Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 895–910

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Dominant speaker identification for multipoint videoconferencing☆,☆☆

Ilana Volfin, Israel Cohen *

Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

Received 8 September 2011; received in revised form 28 February 2012; accepted 7 March 2012
Available online 17 March 2012

Abstract

A multi-point conference is an efficient and cost effective substitute for a face to face meeting. It involves three or more participants placed in separate locations, where each participant employs a single microphone and camera. The routing and processing of the audiovisual information is very demanding on the network. This raises a need for reducing the amount of information that flows through the system. One solution is to identify the *dominant speaker* and partially discard information originating from non-active participants. We propose a novel method for dominant speaker identification using speech activity information from time intervals of different lengths. The proposed method processes the audio signal of each participant independently and computes speech activity scores for the immediate, medium and long time-intervals. These scores are compared and the dominant speaker is identified. In comparison to other speaker selection methods, experimental results demonstrate reduction in the number of false speaker switches and improved robustness to transient audio interferences.

Keywords: Speech processing; Videoconference; Dominant speaker identification; Acoustic signal detection; Acoustic noise; Transient noise

1. Introduction

Multipoint videoconferencing technology has been existent since the early 1960s. Throughout this period it had transformed from an expensive technology restricted for use in large organizations, to cheap and easy to use applications available in almost every home. In multipoint videoconferencing, three or more dispersedly located participants connect for a meeting over telephone or Internet-based networks. Typically the meeting is controlled by a central processing unit, which is in charge of routing signals between participants. The incorporation of video into audioconferencing had significantly raised the amount of information transmitted through the network. In addition to increased bandwidth consumption, it raises the amount of information that is processed by the central processing unit. An effort has been made to offer solutions for reducing the load on the network. Most of these solutions involve the identification of the most active participants through a process referred to as *speaker selection*. Once the active speakers are selected, the remaining audiovisual information may be discarded, thus relieving the network.

☆ This paper has been recommended for acceptance by Hugo Van hamme, PhD.

☆☆ This research was supported by the Israel Science Foundation (grant no. 1130/11).

* Corresponding author. Tel.: +972 4 8294731; fax: +972 4 8295757.
E-mail addresses: ilana.volfin@gmail.com (I. Volfin), icohen@ee.technion.ac.il (I. Cohen).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.03.002

Many works in the field of improving the efficiency of data traffic in audio or videoconferencing rely on speaker selection as a vital component (Shaffer and Beyda, 2004; Howard et al., 2004; Matsumoto and Ozawa, 2010). However, little research attention has been devoted to the speaker selection task itself. The simple methods are based on indicators of the signal level in the channel as measured by its amplitude or mean power (Kwak et al., 2002; Chang, 2001; Kyeong Yeol et al., 1998; Firestone, 2005). In these methods, the most active speakers are selected as the speakers with the highest signal level. Since the selection is based on an instantaneous measure, these methods are known to cause frequent false speaker switches. A method with a more advanced switching mechanism was proposed in Smith et al. (2002). In this method, the active parties are identified by either the signal power or the arrival of silence insertion descriptor (SID) frames. They are then ranked by the order of becoming active speakers. A speaker can be promoted in ranking only if its smoothed signal power exceeds a certain *barge-in* threshold. The ranking list keeps a continuous record of the M most active participants.

An improvement to Smith et al. (2002) is proposed in Xu et al. (2006) by suggesting a more sophisticated method for speech detection. In this method, the speech detection is based on a set of speech specific features and a machine learning technique that classifies each signal frame into either voice or noise. The above-mentioned methods, although constituting an advancement over the level based methods, still concentrate on instantaneous measures for speech activity. No special attention is devoted to long-term properties of dominant speech in the speaker switching mechanism. The barge-in mechanism, that is proposed as the switching mechanism, increases the vulnerability of these algorithms to false switching due to transient interferences.

In this paper, we introduce a novel approach for dominant speaker identification based on speech activity evaluation on time intervals of different lengths. The lengths of the time intervals we use correspond to a single time frame, a few phonemes, and a few words up to a sentence. This mode of operation allows capturing basic speech events, such as words and sentences. Sequences and combinations of these events may indicate the presence of dominant speech activity (or lack of it). Another unique ability offered by the proposed method is a distinction between transient audio occurrences that are isolated and those that are located within a speech burst.

Integration of long-term speech information had already been proven effective in voice activity detection (VAD) applications (Sohn et al., 1999; Ramirez et al., 2004, 2005). Long term information was used in the aforementioned methods in order to determine whether speech is present in a currently observed time-frame. We find this approach well suited to our problem since dominant speech activity in a given time-frame would be better inferred from a preceding time interval than from any instantaneous signal property. Hence we incorporate the approaches from these VAD works into the proposed method. Objective evaluation of the proposed method is performed on a synthetic conference with and without the presence of transient audio occurrences. In addition we test the proposed method on a segment of a real five channel audioconference. Results are compared with existing speaker selection algorithms. We show reduction in the number of false speaker switches and improved robustness to transient audio interferences.

The paper is organized as follows. In Section 2, we formulate the problem of dominant speaker identification. In Section 3, we present the proposed method. We present two approaches for speech activity score evaluation in Section 4, where one is based on a single observation and the other introduces temporal dependence between consecutive time-frames using the score on a sequence of observations. Experimental results are presented in Section 5. This work is concluded in Section 6.

2. Problem statement

A multipoint conference consists of N participants received through N distinct channels. The objective of a dominant speaker identification algorithm is to determine at a given time which one of the N participants is the dominant speaker. We discuss an arrangement where each participant receives a video feed from only one other participant. In the proposed embodiment, the video stream of the dominant speaker is sent to all participants while the dominant speaker himself receives the video stream from the previous dominant speaker. Throughout this paper we use the terms channel, participant, user and speaker interchangeably, as referring to a conference end-point.

We define a *speech burst* as a speech event composed of three sequential phases: initiation, steady state and termination. In the first phase, speech activity builds up. During the second phase speech activity is mostly high, but it may include breaks in activity due to pauses between words. Finally, in the third phase speech activity declines and then stops. Typically, a dominant speech activity is composed of one or more consequent speech bursts. We refer to the point where a change in dominant speaker occurs as a *speaker switch* event.