Digital Signal Processing 23 (2013) 1910–1917

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Constrained temporal structure for text-dependent speaker verification

Anthony Larcher [a,*,1], Jean-Francois Bonastre [a], John S.D. Mason [b]

[a] University of Avignon, LIA-CERI, 84911 Avignon Cedex 9, France
[b] Speech and Image Research, School of Engineering, Swansea University, Swansea SA2 8PP, UK

ARTICLE INFO

Article history:
Available online 1 August 2013

Keywords:
Speaker recognition
Text-dependent
Password
Embedded application

ABSTRACT

In the context of mobile devices, speaker recognition engines may suffer from ergonomic constraints and limited amount of computing resources. Even if they prove their efficiency in classical contexts, GMM/UBM systems show their limitations when restricting the quantity of speech data. In contrast, the proposed GMM/UBM extension addresses situations characterised by limited enrolment data and only the computing power typically found on modern mobile devices. A key contribution comes from the harnessing of the temporal structure of speech using client-customised pass-phrases and new Markov model structures. Additional temporal information is then used to enhance discrimination with Viterbi decoding, increasing the gap between client and imposter scores. Experiments on the MyIdea database are presented with a standard GMM/UBM configuration acting as a benchmark. When imposters do not know the client pass-phrase, a relative gain of up to 65% in terms of EER is achieved over the GMM/UBM baseline configuration. The results clearly highlight the potential of this new approach, with a good balance between complexity and recognition accuracy.

1. Introduction

The speech signal offers several advantages over other biometric signals, with distinct benefits coming from the potential to link together information derived from the context and the content of the message as well as the voice biometric itself. With appropriate classification and fusion, these components can be brought together to enhance any biometric validation process.

There are however practical constraints within the cell-phone scenario, stemming largely from ergonomic factors and the available computing resources typically found within such hand-held devices. Such embedded applications impose constraints and an important one in terms of recognition performance can be the quantity of data, particularly for reference models but also for the subsequent test phase. In certain applications these quantities might prove to have a critical influence on recognition accuracy. Examples include security systems that might well have speech of only a few words spanning just 2 or 3 seconds. The main contributions of this paper address these issues with new computational structures designed to harness maximum information from the temporal structure information (TSI) of speech to reinforce the acoustic modelling.

Classical speaker recognition engines offer a high level of performance as shown for example during NIST evaluations [1]. However such systems, which are invariably founded on the GMM/UBM paradigm [2], exhibit high sensitivity to the quantity of data, particularly the reference model data [3–5]. Their performance degrades strongly while reducing the duration of speech material available [6–8]. For situations where the speech duration is below 30 seconds, recognition performance falls rapidly [9,10]. Text-dependency is well known to compensate for the lack of data by constraining the acoustic content of the spoken utterance [11].

Meaningful comparison of recognition accuracy in text-dependent speaker verification tends to be very difficult due to the lack of controlled evaluations and large scale databases, essential particularly when error rates are very low [12]. Hence the tendency of the community towards the text-independent scenario that benefits from the NIST large scale databases and the independent evaluations [1]. However, two major trends dominate the field of text-dependent speaker verification. Approaches based on dynamic programming have been proposed for tasks where the quantity of speech is limited [13–16]. They provide a precise modelling of the time constraints but lack the generalization power available with hidden Markov model (HMM) approaches [17]. Indeed, HMM and GMM models which are the most common modelling methods [18] are more robust to speaker or environment variabilities and can take advantage of larger amounts of data [11]. Depending on the type of application that is targeted, HMMs can be used to model whole sentences [19,20], word-level units [21,22] or phone-level units [23,24]. In addition to the two major approaches that dominate text-dependent speaker verification, text-dependent and

* Corresponding author.
E-mail addresses: anthony.larcher@univ-avignon.fr, alarcher@i2r.a-star.edu.sg (A. Larcher), jean-francois.bonastre@univ-avignon.fr (J.-F. Bonastre), j.s.d.mason@swan.ac.uk (J.S.D. Mason).
[1] Current address: Institute for Infocomm Research, Singapore.

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2013.07.007

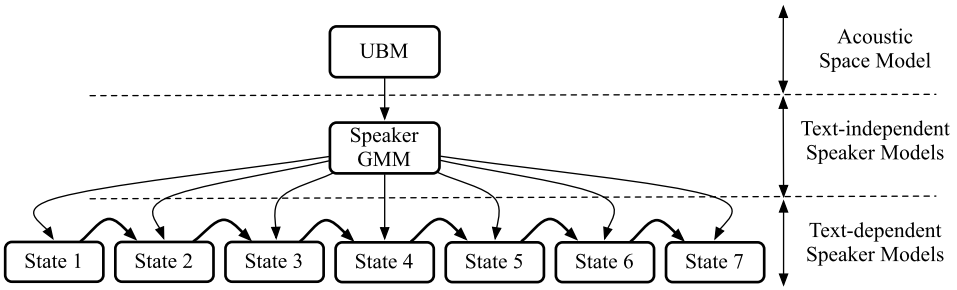

Fig. 1. General view of the EBD architecture.

text-independent speaker verification often cross-pollinate each other. While there have been several attempts to adapt Support Vector Machines [25,26] or i-vector systems [27,28] to take advantages of a lexical constraints, others incorporate text-dependent techniques in text-independent applications such as [29,30].

In this paper, a classification structure that takes advantage of the temporal structure of the speech utterance is examined. The structure utilises text-dependencies derived from a multilayer classifier illustrated in Fig. 1, the foundation of which is the standard GMM/UBM. These first two layers are complemented by a third layer that harnesses temporal structure information extracted from speaker-specific phrases. The approach, described in [31–33] and further developed in this paper, takes advantage of the temporal structure of pass-phrases, an example of which is "Ce petit canard apprend à nager".[2] In order to model the TSI of such a pass-phrase while achieving statistical modelling from the GMM/UBM, we propose to extend the standard paradigm with an HMM/Viterbi approach. Finite-state models aim to incorporate pass-phrase-based information, like temporal organization of acoustic features, not otherwise harnessed by classical GMM/UBM approaches.

A key point here is the inclusion of additional temporal information within the finite-state modelling. This additional information is used to constrain the Viterbi decoding in order to enhance discrimination. It does so by using the temporal structure of the given pass-phrase. A set of N classical HMM nodes is arranged in time sequence with the transitions in time from one node to the next controlled first by the normal acoustic features and then by additional temporal information.

The proposed structure is designed specifically to accommodate the use of two such simultaneous synchronous signals. The roles of the two can be clearly separated: first, variants of the conventional GMM/UBM nodes; and second, additional synchronization control of state transitions. Here, the latter comes from the acoustic signal and divides each pass-phrase into segments overarching several states of the HMM, as shown in Fig. 2. These overarching segments provide constraints at the lexical level that are in addition to those of the finite-state models and that can be harnessed by the recognizer. We refer to these as lexical constraints.

The approach proposed in this work is related to that of others in the literature. For example in [34] Becerra Yoma and Facco Pegoraro constrain the state duration of word-units HMMs. Additional knowledge is included in the HMM topology by training different transition probabilities depending on the position of a given word in the speech segment. In [35], speaker-dependent semi-continuous HMMs are compared to a reference HMM to produce a discriminative representation of the speaker pronouncing a given pass-phrase. In the two previous works, the use of a background HMM to adapt the speaker-dependent models or to model the alternative hypothesis strongly limits the flexibility of the system in terms of lexicon. The architecture proposed in this work takes advantages of the GMM/UBM framework to model the alternative hypothesis and adapt the speaker model and thus, gives more flexibility to the user to choose a specific pass-phrase. Another related work is proposed in [36] where supra-segmental temporal information is used to reinforce the robustness of a Dynamic Time Warping algorithm. By combining the different information sources in a later stage, using a neural network, the system does not take advantage of the temporal synchronization of the different signals as is the case in the work presented here.

Many techniques exist in the literature to compensate for the variabilities due to channel or environment mismatch in the GMM/UBM framework. Some of these techniques like RASTA and Short Term Gaussianization work at the parametrization level [37,38] when others are dedicated to score normalization [39,40]. These techniques have not been applied in this article which focus on the advantages of our approach compared to the GMM/UBM. Nevertheless, most of the techniques that have been developed for the GMM/UBM may be applied to our approach and are expected to provide similar improvement.

The overall system architecture is described in Section 2. The impact of the lexical information in constraining the Viterbi decoding is described in Section 3. Section 4 describes the experimental protocol and results. It includes a description of the MyIdea database [41]. Section 5 summarizes the benefits of this approach and presents future work directions.

2. A three-level acoustic architecture

The architecture presented in Fig. 1 is an extension of the standard GMM/UBM paradigm. Throughout the text we refer to this new structure as the Embedded LIA_SpkDet[3] (EBD) [42]. This architecture is configured to deal with a user-customized speaker recognition task. Each client has a customized pass-phrase, which is unique to that person. Hence some form of text-dependency can be harnessed within the speaker recognition system.

2.1. Training phase

The two first layers of the EBD consist of a classical GMM/UBM speaker recognition system. The upper layer, a standard universal background model (UBM), aims to model the acoustic speech space. This GMM is built off-line using a suitably large amount of data and the classical EM/ML algorithm [2]. A text-independent speaker-specific GMM (2nd layer) is then adapted from the UBM for each client speaker with the client data using the EM algorithm and the maximum a posteriori (MAP) criterion [43].

Finally, a semi-continuous hidden Markov model (SCHMM) [44] is used with the goal of harnessing the TSI of the individual pass-phrase. This third layer introduces text-dependency into the client

[2] This little duck is learning to swim.

[3] Implementation of the EBD is based on the open-source toolkit LIA_SpkDet, part of ALIZE, http://alize.univ-avignon.fr.