Expert Systems With Applications 73 (2017) 27–42

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Spatial features selection for unsupervised speaker segmentation and clustering

Beatriz Martínez-González[a,*], José M. Pardo[a], Julián D. Echeverry-Correa[b], Ruben San-Segundo[a]

[a] Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain

[b] Facultad de Ingenierías.Programa de Ingeniería Eléctrica.Universidad Tecnológica de Pereira.Carrera 27 #10-02, Barrio Alamos, Edificio de Ingeniería Eléctrica, Oficina 1B-136, Pereira, Colombia

ARTICLE INFO

Article history:
Received 7 June 2016
Revised 14 November 2016
Accepted 5 December 2016
Available online 8 December 2016

Keywords:
Speaker diarization
Speaker segmentation
Feature selection
Speaker localization
Speaker clustering

ABSTRACT

The selection of the best features to be used in expert systems is a key issue in obtaining a satisfactory performance. Unsupervised speaker segmentation and clustering is the task of the automatic identification of the number of participants in a meeting and the determination of their speaking turns (also called "diarization"). This is part of an intelligent system that replaces human intervention in several tasks related to automatic language and speech processing. The segmentation and clustering of speakers is crucial if we want to transcribe any audio recording automatically when several people take their turn. It is a task necessary to archive automatically interventions of several people in meetings, broadcast radio, lectures, parliamentary sessions etc. since a simple transcription of what is said without assigning it to a specific speaker makes the information unusable. The automation of this task would save enormous amounts of resources currently spent on human transcribers. When used online it could also be useful to point a video camera automatically to the person talking when a videoconference with multiple speakers is taking place thus replacing a human operator. Furthermore it could also help to scan large amounts of audio automatically in search of crimes or audio interventions of a particular person. In the case of recordings with several distant microphones (MDM), spatial features may and should be used. The most widely used spatial features in diarization are the Time Delay of Arrival (TDOA) features. These delays are extracted from pairs of microphones of unknown location and quality, which makes the selection of the best pairs highly advisable. This paper analyses this issue and proposes and evaluates several methods that significantly improve the performance both in speaker error rate (SER) and in computational time. The methods propose a selection ofTDOA features based on the quality of the cross-correlation of signals coming from different pairs of microphones. We prove that the use of the wrong pairs can be highly detrimental to the overall performance. The methods proposed, based on cross correlation, are compared and combined with other two selection methods, based on the dynamic range of the delay features and the selection of every pair of microphones available followed by a reduction in dimensionality. Although all algorithms achieve some improvements, it is proved that selection methods based on cross correlation have the fewest errors. The improvements on the baseline system for the two best proposed systems are 25.14% and 33.70% for the development set, and 55.06% and 46.09% for the test set. Furthermore the best method for the test set also reduces the computational cost by 20%.

1. Introduction

The selection of the best features to be used in expert systems has been addressed in depth in the literature (Altun & Polat, 2009; Nemati & Basiri, 2011). When several features are available the selection of the most informative could greatly improve the performance of the whole system while reducing the computational cost. Moreover, the use of the wrong features for a particular task would be deeply detrimental to the overall performance. Therefore, the search for new features that better represent the characteristics of the problem to solve has been always an important task. Distinguishing between good and bad features requires the evaluation of the new suggested features.

* Corresponding author.
E-mail addresses: beatrizmartinez@die.upm.es (B. Martínez-González), pardo@die.upm.es (J.M. Pardo), jde@utp.edu.co (J.D. Echeverry-Correa), lapiz@die.upm.es (R. San-Segundo).

http://dx.doi.org/10.1016/j.eswa.2016.12.005
0957-4174/

In this work we focus on the selection of the time delay of arrival (TDOA) features for the unsupervised speaker segmentation and clustering task, also called diarization. Speaker diarization consists of identifying the number of participants in a meeting and creating a list of speech time intervals for each speaker. This task is carried out without knowing either the possible identities of the speakers or the characteristics of the meeting room in advance. Speaker diarization is a pre-processing task necessary as a first step in many speech processing applications such as the automatic speech transcription of meetings, speaker verification or speaker adaptation in speech recognition. In Liu Tian, He, and Liu (2016), speaker recognition in real world data is improved through the application of three different diarization systems as a pre-processing stage. The segmentation and clustering of speakers is crucial when dealing with recordings with several people and we do not know all their identities in advance or we do not have previous data on the speakers. It is a task necessary to transcribe automatically the interventions of several people in meetings, broadcast radio, lectures, parliamentary sessions etc. since a simple transcription of what is said without assigning it to a specific speaker makes the information unusable. When used online it could also be useful to detect new speakers automatically and adapt the system to the new situation, recognising the new speaker or locating him or her. In a conference with multiple participants this would allow, for instance, to point a video camera automatically to the person talking thus replacing a human operator. Furthermore it could also help to scan large amounts of audio automatically in search of crimes or audio interventions of a particular person. The automation of any of these tasks would entail immense savings in human resources.

Speaker diarization was first applied to broadcast news recordings (BN). One of the best published systems can be seen in Barras, Zhu, Meignier, and Gauvain (2004). Subsequently speaker diarization was applied to the meeting domain with a single distant microphone (SDM). The meeting domain differs from BN as the topics are highly diverse, the participants have idiosyncratic relationships and vocabularies, the meetings are highly interactive, and there can be simultaneous speech from multiple speakers. Furthermore, distant microphones are susceptible to reverberation and background noise. On the other hand, the number of speakers in the BN domain tends to be higher. However, since 2002 most efforts have been addressed to the domain of multiple distant microphones (MDM), extending the methods applied to SDM, or previously to BN, and adding new features. An overview of automatic speaker diarization systems is given in Moattar and Homayounpour (2012), Tranter and Reynolds (2006) and Anguera et al. (2012).

Speaker diarization systems usually consist of three main blocks: the Voice Activity Detection module (VAD), which filters out non-speech events, the feature extraction module, where all the necessary information is extracted from the recordings, and the segmentation and clustering module, which uses the previous features to segment the whole recording into clusters, i.e. speakers.

The segmentation and clustering module can use either a bottom-up agglomerative clustering (Wooters & Huijbregts, 2008; Sun, Nwe, Ma, & Li, 2009) or a top down clustering, beginning with a universal background model (UBM) (Fredouille, Bozonnet, & Evans, 2009). A comparison between these approaches with the SDM diarization of meetings is presented in Evans, Bozonnet, Wang, Fredouille, and Troncy (2012), where no significant differences were encountered.

The feature extraction module usually extracts data related to the spectral envelope such as the Mel Frequency Cepstral Coefficients (MFCC) (Vijayasenan, Valente, & Bourlard, 2009a; El-Khoury, Senac, & Pinquier, 2009), as well as data related to energy, the fundamental frequency or the time delays between channels, (Barra-Chicote, Pardo, Ferreiros, & Montero, 2011; Pardo, Barra-Chicote, San-Segundo, Córdoba, & Martínez-González, 2012; Friedland, Vinyals, Huang, & Muller, 2009; Pardo, Anguera, & Wooters, 2007). In this paper we will focus on the selection of the most informative delay features.

MDM speaker diarization, in comparison to SDM diarization, has redundant information available (one signal per microphone). All speech signals are usually combined into one (Anguera, Wooters, & Hernando, 2007) and the acoustic features are extracted from this combined signal. The other source of information usually used in MDM scenarios is the information related to speaker location (Ellis & Liu, 2004) such as the TDOA features (Pardo, Anguera, & Wooters, 2006a). These features permit short-term speaker segmentation but do not provide any speaker identity information. On the other hand, acoustic features provide long-term speaker identity but require minimum durations in which to build reliable acoustic models. In Pardo, Anguera, and Wooters (2006b), it was first demonstrated that TDOA between channels could be mixed with spectral features to obtain improved performance. This TDOA information combined with the MFCC features has been used by all systems in recent years.

The shortcomings of the TDOA methods are the result of the use of distant microphones. There are noises and reverberations in the recordings which can cause errors in the estimation of the delays. The goal of this work is to improve the results of the diarization by improving or optimizing the TDOA values used in segmentation and clustering. The baseline method to calculate TDOAs used in our system is described in Anguera et al. (2007). It first selects the channel with the highest average cross-correlation with the other channels as the reference, and then estimates the TDOAs between this channel and the others for every frame. The set of TDOAs between each microphone and the reference channel will form what we call the TDOA vector, which, therefore, will have a dimension equal to the number of microphones minus one. This vector is used together with the MFCC vector in the subsequent segmenting and clustering procedure.

In the baseline system we do not take advantage of the TDOA information between any two non-reference microphones. We argue that one specific pair of channels could be more suitable for locating one particular speaker than one pair made up of the reference and any remaining microphone. Selecting a set of pairs appropriately could help to better locate all the speakers. The issue is to decide which pairs are the most representative and reliable for the system to choose. In this paper we will present new techniques to select the best pairs appropriately for the target task.

Some preliminary work was presented in Martínez-González, Pardo, Echeverry-Correa, Vallejo-Pinto, and Barra-Chicote (2012). In this paper we complete the results with a more in-depth analysis and discussion of the development techniques and evaluation results and propose new methods that improve general performance.

Speaker diarization task is also very time consuming. Some recent works have tried to address this problem using different approaches such as GPUs (Gonina, Friedland, Cook, & Keutzer, 2011) or binary keys (Anguera & Bonastre, 2011; Delgado, Fredouille, & Serrano, 2014). Both techniques achieve a very high gain in computational time with a low increase in the DER. Although our main goal is to improve performance, our methods usually reduce the number of features to be extracted with the result of a significant reduction in the computational time.

The paper is organized as follows. Section 2 analyses some related work. Section 3 presents the baseline system used. Section 4 details the database used in the experiments together with the evaluation procedure. The methods proposed are presented in section 5. Section 6 evaluates the results. Section 7 discuss these results and their significance and Section 8 ends with the conclusions and future work.