Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 11-110

Binaural voice activity detection for MWF-based noise reduction in binaural hearing aids[1]

Bram Cornelis[2] , Marc Moonen[2], Jan Wouters[3]

Proc. of the 2011 European signal processing conference (EUSIPCO), Barcelona, Spain, August 2011, pp. 486-490

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/bcorneli/reports/EUSIPCO11.pdf*.

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Tel. +32 16 321797, Fax +32 16 321970, WWW: *http://www.esat.kuleuven.ac.be/sista*, E-mail: *bram.cornelis@esat.kuleuven.ac.be*. Bram Cornelis is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This research work was carried out at the Lab Exp ORL and the ESAT Laboratory of Katholieke Universiteit Leuven in the frame of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011), Concerted Research Action GOA-MaNet, research project FWO nr. G.0600.08 ('Signal processing and network design for wireless acoustic sensor networks') and Research Project IBBT. The scientific responsibility is assumed by its authors.

[3]K.U.Leuven, Dept. of Neurosciences, ExpORL, Herestraat 49/721, 3000 Leuven, Belgium.

Abstract

The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) is a powerful multimicrophone noise reduction technique, especially for binaural hearing aids where two devices are connected by a wireless link. As is the case for other single- and multimicrophone techniques, the SDW-MWF relies on a voice activity detection (VAD) algorithm, which classifies frames as noise-only or speech+noise frames. In this paper, two novel binaural fusion VADs are proposed, which are extensions of a fusion VAD originally proposed for a wireless sensor network application. By making use of the exchange of information over the wireless link of the binaural hearing aid, the binaural VADs perform a decision fusion of energy VADs calculated in the left and right device and, if the available bandwidth also allows for transmitting audio signals, a cross-correlation based VAD. The superior performances of the proposed binaural VADs are assessed (in terms of receiver operating characteristics and also by evaluating the impact on the SDW-MWF performance) by experiments with a binaural hearing aid setup.

BINAURAL VOICE ACTIVITY DETECTION FOR MWF-BASED NOISE REDUCTION IN BINAURAL HEARING AIDS

Bram Cornelis[1] *, Marc Moonen*[1]*, Jan Wouters*[2]

[1]ESAT-SCD
Dept. of electrical engineering, K.U.Leuven
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
email: bram.cornelis@esat.kuleuven.be,
marc.moonen@esat.kuleuven.be

[2]ExpORL
Dept. of Neurosciences, K.U.Leuven
Herestraat 49/721, 3000 Leuven, Belgium
email: jan.wouters@med.kuleuven.be

ABSTRACT

The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) is a powerful multimicrophone noise reduction technique, especially for binaural hearing aids where two devices are connected by a wireless link. As is the case for other single- and multimicrophone techniques, the SDW-MWF relies on a voice activity detection (VAD) algorithm, which classifies frames as noise-only or speech+noise frames. In this paper, two novel binaural fusion VADs are proposed, which are extensions of a fusion VAD originally proposed for a wireless sensor network application. By making use of the exchange of information over the wireless link of the binaural hearing aid, the binaural VADs perform a decision fusion of energy VADs calculated in the left and right device and, if the available bandwidth also allows for transmitting audio signals, a cross-correlation based VAD. The superior performances of the proposed binaural VADs are assessed (in terms of receiver operating characteristics and also by evaluating the impact on the SDW-MWF performance) by experiments with a binaural hearing aid setup.

1. INTRODUCTION

Noise reduction has been an active area of research for many years, with applications in hearing aids, hands-free communications and teleconferencing. The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) [1] is a powerful multimicrophone noise reduction technique for speech in noise scenarios, in particular for hearing aid applications. In binaural hearing aids, a wireless link allows for exchanging parameters or even microphone signals between a left and a right device. In addition to a better noise reduction performance, binaural hearing aids aim at preserving the so-called binaural cues (interaural time and level differences), by which the brain can localize sound sources. Unfortunately, fixed beamformers and adaptive beamformers such as the Generalized Sidelobe Canceller (GSC) [2], have been shown to distort the binaural cues [3]. The SDW-MWF on the other hand is an excellent choice for binaural noise reduction, as the binaural cues can be preserved in addition to achieving a better noise reduction performance [4].

The SDW-MWF relies on a voice activity detection (VAD) algorithm, which classifies frames as either speech+noise or noise-only frames. In fact, many other single- and multimicrophone noise reduction algorithms require a VAD to detect speech pauses, as a spectral or spatial estimate of the noise component is needed. The GSC also requires a VAD to control the adaptation, as adaptation during speech activity significantly degrades the performance. A robust VAD algorithm is therefore crucial and highly sought for.

In this paper, we focus on a VAD algorithm for binaural hearing aids. The wireless link is utilized to derive a decision fusion of different VADs, in order to obtain a superior fusion VAD. First, the decisions of single-microphone energy VADs, calculated on a reference microphone signal in the left and right device, are fused using only a small-bandwidth exchange of signal-to-noise ratio (SNR) estimates as in [5]. Second, if the available bandwidth also allows for transmitting an audio signal, a two-microphone cross-correlation VAD can also be applied. Two fusion VADs, which fuse the energy and cross-correlation VADs, are proposed in this paper and are shown to further increase the VAD performance. In contrast to statistical model-based VAD algorithms such as [6], the proposed fusion VADs still only require a (relatively simple) SNR estimation in each device.

The paper is organized as follows. In section 2, the energy and cross-correlation VADs used in this paper are described. It should however be noted that the proposed decision fusion algorithms are general in the sense that they could also be applied to other single or multimicrophone VADs instead of the VADs of this section. A brief review of the binaural SDW-MWF is also given in section 2. In section 3, the fusion VAD of [5] (denoted as *fusion-1*) is discussed and two novel binaural fusion VADs (denoted as *fusion-2* and *fusion-opt*) are proposed. In section 4, the performance of the different VAD algorithms is assessed by simulations with a binaural hearing aid setup. Both the VAD performance (in terms of receiver operating characteristics) as the impact on the noise reduction performance of the SDW-MWF is evaluated. Finally, conclusions are given in section 5.

2. VAD ALGORITHMS AND MWF REVIEW

2.1 Notation and configuration

We consider a microphone array consisting of N microphones. The nth microphone signal $Y_n[f]$ can be specified in the frequency domain as

$$Y_n[f] = X_n[f] + V_n[f], \quad n = 1 \ldots N, \tag{1}$$

where f is the frequency-domain variable, $X_n[f]$ represents the speech component and $V_n[f]$ represents the noise component in the nth microphone. The signals $Y_n[f], X_n[f]$ and $V_n[f]$ are stacked in the N-dimensional vectors $\mathbf{y}[f]$, $\mathbf{x}[f]$ and $\mathbf{v}[f]$, with $\mathbf{y}[f] = \mathbf{x}[f] + \mathbf{v}[f]$. The correlation matrix $\mathbf{R}_y[f]$, the speech correlation matrix $\mathbf{R}_x[f]$ and the noise correlation matrix $\mathbf{R}_v[f]$ are then defined as $\mathbf{R}_y[f] = \mathscr{E}\{\mathbf{y}[f]\mathbf{y}^H[f]\}$, $\mathbf{R}_x[f] = \mathscr{E}\{\mathbf{x}[f]\mathbf{x}^H[f]\}$ and $\mathbf{R}_v[f] = \mathscr{E}\{\mathbf{v}[f]\mathbf{v}^H[f]\}$, where $\mathscr{E}$ denotes the expected value operator.

2.2 Single-microphone energy VAD

There are many single-microphone VAD algorithms available, which are for example based on short-term energy, zero-crossing rate, speech and noise probability distributions, or combinations of

Bram Cornelis is supported by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This research work was carried out at the Lab Exp ORL and the ESAT Laboratory of Katholieke Universiteit Leuven in the frame of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011), Concerted Research Action GOA-MaNet, research project FWO nr. G.0600.08 ('Signal processing and network design for wireless acoustic sensor networks') and Research Project IBBT. The scientific responsibility is assumed by its authors.

© EURASIP, 2011 - ISSN 2076-1465

properties (cfr. [7] for a summary of some algorithms). In this paper, the log-energy based VAD of [8] will be used (and denoted by *energy VAD*), as it was shown in [7] that it achieved the best performance (over different scenarios, input SNRs and performance measures) compared to the other single-microphone VADs. The energy VAD tracks the short-term log-energy of a reference microphone signal, on a frame-by-frame basis. A histogram of the log-energy shows two clusters (corresponding to noise-only frames and speech+noise frames) [8], which can be fitted by a bimodal Gaussian distribution. An online approximation is also proposed in [8] and used in the simulations in this paper. In the online method, the noise variance is tracked adaptively. When the short-term frame energy is significantly higher than the noise variance estimate, the frame is classified as speech+noise. As a result, the energy VAD does not perform well at low input SNRs or for nonstationary noise. On the other hand, no prior information about the target speech location is required.

2.3 Two-microphone cross-correlation based VAD

The considered two-microphone VAD (denoted as *cross-corr VAD*) is based on a per-frame time difference of arrival (TDOA) estimation. The instantaneous TDOA at frame-index m can be estimated by finding the delay corresponding to the maximum of the cross-correlation function [9]:

$$\widehat{\tau}_{12}[m] = \underset{\tau}{\operatorname{argmax}} \; r_{12}[\tau, m] \,, \tag{2}$$

where $r_{12}[\tau, m]$ is the instantaneous cross-correlation function between microphones 1 and 2, at frame-index m. $r_{12}[\tau, m]$ is calculated as:

$$r_{12}[\tau, m] = \int_{-\infty}^{\infty} Y_1[f, m] \, Y_2^H[f, m] \, e^{j2\pi f \tau} \, \mathrm{d}f \,, \tag{3}$$

where $Y_1[f, m]$ and $Y_2[f, m]$ are the instantaneous (smoothed) spectra of microphone 1 and 2 at frame-index m. Speech is then detected if the cross-correlation value corresponding to the assumed speech TDOA is sufficiently close to the cross-correlation value corresponding to the estimated TDOA:

---

if $r_{12}[\tau_{assumed}, m] > T_{corr} * r_{12}[\widehat{\tau}_{12}, m]$ **then**
  $\mathrm{VAD}_{corr}[m] \leftarrow 1$
else
  $\mathrm{VAD}_{corr}[m] \leftarrow 0$
end if (4)

Simulations (cfr. section 4.2) indicate that a threshold value $T_{corr} = 0.7$ is a good setting and so this value will be used from now on (unless otherwise specified).

Correlation-based techniques are not suitable for use in a single hearing aid device (with close-spaced microphones), as they assume the intermicrophone distance is sufficiently large (> 7 cm) [10], or only work if the interferers are located in the rear-half plane [11]. However, if the hearing aids are connected by a wireless link, each device can stream a microphone signal to the contralateral device, so that each device can calculate (2) and (3) using one of its own microphone signals and the received microphone signal of the contralateral device. A drawback of the cross-correlation VAD is the fact that the speech source location ($\tau_{assumed}$) has to be known a priori. We will always choose $\tau_{assumed} = 0$ in this paper, i.e. the target speaker is located in the look direction of the hearing aid user (a frequently occurring case).

2.4 Multichannel Wiener Filter and correlation matrix estimation

The (frequency-domain) Multichannel Wiener Filter (MWF) produces a minimum-mean-square-error (MMSE) estimate of the speech component in a reference microphone. To provide a more explicit tradeoff between speech distortion and noise reduction, the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) has been proposed, which minimizes a weighted sum of the residual noise energy and the speech distortion energy [1]. The SDW-MWF is given as:

$$\mathbf{w}_{\mathrm{SDW-MWF}}[f] = \left(\mathbf{R}_x[f] + \mu \mathbf{R}_v[f]\right)^{-1} \mathbf{R}_x[f] \mathbf{e}_{\mathrm{ref}} \,. \tag{5}$$

$\mathbf{e}_{\mathrm{ref}}$ is a vector which selects the column corresponding to the reference microphone out of $\mathbf{R}_x[f]$. The trade-off parameter μ allows putting more emphasis on noise reduction, at the cost of a higher speech distortion. The frequency-domain variable f is now omitted for conciseness.

By assuming that speech and noise are uncorrelated, $\mathbf{R}_x$ can be found by calculating $\mathbf{R}_x = \mathbf{R}_y - \mathbf{R}_v$, where $\mathbf{R}_y$ is the speech+noise correlation matrix. The SDW-MWF thus only needs reliable estimates of $\mathbf{R}_y$ and $\mathbf{R}_v$. To obtain these, frames have to classified as either speech+noise or noise-only frames by a VAD algorithm. The correlation matrix estimates $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_v$ are then recursively updated (per frequency bin) as:

---

if $\mathrm{VAD}[m] == 1$ (speech + noise) **then**
  $\hat{\mathbf{R}}_y[m] \leftarrow \lambda_y \hat{\mathbf{R}}_y[m-1] + (1-\lambda_y)\mathbf{y}[m]\mathbf{y}^H[m]$
  $\hat{\mathbf{R}}_v[m] \leftarrow \hat{\mathbf{R}}_v[m-1]$
else
  $\hat{\mathbf{R}}_y[m] \leftarrow \hat{\mathbf{R}}_y[m-1]$
  $\hat{\mathbf{R}}_v[m] \leftarrow \lambda_v \hat{\mathbf{R}}_v[m] + (1-\lambda_v)\mathbf{y}[m]\mathbf{y}^H[m]$
end if (6)

λ_y and λ_v are exponential forgetting factors (usually chosen close to one). For noise reduction in binaural hearing aids, the SDW-MWF is calculated in each device using its own available microphone signals and the microphone signal(s) received from the contralateral device [4].

3. BINAURAL FUSION VADS

3.1 Decision fusion of energy VADs: Fusion-1

The energy VAD can be applied to the reference microphone signal of each device. If a small bandwidth wireless link is available so that it is feasible to transmit parameters, the energy VAD decisions and estimated (left and right) local SNRs can be exchanged between the devices. As the energy VAD is more robust for higher SNRs, the estimated local SNR is an indication for the reliability of the energy VAD decision, and this information can be used to obtain a superior fusion VAD. The SNR estimates are calculated as in [5]:

---

if $\mathrm{VAD}_{energy}[m] == 1$ (speech + noise) **then**
  $S[m] \leftarrow \beta * S[m-1] + (1-\beta) * E[m]$
  $N[m] \leftarrow N[m-1]$
else
  $S[m] \leftarrow S[m-1]$
  $N[m] \leftarrow \alpha * N[m-1] + (1-\alpha) * E[m]$
end if
$\hat{\mathrm{SNR}}[m] \leftarrow \frac{S[m]}{N[m]}$ (7)

$S[m]$ is the estimated local signal level (speech+noise), $N[m]$ is the estimated local noise level and $E[m]$ is the frame-based energy, i.e.

$$E[m] = \frac{1}{L} \sum_{l=0}^{L-1} (y_{\mathrm{ref}}[mL + l])^2 \,, \tag{8}$$

where L is the frame size and y_{ref} is the (time-domain) reference microphone signal. The forgetting factors α and β will both be set to 0.9.

In the fusion algorithm of [5] (denoted here as *fusion-1 VAD*), the energy VAD decisions and local SNR estimates of different sensor nodes in a wireless network are transmitted to a fusion center, where a weighted sum of the local VAD decisions is calculated and compared to a threshold:

$$w_k[m] = \frac{\widehat{\mathrm{SNR}}_k[m]}{\sum_{k=1}^{K} \widehat{\mathrm{SNR}}_k[m]}\,, \tag{9}$$

if $\sum_{k=1}^{K} \left(w_k[m] * \mathrm{VAD}_{energy,\,k}[m]\right) > T_{fusion-1}$ **then**
  $\mathrm{VAD}_{fusion-1}[m] \leftarrow 1$
else
  $\mathrm{VAD}_{fusion-1}[m] \leftarrow 0$
end if (10)

where K is the number of nodes. The binaural hearing aid application corresponds to a simple two-node ($K = 2$) network. If the local SNR estimates and energy VAD decisions are exchanged between the devices, (9) and (10) can be calculated in each device. Because of the headshadow effect, there is often a best-ear side (i.e. high SNR) and a worst-ear side (low SNR). If the received SNR estimate is larger than the local SNR estimate, the other device is probably at the best-ear side. An alternative fusion strategy is therefore to use the VAD decision of the other device instead of the own VAD decision, if the received SNR estimate is larger than the own SNR estimate. This fusion strategy actually corresponds to a threshold $T_{fusion-1} = 0.5$ in (10), and will be used from now on (unless otherwise specified).

3.2 Decision fusion of energy and cross-correlation VADs: Fusion-2

If microphone signals can be streamed over the wireless link, both devices can calculate the energy VAD and SNR estimates as in the previous section (as now, both devices have access to signals from both sides of the head). In addition, the two-microphone cross-correlation VAD can also be calulated in both devices. We now propose to fuse the two energy VAD decisions with the cross-correlation based VAD, in a similar manner as in the previous approach. As the cross-correlation VAD is not linked to a local SNR estimate, $\widehat{\mathrm{SNR}}_{corr}[m]$ is set equal to the average of the other SNRs:

$$\widehat{\mathrm{SNR}}_{corr}[m] = \frac{\widehat{\mathrm{SNR}}_L[m] + \widehat{\mathrm{SNR}}_R[m]}{2}\,, \tag{11}$$

where $\widehat{\mathrm{SNR}}_L[m]$ and $\widehat{\mathrm{SNR}}_R[m]$ are the estimated local SNRs at the left and right device. As a consequence, the cross-correlation VAD gets an intermediate weight in the decision fusion. The decision fusion rule (denoted as *fusion-2*) is then equal to:

$$w_1[m] = \frac{\widehat{\mathrm{SNR}}_L[m]}{\widehat{\mathrm{SNR}}_L[m] + \widehat{\mathrm{SNR}}_R[m] + \widehat{\mathrm{SNR}}_{corr}[m]}\,, \tag{12}$$

$$w_2[m] = \frac{\widehat{\mathrm{SNR}}_R[m]}{\widehat{\mathrm{SNR}}_L[m] + \widehat{\mathrm{SNR}}_R[m] + \widehat{\mathrm{SNR}}_{corr}[m]}\,, \tag{13}$$

$$w_3[m] = \frac{\widehat{\mathrm{SNR}}_{corr}[m]}{\widehat{\mathrm{SNR}}_L[m] + \widehat{\mathrm{SNR}}_R[m] + \widehat{\mathrm{SNR}}_{corr}[m]}\,, \tag{14}$$

if $w_1[m] * \mathrm{VAD}_{energy,\,L}[m] + w_2[m] * \mathrm{VAD}_{energy,\,R}[m]$
  $+ w_3[m] * \mathrm{VAD}_{corr}[m] > T_{fusion-2}$ **then**
  $\mathrm{VAD}_{fusion-2}[m] \leftarrow 1$
else
  $\mathrm{VAD}_{fusion-2}[m] \leftarrow 0$
end if (15)

A threshold value $T_{fusion-2} = 0.45$ was found to be a good setting, as will be illustrated in section 4.2.

3.3 Decision fusion of energy and cross-correlation VADs based on optimal AND rule: Fusion-opt

An alternative for fusing the energy and cross-correlation VAD is to combine the VAD decisions by an AND rule. In [12], it is shown that the AND rule can be applied in an optimal way, by making use of the Receiver Operating Characteristics (ROC) graphs [13] of the separate VAD algorithms. We propose here to apply the AND rule to the output of the fusion-1 VAD (fusion of energy VADs) together with the output of the cross-correlation VAD. To generate the corresponding ROC graphs, the thresholds $T_{fusion-1}$ and T_{corr} are varied from zero to one in steps of 0.05.

In figure 1, the average ROC of the fusion-1 and cross-correlation VAD are shown. In the ROC graph, the false positive rate (number of noise-only frames erroneously classified as speech+noise, divided by the total number of noise-only frames) is plotted on the X axis and the true positive rate (number of speech+noise frames correctly classified as speech+noise, divided by the total number of speech+noise frames) is plotted on the Y axis. The performance of the VADs was averaged over a range of input SNRs and over several spatial scenarios. Both scenarios with speech in the frontal look direction as in other directions were included. More information about the setup and test stimuli is given in section 4.

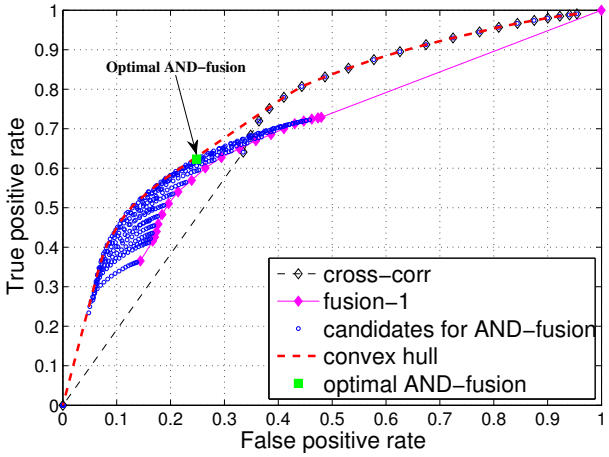


Figure 1: Average ROC graphs for cross-correlation VAD and Fusion-1 VAD, and optimal AND-fusion. The average performance over various spatial scenarios and input SNRs is shown.

By measuring the performance obtained with the AND rule for every possible combination of $T_{fusion-1}$ and T_{corr}, where both parameters are varied from zero to one in steps of 0.05, the points marked as *candidates for AND-fusion* are found. By definition, the optimal AND rule ROC is then the convex hull of these candidate points. It is indeed observed that the area under the convex hull ROC is larger than the area's of the fusion-1 and cross-correlation VADs, so that in principle, a better performance is obtained. From now on, the point closest to the top-left corner is chosen as the optimal fusion VAD (denoted *fusion-opt*). The corresponding thresholds are $T_{corr} = 0.55$ and $T_{fusion-1} = 0.25$.

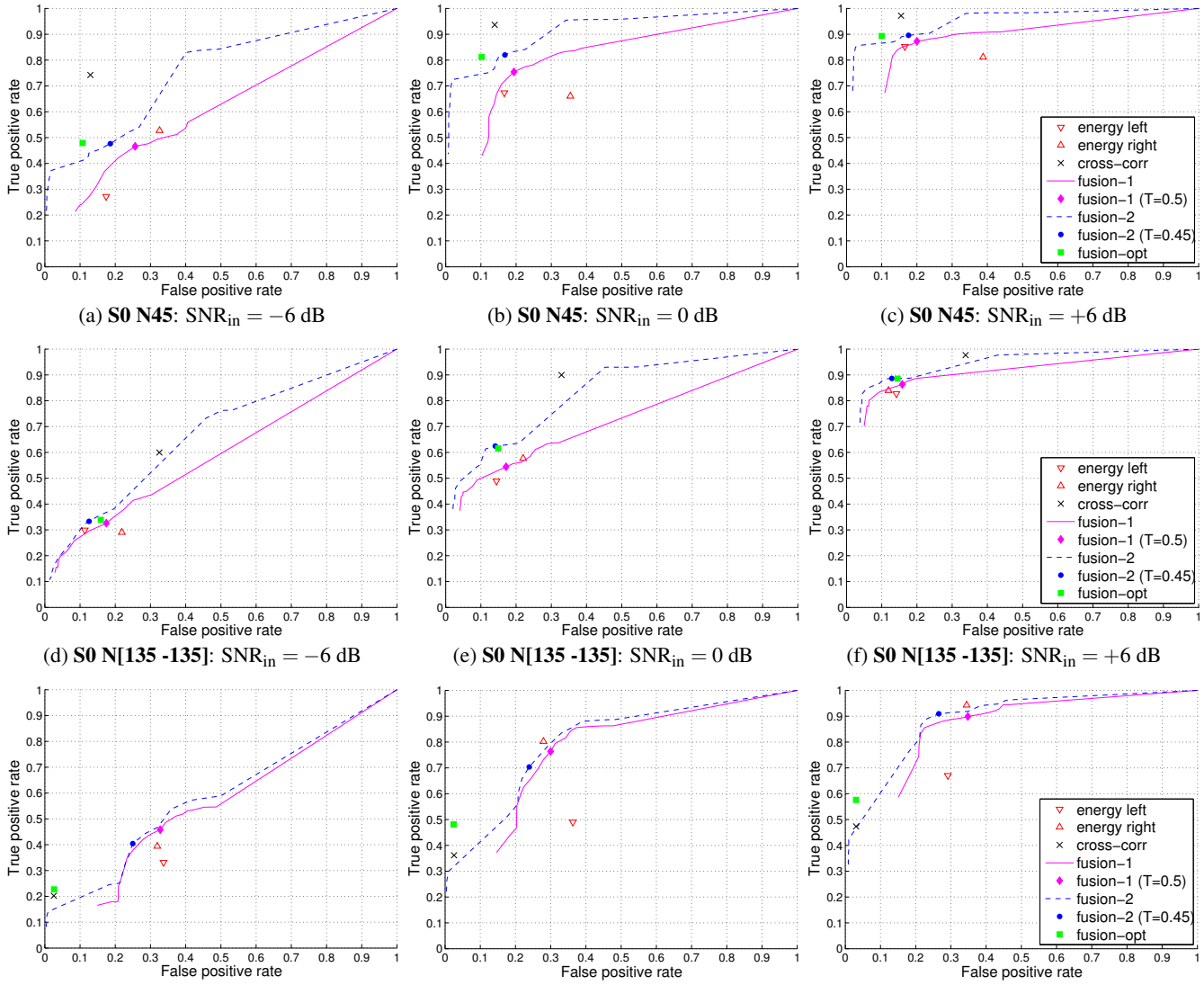
4. EXPERIMENTS

4.1 Setup and stimuli

We consider a binaural setup with two behind-the-ear hearing aids connected by a wireless link. There are two omnidirectional microphones per device, and we assume that the link allows for transmitting one audio signal (i.e. the front microphone signal) to the other device (full-duplex).

Head-related transfer functions (HRTFs) are measured in a reverberant room (reverberation time of 0.62s) on a Cortex MK2 manikin. The considered acoustic scenarios have a target speech (S) source and interfering noise (N) source(s) at specified azimuthal angles (with 0° in front of the head, 90° to the right of the head).

As speech stimulus, 4 consecutive sentences of the Dutch VU sentence material [14] were used. Multitalker babble noise was used as interfering noise signal(s). The signals are sampled at $f_s = 20480$



(a) **S0 N45**: $\mathrm{SNR_{in}} = -6$ dB (b) **S0 N45**: $\mathrm{SNR_{in}} = 0$ dB (c) **S0 N45**: $\mathrm{SNR_{in}} = +6$ dB

(d) **S0 N[135 -135]**: $\mathrm{SNR_{in}} = -6$ dB (e) **S0 N[135 -135]**: $\mathrm{SNR_{in}} = 0$ dB (f) **S0 N[135 -135]**: $\mathrm{SNR_{in}} = +6$ dB

(g) **S90 N-90**: $\mathrm{SNR_{in}} = -6$ dB (h) **S90 N-90**: $\mathrm{SNR_{in}} = 0$ dB (i) **S90 N-90**: $\mathrm{SNR_{in}} = +6$ dB

Figure 2: Receiver Operating Characteristics (ROC) graphs for fusion-1 and fusion-2 VADs, for different spatial scenarios and input SNRs.

Hz and processed by a weighted overlap-add (WOLA) filterbank. The signals are segmented in frames of $L = 128$ samples with 50% overlap, and windowed by a Hann Window.

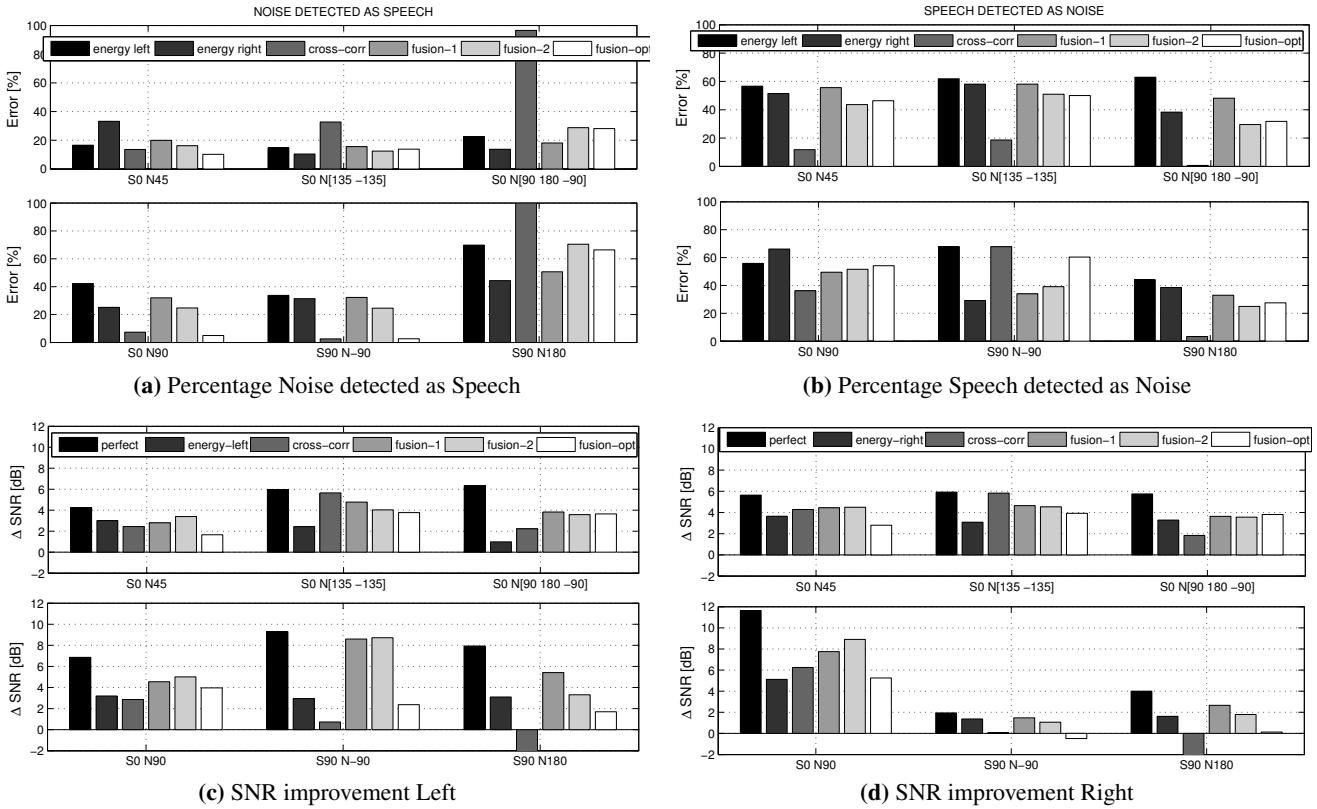
4.2 VAD performance: ROC graphs

In figure 2, the ROC graphs of the VAD algorithms are plotted for three different spatial scenarios, and three different input SNRs per scenario. The input SNR is measured in absence of the head, at the center point of the setup. Because of the head-shadow effect, the actual input SNRs at the left and right device are therefore dependent on the spatial scenario. The ROC graphs were generated by varying the decision thresholds $T_{fusion-1}$ and $T_{fusion-2}$ from 0 to 1. The cross-correlation VAD threshold is set to $T_{corr} = 0.7$ and the optimal AND-fusion VAD (fusion-opt) is fixed as in section 3.3. As the VADs perform better at higher input SNRs, the ROC graphs indeed shift towards the top-left corner (i. e. have a larger area below the ROC graph) as the input SNR increases.

Because of the head-shadow effect, it can be observed that the energy VAD performs better (i.e. more towards the top-left corner in the ROC space) in the left device in the S0N45 scenario (left ear is the best ear), whereas it performs better in the right device in the S90 N-90 scenario (right ear is the best ear). It can be seen that the fusion-1 VAD generally gives an improvement over the individual energy VADs. The ROC point corresponding to $T_{fusion-1} = 0.5$ which is used in subsequent experiments, is also indicated on the curve. The cross-correlation VAD generally performs well if the speech source is located in the assumed direction (i. e. 0°), for S90 N-90 a low true positive rate is however observed. The fusion-2 VAD generally gives a better performance than the fusion-1 VAD, and even for S90 N-90 the performance is not degraded as is the case for the cross-correlation VAD. The ROC point corresponding to $T_{fusion-2} = 0.45$ which is used in subsequent experiments, is also indicated on the curve. The fusion-opt VAD of section 3.3 generally gives a performance similar to the fusion-2 VAD, but is however degraded in a similar way as the cross-correlation VAD for the S90 N-90 scenario. It should also be noted that the three spatial scenarios of this section were not included as training scenarios to derive the optimal AND rule fusion.

4.3 SDW-MWF performance

In figure 3, the SDW-MWF performance (speech intelligibility (SI)-weighted SNR improvement [15] in dB) using five different VADs is shown for six spatial scenarios, for an input SNR (measured in absence of the head) of -2dB. The performance with a perfect VAD is also shown as a reference point. The SDW-MWF is implemented using a decomposed filter expression which was shown [16] to be more robust to estimation errors compared to (5). We assume that a microphone signal can be streamed over the wireless link so that



(a) Percentage Noise detected as Speech

(b) Percentage Speech detected as Noise

(c) SNR improvement Left

(d) SNR improvement Right

Figure 3: VAD errors (a)-(b), left (c) and right (d) SI-weighted SNR improvements, input SNR=-2dB.

each device has access to a total of $N = 3$ microphone signals (two ipsilateral signals and one contralateral signal). The speech distortion parameter was set to $\mu = 5$, the exponential forgetting factors were set to $\lambda_y = \lambda_v = 0.999$. The VAD errors are also shown, where a distinction is made between the ratio of noise-only frames erroneously detected as speech+noise (equal to false positive rate), and the ratio of speech+noise frames erroneously detected as noise-only (equal to one minus the true positive rate).

In general, it can be observed that the binaural fusion VADs achieve a superior perfomance over the individual VADs. The cross-correlation VAD fails in scenarios where speech is not coming from the assumed frontal direction (S90 N-90 and S90 N180) and/or when there is a noise source at 180° (S90 N180 and S0 N[90 180 -90]) as for this location the same TDOA is obtained as for a source coming from 0°. The fusion-2 VAD stays robust in these scenarios, whereas the performance with the fusion-opt VAD also degrades. Overall, it can be concluded that the fusion-1 and fusion-2 VADs give the best overall performance for binaural MWF-based noise reduction.

5. CONCLUSION

In this paper, two binaural fusion VADs (fusion-2 and fusion-opt) have been proposed, which are extensions of a fusion VAD for wireless sensor networks (fusion-1). The binaural fusion VADs make use of the wireless link in a binaural hearing aid to combine energy VAD decisions calculated in the left and the right device and a cross-correlation based VAD decision, in order to obtain a VAD performance which is superior to the individual VAD algorithms. The performance was assessed by ROC graphs and by evaluating the impact on the noise reduction performance of the binaural SDW-MWF.

REFERENCES

[1] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[3] T. Van den Bogaert, T. J. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localisation with bilateral hearing aids: without is better than with," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 515–526, 2006.

[4] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[5] V. Berisha, H. Kwon, and S. Spanias, "Real-time implementation of a distributed voice activity detector," in *IEEE Workshop on Sensor Array and Multichannel Process.*, July 2006, pp. 659–662.

[6] J. H. Park, M. H. Shin, and H. K. Kim, "Statistical model-based voice activity detection using spatial cues and log energy for dual-channel noisy speech recognition," in *Communication and Networking*, ser. Communications in Computer and Information Science, T.-h. Kim, T. Vasilakos, K. Sakurai, Y. Xiao, G. Zhao, and D. Izak, Eds. Springer Berlin Heidelberg, 2010, vol. 120, pp. 172–179.

[7] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, ESAT, Katholieke Universiteit Leuven, Belgium, May 2003.

[8] S. Van Gerven and F. Xie, "A Comparative Study of Speech Detection Methods," in *Proc. EUROSPEECH*, vol. 3, Rhodos, Greece, Sept. 1997, pp. 1095–1098.

[9] Y. Huang, J. Benesty, and J. Chen, *Time delay estimation and source localization*. Springer-Verlag, 2007, ch. 51 part I in "Springer handbook of speech processing" (Benesty, J. and Sondhi, M. and Huang, Y., Eds.), pp. 1043–1064.

[10] A. Koul and J. Greenberg, "Using intermicrophone correlation to detect speech in spatially separated noise," *EURASIP J. Appl. Signal Process.*, pp. 91–14, 2006.

[11] S. Srinivasan and K. Janse, "Spatial audio activity detection for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, April 2008, pp. 4021–4024.

[12] Q. Tao, R. van Rootseler, R. Veldhuis, S. Gehlen, and F. Weber, "Optimal decision fusion and its application on 3D face recognition," in *Proc. of the Spec. Int. Group on Biometr. and Electr. Signatures*, Darmstadt, Germany, July 2007, pp. 15–24.

[13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

[14] N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast, "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1671–1684, 2000.

[15] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[16] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, July 2011.