Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 314–332

www.elsevier.com/locate/csl

Cepstral distance based channel selection for distant speech recognition☆

Cristina Guerrero Flores[a,b], Georgina Tryfou[a,b], Maurizio Omologo[b,*]

[a] University of Trento Via Sommarive, 14, 38123 Trento, Italy
[b] Fondazione Bruno Kessler-irst Via Sommarive 18, 38123 Trento, Italy

Received 9 November 2016; received in revised form 9 August 2017; accepted 18 August 2017
Available online 25 August 2017

Abstract

Shifting from a single to a multi-microphone setting, distant speech recognition can be benefited from the multiple instances of the same utterance in many ways. An effective approach, especially when microphones are not organized in an array fashion, is given by channel selection (CS), which assumes that for each utterance there is at least one channel that can improve the recognition results when compared to the decoding of the remaining channels. In order to identify this most favourable channel, a possible approach is to estimate the degree of distortion that characterizes each microphone signal. In a reverberant environment, this distortion can vary significantly across microphones, for instance due to the orientation of the speaker's head. In this work, we investigate on the application of cepstral distance as a distortion measure that turns out to be closely related to properties of the room acoustics, such as reverberation time and direct-to-reverberant ratio. From this measure, a blind CS method is derived, which relies on a reference computed by averaging log magnitude spectra of all the microphone signals. Another aim of our study is to propose a novel methodology to analyze CS under a wide set of experimental conditions and setup variations, which depend on the sound source position, its orientation, and the microphone network configuration. Based on the use of prior information, we introduce an informed technique to predict CS performance. Experimental results show both the effectiveness of the proposed blind CS method and the value of the aforementioned analysis methodology. The experiments were conducted using different sets of real and simulated data, the latter ones derived from synthetic and from measured impulse responses. It is demonstrated that the proposed blind CS method is well related to the oracle selection of the best recognized channel. Moreover, our method outperforms a state-of-the-art one, especially on real data.

Keywords: Distant speech recognition; Channel selection; Cepstral distance; Reverberation; Direct to reverberant ratio; T60

1. Introduction

Despite the extensive efforts that have been made for reliable automatic speech recognition (ASR), the performance of many voiced based systems is still inadequate under certain conditions. For example, ASR is seriously affected by the presence of reverberation, background noise, and overlapping speakers. In order to overcome these

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author at: SHINE - Fondazione Bruno Kessler. Via Sommarive, 18 I-38123 Trento, Italy.
E-mail address: cristinaguerreroflores@gmail.com (C. Guerrero Flores), tryfou@gmail.com (G. Tryfou), omologo@fbk.eu (M. Omologo).

http://dx.doi.org/10.1016/j.csl.2017.08.003
0885-2308/

limitations in distant-talking scenarios, some of the most effective strategies adopt the use of multiple microphones (Wölfel and McDonough, 2009; Brandstein and Ward, 2001). There are many applications, *e.g.,* in domestic environments, for which a significant improvement in terms of speech recognition rate can be obtained by deploying a large number of microphones, clustered in arrays with specific geometries, and distributed in such a way to cover the whole environment. A sparse distribution of single microphones in space, combined with an automatic channel selection (CS), represents a simple and effective solution to limit the overall complexity of a distant speech recognition (DSR) system.

CS makes the reasonable assumption that among the acquired microphone signals there is one that can lead to a better recognition performance than the others. In order to identify the related microphone, it is worth addressing the attributes of the signal and the characteristics of the communication *channel* that shaped the uttered speech from the source to the sensor, and depends on the speaker location, the head orientation, and the room acoustics. The latter variabilities determine the overall reverberation effects that can be observed in the distant microphone signal. Environmental noise, although it is not the main focus of this work, also represents a relevant issue, in particular when it is more concentrated in some areas, *i.e.,* when it introduces more distortion into a subset of the available microphones.

Various CS methods have been presented in the literature, as reported in the following. Some of them rely on measures that quantify the effect of the channel on the speech signals. Examples of these measures are the envelope variance (EV) (Wolf and Nadeu, 2014) and the modulation spectrum ratio (Himawan et al., 2015). Also energy-based techniques can be applied to CS, in particular under controlled conditions as when a calibrated set of microphones is available (Wolf and Nadeu, 2010).

In a previous work, we presented an initial study of how objective signal quality measures, in particular the cepstral distance (CD), can be successfully applied to CS problem (Guerrero et al., 2016). However, we believe that an important requirement, for a more effective application of these quality measures to our problem, is an in-depth understanding of their relationship with DSR performance. In order to address this missing link between CS and DSR, this work aims to provide a novel methodology for assessing the performance and limitations of CS methods, as far as reverberation effects are concerned. To the best of our knowledge, this represents the first empirical study that characterizes, from a quantitative standpoint, the overall system behavior under parameters such as the distance between the speaker and microphones, the speaker orientation, and the microphone network configuration. Additionally, this work constitutes an extensive and deeper investigation of the CD based technique outlined in Guerrero et al. (2016). We discuss the effectiveness of CD to characterize the reverberation in a room e.g., relating it to the direct-to-reverberant ratio (DRR) feature, supporting its application to CS for DSR. Also, we present evidence that shows that CD based CS is strongly related to an oracle selection of the best recognized channels. Then, the investigated approach is analyzed under variations on the setup that regard the speaker position and orientation, and the microphone network configuration. Finally, we extend our findings and confirm the benefits of applying CS to DSR with the use of real data, on which the proposed method achieves a better performance than an EV based state-of-the-art method.

The remaining of this paper is organized as follows. In Section 2 multi-microphone processing for DSR is discussed. Specific parameters of the room acoustics are presented in Section 3. An overview of the most relevant CS methods is given in Section 4. CD based CS is elaborated in Section 5. In Section 6, details about the experimental framework are provided. The activities and analysis performed on the different experimental settings, and their corresponding results, are presented in Sections 7 and 8. Finally, in Section 9 the conclusions of the study and possible directions for future activities are discussed.

2. Multi-microphone processing for DSR

The problem of DSR in a multi-microphone setting comprises, on one hand, the techniques used for multi-microphone speech processing and, on the other hand, the acoustic properties of the reverberant environments.

Multi-microphone speech processing approaches have proved their potential to significantly improve DSR performance in comparison to single channel solutions. Various architectures can be adopted to process the multiple inputs in order to derive a single recognition output of a spoken utterance (Wölfel and McDonough, 2009; Kinoshita et al., 2013).