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.

The challenges in the dominant speaker identification problem arise from equipment, surrounding and personal characteristics of the different users. The type of equipment used may introduce noise, such as crosstalk (speakers) or reverberations (far-talking microphone). The quality of the sensor affects the SNR of the incoming signal. The type and level of noise in the surrounding of the speaker influence the ability of the system to identify each speaker as dominant. The presence of transient noises, characterized by short duration and high energy signals, may distract the decision regarding the dominant speaker. Finally, personal characteristic of the speaker, such as loudness or quality of voice, may also affect the identification of the dominant speaker.

The desired behavior of a dominant speaker identification algorithm is as follows.

- No false switching should occur during a dominant speech burst. Both transient noise occurrences and single words that are said in response to or in agreement with the dominant speaker are considered transient occurrences. These should not cause a speaker switch.
- A speaker switch event cannot occur during a break in speech between two dominant speakers. It has to be triggered by a beginning of a speech burst.
- A tolerable delay in transition from one speaker to another, in a speaker switch event, is up to 1 s.
- When simultaneous speech occurs on more than one channel, the dominant speaker is the one who began speaking first.
- The relative loudness of the voice of a speaker should not influence his chance to be identified as the dominant speaker.

3. Dominant speaker identification based on time intervals of variable lengths

The proposed method for dominant speaker identification consists of two stages, a *local processing* and a *global decision*, as depicted in Fig. 1. In the first stage (Fig. 1, blocks 1 and 2), speech activity scores are evaluated for the immediate, medium, and long time-intervals. The lengths of the chosen time-intervals correspond with the lengths of one time-frame, a few phonemes, and several words. In the second stage (Fig. 1, block 3), the dominant speaker is identified based on the speech activity scores obtained in the first stage. This stage is designed to detect speaker switch events. We assume that a speaker switch event can be inferred from a rise in the three speech activity scores on a certain channel, relatively to scores of the dominant channel. The rationale of speaker switch event detection is discussed in Section 3.2. The implementation of the proposed algorithm is summarized in Fig. 2.

3.1. Local processing

In this stage, the signal in each channel is processed separately. The objective of our approach is to place each signal frame into a broader context than its instantaneous audio activity. This is accomplished by processing the currently observed frame by itself in addition to a medium-length preceding time interval and in addition to a long time interval that precedes it. Thus each time we move up to a longer time interval, the speech activity obtained in the previous step is analyzed again in a broader context.

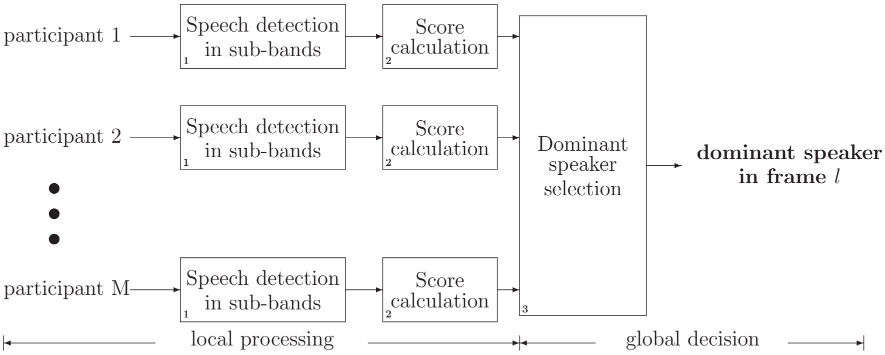

Fig. 1. A flow-chart describing the proposed method. Block 2 is described in more detail in Fig. 3.

```
For all channels i
    For all time frames $l$
        For frequency bins $k_1 < k < k_2$
            Compute $\underline{\xi}_{i,l} = \{\xi_i(k,l) | k \in [k_1, k_2]\}$, using the OMLSA algorithm [13].
        Count the number of active sub-bands in $\underline{\xi}_{i,l}$, using:
        $a_{i,1}(l) = \sum_{k=k_1}^{k_2} \text{sign}[\max(\xi_i(k,l) - \xi_{th}, 0)]$
        where sign$(x)$ is the Signum function.
        Compute the score $\Phi_{i,l}^{\text{immediate}}$ using one of the methods described in Section IV.
        Construct the vector $\underline{\alpha}_{i,l} = \{a_{i,1}(l-m) | m \in [0, N_2 - 1]\}$
        Compute $a_{i,2}(l) = \sum_{m=0}^{N_2-1} \text{sign}[\max(\alpha_{i,l}(m) - \alpha_{th}, 0)]$
        Compute score $\Phi_{i,l}^{\text{medium}}$ using one of the methods described in Section IV.
        Construct the vector $\underline{\beta}_{i,l} = \{a_{i,2}(l - mN_2) | m \in [0, N_3 - 1]\}$
        Compute $a_{i,3}(l) = \sum_{m=0}^{N_3-1} \text{sign}[\max(\beta_{i,l}(m) - \beta_{th}, 0)]$
        Compute score $\Phi_{i,l}^{\text{long}}$ using one of the methods described in Section IV.
    Perform comparison of scores $\{\Phi_{i,l}^{\text{immediate}}, \Phi_{i,l}^{\text{medium}}, \Phi_{i,l}^{\text{long}}\}$ across channels, as described in
    Section III-B and determine which of the channels contains dominant speech.
```

Fig. 2. Dominant speaker identification algorithm based on speech activity information from time intervals of different lengths.

The motivation for this mode of processing is the nature of the signals we expect to receive through the channels. The expected types of signals during a multipoint conference are:

(1) silence or stationary noise at different power levels.
(2) transient audio occurrences such as knocks, coughing, and sneezing.
(3) fluent and continuous speech consisting of words and sentences.

Each of these signal types would yield a typical combination of score values. This would allow us to discriminate between the different types of signals. Specifically, in the *global decision* stage it would enable the discrimination of dominant speech activities on a certain channel from non-dominant activity on other channels.

In the proposed approach, we relate to each time interval as composed of smaller sub-units. The speech activity in each time interval is determined according to the number of *active* sub-units by attributing a *speech activity* score to this number. The score is obtained from the likelihood ratio between hypothesis of speech presence and hypothesis of speech absence. The score evaluation method is fully described in Section 4. The speech activity evaluation process consists of three sequential steps, referred to as *immediate, medium* and *long*. The input into each step is a sequence of the number of active sub-units acquired in the previous step.

For the step of immediate speech activity evaluation we use a frequency representation of the frame to test for speech activity in sub-bands. We operate on the frequency range that corresponds to the range of voiced speech. Let this range be denoted by $k \in [k_1, k_2]$ and the total number of sub-bands in this range by N_1. As the frequency representation we use the SNR value in each sub-band. The SNR is obtained from the OMLSA algorithm (Cohen, 2002), Matlab implementation is available at Cohen (2012). Let it be denoted by $\underline{\xi}_l = \{\xi(k, l) | k \in [k_1, k_2]\}$, where l is a discrete time index. Next we find the number of *active* sub-bands in the frame. A sub-band is considered active if its SNR is higher than a threshold ξ_{th}. The threshold value was obtained from a speech training set. The number of active sub-bands in a time-frame l is denoted by $a_1(l)$. Specifically,

$$a_1(l) = \sum_{k=k_1}^{k_2} \text{sign}[\max(\xi(k,l) - \xi_{th}, 0)] \quad (1)$$

where sign(x) is the Signum function.

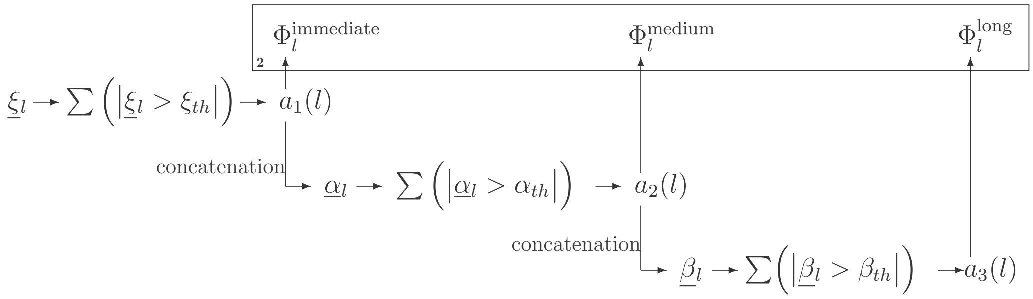
$\Phi_l^{\text{immediate}}$
Φ_l^{medium}
Φ_l^{long}
2
$\underline{\xi}_l \rightarrow \sum\left(\left|\underline{\xi}_l > \xi_{th}\right|\right) \rightarrow a_1(l)$
concatenation
$\underline{\alpha}_l \rightarrow \sum\left(\left|\underline{\alpha}_l > \alpha_{th}\right|\right) \rightarrow a_2(l)$
concatenation
$\underline{\beta}_l \rightarrow \sum\left(\left|\underline{\beta}_l > \beta_{th}\right|\right) \rightarrow a_3(l)$

Fig. 3. Speech-activity-score evaluation process.

The thresholding approach serves several purposes in the discussed problem. It allows measuring the amount of speech activity while suppressing isolated high-energy noise spikes. One advantage is avoiding the masking effect that is caused when using the absolute value of isolated high-energy spikes. Another advantage is an equalization effect between loud and quiet activity of comparable quality speech segments. For example, in the immediate time processing, high amplitude noise in a few isolated frequency bins would not be identified as speech. Whereas a measure that relies on the absolute value of the SNR would cause such activity to be identified as speech activity. We proceed with the thresholding approach in the next steps of speech activity detection motivated by these advantages.

The number of active sub-bands is provided as an input into the *score calculation* block, see Fig. 1, block 2. In this block, two additional thresholding steps are carried out for time intervals of medium and long lengths. The input into the speech activity evaluation step for the medium length time interval is a sequence of the number of active sub-bands in the last N_2 frames. We denote it by $\underline{\alpha}_l = \{a_1(l-m)|m \in [0, N_2-1]\}$. Next, $\underline{\alpha}_l$ is thresholded by α_{th} and the number of active frames in the medium length time interval that precedes the time-frame l is obtained. The number of active frames is denoted by $a_2(l)$, where $0 \leq a_2(l) \leq N_2$. Specifically,

$$a_2(l) = \sum_{m=0}^{N_2-1} \text{sign}[\max(a_1(l-m) - \alpha_{th}, 0)] \tag{2}$$

The number $a_2(l)$ indicates the amount of independent instantaneous activities in a group of N_2 sequential frames. Let us assume that a certain frame p exhibited a high number of active sub-bands, $a_1(p)$. The amount of activity in its neighbors would determine if the frame p contains an isolated noise spike or that it is part of a longer audio activity. The amount of activity in the medium time-interval is indicated by $a_2(p)$.

Finally, a sequence of speech activity indicators for medium time-intervals is provided as an input into the step of evaluating speech activity in a long time interval. This sequence is denoted by $\underline{\beta}_l = \{a_2(l-mN_2)|m \in [0, N_3-1]\}$, where N_3 is the number of medium-length blocks constituting the long time-interval. The number of active medium-length blocks, denoted by $a_3(l)$, is obtained by thresholding $\underline{\beta}_l$ by β_{th} and counting the number of non-zero elements. Specifically,

$$a_3(l) = \sum_{m=0}^{N_3-1} \text{sign}[\max(a_2(l-mN_2) - \beta_{th}, 0)]. \tag{3}$$

An active block of medium length indicates a short transient occurrence. According to the rationale in the previous steps, a low $a_3(l)$ value indicates that the transient is isolated, while a high $a_3(l)$ value indicates a part of a speech burst.

After obtaining $a_1(l)$, $a_2(l)$ and $a_3(l)$, we have a good representation of the speech activity history in time-frame l. We also achieved here a *backward inference* by each of these values on the shorter time-intervals. A speech activity score is now attributed to the speech activity indicators, $a_1(l)$, $a_2(l)$ and $a_3(l)$. The score evaluation process is illustrated in Fig. 3, and detailed in Section 4. We denote the set of scores in frame l by $\Phi_l^{\text{immediate}}$, Φ_l^{medium} and Φ_l^{long}. The scores from the distinct channels are provided into the *dominant speaker selection* block (Fig. 1), where this information is translated into a dominant speaker identification.

```
IF ($l$ mod *decision interval* $== 0$) DO:
    COMPUTE
        $\underline{c}_1 = \log\left(\frac{\Phi_l^{\text{long}}(\text{everyone})}{\Phi_l^{\text{long}}(\text{dominant})}\right)$
        $\underline{c}_2 = \log\left(\frac{\Phi_l^{\text{medium}}(\text{everyone})}{\Phi_l^{\text{medium}}(\text{dominant})}\right)$
        $\underline{c}_3 = \log\left(\frac{\Phi_l^{\text{immediate}}(\text{everyone})}{\Phi_l^{\text{immediate}}(\text{dominant})}\right)$
    IF exists $\{j : c_1(j) > C_1 \ \& \ c_2(j) > C_2 \ \& \ c_3(j) > C_3\}$,
        $j^* = \arg\max_j \{c_2(j) : c_2(j) > C_2\}$
        $\text{Dominant}(l) = j^*$
    ELSE
        $\text{Dominant}(l) = \text{Dominant}(l-1)$
ELSE
    $\text{Dominant}(l) = \text{Dominant}(l-1)$
```

Fig. 4. The dominant speaker identification algorithm.

3.2. Global decision

The objective of this stage is to identify the channel associated with the dominant speaker. This stage is activated in time steps of a certain interval, which is referred to as the *decision-interval*. It is designed to utilize the scores that are obtained in the local processing stage for dominant speaker identification. The approach we take in this stage is detecting *speaker switch* events, rather than selecting a dominant speaker in every decision-interval. Once a dominant speaker is identified, he remains dominant until the speech activity on one of the other channels justifies a speaker switch. In the following, we refer to the non-dominant channels as *competing* (for dominance).

The expected score behavior of the dominant speaker and a channel that justifies a speaker switch are described as follows. The type of channels (dominant or competing) referred to by the score is indicated in brackets. During the dominance period of the dominant speaker, the score Φ^{long}(dominant) is expected to be high. The scores for the immediate and medium time-intervals, Φ^{medium}(dominant) and $\Phi^{\text{immediate}}$(dominant), are allowed to be low for periods that correspond to breaks between words. As for the speech activity on the competing channels that justifies a speaker switch, all three scores are expected to be high. The speech activity in the long time-interval, Φ^{long} (competing) has to be high to demonstrate prolonged speech activity. In addition, both Φ^{medium} (competing) and $\Phi^{\text{immediate}}$ (competing) are expected to be high to indicate an onset of speech.

The speaker selection algorithm is illustrated in Fig. 4. We propose the following realization for detecting speaker switch events in accordance with aforementioned expected behavior of scores. For each decision-interval three vectors, $\underline{c}_1$, $\underline{c}_2$ and $\underline{c}_3$, are calculated. These vectors contain the information about the ratio between speech activity in the competing channels and speech activity in the dominant channel for the long, medium and immediate time-intervals, respectively. The number of elements in each vector is the number of conference participants. In the next step, a set of channels demonstrating speech activity that might justify a speaker switch is found by comparing the relative speech activities $\underline{c}_1$, $\underline{c}_2$ and $\underline{c}_3$, with a set of respective thresholds, C_1, C_2 and C_3. The set of indices representing these channels is stored in j. In case j contains a reference to more than one channel, a dominant channel, denoted by j^*, is selected based on the highest speech activity score for the medium time-interval.

4. Speech-activity-score evaluation

In this section, we formulate the speech-activity-score evaluation method. As discussed in Section 3, the speech activity score for a certain time-interval is determined by the number of active sub-units in a representative vector. We consider the log-likelihood ratio of the number of active sub-units as the respective speech activity score. In order to determine the likelihood ratio of the number of active sub-units, we assume a certain model under hypotheses of speech presence and absence, denoted by H_1 and H_0, respectively. We propose two approaches for the score calculation method: the first approach uses the information from a single observation, and the second approach makes use of a sequence of observations in the likelihood-ratio formulation process. We also discuss the difference between the two approaches and the difference in performance.

4.1. Modeling the number of active sub-units

Let the representative vector for the immediate, medium or long time intervals be denoted by $\underline{\nu}_l = [\nu(l), \nu(l-1), \ldots, \nu(l-N_R+1)]$. The length of this vector is denoted by the parameter N_R. The elements of the representative vector for the immediate time interval are $\xi(k, l)$, where $k_1 \le k \le k_2$. For the medium and long time intervals the elements are $a_1(l)$ and $a_2(l)$, respectively. Accordingly, the lengths of the representative vectors are N_1, N_2 and N_3.

The vector $\underline{\nu}_l$ is thresholded by the threshold value ν_{th}, resulting in a binary vector $\underline{\nu}_{l,\text{binary}}$. Then the elements of the vector $\underline{\nu}_{l,\text{binary}}$ are summed

$$v(l) = \sum_{m=1}^{N_R} \nu_{\text{binary}}(m). \tag{4}$$

The value $v(l)$ represents the number of active sub-units out of the total number of entries N_R in the original vector $\underline{\nu}_l$. We propose to model this number as follows:

(1) *Given H_1 : speech is present*

We regard every active sub-unit as a success in a Bernoulli trial, where $P(x) = p^x(1-p)^{(1-x)}$ with $x \in \{0, 1\}$ and p is the probability of success, which remains the same for all vector entries. The vector length is N_R, thus we compute the probability of $v(l)$ successes out of N_R experiments. Hence, we assume this number follows a Binomial distribution

$$P(v(l)|H_1) \sim \text{Bin}(N_R, p) = \begin{pmatrix} N_R \\ v \end{pmatrix} p^{v(l)}(1-p)^{N_R - v(l)}. \tag{5}$$

(2) *Given H_0: speech is absent*

When speech is absent, we expect a lower probability for a higher number of active sub-units. Hence we assume an Exponential distribution

$$P(v(l)|H_0) \sim \exp(\lambda) = \lambda e^{-\lambda v(l)}. \tag{6}$$

4.2. Score evaluation

Given an observation vector X_l in time-frame l and two possible classes of its origin, H_0 and H_1, the likelihood of the observation to belong to each class $i \in \{0, 1\}$ is given by $p(X_l|H_i)$. Accordingly, the likelihood ratio is given by (Fukunaga, 1990)

$$\Lambda_l = \frac{p(X_l|H_1)}{p(X_l|H_0)} \tag{7}$$

We define the speech activity score as the log-likelihood ratio of the observation vector. It is obtained from (7) by

$$\Phi_l = \ln\left(\frac{p(X_l|H_1)}{p(X_l|H_0)}\right). \tag{8}$$

In the following sections, two approaches for score evaluation are proposed. In the first approach, X_l is the number of active sub-units in the current representative vector. In the second approach, X_l is a vector consisting of a sequence of the number of active sub-units on a sequence of the respective representative vectors. The scores are hereafter referred to as *Binomial* and *Binomial-sequential* respectively.

4.2.1. Single observation approach

In this approach the score is based on the number of active sub-units $v(l)$ in the representative vector $\underline{v}_l$. Substituting the model assumptions from (5) and (6) into (7), we have the likelihood ratio for a single observation

$$\Lambda_l^{\text{single}} = \frac{\begin{pmatrix} N_R \\ v \end{pmatrix} p^{v(l)}(1-p)^{N_R - v(l)}}{\lambda e^{-\lambda v(l)}} \tag{9}$$

According to (8), the speech activity score based on a single observation is:

$$\begin{aligned}\Phi_l^{\text{single}} &= \ln\left(\frac{p(X_l = v(l)|H_1)}{p(X_l = v(l)|H_0)}\right) = \ln\left(\frac{\begin{pmatrix} N_R \\ v \end{pmatrix} p^{v(l)}(1-p)^{N_R - v(l)}}{\lambda e^{-\lambda v(l)}}\right) \\ &= \ln\begin{pmatrix} N_R \\ v(l) \end{pmatrix} + v(l)\ln\, p + (N_R - v(l))\ln(1-p) - \ln\,\lambda + \lambda v(l).\end{aligned} \tag{10}$$

4.2.2. Multiple observation approach

We follow the approach proposed in Sohn et al. (1999) for a VAD application. In this method, an HMM is used to recursively update the likelihood ratio of frame l using all previous frames.

We base the score in this approach on a sequence of N sequential values of active sub-units $\underline{v}_l^N = [v(l-N+1), v(l-N+2), \ldots, v(l)]$ taken from N preceding and including the observed time-frame, l. The hidden Markov model here consists of two states:

$$\zeta_l = \begin{cases} H_{1,(l)}, & \text{speech is present in frame}\, l, \\ H_{0,(l)}, & \text{speech is absent in frame}\, l, \end{cases}$$

with the state dynamics described by $a_{ij} = p(\zeta_l = j|\zeta_{l-1} = i)$. For speech signals, it is more likely that a frame of speech would be followed by a frame of speech rather than by a frame of silence. This notion is fulfilled by setting a_{00}, $a_{11} > a_{01}$, a_{10}.

The likelihood-ratio in this approach is

$$\Lambda_l^{\text{sequential}} = \frac{p(X_l = \underline{v}_l^N|H_1)}{p(X_l = \underline{v}_l^N|H_0)} = \frac{p(\underline{v}_l^N, H_1)}{p(\underline{v}_l^N, H_0)} \cdot \frac{P(H_0)}{P(H_1)} \tag{11}$$

where $P(H_0)$ and $P(H_1)$ are steady state probabilities that are determined by $P(H_0) = a_{10}/(a_{10} + a_{01})$ and $P(H_1) = a_{01}/(a_{10} + a_{01})$.

Introducing the notations $\alpha_l(1) \triangleq p(\underline{v}_l^N, H_1)$ and $\alpha_l(0) \triangleq p(\underline{v}_l^N, H_0)$, $\alpha_l(j)$ for $j \in \{0, 1\}$ is recursively computed using the forward procedure (Shimkin, 2009):

$$\alpha_k(j) = \begin{cases} P[X_k = v(k)|H_{j,(k)}] \cdot [\alpha_{k-1}(0)a_{0j} + \alpha_{k-1}(1)a_{1j}], & l-N+1 < k \le l \\ P[X_k = v(k)|H_{j,(k)}]p(H_{j,(k)}), & k = l-N+1. \end{cases}$$

Denote the recursively computed term

$$L_l = \frac{\alpha_l(1)}{\alpha_l(0)} = \frac{p(v(l)|H_1)}{p(v(l)|H_0)} \cdot \frac{\alpha_{l-1}(0)a_{01} + \alpha_{l-1}(1)a_{11}}{\alpha_{l-1}(0)a_{00} + \alpha_{l-1}(1)a_{10}} = \Lambda_l^{\text{single}} \frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}}. \tag{12}$$

It is important to note for future discussion that the term L_l is large when speech is present, and small in the absence of speech.

Substituting (12) into (11), we have

$$\Lambda_l^{\text{sequential}} = L_l \cdot \frac{P(H_0)}{P(H_1)} = \Lambda_l^{\text{single}} \cdot \frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}} \cdot \frac{P(H_0)}{P(H_1)} \tag{13}$$

where a_{10} and a_{01} are determined a priori. Finally, we have the speech activity score formulation for the approach based on a sequence of observations

$$\Phi_l^{\text{sequential}} = \Phi_l^{\text{single}} + \ln\left(\frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}}\right) + \ln\left(\frac{P(H_0)}{P(H_1)}\right). \tag{14}$$

The intuitive advantage of the sequential approach may be observed in the structure of the sequential score (14). The sequential score (14) differs from the single score Φ_l^{single} by the addition of the terms $\ln((a_{01} + L_{l-1}a_{11})/(a_{00} + L_{l-1}a_{10}))$ and $\ln(P(H_0)/P(H_1))$. The term $\ln(P(H_0)/P(H_1))$ is a constant bias term added to the score in all cases. The term $\ln((a_{01} + L_{l-1}a_{11})/(a_{00} + L_{l-1}a_{10}))$, on the other hand, influences the score differently in the presence or absence of speech. With the constant values a_{00}, $a_{11} > a_{01}$, a_{10} set a priori:

(1) In the presence of speech, L_{l-1} is large so the term $\ln((a_{01}+L_{l-1}a_{11})/(a_{00}+L_{l-1}a_{10})) \rightarrow \ln(a_{11}/a_{10}) > 0$, hence $\Phi_l^{\text{sequential}} > \Phi_l^{\text{single}}$.

(2) When speech is absent, L_{l-1} is small and the term $\ln((a_{01}+L_{l-1}a_{11})/(a_{00}+L_{l-1}a_{10})) \rightarrow \ln(a_{01}/a_{00}) < 0$. So that in this case, $\Phi_l^{\text{sequential}} < \Phi_l^{\text{single}}$.

Hence, the *Binomial-sequential* score achieves better separability between speech presence and absence in comparison to the *Binomial* score. On the other hand, an integration of more observations from the past introduces a delay into the speaker switching process. This happens because the new information regarding the *speaker switch* is masked by the old information of dominance. Both these differences are exhibited in the experimental results in Section 5.

5. Experimental results

In this section, we compare the performance of the proposed method to other dominant speaker identification methods. Since a standard evaluation framework for the task of dominant speaker identification does not exist, we propose several experiments and objective error measures to test the basic requirements for this type of system.

Throughout this section, we denote the proposed method when used with the single-observation approach for the score evaluation as the *Binomial* method and when used with the sequential-observation approach as the *Bin-Seq* method. In case the distinct name of the method is not specified, the performances of the two methods were similar.

According to the speech activity score evaluation method described in Section 3, speech activity in each time interval is determined by the number of its active sub-units. Hence, the most influential parameters are the length of each time interval and its corresponding threshold. The number of sub-units was chosen in correspondence with the speech structure it represents. The immediate time interval is represented by the sub-bands that correspond to voiced speech. The medium time interval consists of the number of immediate time sub-units that is equivalent to several tenths of a second. The long time interval consists of a number of medium time sub-units that sum up to approximately 1 s. The thresholds in this method were set so they facilitate the separation between speech and noise. The value of the threshold represents the trade-off between the false detection of noise as speech and the false rejection of speech sub-units. In the immediate time interval, the SNR threshold separates between speech sub-bands (high SNR) and noise (low SNR). In the medium and long time intervals, the threshold provides the distinction between dense and sparse speech activities in the corresponding time intervals.

The likelihood model parameters for the immediate time are based on a training set taken from the TIMIT database (Garofolo, 1993). The remaining parameters were manually tuned on both simulated and real conference data. The parameter set that was used in the experiments is provided in Fig. 5.

Fs $= 16$ KHz	window length $= 64$ samples	overlap $= 50\%$	
$k_1 = 2$, $k_2 = 12$	$N_2 = 33$	$N_3 = 16$	
$\xi_{th} = 3$	$\alpha_{th} = 5$	$\beta_{th} = 32$	
$p_{\text{immediate}} = 0.5$	$p_{\text{medium}} = 0.5$	$p_{\text{long}} = 0.5$	
$\lambda_{\text{immediate}} = 0.78$	$\lambda_{\text{medium}} = 24$	$\lambda_{\text{long}} = 47$	
	$P(H_0) = P(H_1) = 0.5$		
$a_{10} = 0.1$	$a_{11} = 0.9$	$a_{01} = 0.2$	$a_{00} = 0.8$
$C_1 = 3$	$C_2 = 2$	$C_1 = 0$	

Fig. 5. Algorithm parameters that were used in all experiments.

In the first experiment, the dominant speaker identification algorithms are evaluated in a simple task of switching to the dominant speaker in the presence of stationary noise. For this purpose, a synthetic multipoint conference was simulated by concatenating speech segments taken from the TIMIT database. Three speakers were randomly chosen from the database and several speech bursts were concatenated on a distinct channel for each speaker. The speech bursts in each channel were spread along the conference length, such that each speech burst requires a switch in the dominant speaker. For the purpose of qualitative evaluation we assume there is no speech overlap between participants, i.e., "barging-in" is not allowed. White noise in the range of -2 to 5 dB SNR was added to all signals. The algorithms were applied to the signals and the dominant speaker was identified once in every time-period, denoted by a *decision interval*. Quantitative analysis was performed on the synthetic test set to examine whether an identification method fulfills the expectations as stated in Section 2.

The analysis includes the following measures:

- *False speaker switches* – number of false switches to a non-dominant speaker.
- *Front end clipping* (FEC) – error in detecting the beginning of a speech burst. The signal on each channel consists of several isolated speech bursts. The FEC error for each speech burst is obtained. Then the mean FEC error for the whole conference is computed as the mean value of individual FEC errors of all speech bursts in all channels. The FEC error in the discussed realization is a delay in the switching of the video to the new dominant speaker, while the audio signal from all speakers is available at all times. Thus we assume that a tolerable delay in switching to the dominant speaker in terms of mean FEC is 1 s. All tested methods fulfilled this requirement.
- *Mid sentence clipping* (MSC) – clipping occurring in the middle of a speech burst. This is the most disturbing type of error since it causes a switch of speaker in the middle of a dominant speech burst. It is computed as the ratio between the undetected mid section of a speech burst [samples] (UMSB) to the length of the speech burst [samples] (LOSB) in percent.

$$\text{MSC} = 100 \cdot \frac{\text{UMSB}}{\text{LOSB}}$$

The illustration of the MSC error is depicted in Fig. 6.

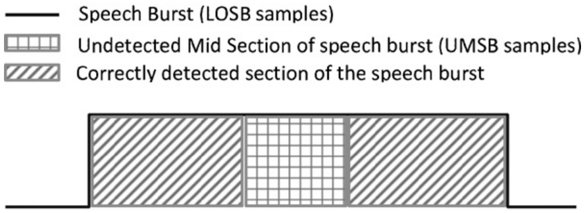
Speech Burst (LOSB samples)
Undetected Mid Section of speech burst (UMSB samples)
Correctly detected section of the speech burst

Fig. 6. Illustration of the MSC error. This error represents the undetected portion from the middle of a speech burst.

The performances of the two proposed methods were compared to the following methods:

- Three methods that identify the dominant speaker by applying a VAD to each channel and identifying the speaker with the highest VAD score as dominant. The VAD methods used for the comparison are denoted by *Ramirez*, *Sohn*, and *GARCH*, and are described in Ramirez et al. (2005), Sohn et al. (1999) and Mousazadeh and Cohen (2011), respectively.
- A method that identifies the dominant speaker as the one with highest signal power. It is referred to as the *POWER* method throughout the comparison.
- A method that identifies the dominant speaker as the one with the highest SNR. It is referred to as the *SNR* method throughout the comparison.

Since the comparison is made once in every decision interval, the instantaneous values of the VAD score, SNR and POWER, are not necessarily representative values of the decision interval. Hence the maximal values of the VAD score, SNR and POWER in the decision interval were used for the comparison. The results of this experiment are displayed in Fig. 7, where the false switching and MSC errors are plotted as a function of the decision interval. In Fig. 7(a), *POWER*, *SNR* and *VAD* based methods show frequent false speaker switching. For the proposed method, both the false switching and the MSC errors are zero.

In the second experiment, we test the robustness of the algorithms to transient noise. Transient noise occurrences of door knocks and sneezing were added to the signals in the synthetic conference of the first experiment. The quantitative influence of the transient occurrences is presented in Fig. 8 and can be compared to the results in Fig. 7. There is a rise both in the number of false switches and a respective rise of the MSC error for all methods. The proposed method is affected by the transient occurrences when a very short decision-interval (0.05–0.2 s) is used (Fig. 8(a)). The false switching that occurs with the proposed method is of shorter duration in comparison to the other methods. This can be observed in the relative rise of the MSC errors in Fig. 8(b) in comparison to Fig. 7(b) for decision interval in the range 0.05–0.2 s.

The difference between the two proposed methods, as discussed in Section 4.2.2, is shown in Fig. 8, where false switching occurs for decision-intervals of 0.05–0.2 s. The *Bin-Seq* method has a slight advantage over the *Binomial* method in the number of false speaker switches, as displayed in Fig. 8(a). This is expected because the *Bin-Seq* score is based on a sequence of observations. Thus when more information from the past is integrated into the decision, it takes more time for the dominance of a speaker to dissipate or to build up. In case of a transient, its duration is too short for building up speech activity that would indicate dominance. The same sequential processing causes a delay in switching back to the actual dominant speaker, as exhibited in Fig. 8(b) by the higher value of the MSC error.

In Fig. 9 we present a qualitative comparison in the presence of transient noise. The proposed method is compared to the *POWER* and two VAD based identification methods, *Ramirez* and *GARCH*. The decision interval in this comparison is 0.3 s. In general, it is noticeable that the algorithms we are comparing to are more responsive to a dominant speaker switch. This is due to the instantaneous nature of their scores. This responsiveness causes false switching to occur when there is a significant rise in energy on another channel, disregarding the cause for this rise. For the proposed method, the use of long-term information facilitates distinction between speech and transient audio occurrences.

The third experiment is a qualitative experiment where we compare the algorithms on a segment of a real 5 channel multipoint conference depicted in Fig. 10. On this segment, only channels 2 and 4 contain speech. Channel 1 contains a high level of stationary noise and some crosstalk from the other channels. Channels 3 and 5 contain only crosstalk from other channels. The *y* axis in Fig. 10 was scaled to the amplitude of the signals in the channel. Taking the maximal signal amplitude in channel 2 as a reference 1, the maximal signal amplitudes in channels 1, 3, 4 and 5 are 0.4, 0.5, 0.5 and 0.1, respectively. In this experiment, the proposed method was compared to the *POWER*, *RAMIREZ* and *SOHN* methods.

The proposed method switches correctly from channel 2 to channel 4. It ignores a high energy transient that occurs during the dominant speech burst in channel 4. From channel 4, the proposed algorithm switches back to channel 2 and stays on this channel. It remains on the dominant speaker in spite of an utterance of the words “yes yes” on channel 4 between 15th and 16th seconds. It also ignores a noise transient on channel 3 on the 17th second of the conference. Thus the proposed method behaves according to the requirements stated in Section 2. The *POWER*, *RAMIREZ* and *SOHN* methods, on the other hand, switch frequently to channel 1 that contains stationary noise, during the first dominant

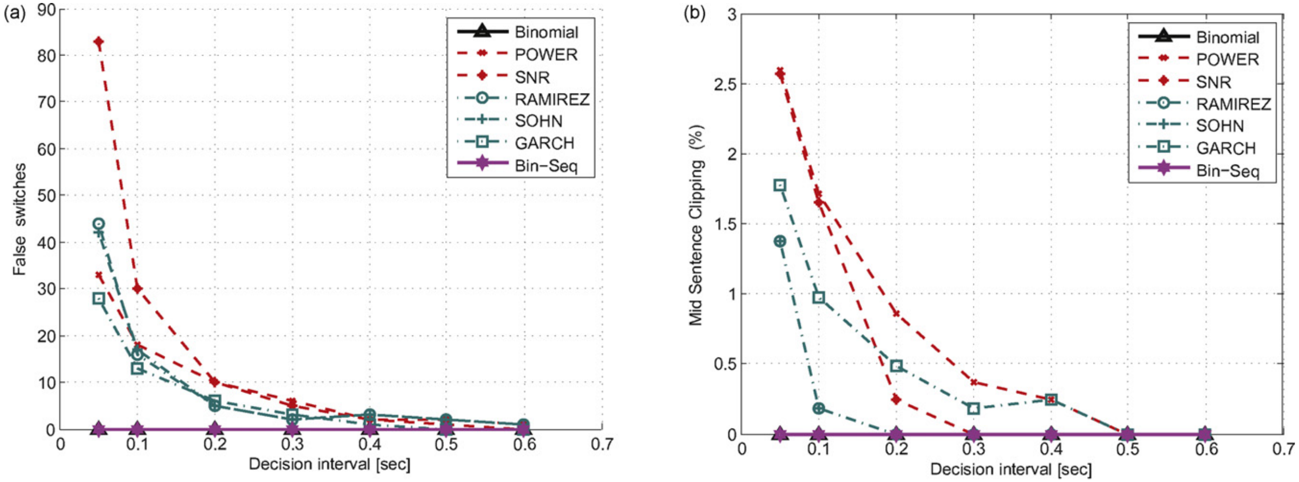

Fig. 7. Evaluation of the algorithms in the task of switching to the dominant speaker in the presence of stationary noise. The test data of this experiment consists of speech bursts concatenated such that each burst causes a speaker switch. The proposed methods are denoted by *Binomial* and *Bin-Seq*: (a) number of false speaker switches; (b) mid sentence clipping.

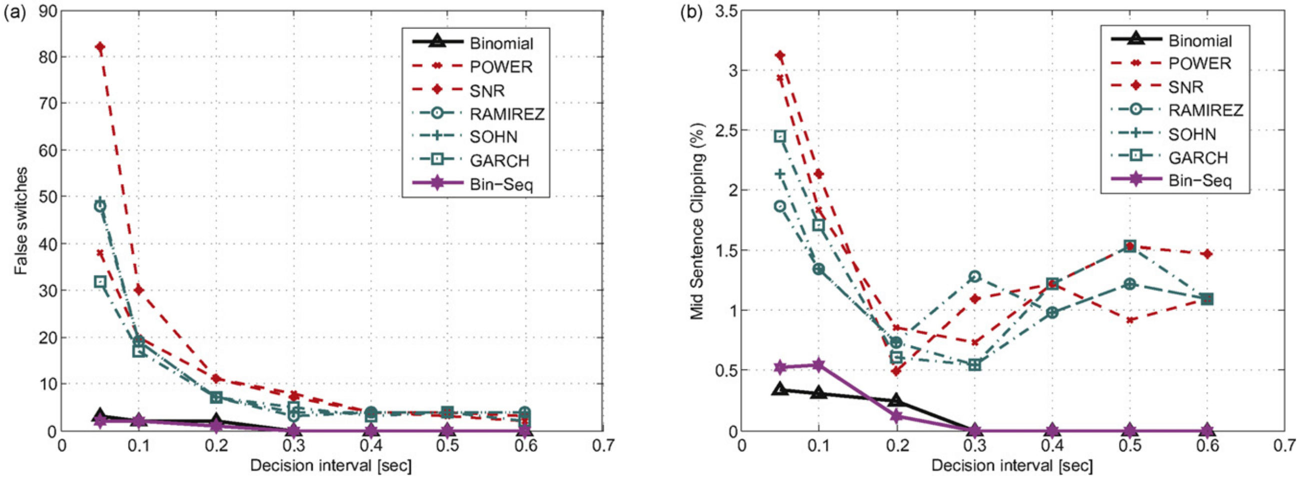

Fig. 8. Synthetic experiment with a presence of transient noise: (a) false speaker switches; (b) mid sentence clipping.

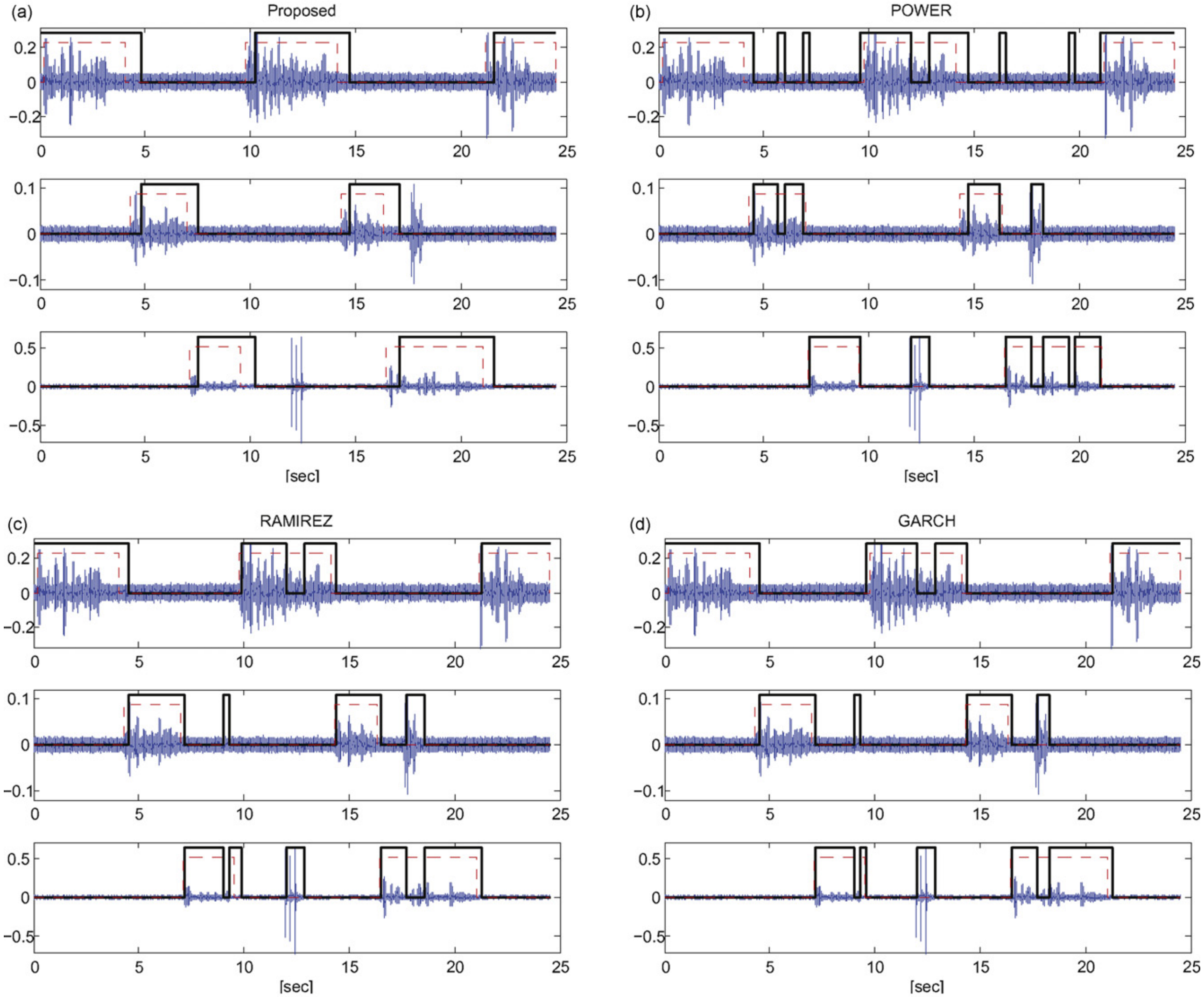

Fig. 9. Results of dominant speaker identification, for a decision-interval of 0.3 s: (a) dominant speaker identification by the proposed method; (b) dominant speaker identified by *POWER* method; (c) dominant speaker selected by the method based on Ramirez VAD; (d) dominant speaker selected by the method based on GARCH VAD; the decision of the algorithm is marked by the higher *solid bold* line and the hand marked decision is marked by the low *dashed* line.

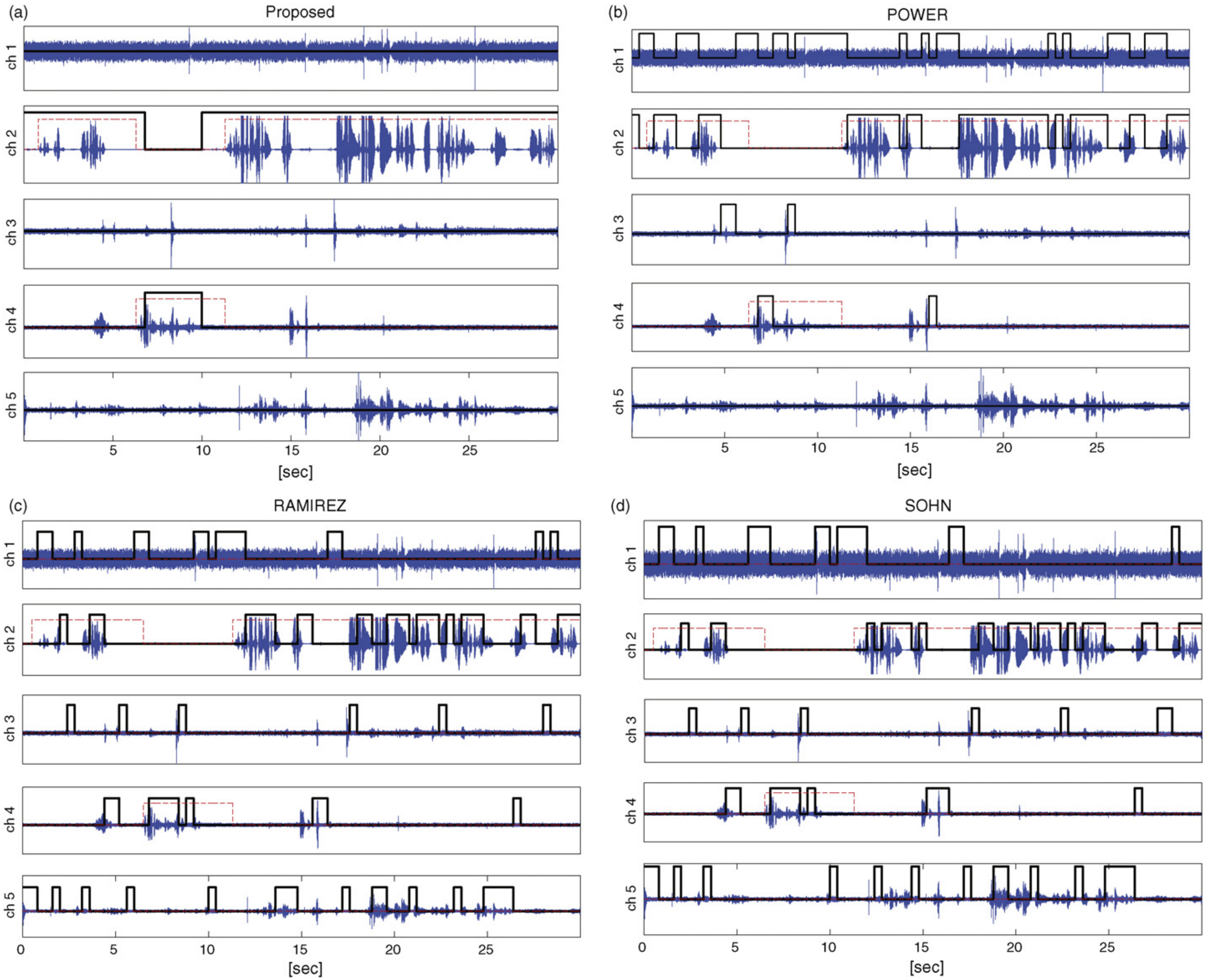

Fig. 10. Experimental results on a real 5 channel multi-point conference for a decision interval of 0.4 s; (a) dominant speaker is selected by the proposed method; (b) dominant speaker is selected by the *POWER* method; (c) dominant speaker is selected by the *RAMIREZ* method; (d) dominant speaker is selected by the *SOHN* method. The decision of the algorithm is marked by the high solid line and the hand marked decision is marked by the low dashed line.

speech burst in channel 2. They also switch to the noisy channels during the dominant speech burst in channel 4 and throughout the second dominance period in channel 2.

6. Conclusion

We have presented a novel dominant speaker identification method for multipoint videoconferencing. The proposed method is based on evaluation of speech activity on time intervals of different lengths. The speech activity scores for the immediate, medium and long time-intervals are evaluated separately for each channel. Then, the scores are compared and the dominant speaker in a given time-frame is identified based on the comparison. We proposed two approaches for the score evaluation method. A single observation based approach, for which the scores enable a faster reaction of the identification algorithm to speaker switches. In the second approach, the score is based on a sequence of observations. This makes the algorithm more robust to transient audio occurrences, but is slower in responding to changes. The information from time intervals of different lengths enables the proposed method to distinguish between speech and non-speech transient audio occurrences. Experimental results have demonstrated the improved robustness of the proposed method to transient audio interferences and frequent speaker switching in comparison to other speaker selection methods.

Acknowledgments

The authors thank Prof. Roger K. Moore and the anonymous reviewers for their constructive comments and useful suggestions.

References

Chang, Y.-F., 2001, October. Multimedia conference call participant identification system and method. US Patent No. 6,304,648 B1.

Cohen, I., 2002. Optimal Speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. 9 (April (4)), 113–116.

Cohen, I., 2012. Matlab Implementation of the OMLSA Algorithm. Available from: http://webee.technion.ac.il/people/IsraelCohen.

Firestone, S., 2005, November. Non-causal speaker selection for conference multicast. US Patent No. 6,963,353 B1.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd ed. Academic Press.

Garofolo, J.S., et al., 1993. TIMIT Acoustic–Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia.

Howard, M., Burns, R., Lee, C., Daily, M., 2004, August. Teleconferencing system. US Patent No. 6,775,247 B1.

Kwak, W., Gardell, S., Mayne Kelly, B., 2002, September. Speaker identifier for multi-party conference. US Patent No. 6,457,043 B1.

Kyeong Yeol, Y., Jong Hoon, P., Jong Hyeong, L., 1998. Linear PCM signal processing for audio processing unit in multipoint video conferencing system. In: Proc. IEEE Third Symposium on Computers and Communications (ISCC‘98), Athens, Greece, June 1998, pp. 549–553.

Matsumoto, K., Ozawa, K., 2010, February. Multi-point conference system and multi-point conference device. US Patent No. 7,667,729 B2.

Mousazadeh, S., Cohen, I., 2011. AR-GARCH in presence of noise: parameter estimation and its application to voice activity detection. IEEE Trans. Audio, Speech, Language Process. May, 916–926.

Ramirez, J., Segura, J.C., Benítez, C., de la Torre, Á., Rubio, A., 2004. Efficient voice activity detection algorithms using long-term speech information. Speech Commun. 42 (April), 271–287.

Ramirez, J., Segura, J., Benitez, C., Garcia, L., Rubio, A., 2005. Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process. Lett. 12 (October), 689–692.

Shaffer, S., Beyda, W., 2004, August. Reducing multipoint conferencing bandwidth. US Patent No. 6,775,247 B1.

Shimkin, N., 2009. Estimation and identification in dynamical systems (048825). Lect. Notes Fall.

Smith, P., Kabal, P., Rabipour, R., 2002. Speaker selection for tandem-free operation VoIP conference bridges. In: Proc. IEEE Workshop on Speech Coding, Tsukuba, Japan, October 2002, pp. 120–122.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (January), 1–3.

Xu, X., wei He, L., Florencio, D., Rui, Y., 2006. Pass: peer-aware silence suppression for internet voice conferences. In: IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006, pp. 2149–2152.