Voice Activity Classification for Automatic Bi-Speaker Adaptive Beamforming in Speech Separation

Thuy N Tran, William Cowley and André Pollok

Institute for Telecommunications Research, University of South Australia

thuy.tran@mymail.unisa.edu.au, {Bill.Cowley,Andre.Pollok}@unisa.edu.au

10.21437/Interspeech.2013-234

Abstract

A simple and low computational complexity system for bi-speaker speech separation is proposed in this paper. The system is constructed of a voice activity classification (VAC) module and an adaptive bi-beamformer module for speech separation using a microphone array. The first module identifies active speaker(s) and allows the system to control the adaptation of the second module automatically. The VAC is based on a novel classification method containing two steps. The first step uses a robust VAC method based on our previous work on beamformer-output-ratio of a bi-beamforming system. The second step refines the VAC results using a novel method derived from an analytical result on the output power of an adaptive beamformer. The system is tested in reverberant environments with both synthesized and real recordings. The synthesized recordings contain two speakers, a background speech and noises. The real recording contains two speakers speaking spontaneously. The VAC results satisfy a conservative classification scheme to avoid the signal cancellation problem. The final separation outputs are compared with the ideal outputs provided by genie-aided adaptive beamformers which have perfect VAC knowledge. The results show that the propose automatic system achieves high performance close to the ideal system.

Index Terms: adaptive beamforming, speech separation, voice activity classification, beamformer output ratio

1. Introduction

In a multi-speaker environment, such as a meeting room, speech separation systems target segregating individual speech signals from mixture recordings. In the last decade, significant advances in this field have been achieved with a boom in the number of more powerful algorithms [1–6], which can be categorized into adaptive optimum beamforming [5], high-order statistical optimum beamforming [2], independent component analysis based methods [6]. These methods show promising results toward practical products [1]. However, as shown in most speech separation campaigns such as PASCAL Challenge II [7], CHiMe [8], SASSEC and SiSEC [1], critical prerequisites are needed for such practical realizations of these methods. While the automatic source localization (or *blind separation*) challenge is commonly addressed, the prerequisite of knowing the number of active sources in each recording period often remains as an assumption [1]. To create a truly blind separation system, this assumption needs to be removed and this is the motivation of the work in this paper.

The active speaker identification task (or voice activity classification (VAC)) provides information about the number of currently active sources and the status of the speakers of interest. In the nature of human conversation, this status changes and appears to be unpredictable, therefore if a speech separation system requires this knowledge, this information should be continuously updated during the separation process. Especially, for optimum beamformers such as minimum variance distortionless response (MVDR) [5, 9], the status of the wanted speaker is essential for controlling the beamformer adaptation. MVDR can sufficiently cancel interference and noise if the covariance matrix of the unwanted signals can be estimated. Due to long reverberation time and the non-stationarity of speech signals, this approximation requires the absence of the wanted signal during the adaptation. Hence the adaptation is often halted when the wanted speaker is active to avoid the signal cancellation problem, i.e. the signal of interest is unexpectedly suppressed [2, 10, 11]. This halting strategy is sensible considering that in a conversation, overlapping periods can be expected to be considerably less frequent than single speaker periods.

Addressing such practical aspects for speech separation, a number of *multimodal* systems have been proposed. In general, multimodal systems employ other processing techniques such as video signal processing, source localization, or speaker recognition to automatically provide voice activity information for a speech separation module [12–15]. The extra modules of these system are often computationally expensive or unoptimized for VAC when the number of active sources varies.

Through solving the VAC problem, we propose a low computational complexity, yet effective system to separate speech for two main speakers in an automatic fashion. Our focus on two speakers stems from the expectation that overlapping speech of more main speakers rarely occurs in practice. Using a well-known adaptive beamforming technique such as MVDR, two beamformers are constructed to simultaneously extract speech from each speaker. By continuously identifying the active status of the two speakers in the input, the adaptation of these beamformers is controlled automatically. This requirement is satisfied by employing a novel VAC method that contains two main steps, 1) robust VAC, and 2) refined VAC. The first step, Section 3.2, is an extension from our previous work in [16] based on the relative power of beamformer outputs of two parallel beamformers. The second step, introduced in Section 3.3, is derived from the output power behaviour of MVDR beamformers under different adaptation conditions. The processing of the whole system is given in Section 3.4. The performance of the proposed automatic system in comparison with genie-aided beamforming systems is shown in Section 4 with high VAC classification and competitive separation results. To simplify the problem, speaker locations are assumed stationary, thus they can be estimated only once.

2. Signal Model

Considering an indoor recording environment, an M element microphone array records multi-path propagation signals from P speakers ($P \geq 2$), and noise sources. Notice that beside the two main speakers, signals from the other $P - 2$ speak-

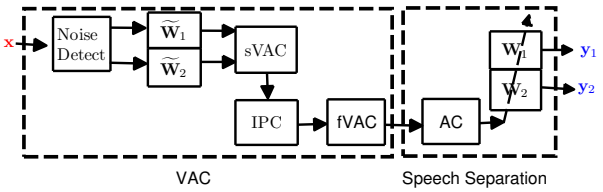

Figure 1: *Automatic adaptation BiBeam speech separation*

ers are called *background speech*. At each microphone, the recorded time domain signals are segmented into *frames* and transformed into the frequency domain (FD) via an N_f-point fast Fourier transform [17]. The corresponding FD signal model is $X_m(k,q) = \sum_{p\in\{1,2,..P\}} X_{p,m}(k,q) + V_m(k,q)$ where X_m are the FFT representation of the recorded signal at microphone m, $X_{p,m}$ is the FFT representation of the signal from speaker p ($1 \le p \le P$) received at microphone m, and V_m is the noise, k ($k \in \{1,2,..,N_f\}$) is the frequency bin index, and q ($q \in \{1,2,...\}$) is the frame index. Using vector notation, the FD multi-channel signal model is

$$\boldsymbol{X}(k,q) = \sum_{p\in\{1,2,..P\}} \boldsymbol{X}_p(k,q) + \boldsymbol{V}(k,q). \qquad (1)$$

In this paper, the term *segment* indicates a set of consecutive frames. Two consecutive segments can overlap.

3. Automatic Adaptive Bi-Beam System

Considering a recording with two speakers with known locations, one wishes to separate speech signals of each individual. Figure 1 shows the diagram of the proposed system for an automatic adaptation beamforming system for speech separation for two target speakers. The system contains two modules 1) VAC module, and 2) a speech separation module. Given a segment of a multiple channel input signal $\boldsymbol{x}$, firstly the VAC module identifies the voice activity status of the two target speakers. This VAC result is then given to the second module so that the system can turn on/off the adaptation of the two adaptive beamformers $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. Finally, each adaptive beamformer can separate the desired signal from the input. The system provides two simultaneous output signals $\boldsymbol{y}_1, \boldsymbol{y}_2$. In this work, we use MVDR beamforming method for this separation step.

We briefly review the well-known adaptive beamforming method MVDR in Section 3.1, then introduce the novel VAC method in Sections 3.2 and 3.3. Processing details of the whole system are presented afterward.

3.1. MVDR Beamforming

Assume that one is interested in separating speech of one speaker, for example speaker 1, from a multiple channel input signal given by Eq. 1. For compact notation, the frequency index is neglected in this section. The MVDR beamforming method offers an effective solution by designing a beamformer $\boldsymbol{W}$ minimizing the power output of the interference and noise while maintaining a distortionless response constraint at the direction of the wanted speaker [5]. Technically, this weighting vector $\boldsymbol{W}$ is found by solving the optimization problem

$$\min_{\boldsymbol{W}} \boldsymbol{W}^H \boldsymbol{R}_{\text{IpN}} \boldsymbol{W} \quad \text{subject to} \quad \boldsymbol{A}_1^H \boldsymbol{W} = 1, \qquad (2)$$

where $\boldsymbol{R}_{\text{IpN}}$ is the covariance matrix of the interference and noise, $\boldsymbol{A}_1$ is the *steering vector* (or the transfer function) toward speaker 1 [18], [5, Chapter 6]. This optimization problem can be solved by Lagrange multipliers, yielding the solution

$$\boldsymbol{W} = \frac{\boldsymbol{R}_{\text{IpN}}^{-1}\boldsymbol{A}_1}{\boldsymbol{A}_1^H \boldsymbol{R}_{\text{IpN}}^{-1}\boldsymbol{A}_1}. \qquad (3)$$

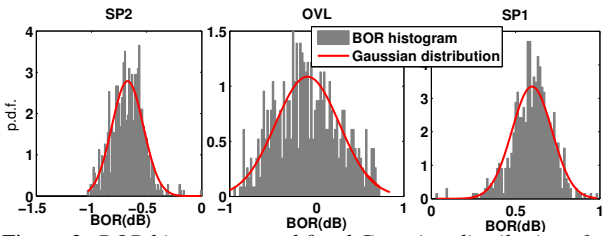

Figure 2: *BOR histograms and fitted Gaussian distributions for SP1,OVL,SP2 at about* 370 *Hz. Synthesized data with* 300 *ms reverberation, PASCAL Speech Separation Challenge II recording setup [20].*

To maintain a high speech separation performance, the MVDR weighting vector should be adapted regularly with updated covariance matrix $\boldsymbol{R}_{\text{IpN}}$ using new data. In practice, for each frame q ($q = \{1,2,3...\}$), the matrix $\boldsymbol{R}_{\text{IpN}}$ should be updated if the input signal contains only interference and noise [5]

$$\boldsymbol{R}_{\text{IpN}}(q) = \boldsymbol{X}(q)\boldsymbol{X}^H(q) + \mu \boldsymbol{R}_{\text{IpN}}(q-1), \qquad (4)$$

where μ ($\mu \in (0,1)$) is the *forgetting rate* to exponentially reduce the impact of the data in the past. Under the presence of the wanted speaker in the input signal, the adaptation should be halted to avoid the signal cancellation problem.

3.2. Gaussian BOR for Segment-level VAC

Given a segment of the input signal, VAC is done for this segment and this step is called segment-level VAC (sVAC). Considering two target speakers, four possible voice activity cases are 1) noise-only, 2) only speaker 1 is active (SP1), 3) overlapping speech (OVL), and 4) only speaker 2 is active (SP2). Assuming that the noise only case can be detected in a pre-processing step by a noise detector, Figure 1, the last three cases can be classified using beamforming-output-ratio (BOR) as introduced in our work at [16, 19]. This method is called BOR-VAC.

Let $\{\widetilde{\boldsymbol{W}}_1(k), \widetilde{\boldsymbol{W}}_2(k)\}$ be two fixed weighting vectors of a BiBeam at frequency k, the BOR for a segment is [16]

$$r(k,q) = \frac{\frac{1}{|\mathcal{Q}_l|}\sum_{q\in\mathcal{Q}_l} |\widetilde{\boldsymbol{W}}_1(k)^H \boldsymbol{X}(k,q)|^2}{\frac{1}{|\mathcal{Q}_l|}\sum_{q\in\mathcal{Q}_l} |\widetilde{\boldsymbol{W}}_2(k)^H \boldsymbol{X}(k,q)|^2}, \qquad (5)$$

where $l \in \mathbb{Z}^+$ is the segment index, $\mathcal{Q}_l$ is the set of frame indexes in segment l, and $|\mathcal{Q}_l|$ is the size the set. The *log BOR* is calculated as $r_{\text{L}}(k,q) = 10\log_{10}(r(k,q))$.

When the two beamformers $\{\widetilde{\boldsymbol{W}}_1, \widetilde{\boldsymbol{W}}_2\}$ can enhance speech of speaker 1 and 2 respectively, such as by using conventional beamforming (CVBF) [5], one can expect that the output ratio follows the ratio of the signals of the two speakers in the input. Hence, BOR can indicate active speaker(s). Technically, given an input segment, BORs are calculated for selected bins and are compared with pre-specified thresholds to identify the active speaker(s). In [16], theoretical probability distributions of BORs have been derived. Based on these statistics, frequency bins and thresholds can be automatically chosen for a given error probability pre-specified for a VAC system. However, this method requires perfect knowledge about the signal propagation, which is often unavailable in practice.

To provide a practical realization for this method, Gaussian distributions are assumed for the log BOR of two CVBF beamformers. Figure 2 shows examples of fitted Gaussian distributions matching the BOR histograms of the three voice activity cases. Also notice that the histograms of the SP1/SP2 cases are well separated, yet each overlaps with the OVL histogram. In general, Gaussian approximation may not always

have a good fit and lead to degradation in the classification. However, through experimental results, it appears to be a reasonable and sufficient compromise for practical systems. In the proposed system, this approach is used for robust VAC. The second method will be developed later to refine the VAC outputs.

Using the Gaussian distribution approximation for log BOR, a bin and threshold selection scheme can be derived in the same fashion as the scheme in [16]. The classification for SP1-OVL-SP2 is constructed as two sub-classification problems: SP1-OVL and OVL-SP2, and each can be solved using one frequency bin and one threshold. Firstly we consider the SP1-OVL classification. Letting OVL be the null hypothesis, the classification rule is to select SP1 if $r(\bar{k}_1,,q) > \theta_1$ and OVL otherwise, where $\bar{k}_1$ and θ_1 are the bin and threshold one needs to select. The selection of these parameters is as follows.

From a training set, the mean and the variance of fitted Gaussian distributions are obtained for the SP1 and OVL BORs on each frequency bin, denoted as $\mu_z(k)$ and $\sigma_z^2(k)$ where $z \in \{1, 0, 2\}$ denotes the voice activity type, ($\{$SP1,OVL,SP2$\}$ respectively). We recall the Gaussian cumulative distribution function

$$g_{z,k}(\theta) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\theta - \mu_z(k)}{\sqrt{2\sigma_z^2(k)}}\right)\right], \quad (6)$$

where $\text{erf}(.)$ is the error function [21]. Hence, on frequency bin k, if a threshold $\theta(k)$ is used for the classification, the type I error probability (mis-detecting OVL) is $\epsilon_{1,I}(\theta(k)) = 1 - g_{0,k}(\theta(k))$ [22]. The type II error probability (mis-detecting SP1) is $\epsilon_{1,II}(\theta(k)) = g_{1,k}(\theta(k))$.

The selection idea is that by pre-specifying a type I error probability $\bar{\epsilon}_{1,I}$, the frequency bin and threshold are chosen to minimize the type II error probability. This is done via four steps. Firstly, on each frequency bin, a potential threshold is calculated to satisfy the type I error by solving the Eq. $\epsilon_{1,I}(\theta(k)) = \bar{\epsilon}_{1,I}$ for θ, yielding the solution

$$\theta(k) = \sqrt{2\sigma_0^2(k)}\text{erf}^{-1}(2\bar{\epsilon}_{1,I}(k) - 1) + \mu_0(k). \quad (7)$$

Secondly, the type II probabilities $\epsilon_{1,II}(\theta(k))$ corresponding to the potential thresholds are calculated for all bins.

Then, a best bin is selected

$$\bar{k}_1 = \arg\min_k\{\epsilon_{1,II}(\theta(k))\}. \quad (8)$$

Finally, the corresponding threshold is chosen $\theta_1 = \theta(\bar{k}_1)$.

The scheme for OVL-SP2 is derived similarly. After the thresholds and bins are selected, when receiving a new segment of input signal, one can calculate the log BOR, then carries out the two classifications. If SP1-OVL classifies the voice activity of a segment as SP1, the sVAC result is SP1. Similarly, if the SP2-OVL output is SP2, the sVAC result is SP2. Otherwise, the sVAC result of segment is OVL.

3.3. Inverse Power Check

The proposed selection scheme allows suitable thresholds be chosen for the BOR-VAC method given the VAC error probabilities. For adaptation control in speech separation, VAC aims to provide a low probability of OVL misdetection to avoid the wanted-signal cancellation problem. One approach is to accept large type II error probabilities which eventually leads to avoiding adaptation. A more effective approach is to use a checking method that can fix VAC mis-detection errors created at the BOR-VAC step, so that the type II error probabilities remain low while the type I errors are reduced.

To derive the checking method, we firstly derive the following result on the output power of a MVDR beamformer under different adaptation conditions. In a fixed frequency bin, we consider an MVDR beamformer $\boldsymbol{W}$ that always adapts using the coming signal, i.e. $\boldsymbol{W}$ is calculated using Eq. (2) with the interference and noise correlation matrix $\boldsymbol{R}_{\text{IpN}}$ is replaced by the correlation matrix of the input signals $\boldsymbol{R}$. Hence, the beamformer output power is [5, 23]

$$|\boldsymbol{W}^H\boldsymbol{X}|^2 = 1/(\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}). \quad (9)$$

We define the *inverse power function* as

$$\zeta(\boldsymbol{A}, \boldsymbol{R}) \triangleq 1/|\boldsymbol{W}^H\boldsymbol{X}|^2 = (\boldsymbol{A}^H\boldsymbol{R}^{-1}\boldsymbol{A}). \quad (10)$$

Note that this function is calculated using the CVBF $\widetilde{W} = \boldsymbol{A}$ instead of the beamformers $\boldsymbol{W}$.

With correct adaptation, i.e. the input contains only the interference and noise, $\boldsymbol{R} = \boldsymbol{R}_{IN}$, the output inverse power is

$$\zeta(\boldsymbol{A}, \boldsymbol{R}_{IN}) = (\boldsymbol{A}^H\boldsymbol{R}_{IN}^{-1}\boldsymbol{A}). \quad (11)$$

When adaptation is incorrect, i.e. the input also contains the wanted speech, $\boldsymbol{R} = \boldsymbol{R}_W + \boldsymbol{R}_{IN}$ where $\boldsymbol{R}_W$ is the matrix of the wanted speech. Using the matrix inversion lemma [5, 11], the inverse of the matrix $\boldsymbol{R}$ can be written as $\boldsymbol{R}^{-1} = \boldsymbol{R}_{IN}^{-1} - \boldsymbol{C}$ with $\boldsymbol{C}$ being positive definite. Thus, the output inverse power is

$$\zeta(\boldsymbol{A}, \boldsymbol{R}_W + \boldsymbol{R}_{IN}) = \zeta(\boldsymbol{A}, \boldsymbol{R}_W) - \boldsymbol{A}^H\boldsymbol{C}\boldsymbol{A}. \quad (12)$$

Note that $\boldsymbol{A}^H\boldsymbol{C}\boldsymbol{A} > 0$ as $\boldsymbol{C}$ is positive definite and hence

$$\zeta(\boldsymbol{A}, \boldsymbol{R}_W + \boldsymbol{R}_{IN}) < \zeta(\boldsymbol{A}, \boldsymbol{R}_{IN}). \quad (13)$$

In other words, the inverse output power reduces when the input is OVL. Based on this result with an assumption that the signal powers in the input do not severely fluctuate, the checking method can be used as follows. If an input segment is identified as SP1 by the BOR-VAC method, on a selected frequency bin dedicated for the SP1-OVL sub-classification, the inverse power function calculated for the segment is compared with a pre-defined threshold. If the result is lower than the threshold, the VAC classification is changed into OVL. A similar checking scheme is applied for the SP2 case. The thresholds and the bins are selected in a training stage in a similar fashion as the selection scheme for BOR-VAC.

3.4. Automatic Speech Separation

By integrating the proposed VAC method into the adaptation for MVDR beamforming, the proposed system becomes feasible to automate. In this section, processing details of the whole system, as shown in Figure 1, are presented. Given a segment of a multiple channel input signal $\boldsymbol{x}$, the target of the system is to simultaneously provide two outputs $\boldsymbol{y}_1, \boldsymbol{y}_2$ which are the separated speech signals for speaker 1 and speaker 2. The overall processing of the system has the following steps:

1. A noise detector detects if the input contains only noise
2. If not only noise in the input, beamform using the CVBF beamformers $\{\widetilde{\boldsymbol{W}_1}, \widetilde{\boldsymbol{W}_2}\}$
3. Segment-level VAC via two steps: 1) robust VAC using Gaussian BOR-VAC and 2) refined VAC using the inverse power check (IPC) method
4. Frame-level VAC (fVAC) to identify a voice activity case for each frame of the input
5. The adaptation control (AC) module decides suitable adaptation modes for the two MVDR beamformers $\{\boldsymbol{W}_1, \boldsymbol{W}_2\}$
6. Speech separation using $\{\boldsymbol{W}_1, \boldsymbol{W}_2\}$ for two speakers.

ID	$\bar{\epsilon}_{1,I}$ $(= \bar{\epsilon}_{2,I})$	iSINRs	VAC Results			ABS - Final Results		Genie-aided Results	
			ϵ_1	ϵ_0	ϵ_2	oSINRs	LSDs	oSINRs	LSDs
1	0.01	$(-1, -2)$	0.2	0	0.1	$(17.4, 13.3)$	$(0.87, 1)$	$(17.4, 14.8)$	$(0.87, 0.95)$
2	0.005	$(-2.6, 0)$	0.1	0	0.1	$(13.8, 19)$	$(0.86, 0.93)$	$(14.5, 19.5)$	$(0.86, 0.93)$
3	0.002	$(-3, -0.3)$	0.1	0	0.13	$(12, 17.2)$	$(0.9, 1)$	$(12, 17.3)$	$(0.981.02)$
4	0.1	-	0.01	0.07	0.24	-	$(1.17, 1.26)$	-	$(1.16, 1.26)$

Table 1: Input parameters and ABS final outputs, VAC outputs in comparison with ideal outputs.

fVAC is the last step in the VAC module, see Figure 1. It makes classification decisions for individual frames based on the sVAC of the segments containing the current frame. To avoid the signal cancellation problem, the fVAC decision is a conservative scheme favouring the OVL case by only approving the single-speaker cases if the majority of the involved segments have that result. This conservative scheme may result in less accurate VAC, however it is beneficial for speech separation as the adaptation is halted during OVL periods.

The final outputs of the VAC module are sent to the AC component in the speech separation module. AC can simply turn on/off the adaptation of each beamformer frame by frame. The adaptive BiBeam simultaneously implements the separation process in the two beamformers and provides two outputs.

4. Simulation Results

This section presents simulation results for the performance of the proposed system in both synthesized and real recordings. The results include VAC accuracy, and quantitative assessment of the final outputs. Audio samples, time domain plots for the inputs/outputs and details of the room setups can be accessed at our webpage [24].

The synthesized recordings use signals and the room setup of the PASCAL Speech Separation Challenge II [20] with an eight-element circular microphone array (10 cm radius) at the center of a table. Speakers sit around the table with the distances to the arrays being about 80 cm. Angle separation between speakers is from $50^\circ - 180^\circ$. The signal propagation is synthesized with 0.3 s reverberation time, using software from [25]. Each recording includes two speakers, a background speech, computer fan and white noise. Each recording contains about three minutes of mixture of single speaker and overlapping periods. Each type is about one minute in total.

The real recording uses a uniform linear array of six microphones in an $9.5\text{m} \times 12$ m laboratory. The inter-element spacing of the array is about 14cm. Two speakers are located at about 1.3m from the array, and are asked to answer two questions without preparation. Noises are mostly from air conditioner vents in the ceiling. Two close-talking microphones recorded reference signals from each speaker. The recording is two minutes long containing all voice activity types.

CVBF and MVDR beamformers are initiated with given speaker locations. However the real recording has location errors ($\approx 20 - 30$ cm). The MVDR beamformers use a forgetting rate of 0.85 [3]. BOR-VAC segment length is 60 frames for the 300 ms reverberation length, and 90 frames for the real recording. In both cases, two consecutive segments overlap about 80%. The frame length is about 30 ms with 50% overlap for BOR-VAC and about 60 ms for the MVDR beamformers. For each setup, BOR-VAC is trained using one minute of data of each BOR type. The BOR-VAC thresholds are selected using the proposed scheme in Section 3.2. The noise-only detection for the real recording uses a simple signal-to-noise threshold as the noise is stationary.

The two objective assessments, output-input signal-to-interference-and-noise-ratio (oSINR-iSINR) and log-spectral-distortion (LSD) are used [3] for the separation outputs. LSD measures the distortion between the outputs and the ideal signals, which are the reference signals. SINRs and LSD are separately calculated for each speaker, i.e. when speaker 1 is of interest, speaker 2 is considered as an interferer and vice versa. In the real recording, only LSD is available. The outputs of the proposed automatic system are compared with the ideal outputs provided by genie-aided adaptive beamformers in which perfect knowledge about voice activity is given.

Besides, VAC performance has been examined via the mis-ditection errors. The mis-detection error of each voice activity type is $\epsilon_z = \frac{\widetilde{N}_z}{N_z}$ where N_z is the number of frames that have voice activity z, and $\widetilde{N}_z$ is the number of frames with incorrect VAC result for voice activity type z.

The results are shown in Table 1. The SINRs and LSDs are written in pairs of (beamformer 1, beamformer 2) outputs. The first part of the table shows the input SINR and the specified error probabilities used for the threshold selection. The first three tests use synthesized signals and the setup 4 is the real recording. The VAC outputs have no OVL mis-detection and the SP1/SP2 mis-detections are from $0 - 0.24$. These errors are expected as the type I error probabilities are set to low values, i.e. 0.002 to 0.01. Overall, the VAC results satisfy the conservative classification to avoid signal cancellation.

The last four columns of the table show the output SINR and LSDs of the automatic systems and of the genie-aided systems. The automatic systems perform close to the ideal results. The differences are due to the mis-detection for the single speaker cases (SP1,SP2) as these mis-detections lead to missing adaptation opportunities. In the real recording case, the higher LSD for the output of speaker 1 is also due to errors of the noise detection step (0.001% error).

In general, the experimental results show that the use of the IPC module to refine VAC becomes essential in real recordings. Without the inverse power checking method, the outputs in the setups $1 - 3$ are slightly affected. However, the performance in test 4 drops significantly (LSD of the first output increases to 1.24) as the VAC errors increase ($\epsilon_0 = 0.07$). Besides, IPC helps by relaxing the selection for the pre-specified type I error probability since it can fix OVL-misdetection errors.

5. Conclusion

In this paper, an automatic adaptive speech separation system for two speakers has been derived. The system relies only on beamforming techniques with low computational complexity. The Gaussian BOR-VAC combined with the inverse power checking method provides reliable VAC results at a segment level, and the final results are at a frame-level using multiple segments. These novel methods are integrated with the adaptive beamforming system and follow a conservative VAC decision scheme to avoid the signal cancellation problem for speech separation. The effectiveness of the whole system is confirmed for spatially-stationary scenarios by both synthesized and real recordings. The outputs show that the proposed automatic system can perform closed to the equivalent genie-aided systems.

6. References

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, Aug. 2012.

[2] D. M. Woelfel and D. J. McDonough, *Distant Speech Recognition*. Wiley, Jun. 2009.

[3] Thuy N. Tran, W. Cowley, and A. Pollok, "Adaptive blocking beamformer for speech separation," in *12th Annual Conference of the International Speech Communication Association*. ISCA, 2011, pp. 577–580.

[4] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[5] H. L. V. Trees, *Optimum Array Processing Part IV of Detection, Estimation, and Modulation Theory*, 1st ed. Wiley-Interscience, 2002.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Springer, 2008, pp. 1687–1692.

[7] "PASCAL speech separation challenge part II," http://homepages.inf.ed.ac.uk/mlincol1/SSC2/index.htm.

[8] "Chime challenge 2011," http://spandh.dcs.shef.ac.uk/projects/chime/workshop/.

[9] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 8, pp. 2527–2541, 2007.

[10] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. McGraw-Hill Science/Engineering/Math, 1999.

[11] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *The Journal of the Acoustical Society of America*, vol. 54, no. 3, pp. 771–785, 1973.

[12] S. Mohsen Naqvi, W. Wang, M. Khan, M. Barnard, and J. Chambers, "Multimodal (audiovisual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking," *IET Signal Processing*, vol. 6, no. 5, pp. 466 –477, Jul. 2012.

[13] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Sig. Proc.*, vol. 15, no. 2, 2007.

[14] N. Madhu, "Acoustic source localization : Algorithms, applications and extensions to source separation," Ph.D. dissertation, Ruhr University Bochum, 2009.

[15] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 1st ed. Addison-Wesley Publishing Company, 1987.

[16] Thuy N. Tran, W. Cowley, and A. Pollok, "Multi-speaker beamforming for voice activity classification," in *submitted to Australian Communications Theory Workshop, Feb 2013*.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[18] I. McCowan, "Robust speech recognition using microphone array," Ph.D. dissertation, Queensland University of Technology, Australia, 2001.

[19] Thuy N. Tran, W. Cowley, and A. Pollok, "Voice activity classification using Beamformer-Output-Ratio," in *2012 Australian Communications Theory Workshop*. IEEE, 2012, pp. 105–110.

[20] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 357–362.

[21] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, 3rd ed. Addison Wesley, 2002.

[22] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*, 1st ed. Prentice Hall, Feb. 1998.

[23] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Trans. Sig. Proc.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[24] "Automatic adaptive bibeam speech separation demo," http://www.itr.unisa.edu.au/itrusers/tratn014/public/InterSpeech13Homepage.html.

[25] E. Habets, "Room impulse response generator," http://home.tiscali.nl/ehabets/rir_generator.html.