An Acoustic Front-End for Interactive TV
Incorporating Multichannel Acoustic Echo Cancellation and Blind Signal Extraction

Klaus Reindl, Yuanhang Zheng, Anthony Lombard, Andreas Schwarz, and Walter Kellermann
Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {reindl, zheng, lombard, schwarz, wk}@lnt.de

Abstract—In this contribution, an acoustic front-end for distant-talking interfaces as developed within the European Union-funded project DICIT (Distant-talking interfaces for Control of Interactive TV) is presented. It comprises state-of-the-art multichannel acoustic echo cancellation and blind source separation-based signal extraction and only requires two microphone signals. The proposed scheme is analyzed and evaluated for different realistic scenarios when a speech recognizer is used as back-end. The results show that the system significantly outperforms simple alternatives, i.e., a two-channel Delay & Sum beamformer for speech signal extraction.

I. INTRODUCTION

The project DICIT (see [1]) focused on the problems of acoustic scene analysis and speech interaction in noisy and reverberant environments by means of microphone networks. The goal was to provide a user-friendly multi-modal interface that allows voice-based access to a virtual smart assistant for interacting with TV-related digital devices and infotainment services, such as digital TV, HiFi audio devices, etc., in a typical living room. Multiple, possibly moving users should be able to comfortably control the TV set via voice, e.g., requesting program information or scheduling desired recordings without using a hand-held or head-mounted control. For this, real-time-capable acoustic signal processing is necessary that can compensate for the impairments of the desired speech signal which may result from interfering speakers, ambient noise, reverberation, and acoustic echoes from the TV loudspeakers. Therefore, in the DICIT project, a combination of state-of-the-art multichannel acoustic echo cancellation (MC-AEC), beamforming (BF), and multiple source localization were evaluated and realized in a prototype (see [1]). As an alternative to the large microphone array in [1], the front-end proposed here requires only two microphone signals.

II. SIGNAL MODEL

The proposed human-machine interface for interactive TV shown in Fig. 1 is based on stereo sound reproduction and two-channel audio capture and it combines MC-AEC and blind signal extraction (BSE). The acquired microphone signals $x_p$, $p \in \{1, 2\}$, contain the signals of $Q$ simultaneously active point sources, where only one signal (here: $s_1$) is considered as desired signal to be extracted, and the remaining $Q-1$ source signals are regarded as interfering signals. Moreover, acoustic echoes from the TV loudspeakers and background noise denoted by $n_p$, $p \in \{1, 2\}$, are present in the observed microphone signals. For a speech recognizer it is important that the target speech components (here: $s_1$) are properly extracted from the acquired microphone signals. Therefore, first of all, the microphone signals are fed into an MC-AEC that compensates for the acoustic coupling between the loudspeakers and the microphones. As the stereo channels of the TV audio are usually very similar and therefore not only highly auto-correlated but also strongly cross-correlated, the so-called non-uniqueness problem of MC-AEC arises. To alleviate this issue, the loudspeaker signals need to be mutually decorrelated without affecting the perceived sound quality. The output signals of the MC-AEC are then fed into a two-channel blind signal extraction (BSE) unit. As the MC-AEC cancels only the echo components contained in $x_p$, $p \in \{1, 2\}$, the signals $y_p$, $p \in \{1, 2\}$, still contain all noise and interference signals ($n_p$, $p \in \{1, 2\}$ and $s_q$, $q = (2, \ldots, Q)$). Therefore, the subsequent BSE concept extracts the desired speech signal components from the MC-AEC output signals by suppressing all noise and interference components.

In principle, BSE could be combined in two different ways with AEC. The AEC can be performed directly on the microphone inputs $x_p$, $p \in \{1, 2\}$, or it can be applied at a later stage to the BSE output. Taking into account considerations described in [2], [3], we concentrate on the AEC-first configuration. If the AEC is applied after the BSE unit, besides the loudspeaker-enclosure-microphone (LEM) system, the AEC has to model the BSE system as well. As the BSE scheme is time-varying, the AEC cannot converge to a stable solution and to this end, this alternative is not considered. For the front-end presented in [1], the echo cancellation scheme is applied after a beamformer. This is possible as the applied beamformer is a linear combination of time-invariant beams, so that the echo canceller does not have to model a rapidly time-varying beamformer. Moreover, due to the fact that thirteen microphone signals were used, the AEC-first configuration would require thirteen MC-AEC’s which is highly undesirable.

III. MULTICHANNEL ACOUSTIC ECHO CANCELLATION

The multichannel AEC applied to the proposed two-channel acoustic human-machine interface is shown in Fig. 2. As discussed in [4],

![Fig. 1. Signal model of the proposed acoustical front-end](image1)

![Fig. 2. Realization of the multichannel AEC and the preceding decorrelation unit](image2)
[5], the integrated acoustic echo cancellation solution is based on a class of efficient and robust adaptive algorithms in the frequency domain. As the robustness issue during double talk is particularly crucial for fast-converging algorithms, the concept of robust statistics is applied to the frequency-domain approach [5]. Correspondingly, the algorithm becomes inherently less sensitive to outliers, i.e., short bursts that may be caused by inevitable detection failures of a double-talk detector. Exploiting the computational efficiency of the FFT for minimizing computational load, it also accounts for the cross-correlations among the different reproduction channels to accelerate bursts that may be caused by inevitable detection failures of a double-talk detector.

A. Blocking Matrix

The blocking matrix that performs time-frequency filtering as well as spatial filtering is based on the TRINICON (TRIple-N-Independent component analysis for CONvolutive mixtures) optimization criterion (introduced in [7], [8]). The TRINICON cost function is given by the Kullback-Leibler divergence (KLD) between the estimated $P$-variate joint probability density function (PDF) $f_{x_1,P}(x_1,\ldots,x_P)$ of the output signals of the demixing system and the product $\prod_{p=1}^P f_{x_p}(x_p)$ of the estimated $P$-variate marginal output PDFs:

$$J_\text{BSS}(n) = \sum_{i=0}^{\infty} \beta(i,n) \sum_{j=1}^{N} \left\{ \log \left( \frac{f_{x_1,P}(x_1,\ldots,x_P)}{\prod_{p=1}^P f_{x_p}(x_p)} \right) \right\}.$$  

where $i$ and $n$ denote block indices and the vectors $x_p$ contain $D$ consecutive output samples each. $\beta(i,n)$ denotes a window function that allows for offline, online, and block-online algorithms. In general, the KLD involves the expectation operator. This operator has been replaced by a short-time average $\hat{J}_\text{BSS}$ over $N$ blocks of length $D$. If and only if the BSS outputs are statistically independent, i.e., for perfect separation and assuming mutually independent source signals, (2) becomes zero. A natural-gradient-descent approach is applied for iterative optimization of the BSS filter coefficients. For our signal extraction approach an efficient second-order-statistics (SOS) realization of the TRINICON update rule was derived based on multivariate Gaussian probability density functions.

As there is no determined solution for a demixing matrix to separate the individual sources in an underdetermined case (more active sources than available microphone signals), the generic TRINICON cost function is modified so that the noise and interference components can be separated from the target signal when only two microphone signals are available. The cost function of this ‘directional BSS’ concept [9] is given by

$$J_{\text{DirBSS}} = J_{\text{BSS}} + \eta_C J_{\overline{C}},$$  

where $J_{\overline{C}}$ represents a geometrical constraint and is given by

$$J_{\overline{C}} = ||b_{11}(k) + b_{21}(k - \tau_v)||^2.$$  

The weight $\eta_C$, typically in the range $0.4 < \eta_C < 0.6$, indicates the relative importance of the geometrical constraint [9]. Owing to the property of BSS to produce independent output signals, directional BSS also suppresses correlated components arriving from other directions, i.e., reflections and reverberation will also be suppressed to the greatest extent possible. To this end, directional BSS is superior to conventional beamforming techniques, e.g., null-beamformers, in suppressing the target signal located at $\phi_{\text{tar}}$, especially in reverberant environments (see [9]). Moreover, in contrast to many beamformer techniques, e.g., [10], [11], no voice-activity detector is needed and no prior knowledge on the microphone positions is required.

The directional constraint as given in (4) forces a spatial null towards the desired source location which has to be estimated or known a priori in real applications. $\tau_v$ describes the time difference of arrival (TDOA) of the target source between the two sensors. It has to be noted that in real applications, this can be any fractional delay. If a-priori information about the target angular position is missing, the localization concept as discussed in [1] can be applied. Throughout this paper it is assumed that the target source is located in front of the microphone array in a predefined angular range of $-20^\circ \leq \phi_{\text{tar}} \leq 20^\circ$ (same assumption as in [9]). Finally, the output
signal of directional BSS can be approximated by
\[ \hat{n}(k) = b_{11}(k) \ast y_1(k) + \ldots \]
and is discussed with respect to speech recognition results. For all experimental results the SER as defined in (7) is set relative importance of the directional constraint for directional BSS.

B. Noise and interference suppression

In order to extract the desired speech signal components, either single-channel or multichannel noise reduction techniques can be applied. However, multichannel techniques require reliable estimates of the noise and interference components in all available microphones. Since, in practice, it is almost impossible to obtain these separate noise and interference estimates in highly non-stationary scenarios, the combination of BSS methods with single-channel Wiener filtering techniques to obtain an estimate of the desired speech signal components \( \hat{s}_i \) is investigated. To this end, the single noise and interference reference \( \hat{n} \) obtained by directional BSS is used to control spectral enhancement filters \( w_p \), \( p \in \{1, 2\} \), as shown in Fig. 4. The spectral weights of the applied Wiener filtering strategy are given by
\[ w_p = \max \left[ 1 - \mu \frac{P_{k_\text{in}}}{P_{k_p}}, w_{\text{min}} \right], \quad p \in \{1, 2\}, \]
where \( \mu \) and \( w_{\text{min}} \) denote a gain factor and the spectral floor, respectively. These parameters are real-valued constants and are used to achieve a trade-off between noise reduction and speech distortion. \( P_{k_\text{in}} \) and \( P_{k_p} \), \( p \in \{1, 2\} \), represent power spectral density (PSD) estimates of the noise and interference reference \( \hat{n} \) and the filtered microphone signals \( v_p \) (see Fig. 4), respectively.

V. EXPERIMENTS

Experimental results are illustrated and discussed in order to show the effectiveness of the proposed two-channel acoustic front-end. The experiments are performed in a living-room-like environment with a reverberation time of \( T_{60} \approx 300 \text{ms} \). The setup in this environment is illustrated in Fig. 5. The two-channel microphone array (microphone spacing \( d = 16 \text{cm} \)) is located in front of the TV screen. The distance between the microphone array and the user’s mouth is about 3m. I1 to I5 show interferer positions as considered for the following evaluations. For all experiments, the user is represented by a real person whereas all interfering speech signals (I1 to I5) are simulated by loudspeaker signals. The sampling frequency is set to \( f_s = 16 \text{kHz} \). The filter length for the MC-AEC is set to \( L_{\text{AEC}} = 4096 \), for directional BSS a filter length of \( L_{\text{dirBSS}} = 1024 \) is used, and for the Wiener filtering concept the filter length is set to \( L_{\text{Wiener}} = 512 \). The relative importance of the directional constraint for directional BSS \( \eta_z \) is equal to 0.5. In order to achieve a trade-off between noise and interference suppression and speech distortion of the Wiener filtering concept, the parameters \( \mu \) and \( w_{\text{min}} \) are set to 1.5 and 0.05, respectively. The Wiener filtering concept is implemented with a polyphase filterbank. The filter length of the prototype lowpass filter is 1024, the number of subbands (complex-valued) equals 512, and the downsampling rate is set to 128. In contrast to many noise reduction techniques, e.g., [12], [13], we focus on a suppression of highly non-stationary signals, i.e., speech signals.

Before the individual results are discussed, the different performance measures that are used to evaluate the performance of the proposed front-end are introduced. In order to characterize the individual scenarios, a segmental signal-to-echo ratio (SER) as well as a segmental signal-to-interference ratio at the microphones (SIRin) is defined. These measures are given by
\[ \text{SER} = \frac{1}{2N} \sum_{p=1}^{2} R_{k_\text{in}}(k + nK), \quad (7) \]
\[ \text{SIRin} = \frac{1}{2N} \sum_{p=1}^{2} \frac{R_{k_\text{in}}(k + nK)}{R_{k_\text{in}}(k + nK)} \]
where \( s_{1,p}(k), p \in \{1, 2\} \), denotes the desired speech signal contained in the microphone signals \( s_{1,p}(k) \). \( e_{p}(k) \) represents the echo signal contained in the microphone signals, and \( n_{\text{g},p}(k) \) denotes all noise and interference components contained in the microphones and is given by \( n_{\text{g},p}(k) = \sum_{q=0}^{Q} s_{q,p}(k) + e_{p}(k) \). \( k \) and \( N \) represent the discrete time index and the total number of data blocks, respectively. The block length is denoted by \( K \) and is set to 1024 samples. In order to evaluate the MC-AEC performance, the echo-return-loss-enhancement (ERLE) is evaluated and is defined as
\[ \text{ERLE}(n) = \frac{1}{2} \sum_{p=0}^{2} 10 \log_{10} \left( \frac{K-1 \sum_{k=0}^{K-1} s_{2,p}^2(k + nK)}{K-1 \sum_{k=0}^{K-1} e_{2,p}^2(k + nK)} \right), \quad (9) \]
where \( e_{\text{r},p}(k) \) denotes the residual echo components contained in the output signals of the MC-AEC \( y_{\text{p},k}(k) \), \( p \in \{1, 2\} \). For evaluating the BSE performance, the SIR gain (SIRgain) as well as speech distortion (SD) at the output of BSE are studied. The SIR gain SIRgain is defined as follows:
\[ \text{SIRgain} = \text{SIRout} - \text{SIRin}, \quad (10) \]
\[ \text{SIRout} = \frac{1}{N} \sum_{n=0}^{N-1} 10 \log_{10} \left( \frac{K-1 \sum_{k=0}^{K-1} s_{2,n}^2(k + nK)}{K-1 \sum_{k=0}^{K-1} n_{\text{g},n}^2(k + nK)} \right), \quad (11) \]
where \( n_{\text{r},n}(k) \) denotes all residual noise and interference components contained in the output of BSE. Speech distortion SD is calculated as
\[ \text{SD} = \frac{1}{N} \sum_{n=0}^{N-1} 10 \log_{10} \left( \frac{1}{K-1} \sum_{k=0}^{K-1} \left( \hat{s}_1(k + nK) - s_{1,n}(k + nK) \right)^2 \right), \quad (12) \]
with \( \hat{s}_1(k) \) denoting an estimate of the desired source signal components obtained by BSE (here, the signal delay introduced by the entire acoustic front-end is already compensated) and \( s_{1,n}(k) \) is given by \( s_{1,n}(k) = \frac{1}{K} \sum_{k=0}^{K-1} s_{1,n}(k) \). In the following, the individual building blocks (MC-AEC and BSE) are evaluated and finally the overall performance is discussed with respect to speech recognition results. For all experimental results the SER as defined in (7) is set.
to SIR$_{\text{SER}} = -3$dB and the SIR at the microphones as defined in (8) is equal to SIR$_{\text{mic}} = 0$dB.

A. MC-AEC performance

The MC-AEC performance is evaluated for a typical TV control scenario, where the desired speaker (user in Fig. 5) utters several commands. The desired signal is superimposed by loudspeaker echoes and interfering signals. Here, the interferers I$_2$ to I$_5$ as shown in Fig. 5 are active. The performance of the MC-AEC in terms of the echo-return-loss-enhancement (9) is shown in Fig. 6. This result shows that after a certain convergence phase of the MC-AEC, a stable gain of 20 – 25dB can be obtained. The convergence phase strongly depends on the double-talk detection, i.e., it strongly depends on the activities on the individual sources. However, after a certain convergence phase of the AEC, a stable gain of at least 20dB can always be ensured for a typical TV control scenario.

B. BSE performance

In the following, the performance of the blind signal extraction scheme is analyzed. As a first step, the behavior of the proposed BSS-based blocking matrix (directional BSS) is studied along with Fig. 7. The preceding AEC is not considered for this analysis. The signals captured by two microphones at a distance of $d = 10$cm (the same microphone array as shown in Fig. 5) are fed into a blocking matrix, i.e., the microphone signals are filtered by the filters $b_p$, $p \in \{1, 2\}$, the resulting signals are summed up and finally yield a reference of all noise and interference components denoted by $\hat{n}$. In order to evaluate the behavior of this system, the spatiotemporal frequency response associated with the blocking matrix is analyzed. Therefore, a source signal $s$ is located in front of the microphone array at a certain position $\phi (-50^\circ \leq \phi \leq 50^\circ)$. The blocking matrix is steered towards $0^\circ$ and this steering direction is fixed for this analysis. The distance between the source and the center of the microphone array is $3m$. The analysis is performed in a living-room-like environment with a reverberation time $T_{\text{eq}} \approx 300$ms. In this case, the direct-to-reverberation ratio (DTRR) is about DRR $\approx 2.7$dB.

In general, the spatiotemporal frequency response for DFT bin $\mu$ and angle $\phi$ in the horizontal plane associated with the blocking matrix $H_{\text{BM}}$ is given by

$$H_{\text{BM}}(\mu, \phi) = \frac{\hat{N}(\mu, \phi)}{S(\mu, \phi)} = H_1(\mu, \phi)B_1(\mu) + H_2(\mu, \phi)B_2(\mu),$$  \hspace{1cm} (13)

where $H_p(\mu, \phi)$, $p \in \{1, 2\}$, represents the angular and frequency-dependent frequency response from the source to the $p$-th microphone, and $B_p(\mu)$, $p \in \{1, 2\}$, denote the spectral weights of the blocking matrix. A blocking matrix should separate all noise and interference components from the desired source signal components. To this end, a pronounced spatial null needs to be steered towards the direction of the target source whereas all other directions need to be well preserved. For this analysis this means that a spatial null is forced to $0^\circ$:

$$|H_{\text{BM}}(\mu, 0^\circ)| = 0.$$  \hspace{1cm} (14)

The spatiotemporal frequency response (13) of directional BSS is compared with the response of a simple alternative, a Delay & Subtract beamformer. In the considered case, the coefficients of the Delay & Subtract beamformer are given by $b_1 = 1$, $b_2 = -1$. The magnitude responses of both blocking matrices for the setup shown in Fig. 7 are depicted in Fig. 8. These results show that directional BSS is able to force a pronounced spatial null towards the $0^\circ$ steering direction (see Fig. 8a) even in reverberant conditions as considered here. This demonstrates that directional BSS cannot only suppress the direct path but also reflections of the source signal impinging from other directions. In contrast to directional BSS, the Delay & Subtract beamformer can only suppress the direct path whereas reflections cannot be suppressed and correspondingly no spatial null can be steered towards the steering direction (see Fig. 8b). To this end, directional BSS performs significantly better than a simple Delay & Subtract beamformer serving as blocking matrix, even when only two microphone signals are available.

In the following, the performance of the BSE scheme as discussed in Section IV is analyzed. For this evaluation, three different scenarios are discussed (see Fig. 5):

- Scenario 1: Only interferer I$_1$ is active
- Scenario 2: Only interferer I$_2$ is active
- Scenario 3: Interferers I$_2$ to I$_5$ are active

The overall performance is discussed in terms of the SIR gain and speech distortion as defined in (10) and (12), respectively. These measures are always calculated after convergence of the AEC and the BSE unit. The results for the scenarios defined above are illustrated in Fig. 9 and show that even for the discussed adverse conditions, an SIR gain of at least 7dB can be obtained (SIR$_{\text{gain}} \geq 7$dB). Moreover, for all scenarios, the distortion for the desired signal is very low, i.e., always lower than $-10$dB (SD $< -10$dB). This shows that the proposed two-channel BSE concept leads to a very good performance in terms of noise and interference suppression even in adverse conditions. From that it can be expected that speech recognition results can be significantly improved. Hence, in the following, the proposed two-channel acoustic front-end is applied to a speech recognizer as back-end and the overall performance is analyzed.

C. Speech recognition results

Finally, the performance of the proposed two-channel acoustic human-machine interface is evaluated. Therefore, the output signal obtained from the acoustic front-end ($\hat{s}_1$ in Fig. 1) is applied to
The speech recognizer [14]. For this analysis, a restricted language model is used. The recognizer is trained with a general purpose model based on broadcast speech (default training) and is re-adapted using a scenario-related training set. The training set consists of 200 commands uttered by the actual user in the above-mentioned living-room-like environment and no interfering source or background noise was active. In order to evaluate the overall performance, the command error rate (CER) was calculated which is defined as

$$\text{CER} = \frac{\# \text{correctly recognized commands}}{\# \text{training set}} \times 100\%,$$  

where ‘\# correctly recognized commands’ denotes the number of correctly recognized commands and ‘\# training set’ denotes the total number of available commands (equal to 200). For evaluating the overall performance of the proposed concept, two different scenarios are considered (see Fig. 5):

- **Scenario 1:** Only interferer I is active
- **Scenario 2:** Interferers I2 to I5 are active

In order to evaluate different conditions that might occur in realistic TV control scenarios, the temporal overlap between the spoken command and the interfering signal is varied from 0% to 100%. As in reality the loudspeaker signals are always present, the residual echo signal after the AEC was also always present. This temporal overlap of the individual signal components is illustrated in Fig. 10 for overlaps of 50% and 100% of the desired speech signal and the interfering components. The signals, as the examples shown in Fig. 10, are processed by the BSE concept as discussed in Section IV and then the resulting output signals are used for the speech recognizer. The proposed two-channel BSE concept is compared with a simple alternative, a two-channel Delay & Sum beamformer. The obtained results in terms of the CER (15) are illustrated in Fig. 11. Fig. 11a shows the results obtained for both concepts when only a single interferer (I1) is active and Fig. 11b depicts the results, when the interfering sources I2 to I5 are active. From the results obtained for Scenario 1 (Fig. 11a) it can be seen that only a slight improvement of the proposed BSE concept over a Delay & Sum beamformer can be obtained if the overlap between the spoken command and the interfering signal is lower than or equal to 25%. However, as soon as the Scenario becomes more difficult, i.e., if the temporal overlap increases (overlap > 25%), then the proposed BSE concept clearly outperforms the Delay & Sum beamformer and the CER can be reduced by 10% to 20%. If more difficult conditions are considered as given for Scenario 2 (Fig. 11b), then the recognition results are already significantly reduced for small overlaps (temporal overlap < 25%) over a Delay & Sum beamformer. Besides, no performance degradation is obtained for the proposed concept if no interfering source is active (temporal overlap = 0%) for both scenarios. In fact, the CER is slightly reduced which might be caused by a slight dereverberation effect of the proposed BSE concept. These results show that the proposed two-channel acoustic human-machine interface is well-suited for a ‘natural’ voice dialogue system especially if very adverse conditions are considered.

### VI. Conclusions

In this work, an acoustic human-machine interface for ‘natural’ voice dialogue systems was presented. This concept comprises MC-AEC and a blind signal extraction scheme based on BSS and Wiener filtering strategies. In contrast to previous works (the DICIT project discussed in [11]) where thirteen microphones were foreseen for beamforming, here, only two microphones are used. Experimental results discussed in Section V showed that the proposed scheme can significantly improve the command error rate of a speech recognizer used as back-end over a Delay & Sum beamformer especially for very adverse conditions (low SIR conditions and interfering speech signals). It was also shown that the performance does not degrade when no noise and interference signals are present. Accordingly, the proposed two-channel concept shows a great potential for ‘natural’ voice dialogue systems.

### VII. Acknowledgment

This work was partially funded by the European Commission, Information Society Technologies (IST), FP6 IST-034624, under DICIT.

### References


