Pattern Recognition Letters 30 (2009) 589–594

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

α-Gaussian mixture modelling for speaker recognition

Dalei Wu *, Ji Li, Haiqing Wu

Department of Computer Science and Engineering, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3

ARTICLE INFO

Article history:
Received 16 December 2007
Received in revised form 15 December 2008
Available online 9 January 2009

Communicated by O. Siohan

Keywords:
Gaussian mixture model (GMM)
α-GMM
Speaker recognition

ABSTRACT

Gaussian mixture model is the conventional approach employed in speaker recognition tasks. Although it is efficient to model specific speaking characteristics of a speaker, especially in quiet environments, its performance in noisy conditions is still far from the human cognitive process. Recently, a new method of α-integration of stochastic models has been proposed based on psychophysical experiments that suggests α-integration is used in a human brain. In this paper, we proposed a method to extend the conventional GMM to the α-integrated GMM (α-GMM) to model personal speaking traits. Model parameters were re-estimated recursively based on a given data set. The experiments showed that the new approach significantly outperforms the traditional method, especially on telephony speech.

1. Introduction

The traditional Gaussian mixture model (GMM) has been proved to be effective to model personal speaking traits, particularly in clean speech. However, like most of the other modelling techniques, its performance significantly degrades in low band and noisy conditions (Wu et al., 2005a,b,c; Reynolds, 1995c), primarily due to a mismatch between training and test conditions. On the contrary, a human auditory and cognitive system is able to handle a variety of difficult conditions more gracefully and intelligently.

In order to deal with the recognition problem in noisy conditions for automatic speaker recognition, a series of methods have been proposed in last decades primarily from three directions: processing at the feature level, processing at the model level and processing at the score level. Processing at the feature level is often referred to as preprocessing techniques at the feature level that is the earliest stage of a recognition system. The most representative methods include some normalisation methods at the feature level, such as cepstral mean subtraction (CMS), H-norm, etc. (Wu et al., 2008) and transformation techniques at the feature level, such as LDA-based and NLDA-based feature transformation (Wu et al., 2005a,b,c, 2008). Processing at the model level then uses a variety of normalisation and transformation techniques at the model level that is a middle stage of a standard recognition system. The most representatives are model adaptation (Reynolds et al., 1995a,b,c) and model combination (Falthhauser and Ruske, 2001). The third class of methods are focused on carrying out some normalisation or transformation at the score level among which are Z-norm, T-norm and so on. For more details of the discussion of these methods at three different levels, the readers can refer to an overview paper on this topic (Wu et al., 2008).

The method proposed in this paper can in fact be classified into the second group, i.e., processing at the model level. As mentioned above, in this group, the most popular method is model adaptation (Reynolds et al., 2000). In this method, model parameters are re-estimated using new data obtained under an incoming environment so as to reduce a mismatch between the training and test conditions. However, α-GMM proposed in this article is completely different from the method of model adaptation, though they belong to the same group. The essence of the idea of α-GMM is its extension of the classical GMM in modelling capacity. By introducing a new factor of α, more powerful integrated models can be used. As we mentioned in Wu (in press), the most powerful feature of α-GMM is its property of α-warping, i.e., by using negative α-values, the integrated α-GMM de-emphasises small values but emphasises large-values and this property is beneficial to tolerate a mismatch between training and testing.

α-GMM re-considers the procedure of information integration. Information integration is one of the commonest phenomena in human brains. It is very crucial to understand this process in order to develop realistic artificial intelligence, pattern recognition and in particular speaker recognition. Among all the integration methods that could be used in a brain, the simplest method is linear combination, which is in fact the way employed in the conventional GMM approach. Even though linear combination might be used in an integration process occurring in a human brain, there must be other

* Corresponding author. Tel.: +1 416 637 5275; fax: +1 416 736 2100.
E-mail addresses: daleiwu@gmail.com (D. Wu), jili@gmail.com (J. Li), haiqing-wu@gmail.com (H. Wu).
URL: http://www.cse.yorku.ca/~daleiwu/ (D. Wu).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.12.013

more complex ways of information integration employed by a human brain to fulfil advanced human intelligence. Among these, α-integration might be worth to investigate. Meanwhile, psychophysical experiments have recently suggested that α-integration is used in human brains, instead of the simple way of linear combination (Amari, 2007). Furthermore, α-integration was found to be optimum in the sense of minimising α-divergence between the integration channel and its multiple component channels of information sources (Amari, 2007). Inspired by this, we consider that if α-integration is used to replace the linear combination in the conventional Gaussian mixture model, the new model, which is therefore referred to as α-Gaussian mixture model (α-GMM), could take some advantages over the conventional GMM in modelling multiple channels of information sources, since the integration process emulated by the new method is more similar to that occurring in a human cognition system. This is the essential idea for this article.

The rest of this paper is organised as follows: in Section 2, we introduced the concept of α-integration. In Section 3, α-GMM was proposed for addressing speaker recognition. Parameter reestimation formulae were used to train α-GMM recursively. In Section 4, α-GMM based speaker recognition was described. In Section 5, experiments and results were presented. The discussions and conclusions were summarised in Section 6 and 7.

2. α-integration

Given a number of probability density functions (p.d.f.) $p_i(s),\ i = 1, \ldots, K$, the α-integration $q(s)$ is defined by

$$q(s) = cf_\alpha^{-1}\left\{\sum_{i=1}^{K} w_i \cdot f_\alpha[p_i(s)]\right\}, \tag{1}$$

where $f_\alpha[p_i(s)]$ is referred to as the α-representation of each $p_i(s)$ and c is a constant to make the integrated function as a p.d.f.

$$c = \frac{1}{\int f_\alpha^{-1}\{\sum_{i=1}^{K} w_i f_\alpha[p_i(s)]\} ds}. \tag{2}$$

$f_\alpha[p_i(s)]$ is defined by

$$f_\alpha[p_i(s)] = \begin{cases} \frac{2}{1-\alpha} p_i(s)^{(1-\alpha)/2}, & \alpha \neq 1 \\ log(p_i(s)), & \alpha = 1. \end{cases} \tag{3}$$

The inverse function $f_\alpha^{-1}(y)$ can be easily obtained as follows

$$x = f_\alpha^{-1}(y) = \begin{cases} (\frac{1-\alpha}{2} y)^{\frac{2}{1-\alpha}}, & \alpha \neq 1 \\ e^y, & \alpha = 1. \end{cases} \tag{4}$$

From Eqs. (1), (3) and (4), we can see that in the case of $\alpha = -1$, the α-integration degenerates to the simple linear combination and in the case of $\alpha = 1$, the α-integration is referred to as the exponential integration.

It is worth to note in Eq. (1), as α increases, the α-integration relies more on the smaller elements of mixtures, while as α decreases, the larger ones are taken into account more seriously. Thus, it is said that a small-value of α represents a pessimistic attitude and a larger one of α represents a more optimistic attitude (Amari, 2007).

Moreover, it was found in Amari (2007) that α-integration $q(s)$ (see Eq. (1)) of probability density functions $p_1(s), \ldots, p_K(s)$ with weights $w_1, \ldots, w_K$ is optimal under the α-divergence criterion of minimising the integrated divergence $R_\alpha[q(s)]$ between the integrated p.d.f $q(s)$ and all its components $p_i(s)$, i.e.,

$$R_\alpha[q(s)] = \sum_{i=1}^{K} w_i D_\alpha[p_i(s) : q(s)]. \tag{5}$$

The α-divergence $D_\alpha[p(s) : q(s)]$ is defined as

$$D_\alpha[p(s) : q(s)] = \begin{cases} \int p(s) \log \frac{p(s)}{q(s)} ds, & \alpha = -1 \\ \int q(s) \log \frac{q(s)}{p(s)} ds, & \alpha = 1 \\ \frac{4}{1-\alpha^2}\left\{1 - \int p(s)^{\frac{1-\alpha}{2}} q(s)^{\frac{1+\alpha}{2}} ds\right\}, & \alpha \neq \pm 1 \end{cases} \tag{6}$$

α-divergence is an extension to the well-known Kullback–Leibler distance, i.e.,

$$D_{-1}[p(s) : q(s)] = KL[p(s) : q(s)] \tag{7}$$

$$D_1[p(s) : q(s)] = KL[q(s) : p(s)], \tag{8}$$

where KL is the Kullback–Leibler divergence and

$$D_0[p(s) : q(s)] = 2\int \left(\sqrt{p(s)} - \sqrt{q(s)}\right)^2 ds \tag{9}$$

is the square of the well-known Hellinger distance. The divergences satisfy

$$D_\alpha[p(s) : q(s)] \geqslant 0, \tag{10}$$

with equality if and only if $p(s) = q(s)$. They are not symmetric,

$$D_\alpha[p(s) : q(s)] \neq D_\alpha[q(s) : p(s)], \tag{11}$$

except for the case $\alpha = 0$.

In summary, the α-integration is an optimal integration approach in the sense of minimising an extension of KL-distance between each component and their integration.

3. α-Gaussian mixture model

Bearing in mind the concept of the α-integration, we can easily define the α-GMM for speaker recognition.

3.1. Definition of α-GMM

The conventional GMM is defined as

$$p(\boldsymbol{X}|S) = \sum_{i=1}^{K} w_i \mathcal{N}(\boldsymbol{X}, \boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i^{-1}|S), \tag{12}$$

where $\mathcal{N}(\boldsymbol{X}, \boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i^{-1})$ is the ith normal (or Gaussian) distribution for a given utterance $\boldsymbol{X}$ and speaker S with the mean $\boldsymbol{\mu_i}$ and the variance $\boldsymbol{\Sigma}_i^{-1}$ and K is the number of Gaussians.

By using α-integration (Eq. (1)), we can define α-GMM as follows

$$p_\alpha(\boldsymbol{X}|S) = cf_\alpha^{-1}\left\{\sum_{i=1}^{K} w_i \cdot f_\alpha[\mathcal{N}_i(\boldsymbol{X}|S)]\right\}, \tag{13}$$

where c is a normalisation constant.

Explicitly, Eq. (13) can be rewritten as

$$p_\alpha(\boldsymbol{X}|S) = \begin{cases} c(\sum_{i=1}^{K} w_i \mathcal{N}_i(\boldsymbol{X}|S)^{\frac{1-\alpha}{2}})^{\frac{2}{1-\alpha}}, & \alpha \neq 1 \\ ce^{\sum_{i=1}^{K} w_i \log \mathcal{N}_i(\boldsymbol{X}|S)}, & \alpha = 1. \end{cases} \tag{14}$$

Obviously, when $\alpha = -1$, Eq. (14) is a conventional GMM.

Till now we have defined the concepts for the α-GMM, however, there is still an open question about how to estimate parameters related to a given α-GMM. The concerned parameter set is defined as $\Theta = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i^{-1}\}$. In the next section, we will particularly address this question.

3.2. Model parameter estimation

An adapted expectation maximisation (EM) algorithm can be used to estimate parameters of α-GMM for the case of $\alpha \neq 1$. The essence for the EM algorithm is maximising the expectation of log-