MULTICHANNEL SPEAKER ACTIVITY DETECTION FOR MEETINGS

Patrick Meyer, Rolf Jongebloed, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
38106 Braunschweig, Germany
{patrick.meyer, t.fingscheidt}@tu-bs.de

ABSTRACT
Multichannel recordings of meetings with a (wireless) headset for each person deliver commonly the best audio quality for subsequent analyses. However, still speech portions of other participants can couple into the microphone channel of the associated target speaker. Due to this crosstalk, a speaker activity detection (SAD) is required in order to identify only the speech portions of the target speaker in the related microphone channel. While most solutions are either complex and need a training process, or achieve insufficient results in multi-talk situations, we propose a low complexity method, which can handle both crosstalk and multi-talk situations. We investigate single- and multi-talk in a wide range of different crosstalk levels, and improved the detection accuracy towards a standardized voice activity detection overall by 12.89 % absolute, whereas a state-of-the-art multichannel SAD was exceeded even by 13.76 % absolute.

Index Terms— multichannel speaker activity detection, voice activity detection, meeting, social signal processing, crosstalk

1. INTRODUCTION
Automatic analyses of meetings for empirical (psychological) purposes have meanwhile become a vital research field with diverse tasks in recognition and signal processing. It is an important topic in social signal processing, which in general deals with the analysis of nonverbal speech [1–3], but also aims at automatic speech recognition (ASR) [4]. Meetings are a natural form of communication and thus, they entail a number of challenges for an automatic analysis, which is emphasized by Morgan et al. [5], who take the view that "nearly every problem in spoken language recognition (and understanding) can be explored in the context of meetings".

The acquisition of audio data in this context takes commonly place with a single table-top microphone, a microphone array (both placed in the middle of the table), or with personalized close-talk microphones like headsets (see Fig. 1), and lapel microphones for each person [6, 7]. It is obvious that a multichannel close-talk microphone recording promises the best speech quality for nonverbal and verbal investigations. Hence, this is the method of choice in this paper.

In order to be capable of analyzing nonverbal or verbal aspects of such recorded meetings, first, a speaker activity detection (SAD) identifying all speech portions of each speaker is necessary. Although each target speaker has a separate microphone channel with a very good speech signal quality, a simple single-channel voice activity detection (VAD) delivers insufficient results, which was shown in various studies [5, 8, 9]. Pfau et al. [8] named crosstalk (speech portions with a significant level of an interfering speaker couples into the microphone channel of the target speaker), breath noise (high level breath or contact noise is induced by the target speakers), and channel variations (the absolute active speech level (ASL), as well as the relative background noise level vary over time and meetings) as the main reasons for this problem. Additionally, Shriberg et al. [10, 11] reported that 6 to 14 % of spoken words in meetings are overlapped (multi-talk situations like double-talk), which makes the SAD problem a difficult one, decreasing the performance of their ASR system along with crosstalk significantly.

For that reason, several previous works have proposed methods of a multichannel SAD (MSAD), which takes all microphone channels into account to improve the performance of each person’s speech segmentation regarding these aspects. State-of-the-art approaches typically use hidden Markov models [8, 9, 12, 13], multi-layer perceptron classifiers [14], or approaches with deep neural networks [15]. All these methods require quite considerable computational complexity and a model training. Simpler methods are based on the cross-correlation [16] or on the characteristics of the signal-to-power ratio [17] between the microphone channels in order to be robust against crosstalk. These concepts do not need a comprehensive training, but a rather simple parameter tuning. However, they are usually incapable to adequately detect multi-talk situations.

On the basis of the approach by Matheja et al. [17], we propose in this paper a new MSAD of low complexity and without any training processes, which is able to deal with both crosstalk and multi-talk. Moreover, we investigate the influence of a wide range of different crosstalk levels during single- and multi-talk to our method and the Matheja baseline. We further disclose the insufficient performance of a standardized single-channel VAD approach in this context as second baseline and verify the robustness of all these three methods against channel variations.

The rest of the paper is structured as follows: Section 2 introduces the considered meeting scenario and our data preparation for simulation purposes. The proposed MSAD algorithm is described in detail in Section 3, and a comparison between our new MSAD
approach and the two baselines is outlined in Section 4. Finally, we conclude this paper with some remarks in Section 5.

2. SCENARIO AND DATABASE SETUP

In order to be able to investigate different speech levels of crosstalk and channel variations, we generated an artificial meeting scenario, illustrated in Fig. 1. It consists of three participants (P1, P2, P3), who are equipped with a headset, sitting around a table and talking to each other. Headsets with an omnidirectional characteristic are chosen to obtain a good and robust close-talk audio quality.

For the purpose of simulations, we recorded a development set and test set of room impulse responses (RIRs) between the three persons in a 6.6 m x 5.75 m x 2.5 m (length x width x height) meeting room. Both sets are kept acoustically different by having the table depicted in Fig. 1 in a development position, and for test in yet another position. Recordings are done with a Yamaha HS80M studio monitor as acoustic source and two HEAD acoustics HMS II, 6 head-and-torso simulators equipped with a Beyerdynamic MLL measurement microphone representing the two other persons with a headset (acoustic sinks).

Based on the recorded RIRs and the NTT multilingual speech database [18], we generated a three-channel meeting scenario with three target speakers and a sampling rate of 16 kHz. We applied for each speaker a different language in the development set (French, Spanish, Chinese) and in the test set (American English, Japanese, German), as depicted in Fig. 1. The respective speakers and sentences (each four seconds long) of the NTT data are chosen randomly and each generated channel has a length of 12 seconds (containing three time slots for a four second sentence). We considered three different conversational scenarios: Single-talk (one single speaker is active in a time slot), double-talk (two speakers are active at the same time), and triple-talk. An example is depicted in Fig. 2.

Each microphone channel \( y_m(n) \) of the associated target speaker \( m \in M = \{1, \ldots, M\} \) is defined as

\[
y_m(n) = \alpha_s \cdot s_m(n) + \alpha_n \cdot n_m(n) + \alpha_d \cdot \sum_{\mu \in M \setminus \{m\}} d_{m,\mu}(n),
\]

with \( s_m(n) = s'_m(n) \ast h_{m,m}(n) \) being the target speech signal with sample index \( n \), convolved with the RIR from the microphone of the target speaker. Here, we consider \( h_{m,m}(n) = 1 \) due to close-talk. Moreover, \( d_{m,\mu}(n) = s'_m(n) \ast h_{m,\mu}(n) \) is the speech signal of interferer \( \mu \in M = \{1, \ldots, M\} \setminus \{m\} \) convolved with the associated RIR from the mouth of interferer \( \mu \) to the microphone of target speaker \( m \). We adjust the active speech level (ASL) of \( s_m(n) \) and \( d_{m,\mu}(n) \) by \( \alpha_s \) and \( \alpha_d \) and \( \alpha_n \) in accordance to ITU-T Recommendation P.56 [19], thereby investigating various crosstalk levels. Finally, white Gaussian noise is added to \( y_m(n) \) as sensor noise \( n_m(n) \), and scaled to -75 dB from by using \( \alpha_s \) in both the development and the test set.

For the development set we adjusted \( s_m(n) \) to -26 dB and attenuated \( d_{m,\mu}(n) \) to 4 different levels from -26 dB to -38 dB with a step size of 4 dB for each conversational scenario (CS). This was repeated for \( V = 5 \) times with randomly picked sentences to obtain a sufficient amount of samples. This results in \( 1 \times 4 \times 1 \times 3 \times 5 = 60 \) samples (\( \alpha_s \times \alpha_d \times \alpha_n \times CS \times V \)).

In the test set we increased the number of files by attenuating \( d_{m,\mu}(n) \) to 13 different levels (same range as development set, but 1 dB step size), choosing \( V = 10 \) and scaling \( s_m(n) \) to -16 dB, -26 dB and -36 dB to investigate channel variations. This results in \( 3 \times 13 \times 1 \times 3 \times 10 = 1170 \) samples.

3. MULTICHANNEL SPEAKER ACTIVITY DETECTION

Matheja et al. [17] proposed an MSAD along with an extra multi-talk detection (MTD) method. On this basis we formulate in the following an overall consistent and robust MSAD method.

The basic idea of the MSAD in accordance with [17] is a comparison of the power spectral densities (PSDs) among all microphone channels of all speakers \( m \in M \). The PSD of each channel \( m \) is estimated in each frame with index \( \ell \) and frequency bin \( k \in K = \{0, 1, \ldots, K-1\} \) by

\[
\hat{\Phi}_{\Sigma,S,m}(\ell,k) = \max \left\{ \Phi_{YY,m}(\ell,k), \hat{\Phi}_{NN,m}(\ell,k), \epsilon \right\},
\]

with \( \Phi_{YY,m}(\ell,k) \) being the microphone signal PSD estimate, obtained by a temporal smoothing of the squared magnitude spectrum \( |Y_m(\ell,k)|^2 \) of the microphone signal, and \( \epsilon \) is a small positive number. We computed the noise signal PSD estimate \( \hat{\Phi}_{NN,m}(\ell,k) \) following a simple, but in this context sufficient and effective 5-state approach [20], instead of the improved minimum recursive averaging approach [21] with higher complexity as applied in the baseline [17]. Frames with index \( \ell \epsilon \{1, 2, \ldots \} \) are obtained by applying a Hann window with frame length \( K \) and frame shift \( R \). The comparison of the PSDs for each channel \( m \) takes then place by calculating a logarithmic speech power ratio (SPR) between the target and the interferer microphone channels by

\[
SPR_m(\ell,k) = 10 \log_{10} \left( \frac{\hat{\Phi}_{\Sigma,S,m}(\ell,k)}{\max_{m' \in M \setminus \{m\}} \{\hat{\Phi}_{\Sigma,S,m'}(\ell,k)\}} \right).
\]

The SPR delivers in each frame \( \ell \) and for each frequency bin \( k \) only for the most dominant channel a positive value. Note that the number of speakers \( |M| \geq 2 \) also represents the minimum number of channels. In order to avoid speaker activity detection in speech pauses, following [22, 23], a modified SPR employing the signal-to-noise ratio (SNR)

\[
\hat{\xi}_m(\ell,k) = \max \left\{ \min \left\{ \Phi_{YY,m}(\ell,k), |Y_m(\ell,k)|^2 \right\}, \hat{\Phi}_{NN,m}(\ell,k), 0 \right\} \Phi_{NN,m}(\ell,k)
\]

of the current frame and frequency bin is determined, whereby \( \hat{\Phi}_{NN,m}(\ell,k) = \lambda_{SNR} \cdot \hat{\Phi}_{NN,m}(\ell,k) \) employs an overestimation factor \( \lambda_{SNR} \). Different to the MSAD determination in [17], we now integrate the formerly isolated multi-talk detection (MTD). For this purpose, we memorize all frequency bin indices with positive SPR (3) and the SNR (4) exceeding some threshold \( \theta_{SNR} \) by

\[
K_m^+(\ell) = \{k \in K \mid SPR_m(\ell,k) > 0, \hat{\xi}_m(\ell,k) \geq \theta_{SNR} \}
\]

Afterwards we compute the fraction of these bins as \( \kappa_m^+(\ell) = \frac{K_m^+(\ell)}{K} \), respectively, for each channel \( m \). With the aid of this fraction a full-band soft MSAD measure can be expressed by

\[
0 \leq \chi_m^{MSAD}(\ell) = \gamma_m^{LB}_{min}(\ell) \cdot \kappa_m^+(\ell) \leq 1
\]

with \( \gamma_m^{LB}_{min}(\ell) \) being an SNR-dependent but frequency-independent soft weighting function that reduces the relevance of frames with a low SNR. This function will be determined on the basis of SNR values (4) averaged within \( B \) frequency bands according to

\[
\hat{\gamma}_m^{LB}(\ell,b) = \frac{1}{|K_b|} \sum_{k \in K_b} \hat{\xi}_m(\ell,k),
\]

5540
with $K_b$ being the set of frequency bin indices $k$ forming band with index $b \in B = \{1, 2, \ldots, B\}$ (band ranges adopted from [17]). The maximum SNR value

$$\xi_m^{(B)}(\ell) = \max_{b \in B} \{\xi_m^{(B)}(\ell, b)\},$$

(8)
of all frequency bands can be used to define the weighting function to be employed in (6):

$$G_{\text{min}, m}(\ell) = \min \{\alpha \cdot \xi_m^{(B)}(\ell), 1\}, \quad \alpha = 0.1.$$

(9)

Now we define two entities on the basis of (6) and (8), which will finally be combined to our proposed hard-decision MSAD measure. The SNR-based measure is determined by the maximum average SNR value of the frequency bands to

$$\xi_m^{(\text{B}, \text{max}, m)}(\ell) = \max \{\alpha \cdot \xi_m^{(\text{B}, m)}(\ell, \epsilon)\},$$

(10)

with $\epsilon$ being a small positive value, and again $\alpha = 0.1$. The second entity first requires smoothing of $\chi_m^{\text{MSAD}}(\ell)$ from (6) according to

$$\bar{\chi}_m^{\text{MSAD}}(\ell) = \gamma(\ell) \cdot \chi_m^{\text{MSAD}}(\ell) + (1 - \gamma(\ell)) \cdot \bar{\chi}_m^{\text{MSAD}}(\ell - 1),$$

(11)

with an adaptive smoothing parameter

$$\gamma(\ell) = \begin{cases} \gamma_{\text{inc}}, & \text{if } \chi_m^{\text{MSAD}}(\ell) > \bar{\chi}_m^{\text{MSAD}}(\ell - 1) \\ \gamma_{\text{dec}}, & \text{otherwise.} \end{cases}$$

(12)

The distinction between an increasing (inc) and decreasing (dec) soft MSAD measure $\chi_m^{\text{MSAD}}(\ell)$ (6) allows a faster increase and a slower decrease of the smoothed MSAD (11). Thereby, fluctuations during multi-talk can be prevented. The final MSAD decision for each channel results in

$$\text{MSAD}_m(\ell) = \begin{cases} 1, & \text{if } \bar{\chi}_m^{\text{MSAD}}(\ell) > \theta^{\text{MSAD}} \land \bar{\chi}_m^{\text{SNR}}(\ell) > \theta^{\text{SNR}} \\ 0, & \text{otherwise,} \end{cases}$$

with $\theta^{\text{MSAD}}$ and $\theta^{\text{SNR}}$ being thresholds for the two defined measures, respectively.

4. EXPERIMENTAL VALIDATION

4.1. Experimental Setup

All experiments are based on the test set described in Section 2. The analysis of the microphone signals is carried out by a Hann window with a frame length of $K = 512$ samples (32 ms), along with a frame shift of $R = 128$ samples (8 ms) in line with [17].

Besides the proposed MSAD and the baseline MSAD of Math-eja [17], we further take a single-channel VAD in accordance with the ETSI adaptive multi-rate wideband (AMR-WB) speech codec recommendations [24, 25] into account. Since Math-eja distinguishes between MSAD and MTD, for fair comparison we here combined the detection results of both of his methods with a logical OR operator. Furthermore, his original MTD does not support a specific channel allocation, but this can be obtained by means of the lower threshold $\theta_{\text{MTD}}^{\text{low}}$ in [17, eq. (19)]. While the AMR-WB VAD needs no further parameter tuning, both the baseline in [17] and our MSAD require several parameters to be optimized. Since the baseline MSAD is also published for 16 kHz signals, we applied the proposed parameter values of $\chi^{\text{SNR}} = 4$, $\theta^{\text{SNR}} = 0.25$, and $B = 10$ [17, 26]. For the remaining parameters of our method, we applied a pattern search optimization [27] on the development set (cf. Sec. 2), yielding $\gamma_{\text{inc}} = 2.5$, $\gamma_{\text{dec}} = 0.005$, $\theta^{\text{MSAD}} = 0.25$ and $\theta^{\text{SNR}} = 17$. In order to obtain comparable results, note that all parameters which were used in both methods follow the recommended values in [17, 26].

Fig. 2. Example results of the proposed MSAD for $|M| = 3$ speakers. ASLs of target speakers are $-26$ dBov and interferer speakers are adjusted to $-32$ dBov. Top down the microphone signals $y_1(n)$, $y_2(n)$, and $y_3(n)$ are depicted in black color, whereby the respective target speech components $s_1(n)$, $s_2(n)$ and $s_3(n)$ are marked in blue. Solid red and green curves show the results of $\chi_m^{\text{MSAD}}(\ell)$ in (11) and $\chi_m^{\text{SNR}}(\ell)$ in (10), respectively. The corresponding dashed red and green lines indicate the thresholds $\theta^{\text{MSAD}}$ and $\theta^{\text{SNR}}$, respectively, as used in (13). Background colors denote true accept (green), true reject (white), false accept (gray), and false reject (red). $F_{\beta=2} = 95.79\%$
The evaluation of the MSAD performance exploits a ground truth, which we acquired during the signal generation of the clean speech signals \( s_m(n) \) for each microphone channel with the aid of an energy threshold. In order to provide a good impression of the MSAD performance, we applied the \( F_{\beta=2} \)-measure [28, 29], which allows to weight the importance of detecting or not detecting speech frames. Here, we choose \( \beta = 2 \) in order to give some penalty to non-detecting speech frames, since we do not wish to reject utterances during analysis.

4.2. Results and Discussion

Fig. 2 illustrates the collaboration of two important entities. While \( \hat{E}^{(1)}_{\text{MSAD}}(\ell) \) (green color, cf. (10)) is a robust single-channel VAD with a short reaction time, \( \hat{\chi}^{\text{MSAD}}(\ell) \) (red color, cf. (11)) is a kind of channel activity detection (CAD) with a short reaction time as well, but a long decay time, which is induced by a time smoothing to obtain higher robustness. It is easy to see that the slow \( \hat{\chi}^{\text{MSAD}}(\ell) \) and the agile \( \hat{E}^{(1)}_{\text{MSAD}}(\ell) \) complement each other very well, since our MSAD votes only for the presence of speech activity, if both measures exceed their threshold \( \theta^{\text{MSAD}} \) and \( \theta^{\text{SNR}} \) (depicted with dashed red and green lines), respectively. Due to this collaboration, our MSAD is robust against crosstalk and simultaneously capable of detecting multi-talk situations.

The comparison between the proposed and the two baseline approaches regarding different crosstalk levels during single- and multi-talk situations is depicted in Fig. 3. The ASL of the target speaker signal \( s_m(n) \) is adjusted to \(-26 \, \text{dBoV}\), hence no channel variations are considered in this experiment. The \( x \)-axis denotes the input speech-to-interferer ratio (iSIR) in [dB], whereas the \( y \)-axis characterizes the accuracy regarding \( F_{\beta=2} \) in [%]. Starting with the AMR-WB VAD, adequate results for multi-talk are obtained, interestingly, however, not for single-talk. This is because the VAD was developed for telecommunication and hence, even the smallest and quietest portions of speech have to be detected, neglecting crosstalk. Consequently, the AMR-WB VAD detects each speech portion in the microphone channels, whether it is target or interfering speech, and since all speech portions in the multi-talk scenario are overlapped, the error is accordingly low. Furthermore, the results for each scenario are practically iSIR-independent in the considered range. Matheja’s MSAD [17] achieves good results for single-talk, which shows its ability to handle crosstalk and also the double-talk and triple-talk results with up to 85 % and 79 %, respectively for an iSIR of 12 dB are acceptable. However, it is not robust against crosstalk levels close to the target speech level and obtains the worst multi-talk results in our experiment. Our proposed MSAD outperforms both approaches except the AMR-WB VAD in the triple-talk scenario for the reasons given above. It is clearly evident that over the whole iSIR range of all three scenarios, even for single-talk with high crosstalk levels, which is the most difficult task, a sufficient result is obtained. This improvement is due to a joint consideration of the two problems cross-talk and multi-talk. Instead of solving them separately (MSAD plus MTD, cf. [17]), we cover both problems at once by combining a single-channel VAD (10) and a CAD (11), which are individually not even able to solve any of the two problems.

In order to investigate the robustness against channel variations, we repeated the experiment in Fig. 3 and scaled the ASL of \( s_m(n) \) additionally to \(-16 \, \text{dBoV} \) and \(-36 \, \text{dBoV} \). Note that an adjustment of algorithmic parameters was not carried out. The results are listed in Table 1, averaged over all different crosstalk levels (iSIRs) for a more compact presentation. It becomes obvious that all methods are appropriately robust against channel variations with fluctuations up to 5 % of the proposed MSAD, 8 % of the baseline MSAD and 5 % of the AMR-WB VAD. To conclude, the overall average in Table 1 summarizes the results of our experiments and thus it is clear that the proposed MSAD significantly outperforms both the baseline MSAD by 13.76 % absolute and the AMR-WB VAD by 12.89 % absolute.

### Table 1

<table>
<thead>
<tr>
<th>ASL [dBov]</th>
<th>Considered Scenario</th>
<th>( F_{\beta=2} ) [%]</th>
<th>Prop. [17]</th>
<th>[24]</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16</td>
<td>Single Talk</td>
<td>93.80</td>
<td>79.92</td>
<td>62.72</td>
</tr>
<tr>
<td></td>
<td>Double Talk</td>
<td>93.75</td>
<td>82.43</td>
<td>80.88</td>
</tr>
<tr>
<td></td>
<td>Triple Talk</td>
<td>91.10</td>
<td>78.48</td>
<td>90.32</td>
</tr>
<tr>
<td>-26</td>
<td>Single Talk</td>
<td>96.98</td>
<td>81.98</td>
<td>64.87</td>
</tr>
<tr>
<td></td>
<td>Double Talk</td>
<td>94.35</td>
<td>80.78</td>
<td>82.50</td>
</tr>
<tr>
<td></td>
<td>Triple Talk</td>
<td>89.78</td>
<td>73.14</td>
<td>91.58</td>
</tr>
<tr>
<td>-36</td>
<td>Single Talk</td>
<td>93.41</td>
<td>83.58</td>
<td>67.24</td>
</tr>
<tr>
<td></td>
<td>Double Talk</td>
<td>93.27</td>
<td>77.98</td>
<td>83.68</td>
</tr>
<tr>
<td></td>
<td>Triple Talk</td>
<td>86.10</td>
<td>70.35</td>
<td>92.70</td>
</tr>
<tr>
<td>Overall average</td>
<td></td>
<td><strong>92.50</strong></td>
<td>78.74</td>
<td>79.61</td>
</tr>
</tbody>
</table>

### 5. CONCLUSIONS

We proposed in this paper a simple method of a new multichannel speaker activity detection (MSAD) for close-talk microphone recordings in a meeting scenario. In contrast to other published MSADs, our method is not based on a complex classification system and hence, no training processes are required. For evaluation, we investigated the typical challenges in a meeting scenario in detail and pointed out, why a single-channel voice activity detection is unsuitable for this task. We demonstrated that the proposed MSAD is not sensitive to channel variations, and moreover that it is both robust against crosstalk and able to handle multi-talk situations in significantly higher accuracy than former MSAD methods.
6. REFERENCES


