Expert Systems With Applications 90 (2017) 250–271

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Speaker identification features extraction methods: A systematic review

Sreenivas Sremath Tirumala[a,1], Seyed Reza Shahamiri[a,*], Abhimanyu Singh Garhwal[a,1], Ruili Wang[b,2]

[a] Faculty of Business and Information Technology, Manukau Institute of Technology, Auckland, New Zealand
[b] Computer Science and Information Technology, Institute of Natural and Mathematical Sciences (INMS), Massey University, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 12 May 2017
Revised 4 August 2017
Accepted 6 August 2017
Available online 16 August 2017

Keywords:
Feature extraction
Kitchenham systematic review
MFCC
Speaker identification
Speaker recognition

ABSTRACT

Speaker Identification (SI) is the process of identifying the speaker from a given utterance by comparing the voice biometrics of the utterance with those utterance models stored beforehand. SI technologies are taken a new direction due to the advances in artificial intelligence and have been used widely in various domains. Feature extraction is one of the most important aspects of SI, which significantly influences the SI process and performance. This systematic review is conducted to identify, compare, and analyze various feature extraction approaches, methods, and algorithms of SI to provide a reference on feature extraction approaches for SI applications and future studies. The review was conducted according to Kitchenham systematic review methodology and guidelines, and provides an in-depth analysis on proposals and implementations of SI feature extraction methods discussed in the literature between year 2011 and 2106. Three research questions were determined and an initial set of 535 publications were identified to answer the questions. After applying exclusion criteria 160 related publications were shortlisted and reviewed in this paper; these papers were considered to answer the research questions. Results indicate that pure Mel-Frequency Cepstral Coefficients (MFCCs) based feature extraction approaches have been used more than any other approach. Furthermore, other MFCC variations, such as MFCC fusion and cleansing approaches, are proven to be very popular as well. This study identified that the current SI research trend is to develop a robust universal SI framework to address the important problems of SI such as adaptability, complexity, multi-lingual recognition, and noise robustness. The results presented in this research are based on past publications, citations, and number of implementations with citations being most relevant. This paper also presents the general process of SI.

1. Introduction

Speech is a universal form of communication. Speaker Recognition (SR) is the process of identifying the speaker according to the vocal features of the given speech. This is different to speech recognition where the identification process is confined to the content rather than speaker. The process of SR is based on identifying and extracting unique characteristics of the speaker's speech. The characteristics of voices of the person is also known as voice biometrics.

A SR system is used to identify and distinguish speakers and extract unique characteristics that may be used for user verification or authentication. Speaker Identification (SI) is known as the process of identifying the speaker from a given utterance by comparing voice biometrics of the given sample of the speaker. When voice is used for authorization, it is termed as Speaker Verification. The key application area of SR is security and forensic science. SR systems are also used as a replacement for password and other user authentication processes (voiced password). Forensic science applies SR to compare the voice samples of the person claimed to be with other evidences obtained like telephone conversation or other recorded evidence. This process is also referred as speaker detection. The most important aspect of using SI systems is for automating processes like directing clients' mails to the right

* Corresponding author at: MIT Manukau, Cnr of Manukau Station Rd Davies Ave, Private Bag 94006, Manukau 2241, New Zealand.
E-mail addresses: ssremath@aut.ac.nz (S.S. Tirumala), admin@rezanet.com, rshahamiri@gmail.com, rshahamiri@yahoo.com (S.R. Shahamiri), abhimanyu garhwal@gmail.com (A.S. Garhwal), Ruili.wang@massey.ac.nz (R. Wang).
[1] MIT Manukau, Cnr of Manukau Station Rd Davies Ave, Manukau, Private Bag 94006, Manukau 2241, New Zealand.
[2] Room 3.10, IIMS Building, Albany Campus, Massey University, Albany, Auckland, New Zealand.

http://dx.doi.org/10.1016/j.eswa.2017.08.015
0957-4174/

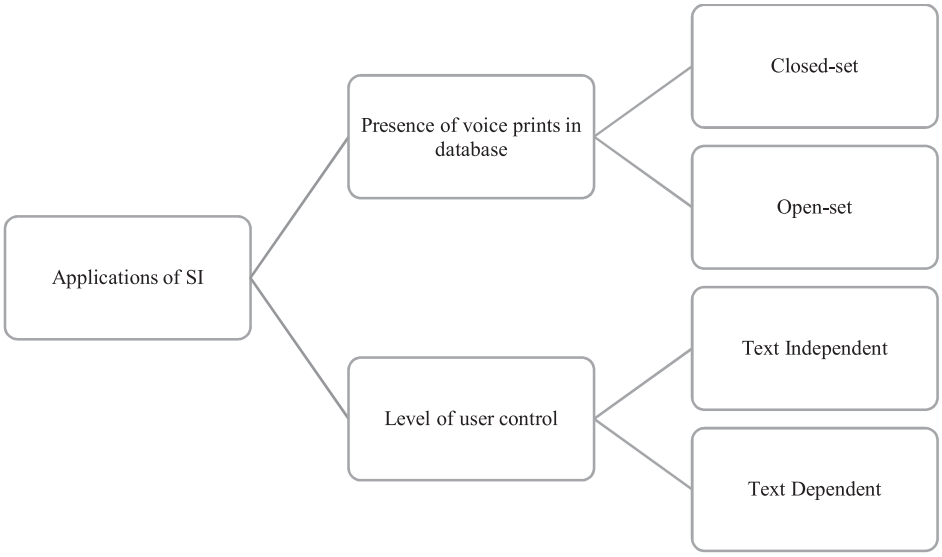

Fig. 1. Speaker identification applications classification.

mailbox, recognizing talkers in discussion, cautioning discourse acknowledgment frameworks of speaker changes, checking if a client is enlisted in the framework as of, and so on. These SI systems may work without the knowledge of a client's voice sample since they rely only on identifying an input speaker from the existing database of speakers.

Our systematic review is confined to SI as one of the primary types of SR systems (Reynolds, 2002). Feature extraction is one of the important SI aspects that significantly influences the quality of SI. In particular, the selection of proper feature extraction approaches plays a vital role since the identification is carried out by comparing unique characteristic features of a voice input. Therefore, the aim of this article is to carry out a systematic literature review on various feature extraction approaches of SI in order to:

(1) Identify significant feature extraction approaches in the last six years,
(2) Present a systematic review on the research of feature extraction approaches for SI,
(3) Classify various feature extraction approaches and provide recommendations based on the research.

The applications of SI can be classified in two types as presented in Fig. 1. The first type depends on the presence of voice prints in the database, which is further classified into two categories namely closed-set and open-set. In closed-set, the test speaker input is compared with the existing speakers' voice prints in the database and the nearest match is found (Dutta, Patgiri, Sarma, & Sarma, 2015). Hence, a closed-set SI guarantees a result although it may not be the exact speaker. On the other hand, in an open-set SI the input speaker voice print is compared with the database for 'exact match'; the input is rejected if the match is not found (Reynolds, 2002).

The second type of SI applications is based on the level of user control, which is also known as speaker verification process. This SI category is also further classified into two categories: text dependent and text independent. In text dependent the speaker must utter the same phrases or words that are previously used for training (Islam & Rahman, 2009; Kekre, Athawale, & Desai, 2011), while in the last category the input voice print content may not exist in the training set (Boujelbene, Mezghanni, & Ellouze, 2009; Revathi & Venkataramani, 2009; Verma, 2011).

The systematic review is carried out using Systematic Literature Review (SLR) methodology proposed by Kitchenham and Charters (2007) which is detailed in the methodology section. In this review, we presented various feature extraction approaches that were used in speaker identification processes and provide a systematic review on the research of these approaches. It is noteworthy to observe the key components of SI systems (which are detailed in the next sections) like parametrization (i.e. feature extraction), speaker modelling, pattern matching and scoring method that are core components for SR as well. This systematic review explained all SI components but more emphasis was put on feature extraction.

Since SR can be considered as a pattern recognition problem, various Artificial Intelligence (AI) approaches are used for SR systems (Rajesh et al., 2012). Deep Learning, which attained state of the art results for complex pattern recognition problems, has also been implemented for SR systems (Ghahabi & Hernando, 2014; McLaren, Lei, & Ferrer, 2015; Richardson, Reynolds, & Dehak, 2015b). Recent deep learning implementations for SR highlights the complexity involved in SR which requires special attention compared with traditional pattern recognition problems in general, and speech recognition problems in particular (Richardson, Reynolds, & Dehak, 2015a).

Existing review works and surveys on speaker recognition can be broadly categorized into three categories. The first category is the comprehensive surveys on SR that review the literature on generic SR processes and different SR categories (the SR categories are explained in the next section). There are numerous works in this aspect, such as El Ayadi, Kamel, and Karray (2011); Lawson, et al. (2011); Saquib, Salam, Nair, Pandey, & Joshi, 2010a). The second category mostly focuses on the types of statistical and machine learning approaches used as SR classifiers, for example Farrell, Mammone, and Assaleh (1994); Larcher, Lee, Ma, and Li (2014); Lippmann (1989). This category of SR reviews mostly falls under classification and machine learning research where there is considerable amount of literature available. In terms of speaker identification, the only work that specifically discussed SI and its processes is a brief survey presented by Sidorov, Schmitt, Zablotskiy, and Minker (2013) in which few generic SI methods were explained and compared.

The third category of speaker recognition surveys deals with feature extraction approaches in SR. One of the most recent SR feature extraction surveys is Dişken, Tüfekçi, Saribulut, and Çevik (2017) that concentrated on methods for extracting robust speaker