Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 39 (2006) 147–155
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communication

Combining classifier decisions for robust speaker identification

Daniel J. Mashao*, Marshalleno Skosan

Speech Technology And Research (STAR), University of Cape Town, Rondebosch 7701, South Africa

Received 23 June 2005; received in revised form 17 August 2005

Abstract

In this work, we combine the decisions of two classifiers as an alternative means of improving the performance of a speaker recognition system in adverse environments. The difference between these classifiers is in their feature-sets. One system is based on the popular mel-frequency cepstral coefficients (MFCC) and the other on the new parametric feature-sets (PFS) algorithm. The feature-vectors both have mel-scale spectral warping and are computed in the cepstral domain but the feature-sets differs in the use of spectral filters and compressions. The performance of the classifier is not much different in recognition rates terms but they are complementary. This shows that there is information that is not captured in the popular mel-frequency cepstral coefficients (MFCC), and the parametric feature-sets (PFS) is able to add further information for improved performance. Several ways of combining these classifiers gives significant improvements in a speaker identification task using a very large telephone degraded NTIMIT database.

Keywords: Speaker identification; Parametric feature sets; Multiple classifier systems; Gaussian mixture model

1. Introduction

Over the years, there has been much interest in using speech as a means of identifying speakers. Speech, as opposed to other biometrics such as fingerprints and face recognition, allows recognition to be performed remotely as it can easily be transmitted over communication channels. It has been shown, however, that the performance of speaker recognition (SR) systems degrades considerably when contaminated by telephone noise in transmission [1]. Hence, the robustness of SR systems has been a major research issue in recent years [2]. For SR tasks, numerous speech parameterisations and classification algorithms have been proposed over the years [3]. However, it is still difficult to implement a single classifier that exhibits sufficiently high performance in practical applications. As a result, several researchers have cited the fusion of multiple information sources as a promising option in SR research [3–5]. A sufficient condition for fusing the outputs of many classifiers is that they make errors that are uncorrelated i.e. they misclassify different patterns of the same data [6]. Altincay and Demirekler [7], have improved speaker identification (SI) performance by fusing the outputs of two classifiers where one used a form of channel compensation and the other did not. They showed that SI performance is very sensitive to the signal processing done when extracting a particular speech features. In this work we develop two high performing baseline SI classifiers that make different errors. Their decisions are then subsequently combined with the aim of improving the robustness and performance of the overall SI system. In particular, large population speaker identification experiments are conducted on the telephone degraded NTIMIT speech database. Large improvements, above the baseline SI classifiers and other systems reported in related literature, are obtained.

2. Features from the speech signals

A speaker identification systems consists of mainly two parts as shown in Fig. 1. In the front-end is the feature-generation part and in the back-end is the classification engine. During the enrollment phase the switch is turned

* Corresponding author. Tel.: 27 216502816.
E-mail addresses: daniel@star.za.net (D.J. Mashao), leno@star.za.net (M. Skosan).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.08.004

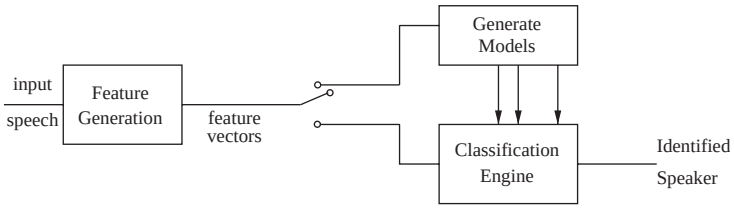

Fig. 1. Speaker identification system.

towards the upper route and the system generates models, and during testing or evaluation phase the models are used to identify a speaker from an unknown speaker's speech. Current state of the art systems uses the mel-frequency cepstral coefficients (MFCC) algorithm for the front end and Gaussian mixture models (GMM) and the classification engine.

An ideal front-end in the SI system is the one that will extract all speaker specific information from the input speech signal without being confused with what the person is saying. Every person has a natural sound quality due to their voice pitch. However, pitch detection has proven challenging and also reliance on it can allow impostors to gain access by changing their own pitch. The other problem with pitch is that it cannot be reliably measured in some speech sounds for example nasals and consonants. As such many front-end algorithms do not seek to use pitch as a specific feature. Instead the design of speech feature-sets seeks to find speaker specific information in other parts of a person's speech.

2.1. Mel-frequency Cepstral Coefficients

The MFCC is the most popular front-end for SI systems and is also used in other speech technology tasks such as speech recognition and SR in general. The MFCC coefficients are generated as shown in Fig. 2. First, the speech signal is acquired in the time domain via sampling and after the application of the discrete Fourier transform (DFT) it is converted into the frequency domain. If the inverse discrete Fourier transform (IDFT) was applied then, the signal would revert back to the time domain, but before the IDFT is applied a log magnitude (taking the logarithms of the magnitude of a complex signal) of the frequency domain signal (the spectrum) is computed. The signal is now said to be in the cepstral (a play on the words spectral) domain and the units in this domain are seconds, same as in the time domain. The signal in the cepstral domain is measured in quefrencies (again a play on the words frequency). The low-order quefrencies contain information that is due to the speech formants and therefore carries information about what is being said and the high quefrencies are due to the pitch and therefore assumed to be speaker dependent.

To obtain the MFCC coefficients, the speech signal is windowed and converted into the frequency domain by using the DFT. In the frequency domain a log magnitude of the complex signal is obtained. A mel-scaling or mel-warping is then performed using filters. The common method of implementing these filters is to use triangular filters that are linear spaced from 0 to 1 kHz and then non-linearly placed according to the mel-scaling approximations. There are several approximations of the mel-scale the most popular is by O'Shaughnessy [8] which is

$$F_{mel} = 2595 \log\left(1 + \frac{F_{in}}{700}\right),$$

where F_{mel} is the frequency in mels and F_{in} is the input frequency in Hertz. This is the scaling used in the design of MFCC coefficients used in this paper. There are many functional approximations of the mel-scale and they all show minor differences in performance as shown by Umesh et al. [9]. The centre frequencies of the triangular filters will be set at the mel-scale frequencies, with the low input frequencies (less than 1 kHz) given a higher profile than the higher frequencies. The resultant signal from the filtering is then transformed using an inverse DFT (usually implemented with a discrete cosine transform) into the cepstral domain. The lower order coefficients are selected as the feature vector. The selection of the lower order coefficients is done on purpose to avoid the higher coefficients which include the pitch. The coefficients are then uniformly scaled and used as the output feature vector for that speech frame.

It is worth noting that these same MFCC feature-vectors are used in both speech recognition and speaker recognition tasks. In speaker independent speech recognition task any speaker information is considered noise but in SR it is actually the kind of information that is sought. The fact that the same feature-vectors can be used for both tasks shows that they contain both semantic and person specific information. The successful application of the low order quefrencies coefficients for SI tasks shows that the person information is still intact.

The MFCC has been used for several years since the late 1990s and has successfully replaced the linear prediction cepstral coefficients (LPCC). The main advantage of the LPCC was computation but with increasing computing power and better performance (and the use of the fast Fourier transform—FFT), the MFCC has completely dominated the designs of the front-ends of speech technology systems. Competition from the auditory based feature-sets has not been successful and mainly due to their lower performance and very high computational cost. Our own research