# TEXT-INDEPENDENT SPEAKER RECOGNITION USING SUPERVECTORS

**Khurrath-Ul-Aien M.R[1], Anitha G[2]**

*[1] Scholar, [2]Assistant Professor, Department of computer science and Engineering,*

*UBDTCE, Davangere (India)*

**ABSTRACT**

*In this gives an overview of automatic speaker recognition technology, with an emphasis on text-independent recognition. Speaker recognition has been studied actively for several decades. We give an overview of both the classical and the state-of-the-art methods. We start with the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling. We elaborate advanced computational techniques to address robustness and session variability. The recent progress from vectors towards supervectors opens up a new area of exploration and represents a technology trend.*

***Keywords: Discriminative models , Feature extraction , Text-independence ,Speaker recognition, Statistical models, Supervectors.***

## I. INTRODUCTION

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. State-of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition. In addition to telephony speech data, there is a continually increasing supply of other spoken documents such as TV broadcasts, teleconference meetings, and video clips from vacations. Extracting metadata like topic of discussion or participant names and genders from these documents would enable automated information searching and indexing. Speaker diarization also known as ‘‘who spoke when”, attempts to extract speaking turns of the different participants from a spoken document, and is an extension of the ‘‘classical” speaker recognition techniques applied to recordings with multiple speakers. In forensics and speaker diarization, the speakers can be considered non-cooperative as they do not specifically wish to be recognized. On the other hand, in telephone-based services and access control, the users are considered cooperative. Speaker recognition systems, on the other hand, can be divided into text-dependent and text-independent ones. In text-dependent systems suited for cooperative users, the recognition phrases are fixed, or known beforehand. For instance, the user can be prompted to read a randomly selected sequence of numbers as described in. In text-independent systems, there are no constraints on the words which the speakers are allowed to use. Thus, the reference (what are spoken in training) and the test (what are uttered in actual use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account. Text-independent

recognition is the much more challenging of the two tasks. In general, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition. Changes in the acoustic environment and technical factors (transducer, channel), as well as ‘‘within-speaker” variation of the speaker him/herself (state of health, mood, aging) represent other undesirable factors. In general, any variation between two recordings of the same speaker is known as session variability.

## II. DESIGN

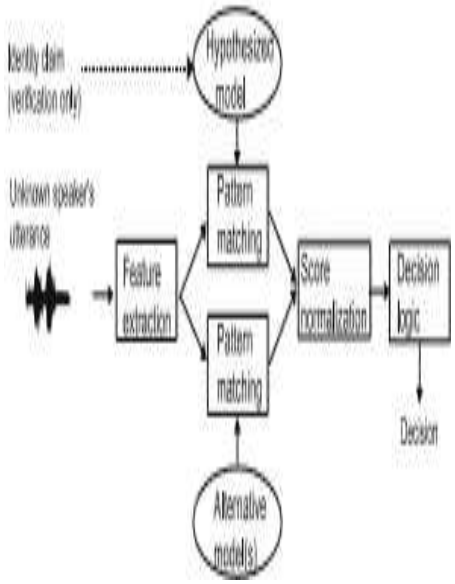


**Fig 1: Speaker verification/identification**

The feature extraction module first transforms the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies suppressed. In the recognition mode, the feature vectors extracted from the unknown person’s utterance are compared against the model(s) in the system database to give a similarity score. The decision module uses this similarity score to make the final decision. Virtually all state-of-the-art speaker recognition systems use a set of background speakers or cohort speakers in one form or another to enhance the robustness and computational efficiency of the recognizer.

## III. FEATURE EXTRACTION

The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short frames of about 20–30 ms in duration. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame. Usually the frame is pre-emphasized and multiplied by a smooth window function prior to further steps. Pre-emphasis boosts the higher frequencies whose intensity would be otherwise very low due to downward sloping spectrum caused by glottal voice source. The window function (usually Hamming), on the other hand, is needed because of the finite-length effects of the discrete Fourier transform. In practice, choice of the window function is not critical. The well-known fast Fourier transform (FFT), a fast implementation of DFT, decomposes a signal into its frequency components. Alternatives to FFT-based signal decomposition such as non-harmonic bases, aperiodic functions. The DFT,

however, remains to be used in practice due to its simplicity and efficiency. Usually only the magnitude spectrum is retained, based on the belief that phase has little perceptual importance. However, provides opposing evidence while described a technique which utilizes phase information. The global shape of the DFT magnitude spectrum known as spectral envelope contains information about the resonance properties of the vocal tract and has been found out to be the most informative part of the spectrum in speaker recognition. A simple model of spectral envelope uses a set of bandpass filters to do energy integration over neighboring frequency bands. Motivated by psycho-acoustic studies, the lower frequency range is usually represented with higher resolution by allocating more filters with narrow bandwidths.

## I V. EXPERIMENTAL WORK

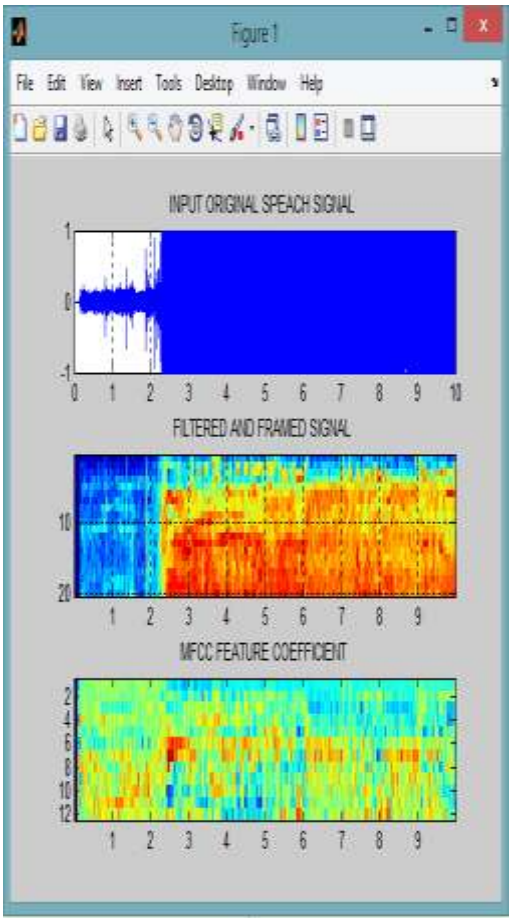


**Fig 2: Feature Extracted from Speech**

## V. CONCLUSION

We have presented an overview of the classical and new methods of automatic text-independent speaker recognition. The recognition accuracy of current speaker recognition systems under controlled conditions is high. However, in practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users.

However, many research problems remain to be addressed, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores.

## REFERENCES

[1].Alexander, A., Botti, F., Dessimoz, D., Drygajlo, A., 2004. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. Forensic Science International 146S, December 2004, pp. 95–99.

[2].Besacier, L., Bonastre, J.-F., 2000. Sub band architecture for automatic speaker recognition. Signal Process. 80, 1245–1259. Besacier, L., Bonastre, J., Fredouille, C., 2000. Localization and selection of speaker-specific information with statistical modeling. Speech Comm. 31, 89–106.

[3].Bimbot, F., Magrin-Chagnolleau, I., Mathan, L., 1995. Second-order statistical measures for text-independent speaker identification. Speech Comm. 17, 177–192.

### Biographical Notes

**Miss. Khurrath Ul Aien M.R** is presently pursuing M. Tech. final year in Computer Science and Engineering Department (Computer Science) from UBDTCE, Davangere, India.

**Mrs. Anitha G**.is working as a Assistant Professor in Computer Science and Engineering Department, from UBDTCE, Davangere, India.