Expert Systems with Applications 36 (2009) 8056–8063

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Speaker identification based on the frame linear predictive coding spectrum technique

Jian-Da Wu *, Bing-Fu Lin

Graduate Institute of Vehicle Engineering, National Changhua University of Education, 1 Jin-De Rd., Changhua City, Changhua 500, Taiwan

ARTICLE INFO

Keywords:
Speaker identification
Linear predictive coding
Gaussian mixture model
General regression neural network

ABSTRACT

In this paper, a frame linear predictive coding spectrum (FLPCS) technique for speaker identification is presented. Traditionally, linear predictive coding (LPC) was applied in many speech recognition applications, nevertheless, the modification of LPC termed FLPCS is proposed in this study for speaker identification. The analysis procedure consists of feature extraction and voice classification. In the stage of feature extraction, the representative characteristics were extracted using the FLPCS technique. Through the approach, the size of the feature vector of a speaker can be reduced within an acceptable recognition rate. In the stage of classification, general regression neural network (GRNN) and Gaussian mixture model (GMM) were applied because of their rapid response and simplicity in implementation. In the experimental investigation, performances of different order FLPCS coefficients which were induced from the LPC spectrum were compared with one another. Further, the capability analysis on GRNN and GMM was also described. The experimental results showed GMM can achieve a better recognition rate with feature extraction using the FLPCS method. It is also suggested the GMM can complete training and identification in a very short time.

1. Introduction

Communications among human beings are simple and common, but the same action between human and computer is difficult beyond our expectation. Till now, many scientists are still working on how to make a machine not only to decipher what people say, but also to understand the meaning implied by the words. Fortunately, due to the great promotion of computing ability in microprocessors, applications of voice signal processing have been increasing rapidly in recent years, such as sound recognition or speaker identification. In speech recognition, the system has to recognize various commands from various speakers. Thus, many algorithms have been developed to extract the vital components hidden in speech signals, for example, spectral or cepstral characteristics. On the other hand, speaker identification originated from the needs for security monitoring in many important buildings or facilities. By applying this mechanism, the approaching people can be observed. In addition, speaker identification is utilized in suspect identification because of its properties of non-contact characteristic.

In this paper, a text-dependent speaker identification system is proposed. The advantage of text-dependent lies in the sentence used for recognition does not need to be very long; it can simply be a word or an utterance. Besides, unlike text-independent system, shorter sentences can increase classification speed. Generally speaking, implementing speaker identification can be divided into two stages: the first is feature extraction and the second is speaker classification based on the extracted features (Sarikaya, Pellom, & Hansen, 1998). At the stage of feature extraction, the extracted features should be capable of separating the speakers from each other in its space. In traditional techniques, the speech features are usually obtained by Fourier transforms and short time Fourier transforms. However, these techniques are unsuitable for speaker identification because they accept stationary signal within a given time frame and may therefore lack the ability to analyze the non-stationary signals or signals in transient state (Avci & Akpolat, 2006). Therefore, many algorithms were developed to find a better representation of a speaker, for example: linear predictive coding (LPC) technique (Adami & Barone, 2001; Haydar, Demirekler, & Yurtseven, 1998; Wutiwiwatchai, Achariyakulporn, & Tanprasert, 1999), Mel frequency cepstral coefficient (MFCC) (Mashao & Skosan, 2006; Sroka & Braida, 2005) and wavelet (Lung, 2006; Wu & Lin, 2009; Wu & Ye, 2009). In this paper, an improved method based on LPC is proposed. In fact, LPC is not a new method, it was developed in 1960s (Atal, 2006), but is popular and widely used till today because LPC coefficients representing a speaker by modeling vocal tract parameters and the data size are very suitable for speech compression through the digital channel. In the present study, the focus will be on modifying LPC coefficients and reducing the size of feature vectors.

* Corresponding author.
E-mail address: Jdwu@cc.ncue.edu.tw (J.-D. Wu).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.051

On the selection of classifier, the general regressive neural network (GRNN) and Gaussian mixture model (GMM) were chosen to be applied in the classification stage. Both two classifiers are popular in many pattern recognition fields because they can achieve good performance and the training time is well-satisfied. In the following section, an experimental investigation was carried to form a comparable result between these two approaches. Both their advantages and disadvantages will be compared with each other.

2. Principles of feature extraction and classification

2.1. Linear predictive coding technique

In modern signal processing, the analysis procedure extracts useful information from the structure of a signal. LPC technique is a developed algorithm used in speech analysis for many years. Its basic idea comes from a model representing the resonances of the human vocal tract. In general, speech sounds are produced by acoustic excitation of the vocal tract. During the production of voiced sounds, the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords. With unvoiced sounds, the excitation is provided by air passing turbulently through constrictions in the tract (Atal & Hanauer, 1971). Fig. 1 shows the speech signal production model in which the speech synthesizer strongly depends on the estimation of a_p. Here, a_p are autoregressive parameters obtained from the linear prediction method, and provide better results to characterize human speech. The use of these parameters assumes the speech signal can be represented as the output signal of an all pole digital filter in which the excitation is an impulse sequence with a frequency equal to the pitch of the speech signal under analysis when the segment is voiced, or with noise when the segment is unvoiced (Perez-Meana, 2007). The steps of acquiring a_p are described as follows:

(a) The input signal is segmented in 20 ms with 10 ms overlapping.
(b) Apply the window function to these segments to avoid distortion of the segmented speech because of the discontinuities introduced during the segmentation process, typically the Hamming window is used. The Hamming window is given by the following equation:

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

where N is the number of samples of the used segment.
(c) Estimate the prediction order and calculate the linear prediction coefficients for each segment.

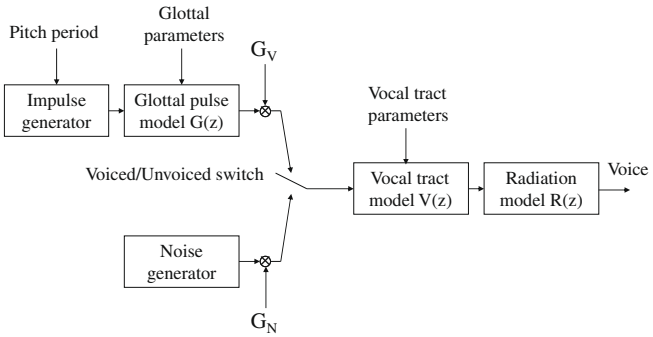

Fig. 1. Flow chart of speech synthesis.

After the speech signal is segmented, the p autocorrelation coefficients are estimated, where p is the linear predictor order. The autocorrelation function can be estimated using the biased or unbiased autocorrelation algorithms (Childers, 2000). Once autocorrelation coefficients p are evaluated from each segment, the signal at time n can be rewritten as a linear combination from the pass samples of the input signal:

$$\widehat{s}(n) = -(a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p(n-p)), \tag{2}$$

or

$$\widehat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \quad k = 1, 2, \ldots, p. \tag{3}$$

Therefore, it is affirmed a filter can be designed to estimate the data at time n only using the previous data at time $n-1$:

$$\widehat{s}(n) = -a_e s(n-1), \tag{4}$$

where a_e is the linear prediction coefficient. To evaluate a_e, the prediction error is minimized expectantly between $s(n)$ and $s(n-1)$:

$$e(n) = s(n) - \widehat{s}(n) = s(n) + a_e s(n-1). \tag{5}$$

After a series of calculations, $s(n)$ is evaluated and rewritten in the Z domain:

$$S(z) = \frac{E(z)}{1 + \left[\sum_{k=1}^{p} a_k z^{-k}\right]} = \frac{E(z)}{A(z)}, \tag{6}$$

where

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}.$$

Eq. (6) denotes the transfer function of an all pole filter shown in Fig. 2. The poles of transfer function are the zeros of the polynomial in the denominator on the right side of Eq. (6). The linear filter thus has a total of p poles which are either real or occur in conjugate pairs. Moreover, for the linear filter to be stable, the poles must be inside the unit circle (Atal & Hanauer, 1971). Once the prediction coefficients a_k are obtained, a more specific algorithm to extract the features for representing a speaker is applied. Therefore, the frame based linear predictive coding spectrum (FLPCS) coefficients is introduced in this paper.

2.2. Frame based linear predictive coding spectrum

LPC coefficients provide good reproduction of human vocal tract in speech synthesis. In this paper, these coefficients were not used as features for speaker identification directly due to the complexity of the data dimension. Moreover, during the procedures of acquiring LPC coefficients, some signal-freed segments were also involved in the LPC computation. This may lead to difficulties in the classification stages because speech signals always have different location on the time axis. Therefore, a new algorithm called FLPCS coefficients is proposed in this paper to solve this dilemma. Basically, FLPCS coefficients are based on LPC but have different ways of extracting features. The mathematical expression is defined as

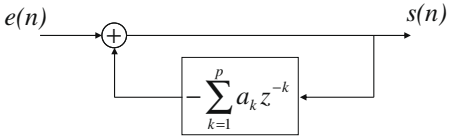

Fig. 2. Transfer function of the all pole filter.