Engineering Applications of Artificial Intelligence 22 (2009) 667–675

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Speaker diarization using autoassociative neural networks

S. Jothilakshmi *, V. Ramalingam, S. Palanivel

Department of Computer Science and Engineering, Annamalai University, Annamalainagar 608 002, India

ARTICLE INFO

Article history:
Received 3 October 2008
Received in revised form
2 January 2009
Accepted 29 January 2009
Available online 9 March 2009

Keywords:
Speaker diarization
Speaker segmentation
Speaker clustering
Mel frequency cepstral coefficients
Autoassociative neural networks

ABSTRACT

This paper addresses a new approach to speaker diarization using autoassociative neural networks (AANN). The speaker diarization task consists of segmenting a conversation into homogeneous segments which are then clustered into speaker classes. The proposed method uses AANN models to capture the speaker specific information from mel frequency cepstral coefficients (MFCC). The distribution capturing ability of the AANN model is utilized for segmenting the conversation and grouping each segment into one of the speaker classes. The algorithm has been tested on different databases, and the results are compared with the existing algorithms. The experimental results show that the proposed approach competes with the standard speaker diarization methods reported in the literature and it is an alternative method to the existing speaker diarization methods.

1. Introduction

Speaker diarization is the process of automatically partitioning a conversation involving multiple speakers into homogeneous segments and grouping together all the segments that correspond to the same speaker. The first part of the process is known as speaker segmentation or speaker change detection while the second one is called as speaker clustering. Hence speaker change detection followed by speaker clustering is known as speaker diarization (Meignier et al., 2006; Tranter and Reynolds, 2006; Sinha et al., 2005).

Nowadays a rapid increase in the volume of recorded speech is manifested which includes television and audio broadcasts, voice mails, meeting and other spoken documents (Solomonoff et al., 1998; Kotti et al., 2007). There is a growing need to apply automatic human language technologies to allow efficient and effective searching, indexing and accessing of these information sources. Diarization can be used for helping speech recognition, facilitating the searching and indexing of audio archives and increasing the richness of automatic transcriptions, making them more reliable and potentially helping with other tasks such as summarization, parsing and machine translation (Tranter and Reynolds, 2006).

Generally, for the task of speaker diarization no prior information is available regarding the number of speakers involved or their identities. So, speaker diarization can be considered as a task of identifying the number of speakers and creating a list of speech time intervals for each speakers. In the literature, various speaker diarization algorithms have been proposed. These algorithms can be categorized into three categories: step by step approaches, integrated approaches and mixed approaches.

Step by step approaches divide the speaker diarization task into number of steps (Siu et al., 1992; Wilcox et al., 1994; Sieglar et al., 1997; Gauvain et al., 1998; Chen and Gopalakrishnan, 1998). First finding the speaker change points using the symmetric Kullback Leibler (KL2), the generalized likelihood ratio (GLR) or the Bayesian information criterion (BIC) distance approaches, then growing the segments during a hierarchical clustering phase and finally determining the number of speakers. In the case of integrated approaches (Meigneir et al., 2001; Ajmera and Wooters, 2003) all the steps involved in speaker diarization are performed simultaneously. Mixed strategies also proposed in Wilcox et al. (1994), Moraru et al. (2004), and Reynolds et al. (2000), where classical step by step segmentation and clustering are first applied and then refined using a re-segmentation process during which the segment boundaries, the segment clustering and sometimes the number of speakers are refined.

Most of the model based speaker diarization systems in the literature use Gaussian mixture model (GMM) or hidden Markov model (HMM) to estimate the probability distribution of the feature vectors of a speaker. While GMMs appear to be general enough to characterize the distribution of the given data, the model is constrained by the fact that the shape of the components of the distribution is assumed to be Gaussian, and the number of mixtures are fixed a priori (Yegnanarayana and Kishore, 2002).

* Corresponding author. Tel.: +91 9894 693493; fax: +91 4144 238080.
E-mail addresses: jothi.sekar@gmail.com, jothi_sekar1993@yahoo.com (S. Jothilakshmi), aucsevr@yahoo.com (V. Ramalingam), spal_yughu@yahoo.com (S. Palanivel).

0952-1976/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2009.01.012

In this context, Yegnanarayana and Kishore (2002) investigated the potential of nonlinear models such as autoassociative neural network (AANN) models, which perform identity mapping of the input space. AANN is a feed forward neural network which can be designed to perform the task of pattern classification or pattern mapping (Yegnanarayana, 1999).

The main contribution of this paper concerns the use of the distribution capturing ability of the AANN for speaker change detection and speaker clustering for speaker diarization. The proposed method relies on a classical two step speaker diarization approach based on a detection of speaker turns followed by a clustering process as shown in Fig. 1. This work formulates a new speaker diarization algorithm and it works without any prior knowledge of the identity of speakers.

The rest of the paper is organized as follows: A brief description about the method of extracting speaker specific information from the speech signal is described in Section 2. AANN model for capturing the distribution of acoustic feature vectors is given in Section 3. The proposed algorithm for speaker diarization is presented in Section 4. In Section 5, the performance measures used for speaker diarization are discussed. Section 6 presents the experimental results and the performance comparison of the proposed method with the existing methods. Section 7 concludes the paper.

2. Feature extraction for speaker segmentation

Mel frequency cepstral coefficients (MFCC) have proved to be one of the most successful feature representations in speech related recognition tasks. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum (Davis and Mermelstein, 1980). In this section we briefly describe the signal processing involved in extracting the MFCC. The selected properties for the speech signals are a sampling rate of 8 kHz, 16 bit monophonic, pulse code modulation (PCM) format in wav audio. The procedure of MFCC computation is shown in Fig. 2 and described as follows:

- *Preemphasis*: The digitized speech signal $s(n)$ is put through a low order digital system to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasis network, $\hat{s}(n)$ is related to the input $s(n)$ by the difference equation

$$\hat{s}(n) = s(n) - \alpha s(n-1) \tag{1}$$

The most common value for α is around 0.95.

- *Frame blocking*: Speech analysis usually assumes that the signal properties change relatively slowly with time. This allows examination of a short time window of speech to extract parameters presumed to remain fixed for the duration of the window. Thus to model dynamic parameters, we must divide the signal into successive windows or analysis frames, so that the parameters can be calculated often enough to follow the relevant changes. In this step the preemphasized speech signal, $\hat{s}(n)$ is blocked into frames of N samples, with adjacent frames being separated by M samples. If we denote the lth frame speech by $x_l(n)$, and there are L frames within the entire speech signal, then

$$x_l(n) = \hat{s}(Ml+n), \quad n = 0, 1, \ldots, N-1, \quad l = 0, 1, \ldots, L-1 \tag{2}$$

We used a frame rate of 125 frames/s, where each frame was 16 ms in duration with an overlap of 50% between adjacent frames.

- *Windowing*: The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of the frame. The window must be selected to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n)$, $0 \leqslant n \leqslant N-1$, then the result of windowing the signal is

$$\tilde{x}_l(n) = x_l(n) w(n), \quad 0 \leqslant n \leqslant N-1 \tag{3}$$

The Hamming window is used for our work, which has the form

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leqslant n \leqslant N-1 \tag{4}$$

- *Computing spectral coefficients*: The spectral coefficients of the windowed frames are computed using Fast Fourier Transform, as follows:

$$X(k) = \sum_{n=0}^{N-1} \tilde{x}_l(n) \exp^{-jk(2\pi/N)n}, \quad 0 \leqslant n \leqslant N-1 \tag{5}$$

- *Computing mel spectral coefficients*: The spectral coefficients of each frame are then weighted by a series of filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters. These filters

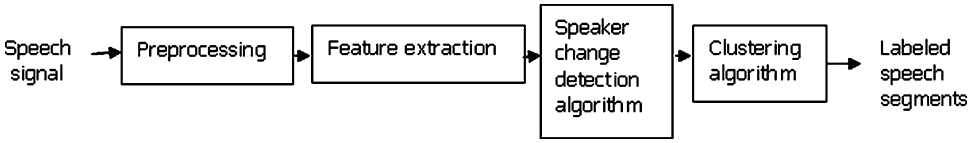

Fig. 1. Block diagram of speaker diarization system.

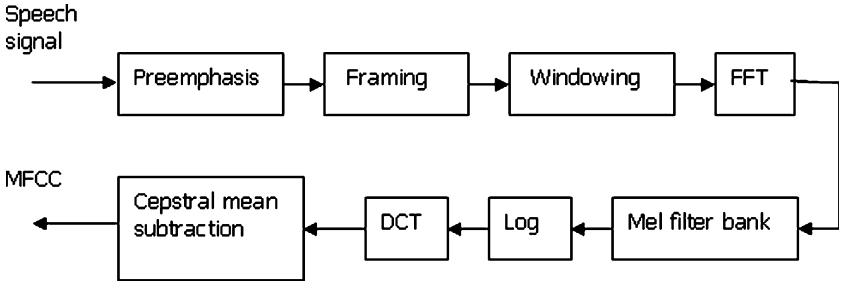

Fig. 2. Extraction of MFCC from speech signal.