Engineering Applications of Artificial Intelligence 35 (2014) 316–323

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Novel third-order hidden Markov models for speaker identification in shouted talking environments

Ismail Shahin *

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

ARTICLE INFO

Article history:
Received 27 October 2013
Received in revised form
9 July 2014
Accepted 9 July 2014
Available online 2 August 2014

Keywords:
First-order hidden Markov models
Second-order hidden Markov models
Shouted talking environments
Speaker identification
Third-order hidden Markov models

ABSTRACT

Speaker identification systems perform almost perfectly in neutral talking environments; however, they perform poorly in shouted talking environments. This work aims at proposing, implementing, and evaluating novel models called Third-Order Hidden Markov Models (HMM3s) to enhance the poor performance of text-independent speaker identification systems in shouted talking environments. The proposed models have been evaluated on our collected speech database using Mel-Frequency Cepstral Coefficients (MFCCs). Our results show that HMM3s significantly improve speaker identification performance in shouted talking environments compared to second-order hidden Markov models (HMM2s) and first-order hidden Markov models (HMM1s) by 12.4% and 202.4%, respectively. The achieved results based on the proposed models are close to those obtained in subjective assessment by human listeners.

1. Introduction

Neutral talking environments can be defined as the talking environments in which speakers utter their speech in a "quiet room" with no task obligations. Stressful talking environments can be defined as the talking environments in which speakers vary their production of speech from neutral talking environments.

Some talking environments are designed to simulate speech generated by different speakers under real stressful talking conditions. Many studies (Cummings and Clements, 1995; Bou-Ghazale and Hansen, 2000; Zhou et al., 2001) used Speech Under Simulated and Actual Stress (SUSAS) database in which eight talking conditions are used to simulate speech produced under real stressful talking conditions and three real talking conditions. The eight talking conditions are neutral, loud, soft, angry, fast, slow, clear, and question. The three talking conditions are 50% task, 70% task, and Lombard. Chen (1988) used six talking environments to simulate speech under real stressful talking environments. These environments are neutral, fast, loud, Lombard, soft, and shouted. Shouted talking environments can be defined as when speakers shout, their intention is to create a very loud acoustic signal, either to increase its range of transmission or its ratio to background noise.

Speaker recognition has two branches: speaker identification and speaker verification (authentication). Speaker identification is the process of automatically determining who is speaking from a set of known speakers. Speaker verification is the process of automatically accepting or rejecting the identity of the claimed speaker. Speaker identification can be used in criminal investigations to determine the suspected persons who uttered the voice recorded at the scene of the crime. Speaker identification can also be used in civil cases or for the media. Speaker verification is heavily used in security access to services via a telephone, including home shopping, home banking transactions using a telephone network, security control for confidential information areas, remote access to computers, and many telecommunication services (Furui, 1991). Based on the text to be spoken, speaker recognition is grouped into text-dependent and text-independent cases. In the text-dependent case, speaker recognition requires the speaker to utter speech for the same text in both training and testing; on the other hand, in the text-independent case, speaker recognition does not depend on the text being spoken.

In this work, we address the issue of enhancing the low performance of text-independent speaker identification in shouted talking environments by proposing, implementing, and testing a novel classifier. This classifier is called Third-Order Hidden Markov Models (HMM3s).

2. Motivation and prior work

The areas of speech recognition and speaker recognition have received considerable interest in the literature. Most studies,

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

http://dx.doi.org/10.1016/j.engappai.2014.07.006
0952-1976/

however, have focused on studying the two areas in neutral talking environments (Furui, 1991; Farrell et al., 1994; Yu et al., 1995; Reynolds, 1995). In fact, these two areas have received less attention in stressful talking environments (Cummings and Clements, 1995; Bou-Ghazale and Hansen, 2000; Zhou et al., 2001; Chen, 1988) (especially shouted (Chen, 1988; Raja and Dandapat, 2010; Zhang and Hansen, 2007; Shahin, 2005, 2006, 2008, 2010)).

Chen (1988) studied talker-stress-induced intraword variability and an algorithm that pays off for the systematic changes observed based on hidden Markov models (HMMs) trained by speech tokens under diverse talking conditions. Raja and Dandapat (2010) studied speaker recognition under stressed conditions to improve the declined performance under such conditions. They used four different stressed conditions of SUSAS database. These conditions are neutral, angry, Lombard, and question. Their study (Raja and Dandapat, 2010) showed that the least speaker identification performance took place when speakers talk in angry talking environments. Angry talking environments are used as alternatives to shouted talking environments since they cannot be entirely separated from shouted talking environments in our real life (Shahin, 2005, 2006, 2008, 2010). Zhang and Hansen, (2007) reported on the analysis of characteristics of speech in five different vocal modes: whispered, soft, neutral, loud, and shouted; and to recognize discriminating features of speech modes. In four of his earlier studies (Shahin, 2005, 2006, 2008, 2010), Shahin focused on enhancing speaker identification performance in shouted talking environments based on each of Second-Order Hidden Markov Models (HMM2s) (Shahin, 2005), Second-Order Circular Hidden Markov Models (CHMM2s) (Shahin, 2006), Suprasegmental Hidden Markov Models (SPHMMs) (Shahin, 2008), and Second-Order Circular Suprasegmental Hidden Markov Models (CSPHMM2s) (Shahin, 2010). The achieved speaker identification performance in such talking environments is 59.0%, 72.0%, 75.0%, and 83.4% based on HMM2s, CHMM2s, SPHMMs, and CSPHMM2s, respectively (Shahin, 2005, 2006, 2008, 2010).

HMMs are powerful models in optimizing the parameters that have been used to model speech signals. This optimization reduces the computational complexity in the decoding procedure and enhances the recognition accuracy (Huang et al., 1990). Most of the works carried out in the fields of speech recognition and speaker recognition based on HMMs have been conducted using First-Order Hidden Markov Models (HMM1s) (Chen, 1988), (Juang and Rabiner, 1991; Dai, 1995; Rabiner, 1989). In HMM1s, the state-transition probability at time $t+1$ depends only on the state of the Markov chain at time t. These models yield extremely high speaker recognition performance in neutral talking environments (Chen, 1988; Shahin, 2005, 2010); however, the models give very low performance in shouted talking environments (Chen, 1988; Shahin, 2005, 2010).

Mari et al. (1996), (1997) proposed, applied, and tested HMM2s in the training and testing phases of a connected word recognition system under the neutral talking condition. In such models, the underlying state sequence is a second-order Markov chain where the state-transition probability at time $t+1$ depends on the states of the Markov chain at times t and $t-1$. Shahin (2005) exploited these models in the training and testing phases of isolated-word text-dependent speaker identification systems under each of the neutral and shouted talking conditions. Based on his work and using HMM2s, Shahin (2005) achieved higher speaker identification performance than that using HMM1s under the shouted talking condition. Hadar and Messer (2009) proposed a simple method based on transforming any high order HMM (including models in which the state sequence and observation dependency are of distinct orders) into an equivalent first order HMM. Chatzis (2013) focused in one of his works on designing infinite-order HMMs to learn from data with sequential dynamics. These models usually depend on the postulation of first-order Markovian dependencies between the consecutive label values y. There are two major advantages of the designed models over the other approaches. The first advantage is that these models allow for capturing very long and complex temporal dependencies. The second advantage is that the models employ a margin maximization paradigm to perform model training, which gives a convex optimization scheme (Chatzis, 2013).

In this work, we focus on further improving (compared to HMM2s) the performance of text-independent speaker identification in shouted talking environments by proposing, implementing, and evaluating novel models called HMM3s. In these new models, the underlying state sequence is a third-order Markov chain where the state-transition probability at time $t+1$ depends on the states of the Markov chain at times t, $t-1$, and $t-2$. Speaker recognition in shouted talking environments can be used in criminal investigations to identify the suspected persons who uttered shouted voice during crimes and in the applications of talking condition recognition. Talking condition recognition can be used in medical applications, telecommunications, law enforcement, and military applications (Hansen et al., 2000). The proposed models have been assessed on our collected speech database. Our approach in this work is different from that in the work of reference (Hadar and Messer, 2009). In the current work, our approach does not depend on transforming HMM3s into equivalent HMM1s. Our present work does not also depend on designing HMM3s that learn from data with sequential dynamics as in the work of reference (Chatzis, 2013).

The remainder of this paper is organized as follows: Brief overview of hidden Markov models is given in Section 3. The details of the proposed third-order hidden Markov models are covered in Section 4. Section 5 describes the collected speech database used in this work and the extraction of features. Section 6 discusses speaker identification algorithm based on HMM3s and the experiments. Section 7 demonstrates the results obtained in the present work and their discussion. Finally, concluding remarks are presented in Section 8.

3. Brief overview of hidden Markov models

HMMs can be described as being in one of the N different states: 1, 2, 3, …, N, at any discrete time instant t. The individual states are denoted as (Huang et al., 1990; Juang and Rabiner, 1991),

$$s = \{s_1, s_2, s_3, \ldots, s_N\}$$

which are generators of a state sequence q_t, where at any time t: $q=\{q_1, q_2, \ldots, q_T\}$, T is the length or duration of an observation sequence O and is equal to the total number of frames of a speech signal.

At any discrete time t, the model is in a state q_t. At the discrete time t, the model makes a random transition to a state q_{t+1}. The state transition probability matrix **A** determines the probability of the next transition between states,

$$\mathbf{A} = [a_{ij}] \quad i,j = 1,2,\ldots,N$$

where a_{ij} denotes the transition probability from a state i to a state j.

The first state s_1 is randomly chosen according to the initial state probability,

$$\pi = [\pi_i] = Prob\ (q_1 = s_i)$$

The states that are unobservable directly are observable via a sequence of outputs or an observation sequence given as,

$$O = \{O_1, O_2, O_3, \ldots, O_T\}$$