Speech Communication 86 (2017) 42–51

Contents lists available at ScienceDirect

Speech Communication

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

Speaker recognition based on Arabic phonemes

Mansour Alsulaiman[a,*], Awais Mahmood[b], Ghulam Muhammad[a]

[a] Computer Engineering Department, King Saud University, Riyadh, Saudi Arabia
[b] College of Computer and Information Sciences, King Saud University (Almuzahmiyah Branch), Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 16 November 2015
Accepted 11 November 2016
Available online 16 November 2016

Keywords:
Speech segmentation
Speaker recognition
Arabic phonemes
Effect of phonemes

ABSTRACT

In this paper, we investigate the effect of Arabic phonemes on the performance of speaker recognition systems. The investigation reveals that some Arabic phonemes have a strong effect on the recognition rate of such systems. The performance of speaker recognition systems can be improved and their execution time can be reduced by utilizing this finding. Additionally, this finding can be used by segmenting the most effective phonemes for speaker recognition from the utterance, using only the segmented part of the speech for speaker recognition. It can also be used in designing the text to be used in high-performance speaker recognition systems. From our investigation, we find that the recognition rates of Arabic vowels were all above 80%, whereas the recognition rates for the consonants varied from very low (14%) to very high (94%), with the latter achieved by a pharyngeal consonant followed by the two nasal phonemes, which achieved recognition rates above 80%. Four more consonants had recognition rates between 70% and 80%. We show that by utilizing these findings and by designing the text carefully, we can build a high-performance speaker recognition system.

1. Introduction

The speech signal carries important information such as message content, language, speaker identity, speaker emotion, speaker personality, and so on (Reynolds, 2002; Gonzalez-Rodriguez, 2014). Speech processing is divided into several areas, including analysis, synthesis, coding, and recognition, the latter of which may be further divided into different types such as speech recognition, speaker recognition (SR), and language recognition (Madisetti and Williams, 1999; Gonzalez-Rodriguez, 2014; Miguel et al., 2008; Lopez-Moreno et al., 2014).

In general, the process of extracting personal identity from the analysis of speech utterances is known as speaker recognition (Madisetti and Williams, 1999; Larcher et al., 2014). SR can be used as a biometric in many applications such as secure access voice control, information structuring, customizing services to individuals, and forensic investigation (Bimbot et al., 2004). SR technologies are expected to make our daily lives more convenient by creating new services through access control applications, including voice dialing, telephone banking, teleshopping, database voice access, reservation services, voicemail, and remote access to personal computers (Singh et al., 2012). With increasing computational power, it is probably only a matter of time before the use of SR technologies in games becomes practical.

1.1. Fundamentals of SR

SR refers to the process of automatically recognizing a person based on the information included in the speech signal. Systems for recognizing a speaker based on his/her speech characteristics are a focus of intense research efforts nowadays. Interest in SR has recently increased due to the growing use of speech technologies in various areas of daily life. Research efforts on SR largely focus on developing practical applications that can be divided into two classes. The first class of research is focused on controlling the access rights to different systems (information and material systems), and the second class is focused on the area of speech forensics (Kinnunen and Haizhou, 2010).

Speakers have different voices because of their vocal tract shapes, larynx sizes, and the other parts of their voice-producing organs (Kinnunen and Haizhou, 2010). The speech of a speaker carries his/her characteristics, which enables us to recognize that speaker. Although other forms of biometrics such as fingerprints and retinal scans are reliable means of identification, speech has the advantage that it is a non-invasive biometric, one which can be collected with or without the person's knowledge. Moreover, unlike other forms of identification, such as passwords or keys, a person's voice cannot be forgotten.

* Corresponding author.
E-mail addresses: msuliman@ksu.edu.sa (M. Alsulaiman), mawais@ksu.edu.sa (A. Mahmood), ghulam@ksu.edu.sa (G. Muhammad).

http://dx.doi.org/10.1016/j.specom.2016.11.004
0167-6393/

SR can be categorized into two classes, speaker identification and speaker verification. Based on the text of the speech, the SR system can then be classified into text-dependent SR and text-independent SR. For text-dependent SR, speakers are allowed to pronounce only specific sentences or words that are known to the system (Hebert, 2008). On the contrary, text-independent SR can process freely spoken speech, which is either user-selected text or conversational speech. Compared with text-dependent SR, text-independent SR is more flexible but also more challenging (Hebert, 2008).

SR is composed of two phases, the enrollment phase and the testing phase. In the enrollment phase, the system is trained by using the given speech examples, whereas in the testing phase, an unknown speech sample is provided to the system and the system then identifies or verifies the speaker.

The major components of SR are feature extraction, or front-end processing, and speaker modeling. Feature extraction converts the input speech signal into a suitable feature space that is fed to the modeling part. Most feature extraction techniques fall into two categories, modeling human voice production and modeling peripheral auditory hearing. In the first category, the most popular feature is linear prediction cepstral coefficients (LPCC), while in the second category, the most popular features are mel frequency cepstral coefficients (MFCC) and relative spectral perceptual linear predictive coefficients (RASTA-PLP). Different features have been proposed over the past decade (Lawson et al., 2011), each with pros and cons. Researchers have used these features for different kinds of speech processing such as SR, speech recognition, and language recognition.

MFCC is the most commonly used speech feature. In our work, we used MFCC and the two speech features introduced in our previous work (Mahmood et al., 2013, 2014): Multi Directional Local Feature (MDLF) and Multi Directional Local Feature with Moving Average (MDLF-MA). A brief description of these two features is given in Sections 2.1 and 2.2 (for a detailed description, see Mahmood et al., 2012, 2013, 2014). These features show better performance than conventional features (Mahmood et al., 2013, 2014) in automatic SR applications.

1.2. Literature review of Arabic SR

Arabic is the fifth most widely spoken language globally (UNESCO, 1995–2012), with approximately 362.5 million speakers (World Bank, 2016). Moreover, due to the status of Arabic as the language of the religion of Islam, many more speakers around the world have at least a passing knowledge of it. Arabic falls into the Semitic subgroup of the Afro-Asiatic languages, stemming from Proto-Semitic. A quite interesting attribute of Arabic is that it has preserved most of the original Proto-Semitic features until now. At present, Arabic is an official language in more than 22 countries. The formal standard language common to all Arabic-speaking countries is Modern Standard Arabic (MSA), which is a form of classical Arabic (El-Imam, 1989). MSA is used in the media, official speeches, educational institutions, and for formal communication. In addition to MSA, colloquial Arabic prevails in everyday conversation between Arabs.

In MSA, utterances must start with a consonant (Alkhouli, 1990; Alghamdi, 2015), and all Arabic syllables must contain at least one vowel. In addition, while Arabic vowels cannot occur in an initial position in the word, they can occur between two consonants or in the final position of a word. The Arabic language consists of syllables, the shortest of which follow the form CV ("C" stands for a consonant, while "V" represents the vowel, so "CV" represents a consonant followed by a vowel). Other longer syllables include CVC and CVCC. The vowel can be short (v) or long (v:). The syllable Cv:CC occurs only within a word. Syllables can also be classified as open or closed; an open syllable ends with a vowel, while a closed syllable ends with a consonant (Holes, 2004; Alghamdi, 2015). In Arabic, a vowel always forms the nucleus of a syllable; there are as many syllables within a word as there are vowels (Alghamdi, 2015).

Table 1
IPA symbols for Arabic phonemes

Arabic transcript	IPA	Arabic transcript	IPA	Arabic transcript	IPA	Arabic transcript	IPA
ء	/ʔ/	د	/d/	ض	/dˤ/	ق	/q/
ب	/b/	ذ	/ð/	ط	/tˤ/	ل	/l/
ت	/t/	ر	/r/	ظ	/ðˤ/	م	/m/
ث	/θ/	ز	/z/	ع	/ʕ/	ن	/n/
ج	/ʒ/	س	/s/	غ	/ɣ/	و	/w/
ح	/ħ/	ش	/ʃ/	ف	/f/	ه	/h/
خ	/x/	ص	/sˤ/	ك	/k/	ي	/j/

The Arabic alphabet consists of 28 consonants and six vowels (three long and three short vowels) (Alghamdi, 2015; Alotaibi and Muhammad, 2010; Versteegh, 2014; Jalt, 2016). The three short vowels are represented by the phonemes /i/, /a/, and /u/, and the three long vowels are represented by the phonemes /i:/, /a:/, and /u:/ (Versteegh, 2014). Table 1 shows the IPA for each phoneme and the corresponding Arabic script (Alghamdi, 2015; Versteegh, 2014; Jalt, 2016). The pharyngeal voiced fricative /ʕ/, although characterized as a "fricative" in the IPA, has variable productions in Arabic, ranging from an approximant, to a stop, to a fricative.

A number of studies on Arabic SR have been published in related literature (Alsulaiman et al., 2009b; Alotaibi et al., 2009); however, we can hardly find any work on phoneme-based Arabic SR. In other words, there is a need to investigate the effect of different phonemes on Arabic SR. Table 2 gives a summary of the research on Arabic SR that we surveyed. All the mentioned work in this table did not investigate the effect of phonemes on Arabic SR.

1.3. Literature review of phoneme-based SR

Antal (2008) used the TIMIT English database and divided all phonemes into five categories: vowels, semivowels, nasals, fricatives/affricatives, and stops. The author found that the highest recognition rate (RR) is achieved with vowels, commenting that some broad phonetic classes are more speaker-specific than others. Savic and Sorensen (1992) showed that the phoneme /i/ achieves the highest SR rate, using the TIMIT database.

Imperl et al. (1996), using the SNABI database, made an SR system for all vowels for 20 speakers and found that /a/ achieves the highest RR compared with the /e/, /i/, /o/, and /u/ vowels. The lowest RR is achieved with /u/. Fattah performed phoneme-based SR using the YOHO speech database and found that \iy\ achieves the highest RR (Fattah et al., 2006). Except for Antal (2008), work on phoneme-based SR has concentrated on vowels. To the best of our knowledge, no work has been carried out on phoneme-based Arabic SR.

In this paper, we look at the effect of Arabic consonants and vowels on the performance of Arabic SR. The methodology is presented in Section 2. Section 3 describes the database and selected utterances, and the results are given in Section 4. We discuss and analyze these results in Section 5 and conclude the paper in Section 6.

2. Methodology

We investigate which phonemes are better at deciding the identity of the speaker, as this may help in designing texts to be used in SR. The proposed technique can be used for both text-dependent