Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 170–180

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Channel selection measures for multi-microphone speech recognition

Martin Wolf*, Climent Nadeu

TALP Research Center, Department of Signal Theory and Communication, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain

Received 8 February 2013; received in revised form 10 September 2013; accepted 30 September 2013
Available online 16 October 2013

Abstract

Automatic speech recognition in a room with distant microphones is strongly affected by noise and reverberation. In scenarios where the speech signal is captured by several arbitrarily located microphones the degree of distortion differs from one channel to another. In this work we deal with measures extracted from a given distorted signal that either estimate its quality or measure how well it fits the acoustic models of the recognition system. We then apply them to solve the problem of selecting the signal (i.e. the channel) that presumably leads to the lowest recognition error rate. New channel selection techniques are presented, and compared experimentally in reverberant environments with other approaches reported in the literature. Significant improvements in recognition rate are observed for most of the measures. A new measure based on the variance of the speech intensity envelope shows a good trade-off between recognition accuracy, latency and computational cost. Also, the combination of measures allows a further improvement in recognition rate.

Keywords: Automatic speech recognition; Channel (microphone) selection; Signal quality; Multi-microphone; Reverberation

1. Introduction

The performance of state-of-the-art automatic speech recognition (ASR) systems tends to decrease when the distance between the speaker's mouth and the microphone grows, due to both noise and reverberation (Wölfel and McDonough, 2009). In many situations the use of close-talking microphones is not possible or practical, so a different solution is required. The use of multiple distant-talking microphones provides several options that may help to solve this problem.

In this work, we assume a practical, cost-effective and unconstrained multi-microphone scenario, where the microphones are arbitrarily located and may show a variety of characteristics. For instance, in a meeting room, some microphones may be hanging on the walls, others standing on the table, or they may be built in the personal communication devices of the meeting participants. Moreover, some of them may be omnidirectional, others directional or noise-canceling, etc. In such situation, where the positions of the microphones are either not known or fixed, the application of commonly used multi-microphone approaches, like array processing (Brandstein and Ward, 2001), becomes difficult.

An alternative is provided by channel selection (CS). Before any processing, the degree of signal distortion differs among the channels, depending on the microphone position and characteristics. Even if speech enhancement is applied, the processed speech signals will not be distorted equally, so some of them may be decoded with less recognition errors than others. Consequently, the ASR system may benefit if signals of higher quality are selected for further processing. To do so, a measure of distortion, or a measure of how well recorded or enhanced signals fit the set of acoustic models of the ASR system is needed.

As the word error rate (WER) is unknown during recognition, the main problem is to develop a measure, that allows to rank the channels in a way as close as possible to the WER based ranking. In this paper, several new measures are presented and compared, in terms of recognition

* Corresponding author. Tel.: +34 93 4016437; fax: +34 93 4017200.
E-mail addresses: martin.wolf@upc.edu (M. Wolf), climent.nadeu@upc.edu (C. Nadeu).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.015

performance, with CS measures found in the literature. For that purpose, we focus on the reverberation problem and take the following approach: only the best channel is selected and its corresponding signal is fed, without undergoing any de-reverberation process, to an ASR system trained with clean speech.

One of the advantages of CS is that it does not require a spatial structure of the microphone set, what simplifies the deployment and reduces the cost of the system. CS may also be combined with beamforming and used to reduce the number of channels. Although, in theory, a higher number of microphones in the array should lead to a better beamforming performance, in practice, it was shown that the use of all possible channels does not always increase the ASR accuracy (Obuchi, 2004, 2006; Kenichi Kumatani et al., 2011).

In the next section both already reported and new CS measures are described categorized, and their application in ASR is discussed. The experimental comparison of all methods is made in Section 3. In Section 4 we evaluate several techniques further and show how their performance depends on the amount of data that is used for the measure estimation.

2. Channel selection measures

The CS measures may be classified into two groups: signal-based and decoder-based measures. The signal-based measures are extracted from the signal or channel characteristics. These CS methods operate in the front-end and the decoder of the ASR system is not involved in the measure extraction. The decoder-based measures do not estimate the degree of signal quality using a signal-processing measure, but their estimation process includes some kind of classification, which may be directly related to the decoding part of the recognition system (e.g. by using likelihood, or posterior probabilities).

2.1. Signal-based measures

In this work, we assume reverberant environment. Reverberation is created in enclosed spaces when acoustic waves reflected by the walls and objects arrive to the microphone attenuated and with different delays, introducing undesirable and unpredictable interferences. The reverberation distortion is usually modeled through a convolution between the room impulse response (RIR) and the clean speech signal. That linear distortion from the acoustic channel can not be easily canceled or attenuated in the feature domain as it is routinely done in ASR for the linear distortions produced in the electric channel (microphone, amplifiers, telephone network, etc.), since the duration of the RIR is usually much longer than the electrical channel impulse responses and encompasses several consecutive phones.

2.1.1. Position and orientation

The information about the relative position and orientation of the speaker and microphone may be used for CS. Speech should be less distorted by reverberation if the microphone is closer to the speaker. The closest microphone may be, for instance, selected by measuring the time of arrival of the waveform. However, it was shown by Wolf and Nadeu (2010) that the information about the orientation is also important. This is mainly due to the attenuation of the signal by the head of the speaker, and the fact that speech used in training is usually recorded by a microphone in front of the speaker. Both position and orientation may be estimated either from multi-microphone audio processing, multi-camera video processing, or a combination of both. In any case, CS would have to rely on the output of another system, that may not always provide accurate measures and the knowledge about the positions of the microphones is needed, what puts additional demands on the system deployment.

2.1.2. Energy and signal-to-noise ratio

Another straightforward way to identify the least distorting channel could be the energy of the signal. A strong signal indicates that the sound was uttered with the speaker close and oriented towards the microphone, so the direct wave is presumably stronger relative to the reverberation. This very simple approach may achieve good results (Wolf and Nadeu, 2010), but one strong assumption must be made. In multi-microphone scenarios, attenuation in the electrical path among microphones varies for reasons like different wire length, varying volume set on preamplifier, etc. If we want to use signal energy as a reliable indicator of the signal quality, a perfect calibration of all microphones is needed, which is not a trivial task.

The problem of calibration could be avoided if the energy of the speech signal was normalized, for example, by the energy of the noise in the silent portions (assuming that some additive noise is present). This leads us to a signal to noise ratio (SNR). CS based on this measure was evaluated by Obuchi (2004) and Wölfel et al. (2006). If speech is recorded by distant-talking microphones, reverberation is often the dominant source of distortion. A problem associated to the use of the SNR is that it does not properly reflect that kind of distortion. Furthermore, an accurate SNR measurement can be hardly obtained, since the boundaries between the speech signal and the silent portions, where the noise power can be estimated, are less clear after the smearing effect of reverberation. Another disadvantage of energy-based measures in general is that they do not consider the specific characteristics of the speech signal (only its energy).

2.1.3. RIR related measures

Assuming constant conditions in the room, the RIR can be used to describe the propagation between the acoustic source and a given microphone. Relations between different parts of the RIR and the WER of the ASR system were investigated by Petrick et al. (2007). Authors showed that there are certain components of the RIR that harm speech recognition more than others. Consequently, if there was a