Chapter 17
Speaker Discrimination on Broadcast News and Telephonic Calls Based on New Fusion Techniques

Halim Sayoud
USTHB University, Algeria

Siham Ouamour
USTHB University, Algeria

ABSTRACT

This chapter describes a new Speaker Discrimination System (SDS), which is a part of an overall project called Audio Documents Indexing based on a Speaker Discrimination System (ADISDS). Speaker discrimination consists in checking whether two speech segments come from the same speaker or not. This research domain presents an important field in biometry, since the voice remains an important feature used at distance (via telephone). However, although some discriminative classifiers do exist nowadays, their performances are not enough sufficient for short speech segments. This issue led us to propose an efficient fusion between such classifiers in order to enhance the discriminative performance. This fusion is obtained, by using three different techniques: a serial fusion, parallel fusion and serial-parallel fusion. Also, two classifiers have been chosen for the evaluation: a mono-gaussian statistical classifier and a Multi Layer Perceptron (MLP). Several experiments of speaker discrimination are conducted on different databases: Hub4 Broadcast-News and telephonic calls. Results show that the fusion has efficiently improved the scores obtained by each approach alone. So, for instance, the authors got an Equal Error Rate (EER) of about 7% on a subset of Hub4 Broadcast-News database, with short segments of 4 seconds, and an EER of about 4% on telephonic speech, with medium segments of 10 seconds.

DOI: 10.4018/978-1-60960-563-6.ch017

INTRODUCTION

While fingerprints (Youssif, Chowdhury, Ray & Nafaa, 2007) and retinal scans (Daugman, 2007) are more reliable means of authentication, speech can be seen as a non-evasive biometric that can be collected with or without the person's knowledge or even transmitted over long distances via telephone. Furthermore, a person's voice cannot be stolen, forgotten, guessed, given to another person or lost. Then, voice based speaker discrimination represents a secure and efficient way in biometry.

Speaker discrimination (Koreman, Wu, & Morris, 2007) consists in checking whether two different pronunciations (speech segments) belong to the same speaker or not. One means used to compare these utterances is to extract the vocal characteristics from each speech signal, in order to detect the degree of similarity between them. Speaker discrimination has several applications such as: speaker verification, audio signal segmentation and speaker based clustering.

In this domain, several classifiers do exist but most of them do not accept very short speech segments as required in some applications (e.g. audio stream segmentation). That is why; we have proposed some techniques of fusion between those classifiers. The principal goal of this investigation is to develop a fusion-based speaker discrimination system for the task of speaker changes detection in multi-speaker audio streams (i.e. SDS system). A second goal will concern the application of this discriminative system in the task of audio document indexing (i.e. ADISDS system). In fact, audio documents indexing represents respectively the process of speech segmentation, which divides the audio flow into homogeneous segments (each segment contains only one speaker) and the process of clustering, which gathers all the speech segments belonging to the same speaker together. Thus, the SDS system will be used firstly for detecting the speaker changes in the audio stream and secondly for gathering all the segments of a same speaker, in order to obtain the overall intervention of each speaker, at the end of the process.

In this research work, we are interested in investigating two classifiers:

- The Statistical classifier based on mono-gaussian model and employing a symmetric measure of similarity (third section);
- The Multi-Layer Perceptron (MLP) using a new characteristic called "RSC" (fourth section), which is developed in order to reduce the neural network input size, minimize the training database and optimize its convergence;

Although it does exist many other classifiers giving high performances in this domain, a lot of them require long speech segments during either the training or the testing step, which makes them not suitable for discrimination applications using short segments, like in speech segmentation (Meignier, 2002).

The particular choice of the statistical mono-gaussian classifier is due to its easy implementation, low cost in computation, and good discrimination on short speech segments: even with only 2 seconds (Sayoud, Ouamour & Boudraa, 2003) unlike the Gaussian Mixture Models (GMM) for example, which require long speech duration at least for one segment (the model). However, the MLP has been chosen for its high discriminative performance, as it is the case for most of the Neural Networks (NNs). That is why we have decided to choose both of these two classifiers.

In order to enhance the discrimination accuracy, we try to combine those two classifiers by a fusion technique. Theoretically, the fusion should get the advantages of the statistical classifier and the advantages of the neural classifier, corresponding respectively to a high resolution (working on short speech segment) and a high discriminative capacity, which is well adapted in speech indexing (Meignier, 2002; Sayoud, Ouamour & Boudraa, 2003).