Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 274–286

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

# A Novel Arabic Text-independent Speaker Verification System based on Fuzzy Hidden Markov Model

Rania M. Ghoniem[a]* and Khaled Shaalan[b]

[a]*Computer Department, Faculty of Specific Education, Mansoura University, Mansoura, Egypt*
[b]*Faculty of Engineering & IT, The British University in Dubai, UAE*

**Abstract**

The most important shortcoming of the current speaker verification methods (based on knowledge or possession) is that this process is not sure whether the holder of the given ID is the entitled one or an imposter. Using biometrics in the verification system is designated for minimizing this problem and decreasing the necessity of carrying the tokens. In this paper, a novel Arabic text-independent speaker verification system is presented. First of all, new speech features are proposed for speaker characterization, which denoted as Wavelet Packet Four-Directional Features (WPFDF). With the objective of speaker verification, the paper proposes a Fuzzy Hidden Markov Model, termed FHMM, where the kernel fuzzy c-means (KFCM) is extended to calculate fuzzy memberships of HMMs training samples. Thus, information loss is reduced as well as recognition rate is increased. The proposed approach reached 98.38% of recognition rate.


Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

*Keywords:* Speech recognition, Arabic text-independent speaker verification, Wavelet packet-four directional features, FHMM;

## 1. Introduction

Recently, using trustworthy authentication systems to recognize identity of legitimate user is becoming greatly influential in commercial applications, military, education, finance personnel security, hospitals, airports, digital

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
*E-mail address:* author@institute.xxx

1877-0509 
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.119

right management systems as well as several other necessary areas [1]. Actually, performance-based biometric systems through which a person is automatically recognized by him carrying out a pre-defined task employing his own biometrics, are favor over knowledge-based (e.g., passwords) or possession-based (e.g., keys) access control procedures [2,3]. Consequently, a number of systems based upon distinct physiological and behavioral traits have been evolved: fingerprint [4], face [5], iris [6], retina [7], and voice [8]. However, a joint drawback of some of such systems is their vulnerability to the potentiality to falsify these features [9]. Unlike other biometrics that employ fingerprints, static bio-signals or pictures to identify a person, human speeches carry not only information concerning the physiological characteristics of a person (correlates to the vocal cords structure) but also the behavioral ones: style of speaking, emotions or mood of the speaker. Furthermore, measuring the voice is entirely non-invasive and socially consented [10].

### *1.1. Automatic speaker verification (ASV)*

The objective of ASV systems is to determine whether a considered speech segment is certainly uttered by a alleged speaker [8]. They can be further classified into: (1) text-independent speaker verification (TISV) systems; and (2) text-dependent speaker verification (TDSV) systems, on the basis of whether we restrict the speech content throughout verification. For TDSV, speakers are permitted to pronounce only certain utterances or sentences that are aware to the system [11]. Such pre-defined sentences and words can be pre-recorded and playback by attacks. On the contrary, TISV can process spoken speech without restrictions, which may be user-chosen text or conversational speech. In contrast to TDSV, TISV is more flexible but also more challenging.

The classic process in majority of proposed ASV systems includes some form of signal pre-processing and analysis, followed by speaker modeling in order to estimate distributions of class-dependent features. Speaker identity can be characterized by: (1) spectral features; (2) prosodic features; and (3) high-level features. As far as the spectral features, they characterize the spectral envelope that is related acoustically to the voice timbre. Accordingly, mel-frequency cepstral coefficients (MFCCs) [12], linear predictive cepstral coefficients (LPCCs) [13] and perceptual linear prediction (PLP) [14] are frequently employed. On the other hand, prosodic features [15], like energy, pitch and duration, are less sensitive to channel influences. Nevertheless, because of their sparsity, extracting them necessitates large quantities of data of training, and algorithms of pitch estimation are usually uncertain in case of noisy environments [16]. Regarding high-level features [17], they are estimated from a lexicon to characterize speaker behavior or lexical cues. They own less sensitivity to noise influences than prosodic or spectral ones. Nevertheless, extracting such features necessitates complicated front-ends [18].

Otherwise, the goal of speaker modeling is characterizing a person who is enrolled to the ASV system in order to define a model. Furthermore, typical speaker models are divided into: (1) template models; and (2) stochastic models [16]. With respect to template models, training and testing phases are contrasted with each other on the basis of the supposition that each one represent an incomplete replica of the other. Vector quantization and dynamic time warping (DTW) [19] are popular examples of template models in case of TDSV as well as TISV. Regarding stochastic models, the speaker is modeled as a probabilistic origin using a probability density function that is unknown and fixed. The model training involves computing the probability density function parameters of the training samples. The process of matching is classically implemented by assessing the likelihood of the tested utterance with regard to the model. Accordingly, the Gaussian mixture model (GMM) [17] and the classical hidden Markov model (HMM) [20] are the most frequently utilized models for TDSV as well as TISV.

### *1.2. Literature review of TISV* systems

This section briefly debates the state-of-the-art on TISV. Summary of the feature extraction together with the modeling methods utilized for TISV are described in Table 1. Actually, there does not yet present entirely ‘‘best” feature extraction or modeling method for TISV. The selection is a trade-off among robustness, speaker discrimination, as well as practicality.