Signal Processing 88 (2008) 2700–2708

Contents lists available at ScienceDirect

Signal Processing

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

Speaker identification in the shouted environment using Suprasegmental Hidden Markov Models

Ismail Shahin *

Electrical and Computer Engineering Department, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates

ARTICLE INFO

Article history:
Received 13 December 2007
Received in revised form
13 March 2008
Accepted 21 May 2008
Available online 28 May 2008

Keywords:
Hidden Markov Models
Second-Order Circular Hidden Markov Models
Shouted environment
Speaker identification
Suprasegmental Hidden Markov Models

ABSTRACT

In this paper, Suprasegmental Hidden Markov Models (SPHMMs) have been used to enhance the recognition performance of text-dependent speaker identification in the shouted environment. Our speech database consists of two databases: our collected database and the Speech Under Simulated and Actual Stress (SUSAS) database. Our results show that SPHMMs significantly enhance speaker identification performance compared to Second-Order Circular Hidden Markov Models (CHMM2s) in the shouted environment. Using our collected database, speaker identification performance in this environment is 68% and 75% based on CHMM2s and SPHMMs, respectively. Using the SUSAS database, speaker identification performance in the same environment is 71% and 79% based on CHMM2s and SPHMMs, respectively.

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individuality information in speech waves. Speaker recognition systems come in two flavors: speaker identification systems and speaker authentication (verification) systems.

Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker identification systems can be used in criminal investigations to determine the suspected persons who produced the voice recorded at the scene of the crime [1]. Speaker identification systems can also be used in civil cases or for the media. These cases include calls to radio stations, local or other government authorities, insurance companies, or recorded conversations, and many other applications.

Speaker authentication is the process of determining whether a speaker corresponds to a particular known voice or to some other unknown voice. The applications of speaker authentication systems involve the use of voice as a key to confirm the identity claim of a speaker. Such services include banking transactions using a telephone network, database access services, security control for confidential information areas, remote access to computers, tracking speakers in a conversation or broadcast, and many other areas.

Speaker recognition systems typically operate in one of two cases: text-dependent (fixed text) case or text-independent (free-text) case. In the text-dependent case, utterances of the same text are used for both training and testing (recognition). On the other hand, in the text-independent case, training and testing involve utterances from different texts.

The process of speaker recognition can be divided into two categories: "open set" and "closed set". In the "open set" category, a reference model for the unknown speaker may not exist; whereas, in the "closed set" category, a reference model for the unknown speaker should be available to the system.

2. Motivation

The majority of researchers who work in the areas of speech recognition and speaker recognition focus their

* Tel.: +971 6 5050967; fax: +971 6 5050877.
E-mail address: ismail@sharjah.ac.ae

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.05.012

work on speech under the neutral talking condition and the minority of the researchers focus their work on speech under the stressful talking conditions. The neutral talking condition is defined as the talking condition in which speech is produced assuming that speakers are in a "quiet room" with no task obligations. The stressful talking conditions can be defined as the talking conditions that cause speakers to vary their production of speech from the neutral talking condition.

Some talking conditions are designed to simulate speech produced by different speakers under real stressful talking conditions. Hansen, Cummings, Clements, Bou-Ghazale, Zhou, and Kaiser used the SUSAS (Speech Under Simulated and Actual Stress) database in which eight talking conditions are used to simulate speech produced under real stressful talking conditions and three real talking conditions [2–4]. The eight talking conditions are: neutral, loud, soft, angry, fast, slow, clear, and question. The three real talking conditions are: 50% task (cond50), 70% task (cond70), and Lombard. The 50% task and the 70% task comprise utterances recorded from subjects engaged in tracking tasks under different levels of workload (the level of workload is higher in 70% than that in 50%). The Lombard effect occurs when speakers vary their speech characteristics in order to increase intelligibility when speaking in a noisy environment. Chen used six talking conditions to simulate speech under real stressful talking conditions [5]. These conditions are: neutral, fast, loud, Lombard, soft, and shouted.

Very few researchers who focus their work on speech under stressful talking conditions consider studying speech under the shouted talking condition [5–8]. Therefore, the number of publications in the areas of speech recognition and speaker recognition under this talking condition is very limited. The shouted talking condition can be defined as when speakers shout, their intention is to produce a very loud acoustic signal, either to increase its range (distance) of transmission or its ratio to background noise.

Speaker identification systems under the shouted talking condition can be used in the applications of talking condition identification systems. Such systems can be used in medical applications where computerized stress classification and assessment techniques can be employed by psychiatrists to aid in quantitative objective assessment of patients who undergo evaluation. These systems can also be used in the applications of talking condition intelligent automated systems in call-centers. It is very important for call-centers to take note of customers' disputes using talking condition intelligent automated systems and successfully respond to these disputes to obtain the customers' satisfaction.

It is well known that the recognition performance of speech recognition and speaker recognition systems is almost perfect under the neutral talking condition. However, the performance is degraded sharply under the shouted talking condition. Many studies show that the performance of speech recognition and speaker recognition systems under this talking condition is deteriorated significantly [4–8].

Speaker identification performance under the shouted talking condition is very low based on Hidden Markov Models (HMMs) [5,7,8]. In previous study, Shahin focused on enhancing the recognition performance of text-dependent speaker identification systems under the shouted talking condition based on each of Second-Order Hidden Markov Models (HMM2s) and second-order Circular Hidden Markov Models (CHMM2s) [7,8].

Our work in this research differs from the work in Ref. [8] is that our work in this research focuses on enhancing the recognition performance of text-dependent speaker identification in the shouted environment based on Suprasegmental Hidden Markov Models (SPHMMs) using each of our collected speech database and the SUSAS database. On the other hand, the work in Ref. [8] focused on enhancing speaker identification performance in the same environment based on CHMM2s. We can claim that this is the first time to use SPHMMs for speaker identification in such an environment.

This paper is organized as follows. Section 3 overviews: Hidden Markov Models (HMMs), First-Order Hidden Markov Models (HMM1s), Second-Order Hidden Markov Models (HMM2s), Circular Hidden Markov Models (CHMMs), and Second-Order Circular Hidden Markov Models (CHMM2s). Section 4 discusses the details of SPHMMs. Section 5 describes the collected speech database used. The algorithm of speaker identification based on each of CHMM2s and SPHMMs is given in Section 6. Section 7 discusses the results that are obtained in this work. Concluding remarks are drawn in Section 8.

3. Overview of HMMs, HMM1s, HMM2s, CHMMs, and CHMM2s

3.1. Hidden Markov Models

The use of HMMs in the fields of speech recognition, speaker recognition, and emotion recognition has become popular in the last three decades. HMMs have become one of the most successful and broadly used modeling techniques in the three fields [3–5,9–13].

Bou-Ghazale and Hansen used HMMs in the study of evaluating the effectiveness of traditional features in recognition of speech under stress and formulating new features which are shown to enhance stressed speech recognition [3]. Zhou et al. applied HMMs in the study of nonlinear feature based classification of speech under stress [4]. Chen studied talker-stress-induced intraword variability and an algorithm that compensates for the systematic changes observed based on HMMs trained by speech tokens in various talking styles [5]. Nwe et al. exploited HMMs in the text independent method of emotion classification of speech [11]. Ververidis and Kotropoulos made use of HMMs in the classification techniques that classify speech into emotional states [12]. Bosch used HMMs to recognize emotion from the speech signal, from the viewpoint of automatic speech recognition (ASR) [13].

HMMs use Markov chain to model the changing statistical characteristics that exist in the actual observations of