2012 IEEE 27-th Convention of Electrical and Electronics Engineers in Israel

Dominant Speaker Identification for Multipoint Videoconferencing

Ilana Volfin and Israel Cohen
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000
Israel
ilana.volfin@gmail.com; icohen@ee.technion.ac.il

***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. 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.**

I. Introduction

In multipoint videoconferencing, three or more dispersedly located participants connect for a meeting over telephone or Internet-based networks. The incoming audiovisual information from each user needs to be processed and routed through the network. The processing and routing of video signals in particular is very demanding. The demand is high both in terms of bandwidth consumption as well as creating a heavy load on the conference processing unit. A considerable amount of work had been dedicated to relieving this load. Most of the 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.

Many works in the field of improving the efficiency of data traffic in audio or videoconferencing rely on speaker selection as a vital component [1], [2]. However, little research attention has been devoted to the speaker selection task itself. The majority of existing methods rely on voice activity detection (VAD) as an indicator for the signal level in the channel [3], [4]. In these methods, the most active speakers are selected as the speakers with the highest voice activity score. Since the selection is based on an instantaneous measure, these methods are known to cause frequent false speaker switches.

In this paper, which summarizes the results in [5], 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. Objective evaluation of the proposed method is performed on a synthetic conference with and without the presence of transient audio occurrences. 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 rest of the paper is organized as follows. In Section II, we describe the conference arrangement that we intend to address. In Section III, we describe our two-stage algorithm for dominant speaker identification. In Section IV, we present the method for speech activity score calculation. Finally, Section V presents some experimental results.

II. 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. 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.

In accordance with this arrangement, we compare the dominant speaker identification algorithms using two objective measures. The first is the number of false dominant speaker switches. The second is the length of the false dominance time interval caused by the false switch.

III. Speech-Activity-Score Evaluation

We consider each of the three time-intervals, immediate, medium or long, as composed of N_R smaller sub-units. An *active* sub-unit is considered as one that is above a respective threshold. The speech activity score for each time-interval is determined by the number of its active sub-units. We take the log-likelihood ratio of this number as the speech activity score. In order to determine the likelihood ratio of the number of active sub-units, a likelihood model is assumed on this number, under the hypothesis that it originates in a speech or non-speech signal segment, denoted by H_1 and H_0 respectively.

For the step of immediate speech activity evaluation we use a frequency representation of the time-frame. As the frequency representation, we use the SNR values in the range of sub-bands that corresponds to voiced speech. This representation is obtained from the OMLSA algorithm [6]. 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. We test for speech activity in sub-bands. Each sub-band with SNR value above the threshold ξ_{th} is considered active. Consequently we construct the medium time representative vector from the number of active sub-bands in a sequence of N_2 immediate-time intervals. This vector is thresholded and summed. The number of active medium-time sub-units is then used to construct the long-time representative vector which consists of N_3 sub-units. This process and the respective speech activity score computation method are described in the following.

Let the representative vector for the immediate, medium or long time intervals be denoted by $\underline{\nu}_l = [\nu(l), \nu(l-1), ..., \nu(l-N_R+1)]$. The length of this vector is denoted by the parameter N_R.

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

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

The value $v(l)$ is 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. Under the hypothesis 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, equal 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:

$$P(v(l)|H_1) \sim \text{Bin}(N_R, p) = \binom{N_R}{v} p^{v(l)}(1-p)^{N_R - v(l)}\,. \tag{2}$$

Under the hypothesis H_0 speech is absent. We expect a lower probability for a higher number of active sub-units. Hence we assume an Exponential distribution of the number of active sub-units:

$$P(v(l)|H_0) \sim \exp(\lambda) = \lambda e^{-\lambda v(l)}\,. \tag{3}$$

IF (l mod *decision interval* $== 0$) **DO**:
  COMPUTE

$$\underline{c}_1 = \log\left(\frac{\Phi_l^{\text{long}}(\text{all})}{\Phi_l^{\text{long}}(\text{dominant})}\right)$$

$$\underline{c}_2 = \log\left(\frac{\Phi_l^{\text{medium}}(\text{all})}{\Phi_l^{\text{medium}}(\text{dominant})}\right)$$

$$\underline{c}_3 = \log\left(\frac{\Phi_l^{\text{immediate}}(\text{all})}{\Phi_l^{\text{immediate}}(\text{dominant})}\right)$$

  IF exists $\{j : c_1(j) > C_1\ \&\ c_2(j) > C_2\ \&\ c_3(j) > C_3\}$,
    $j^* = \arg\max_j\{c_2(j) : c_2(j) > C_2\}$
    Dominant$(l) = j^*$
  ELSE
    Dominant(l) =Dominant$(l-1)$
ELSE
  Dominant(l) =Dominant$(l-1)$

Fig. 1. The dominant speaker identification algorithm.

Given the number of active sub-units $v(l)$ 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 $P(v(l)|H_i)$. We define the speech activity score Φ_l as the log-likelihood ratio. Specifically, the speech activity score for a time interval of a certain length is:

$$\Phi_l = \ln\left(\frac{P(v(l)|H_1)}{P(v(l)|H_0)}\right) = \ln\left(\frac{\binom{N_R}{v} p^{v(l)}(1-p)^{N_R - v(l)}}{\lambda e^{-\lambda v(l)}}\right)\,. \tag{4}$$

IV. Dominant Speaker Identification Based on Time Intervals of Variable Lengths

After processing the signal in each channel separately, we obtain a set of scores $\Phi_l^{\text{immediate}}$, Φ_l^{medium} and Φ_l^{long} for each channel. This set of scores represents the speech activity history in time-frame l. We refer to this stage as *local processing*. Now, the scores from the distinct channels are provided into the *global decision* stage, where this information is translated into a dominant speaker identification.

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. Once a dominant speaker is identified, he remains dominant until the speech activity on one of the other channels justifies a speaker switch. We measure the relative speech activity between channels by looking at the three ratios of speech activity scores for the immediate, medium and long intervals. If the activity on one of the non dominant channels passes the set of thresholds then a speaker switch is granted. The decision algorithm is depicted in Figure 1. In the algorithm, l is a discrete time index. The length of vectors $\underline{c}_i$, $i = 1, 2, 3$ is the number of conference participants. The values C_1, C_2 and C_3 are the set of thresholds for the three activity ratios. The captions *all* and *dominant* in the brackets refer to all channels and the dominant channel, respectively.

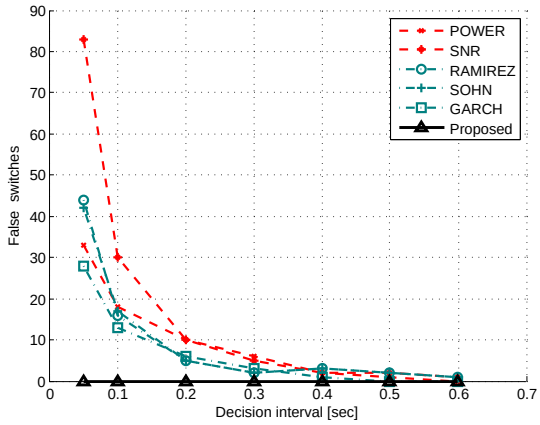
V. Experimental Results

We compared the performance of the proposed method with five speaker selection 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 [7], [8] and [9], respectively. The fourth method identifies the dominant speaker as the one with the highest signal power. It is referred to as the *POWER* method. The fifth method identifies the dominant speaker as the one with the highest SNR. It is referred to as the *SNR* method.

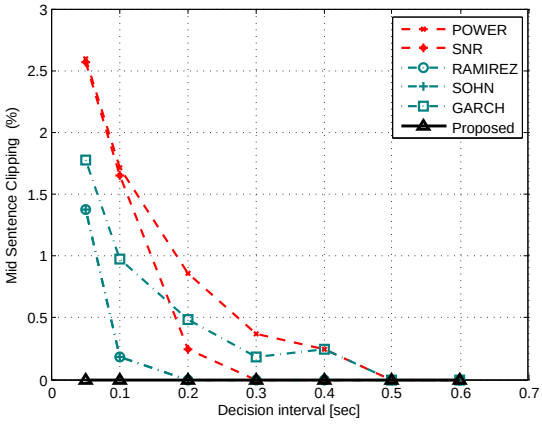
For quantitative comparison between the algotihms, we use the two following measures. The number of *false speaker switches*. Namely, the number of false switches to a non-dominant speaker. The second measure is *Mid Sentence Clipping* (MSC). This error represents the percentage of mid part of the speech bursts that was clipped due to the false switches.

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. 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*. The results of this experiment are displayed in Figure 2, where the false switching and MSC errors are plotted as a function of the decision interval. In Figure 2(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 knocks and sneezing sounds were added at 12^{th} and 17^{th} seconds of the conference respectively. The conference signals and comparison between the proposed and *Ramirez* methods are depicted in Figure 4. The quantitative influence of the transient occurrences is presented in Figure 3 and can be compared to the results in Figure 2. 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 ($0.05 - 0.2$ sec) is used (Figure 3(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 Figure 3(b) in comparison to Figure 2(b) for decision intervals in the range $0.05 - 0.2$ sec.



(a)

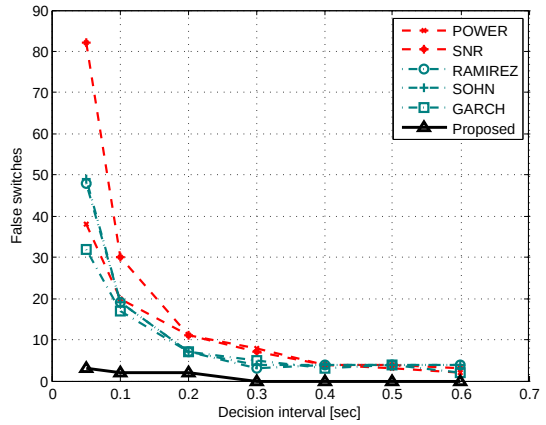


(b)

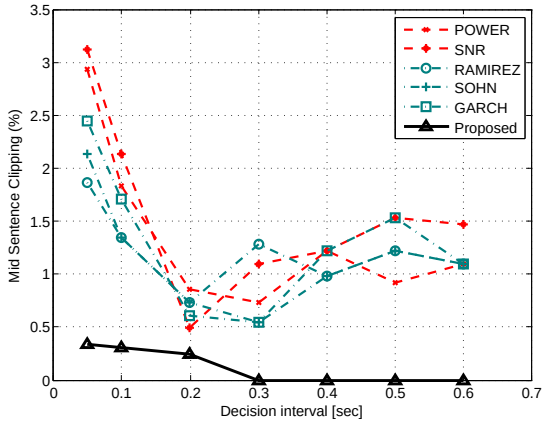
Fig. 2. Evaluation of the algorithms in the task of switching to the dominant speaker in the presence of stationary noise. The test data of this experiment consists of speech bursts concatenated such that each burst causes a speaker switch. (a) Number of false speaker switches; (b) Mid sentence clipping.

VI. 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. 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.

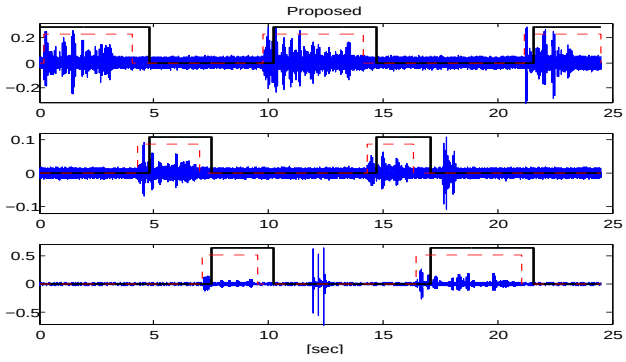


(a)

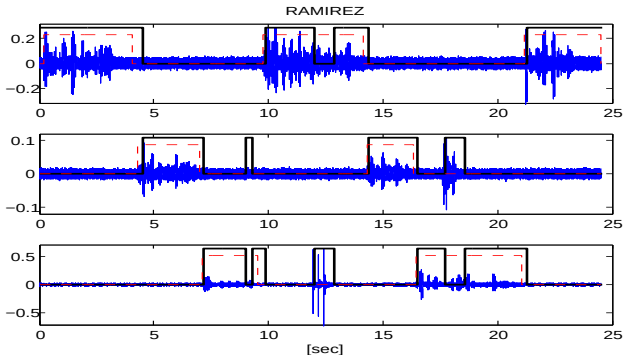


(b)

Fig. 3. Synthetic experiment with a presence of transient noise: (a) False speaker switches; (b) Mid sentence clipping.



(a)



(b)

Fig. 4. Results of dominant speaker identification, for a decision-interval of 0.3 sec: (a) dominant speaker identification by the proposed method; (b) dominant speaker identification by the method based on Ramirez 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.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Shaffer and W. Beyda, "Reducing multipoint conferencing bandwidth," US Patent No.: 6,775,247 B1, Aug. 2004.

[2] M. Howard, R. Burns, C. Lee, and M. Daily, "Teleconferencing system," US Patent No.: 6,775,247 B1, Aug. 2004.

[3] W. Kwak, S. Gardell, and B. Mayne Kelly, "Speaker identifier for multi-party conference," US Patent No.: 6,457,043 B1, Sep. 2002.

[4] Y.-F. Chang, "Multimedia conference call participant identification system and method," US Patent No.: 6,304,648 B1, Oct. 2001.

[5] I. Volfin and I. Cohen, "Dominant speaker identification for multipoint videoconferencing," *Computer Speech and Language*, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.03.002

[6] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113 –116, Apr. 2002.

[7] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689 – 692, Oct. 2005.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1 –3, Jan. 1999.

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