaker from a spoken utterance by the speaker. Also, speaker recognition system must result in accurate identification of the speaker in a short duration. To fulfil this, extraction of the features from sound signal is an important task because speaker recognition systems are largely depend on speaker specific characteristics of a speech signal. The efficiency of this phase is crucial due to its effect on the performance and accuracy of the system. This paper presents a systematic study of the features contained in a speech signal and some techniques for extracting these features from speech signal. There are different feature extraction techniques which are used for speaker recognition like LPC, MFCC, GFCC, RASTA-PLP, etc. But MFCC and GFCC are the most widely used because they outperform other techniques and provides high accuracy rate of speaker recognition.

Keywords: Feature Extraction, GFCC, LPC, MFCC, Speaker Recognition

Paper Code: 16125; **Originality Test Ratio:** 18%; **Submission Online:** 21-May-2017; **Manuscript Accepted:** 25-May-2017; **Originality Check:** 18-June-2017; **Peer Reviewers Comment:** 09-Sep-2017; **Double Blind Reviewers Comment:** 14-Oct-2017; **Author Revert:** 05-Nov-2017; **Camera-Ready-Copy:** 12-Nov-2017)

1. Introduction

Humans have much better ability than computers to recognise face and speech. Speaker recognition is one of the tasks computer outperforms humans. Scientists have done regress research on the human ability to recognize and distinguish voices. By establishing the factors detailing speaker dependent information scientist have been able to design reliable speaker recognition systems for forensic science applications. In this era of digital computers, researchers have developed such automatic speaker recognition systems that could outperform human listeners on similar task.[1] Many limitation and challenging problems remain to be overcome with automatic speaker recognition systems. Speech signal retains information at several levels. In this paper, section-2 presents the characteristics of a feature ideal for speaker recognition and different types of features contained in the speech signal. Section-3 provides details about some of the feature extraction techniques used for speaker recognition. Finally, the paper is concluded in section-4.

2. Types of Features

Second phase is extraction of features from the speech signal. 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[2]:

- It should have large inter-speaker and small intra-speaker variability.
- It could be easily extracted from speech signal.
- It should not be affected by session and age variability.
- It should occur naturally and frequently in speech.
- It should be difficult to impersonate or mimic.
- It should be robust against noise and distortion.

No single feature have all of these characteristics and due to this more than one features have to be used for speaker recognition but simultaneously, the number of features considered for processing and recognition should also be small as techniques like Gaussian mixture model cannot deal with high-dimensional data.[3] The requirement of training samples increases exponentially with features for reliable density estimation. This is called "curse of dimensionality".[4] The choice of features is largely dependent on particular application, size of the available speech database, resources available for computing and type of speaker to be recognized whether cooperative or not. Short-term spectral features have advantage of easy computation and good performance[5] while high-level and prosodic features have robustness against noise. These features have drawbacks also like they are less discrimina-

*Author for correspondence

tive and can be easily mimicked. Also, more complex system is required for high-level features. Hence, it can be concluded that no feature can be considered best for recognition and the choice of feature is a trade-off between robustness, discriminative property, and feasibility of the system implementation.

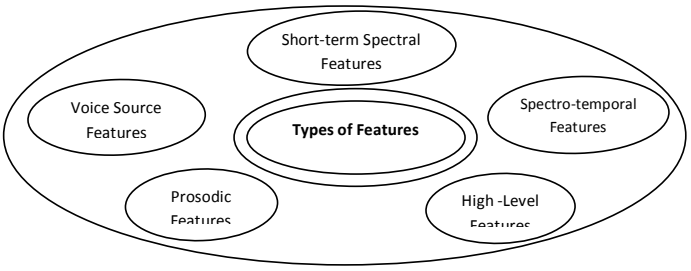

Figure 1. Types of Features.

According to their physical interpretation, features can be categorized as:

- Short-term spectral features,
- Voice source features,
- Spectro-temporal features,
- Prosodic features, and
- High-level features.

2.1 Short-term Spectral Features

Basically, the speech signals are highly non-stationary due to articulatory movements. Hence, to obtain a pseudo-stationary signal, speech signal is divided in the form of short frames. The duration of these short frames is around 20-30 ms and speech signal can be considered stationary for extracting the spectral feature vector in this short duration. The glottal voice source results in downward slopping spectrum which in turn causes the higher frequencies having very low intensity. Therefore, these frequencies have to be pre-emphasized to boost up the higher frequencies. This pre-emphasized signal frame is then required to be passed through a smooth window function, generally the hamming window due to the finite length effect of Discrete Fourier Transform (DFT). Being very simple and efficient, the DFT is generally used technique to decompose a signal into its frequency components. On the basis of the assumption that small perceptual information is contained in phase spectrum, the magnitude spectrum is believed to be more important and generally retained for further processing. A technique utilizing phase information is given in [6]. In speaker recognition, a large amount of information is contained in the "spectral envelop of the spectrum" like resonance property of vocal tract. Spectral envelop is nothing but the global shape of the magnitude spectrum of DFT. A technique utilizing spectral envelop have a set of band-pass filters which integrates the energy of adjacent frequency bands. More band-pass filters are allocated to lower frequencies to represent them with higher resolution. Transformations are used now to reduce the dimensionality of the sub-band energy values which were earlier directly used as features. The feature obtained from this process is known as Mel-frequency cepstral coefficients (MFCC) and introduced in 1980s primarily for speech recognition. But later on MFCCs are adopted for speaker recognition as well.

2.2 Voice Source Features

These features carry speaker-specific information like glottal pulse shape, fundamental frequency etc. the parameters like degree of vocal fold opening and the duration of the closing phase. These parameters determine the quality of the speech, which can be categorized as creaky, breathy, and modal or pressed. The glottal features cannot be measured directly because of the vocal tract filtering effect. The LP model can be used for estimating the vocal tract parameter first and then inverse filtering of original waveform can be done to get an estimation of speech signal from source.[7] If vocal folds are closed, then close-phase covariance analysis can also be used. The estimation of vocal tract is enhanced by this but along-with this, it also requires that the closed phase is detected accurately, which is very difficult in noisy conditions. An auto-associative neural network can be used to extract the features of the signal after inverse filtering. Some other approaches have used parameters like residual phase, glottal flow model, cepstral coefficient, higher order statistics and many more. Voice source features has less dependency on phonetic content while vocal tract features are much more dependent on phonetic factors and hence, need large phonetic coverage which in turn give rise to the requirement of huge amount of training and testing data. This need for large amount of data can be well justified because vocal tract features are more discriminative as compared to voice source features. But it is also worth stating that accuracy can be improved by fusing these two features.

2.3 Spectro-temporal Features

The large amount of speaker-specific information can be extracted from spectro-temporal details like energy modulation and formant transitions. Delta (Δ) and Double-delta (Δ^2) coefficients, which are the first and second order derivative estimates can be used to include temporal information to the features.[8] These coefficients are determined by taking time differences between successive feature vector coefficients and then these coefficients are combined with the original coefficients. For example, if n is the number of original coefficients then with Δ & Δ^2 coefficients total number of coefficients will be 3n. This process is repeated for each frame. Another method fits a regression line to the temporal curves, giving a more robust alternative. But research shows that simple differentiation can also provide equal or better performance. "Time-Frequency principle com-

ponent" and "data-driven temporal filters" can also be used. In speaker recognition, modulation frequency can also be used as feature. Modulation frequency contains the information of the rate at which speaker utters words along with some other stylistic attributes. Modulation frequencies less than 20 Hz are considered for speech intelligibility. In [9], a temporal window of 300 ms was used along with modulation frequencies of less than 20 Hz to achieve highest efficiency using this feature. The number of FFT points and number of frames decides the dimension of the feature vector. To reduce the dimensionality of the spectro-temporal features, DCT can be applied on the temporal trajectories instead of spectrogram magnitudes.

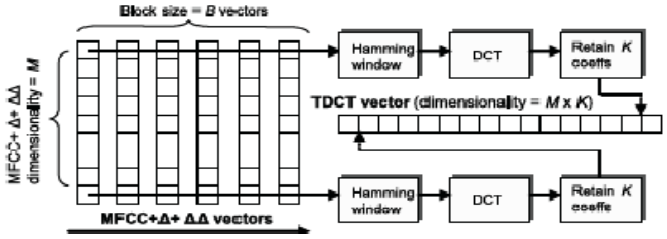

Figure 2. Temporal Discrete Cosine Transform (TDCT)[10].

DCT is advantageous over DFT as it can reduce the dimensionality along with retaining the relative phases of the feature vectors[10] and thus can confine phonetic and speaker-specific both type of the information. By combining cepstral and temporal features, efficiency can be enhanced as compared to the cepstral system alone but the improvement was small and rigorous research is required in this direction before its use in practical applications. One of the probable problems can be that both types of the features have different frame rate and thus, cannot be simply combined at frame level. To overcome this, an entirely different speaker modelling and fusing techniques are needed. To improve speaker recognition system, frequency modulation can be used in place of amplitude based methods. In this, a band-pass filter bank is used to divide speech signal into sub-band signals. Then formant frequency features are extracted using dominant frequency components like frequency centroids. For instance, the dominant frequency can be detected first by using second-order all-pole analysis. Then, a measure of deviation from "default" frequency can be determined by taking the difference of centre frequency and the pole frequency sub-band, which can be used as frequency modulation based feature. This feature when used with MFCC has shown promising results.

2.4 Prosodic Features

Prosody is a major part of the speech perception process and thus, it is very important to use prosodic features for enhancement in speech processing. Basically, the prosodic features are combined with some other acoustic feature for use in speaker recognition system but there are several limitations like, the range of prosodic feature is much wider as compared to that of phonemes and hence, the framework handling segmental features cannot handle these features. This is why these features are also known as supra-segmental features and includes, pause duration, pitch, syllable stress, speaking rate or tempo, intonation patterns and energy distribution.[11] Another problem is to determine speaker differences by processing the prosodic information which can be either instantaneous or long term. Moreover, the features may depend on the aspects which can be changed intentionally by the speaker. The Fundamental frequency (F_o) is generally used prosodic feature. In the noisy environment, the combination of spectral and F_o related features have been most effective and accurate. Besides F_o-related features, energy and duration also have better accuracy than other prosodic features. As F_o is the most important prosodic feature, it is now discussed in detail. The F_o determination is not an easy task. For instance, F_o is generally outside of the normal telephone speech passband (0.3-3.4KHz) and thus, it can be detected using upper harmonics only. Recent methods for F_o estimation include, YIN method and auto-correlation method. F_o contains physiological as well as learned information, both of which are important for speaker recognition. It can be illustrated by means of F_o which can be correlated to the larynx size or the pitch variation which can be correlated to the speaking style. In two broad categories namely text-dependent and text-independent speaker recognition, former uses temporal alignment of pitch contours and later utilizes the long term F_o statistics like mean value. The combination of the mean value and another statistics like variance or kurtosis can also be used for speaker modelling. But the performance of histograms, support vector machines[11] and latent semantic analysis is better. Further, logarithm of F_o results in an improved feature than F_o and this is concluded on the basis of several experiments. Mathematically, F_o is not very discriminative because of the reason that it is a single-dimension feature. Therefore, to enhance the accuracy, the auto-correlation function is used to extract pitch and voice-related features which are multidimensional. Also, local and long-term temporal variations of fundamental frequency, F_o, can be considered simultaneously to enhance the accuracy. Local temporal variations of F_o can be captured by combining the delta feature and instantaneous F_o, while for capturing long-term dynamics, F_o contour is divided into segments and then represented with the help of parameters associated with the individual segment.

2.5 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.[12] The concept behind using High-level

Features for speaker recognition is converting the utterances in to a series of tokens and then the occurrence of similar pattern of tokens is used to discriminate speakers. These tokens can be word[12], phonemes, and prosodic variations like rise or fall in pitch or energy[13], and some articulatory token based on manner and place of articulation. Some of the Gaussian mixture component indices can also be used as tokens. More than one tokenizers can also be used[14] assuming that they can capture the complementary aspects of utterances e.g. two or more phone recognizers trained on different languages can be used or parallel GMM tokenizers can also be used out of which every tokenizer can be trained with a separate clustered group of speakers. This idea was evolved from the promising results of parallel phone recognizers in the field of spoken-language recognition. N-gram modelling is the basic classifier for these token features. The token sequence of utterances can be denoted as {a1, a2, a3,……, ak}, where ai∈ V: a definite vocabulary. The joint probability of the n successive tokens is estimated to construct N-gram model. For example, estimating the probabilities of two consecutive tokens ai, ai+1 gives a bigram model (N=2) and estimating the probabilities of three consecutive tokens ai, ai+1, ai+2 gives a trigram model (N=3). This can be illustrated with the bigrams of speaker recognition as (s,p), (p,e), (e,a), (a,k), (k,e), (e,r), (r,_), (_,r), (r,e), (e,c), (c,o), (o,g), (g,n), (n,i) (i,t), (t,i), (i,o), and (o,n). The probability of each N-gram is ML (Maximum Likelihood) or MAP (Maximum a Posteriori) estimate of the N-gram in the training corpus which is used along with entropy measures to recognize the speaker by assessing similarity between them.

3. Feature Extraction Techniques

Many feature extraction techniques are available now-a-days for extracting speaker-specific information out of sound signal. Some of them are:

- Mel-Frequency Cepstral Coefficients (MFCC)
- Gammatone-Frequency Cepstral Coefficients (GFCC)
- LPC-based Cepstral parameters
- RelAtive spectra filtering coefficients (RASTA)

3.1 Mel-frequency Cepstral Coefficients (MFCC)

To develop robust speaker recognition system, a mechanism is required to not only accurately represent the acoustic signal of a given speaker but it also has to be reliable. Fortunately, huge amount of research is done in this field of signal acoustics. Research has led to a proven method to extract unique characteristics of speakers, the *Mel-Frequency Cepstral Coefficients*. Researches have also shown that humans do not perceive acoustic frequency content of speech signal on a linear scale. Thus for each frequency, *f* in Hz, a subjective pitch is measured on a scale known as 'Mel' scale as given below,

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Where f_{mel} is the subjective pitch in Mel for a frequency *f* in Hz. This is the basis for computation of MFCC, most generally used feature set in speech and speaker recognition. MFCC is used in numerous researches because it provides robust features which lead to high accuracy rate of speaker recognition. MFCC has been the most widely used feature due to high recognition accuracy, lower complexity and capability to capture major characteristics of speech signal but the performance is largely affected by background noise. In this sub-section, this widely used feature extraction method for Speaker Recognition is described. After pre-processing of speech signal, features are extracted in the form of mel-frequency cepstral coefficient. MFCC represents voice signal, on the basis of perception. The features can be obtained using Fourier transform or discrete cosine transform of speech signal. The main difference is that, in the former, frequency ranges are placed on logarithmic scale called mel-scale approximately mimicking auditory system's response better as compared to linearly positioned frequency ranges of FFT\DCT. The most important objective of MFCC processor is mimicking behavior of human ear. The typical procedure for feature extraction is shown in figure 3, with the assumption that it has been processed digitally and properly quantized.

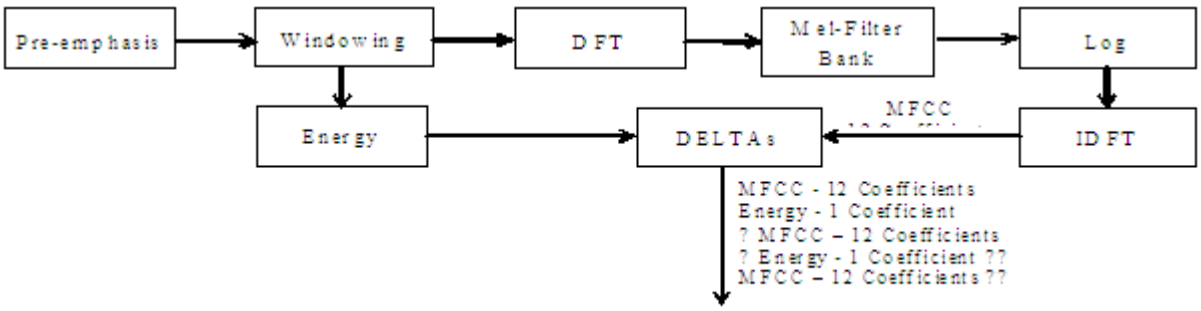

Figure 3. Feature extraction processing workflow.

3.2 Gammatone Frequency Cepstral Coefficients (GFCC)

Noise robustness is one of the biggest challenges in ASR. One of the major disadvantages of MFCC is the sensitivity to additive noise. GFCC is a feature based on Gammatone filter-bank. A frequency-time representation of the signal, called as *Cochleagram,* can be obtained from the output of the Gammatone filter-bank. To compute the GFCC features a cochleagram is needed; the different stages of its computation have similarities with MFCC counterpart.

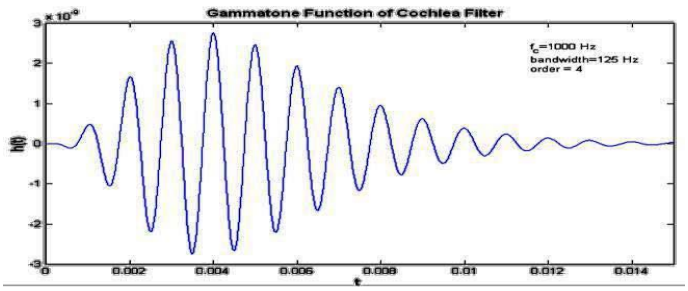

Figure 4. Impulse response of a Gammatone filter.

The Gammatone filter is devised for simulating the human auditory phenomena. A GF with a central frequency, *fc*, has a transfer function as:

$$g(t) = at^{n-1}e^{-j2\pi bt}Cos\ (2\pi f_c + \emptyset) \quad (2)$$

Where, ϕ is phase which often kept equal to zero and constant *a,* decides the gain and the value n gives the order of the filter which is typically set to a value less than 4. The factor *b* is defined as:

$$b = 25.17 * \left(\frac{4.37\ f_c}{1000} + 1\right) \quad (3)$$

3.3 Linear Prediction based Cepstral Parameters

The LPC analysis is based on a linear model of speech production. The model generally used is an auto regressive moving average (ARMA) model, simplified in an auto regressive (AR) model.[15] The speech production apparatus is usually described as a combination of four modules: (1) the glottal source, which can be seen as a train of impulses (for voiced sounds) or a white noise (for unvoiced sounds) (2) the vocal tract (3) the nasal tract and (4) the lips. Each of them can be represented by a filter: a low pass filter for the glottal source, an AR filter for the vocal tract, an ARMA filter for the nasal tract, and an MA filter for the lips. Globally, the speech production apparatus can therefore be represented by an ARMA filter. The principle of LPC analysis is to estimate the parameters of an AR filter on a windowed (pre-emphasized or not) portion of a speech signal. Then, the window is moved and a new estimation is calculated. For each window, a set of coefficients (called predictive coefficients or LPC coefficients) is estimated and can be used as a parameter vector. Finally, a spectrum envelope can be estimated for the current window from the predictive coefficients. These features are based on Linear Prediction which is a spectrum estimation technique like DFT having good intuitive interpretation in time as well as frequency domain.[15] For time domain, linear predictor equation can be given as:

$$S'[n] = \sum_{k=1}^{p} \alpha_k S[n-k] \quad (4)$$

Where, S[n] is original signal, $S'[n]$ is predicted signal and α_k is the predictor coefficient. Residual can be given as e[n] = S[n]- S'[n], also known as prediction error. Levinson-Durbin algorithm is generally used to minimize the residual energy and to predict the coefficient α_k. The predictor coefficients are seldom used as features because Although, LPC based cepstral feature are based on formant structure, but it ignores several useful detail. Another disadvantage is its incapability to capture spectral valleys. Hence, this feature is not so good for speaker identification but it is used in several speech recognition algorithms because of its capability to encode speech at low bit-rate and high computation speed but they can be converted into features which are less correlated and robust, for example, "Linear Predictive Cepstral Coefficient (LPCC)", "Line Spectral Frequencies (LSF)" and "Perceptual Linear Predictive (PLP) coefficient. Some more features are Log Area Ratios (LARs), Partial Correlation Cop-efficient (PARCOREs) and Formant Frequency and Bandwidth. But these features are not as successful as MFCC and GFCC.

3.4 Relative Spectral Analysis Technique (RASTA)

RASTA analysis technique is based on the idea that the rate of change of the short-term spectrum for linguistic and non-linguistic components in speech is different. This means that spectral components of the communication channel vary more quickly or more slowly than the spectral components of the speech and they could be separated by filtering. The core part of RASTA processing is a band-pass filtering of the spectral parameters trajectories by an IIR filter. The convolved (in the time domain) distortions in the communication channel can be reduced by using the RASTA filtering in the logarithmic domain (spectral or cepstral). The RASTA approach can be combined with the Perceptual Linear Prediction method (so called RASTA-PLP approach) or can directly be applied to the cepstral trajectories.[15] The speech

analysis can be made less sensitive to the slowly varying factors of speech signal by replacing a critical band short-term spectrum in Perceptual Linear Prediction method of speech analysis with a spectral estimate obtained by band pass filtering of each frequency channel. This spectral estimate will be less sensitive to the slow variation in the short-term spectrum due to the suppression of slow varying components in each frequency channel by the above explained process. In RASTA-PLP each frame has to undergo following steps:

- Computation of the critical-band power spectrum
- Transformation of spectral amplitude by a compression non-linear transform
- Filter out all transformed spectral component
- Transformation of output of previous step by expansion of static non-linear transform.
- Just like conventional PLP, the power law of hearing is stimulated by multiplying with the equal loudness curve and increasing the power to 0.33.
- Computing the all-pole model of the resulted spectrum

Major difference is suppression of the constant or slow varying characteristics in all spectral components of short-term spectrum before estimation of all-pole model. This technique is utilized in noise robust speaker recognition algorithms due to its capability to lessen the impact of noise in the speech signal and to remove the slow environmental variations.[16]

4. Conclusion and Future Scope

An efficient speaker recognition system is dependent on robust feature extraction from the speech signal. There are different features which can be used for recognizing speaker but all features have certain advantages and disadvantages according to specific applications. Out of several feature extraction techniques like LPC, MFCC, GFCC, RASTA-PLP, PLDA, LPCC etc., MFCC and GFCC are the generally used feature extraction techniques due to their better performance as compared to other feature extraction techniques for speaker recognition. As an extension of this work, the combination of two or more techniques can be evolved for further enhancing the efficiency of speaker recognition systems.

5. References

1. Rosenberg AE. Listener Performance in Speaker Verirfication Tasks. IEEE Trans AE. 1973.
2. Kinnunen T, Li H. An overview of text-independent speaker recognition:From features to supervectors. Speech Communication. 2010; 12–40. https://doi.org/10.1016/j.specom.2009.08.009
3. Reynolds DA, Quatieri T, Dunn R. Speaker verification using adapted gaussian mixture models. DSP. 2000; 19–41.
4. Jain A, Duin R, Mao J. Statistical pattern recognition: A review. IEEE Trans. PAMI. 2000; 4–37.
5. Reynolds DA. Channel Robust Speaker verficiation via feature mapping. ICASP. 2003.
6. Hedge R, Murthy H, Rao G. Application of the modified group delay function to speaker identification and discrimination. ICASSP. 2004.
7. Kinnunen T, Alku P. On separating glottal source and vocal tract information in telephony speaker verification. ICASSP. 2009. https://doi.org/10.1109/ICASSP.2009.4960641
8. Soong F, Rosenberg A. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans ASSP. 1988; 36(6):871–9. https://doi.org/10.1109/ICASSP.2009.4960641
9. Kinnunen T. Spectral Features for Automatic Text-Independent Speaker Recognition. University of Joensuu, Joensuu, Finland. 2004.
10. Kinnunen T, Koh C, Wang L, Li H, Chang E. Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. ISCSLP. 2006.
11. Shriberg E, Ferrer L, Kajarekar S, Venkataraman A, Stolcke A. Modeling prosodic feature sequences for speaker recognition. Speech Communication. 2005; 46(4):455–72. https://doi.org/10.1016/j.specom.2005.02.018
12. Doddington G. Speaker recognition based on idiolectal differences between speakers. Eurospeech. 2001.
13. Adami A, Mihaescu R, Reynolds D, Godfrey J. Modeling prosodic dynamics for speaker recognition. ICASSP. 2003.
14. Campbell W, Campbell J, Reynolds D, Jones D, Leek T. Phonetic speaker recognition with support vector machines. ANIPS. 2004.
15. Nisha V, Jayasheela M. Survey on Feature Extraction and Matching Techniques for Speaker Recognition Systems. IJARECE. 2013; 2(3).
16. Hermansky H, Morgan N. Rasta processing of speech. IEEE Trans. on SAP. 1994; 2(4):578–89. https://doi.org/10.1109/89.326616

Annexure-I

Feature Extraction for Speaker Recognition: A Systematic Study

ORIGINALITY REPORT

18%

SIMILARITY INDEX

PRIMARY SOURCES

1	ijarece.org Internet	331 words — 8%
2	www.ijcsitre.org Internet	55 words — 1%
3	cs.uef.fi Internet	50 words — 1%
4	Awais Mahmood. "Multidirectional Local Feature for Speaker Recognition", 2012 Third International Conference on Intelligent Systems Modelling and Simulation, 02/2012 Crossref	28 words — 1%
5	www.manualslib.com Internet	25 words — 1%
6	C.B. Soh. "Feature Extraction Based on Mel-Scaled Wavelet Transform for Heart Sound Analysis", 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005 Crossref	18 words — < 1%
7	Kinnunen, T.. "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication, 201001 Crossref	17 words — < 1%
8	Qiu, Zuochun. "ICA-based Rasta-PLP feature for speaker identification", The 2nd International Conference on Information Science and Engineering, 2010. Crossref	15 words — < 1%
	M.S. Okun. "Linear Predictive Analysis for Targeting the Basal	
9	Ganglia in Deep Brain Stimulation Surgeries", Conference Proceedings 2nd International IEEE EMBS Conference on Neural Engineering 2005, 2005 Crossref	14 words — < 1%
10	SpringerBriefs in Electrical and Computer Engineering, 2013. Crossref	14 words — < 1%
11	"Gaussian Membership Function-Based Speaker Identification Using Score Level Fusion of MFCC and GFCC", Advances in Intelligent Systems and Computing, 2016. Crossref	13 words — < 1%
12	www.ijcaonline.org Internet	13 words — < 1%
13	Mansour, Asma, and Zied Lachiri. "Emotional speaker recognition in simulated and spontaneous context", 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2016. Crossref	11 words — < 1%
14	ijircce.com Internet	10 words — < 1%
15	WU, Z, and Z CAO. "Improved MFCC-Based Feature for Robust Speaker Identification", Tsinghua Science & Technology, 2005. Crossref	9 words — < 1%
16	uhra.herts.ac.uk Internet	9 words — < 1%
17	epublications.uef.fi Internet	9 words — < 1%
18	Jawarkar, N. P., R. S. Holambe, and T. K. Basu. "Speaker Identification Using Whispered Speech", 2013 International Conference on Communication Systems and Network Technologies, 2013. Crossref	8 words — < 1%

Source: http://www.ithenticate.com/

Citation:
Pardeep Sangwan
"Feature Extraction for Speaker Recognition: A Systematic Study",
Global Journal of Enterprise Information System. Volume-9, Issue-4, October-December, 2017. (http://informaticsjournals.com/index.php/gjeis)
DOI: 10.18311/gjeis/2017/16125
Conflict of Interest:
Author of a Paper had no conflict neither financially nor academically.