JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

A unified beamforming and source separation model for static and dynamic human-robot interaction

Jorge Wuth,[1] Rodrigo Mahu,[1] Israel Cohen,[2] Richard M. Stern,[3] and Néstor Becerra Yoma[1,a)]
[1]Speech Processing and Transmission Laboratory, Department of Electrical Engineering, University of Chile, Av. Tupper 2007, Santiago, Chile
[2]Technion–Israel Institute of Technology, Haifa 3200003, Israel
[3]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA
jwuths@uchile.cl, rmahus@gmail.com, icohen@ee.technion.ac.il, rms@cs.cmu.edu, nbecerra@ing.uchile.cl

Abstract: This paper presents a unified model for combining beamforming and blind source separation (BSS). The validity of the model's assumptions is confirmed by recovering target speech information in noise accurately using Oracle information. Using real static human-robot interaction (HRI) data, the proposed combination of BSS with the minimum-variance distortionless response beamformer provides a greater signal-to-noise ratio (SNR) than previous parallel and cascade systems that combine BSS and beamforming. In the difficult-to-model HRI dynamic environment, the system provides a SNR gain that was 2.8 dB greater than the results obtained with the cascade combination, where the parallel combination is infeasible.

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0025238

Received: 3 November 2023 **Accepted:** 20 February 2024 **Published Online:** 5 March 2024

1. Introduction

Human-robot interaction (HRI), primarily through spoken language, is becoming enormously relevant in all areas where collaborative and social human-robot integration occurs.

1.1 Beamforming and dynamic environments

Speech enhancement is a crucial part of speech communication and speech recognition systems.[1] Unlike single-channel speech enhancement,[2] multichannel speech enhancement can exploit the extra spatial information provided by additional microphones, allowing the separation of the target signal and interference if they come from different directions.[3] The delay-and-sum (D&S) and minimum-variance distortionless response (MVDR) beamformers are two of many techniques that have been proposed. D&S beamforming[4] is accomplished by delaying the microphone signals according to the time difference of arrival (TDOA) to synchronize and add all the channels to enhance the target source, improving the signal-to-noise ratio (SNR). In standard beamforming technology, the TDOA or steering vector is generally estimated from the audio signal, which can be inaccurate in indoor environments due to reverberation. However, social robots are usually equipped with cameras, and the audio sources can also be identified using computer vision.

1.2 Blind source separation

The goal of blind source separation (BSS) is to estimate individual sources given mixtures of those sources.[5] One technique for channel separation is independent component analysis (ICA),[6] which considers the measured signals to correspond to mixtures of different sources. In other words, the observed signals correspond to the sources multiplied by a so-called mixture matrix. ICA seeks to separate the sources by maximizing an objective function that depends on the temporal statistics of the observed mixtures. These techniques have been used to separate speech in noisy and reverberant environments.[7] Nonnegative matrix factorization (NMF)[8] operates on a similar system and seeks to determine the mixture matrix and sources as a decomposition of matrices that are assumed to be nonnegative. These techniques have been used for voice separation in noisy environments with microphone arrays.[9] ICA and NMF require estimating the signal statistics that cannot be obtained reliably in dynamic environments.

1.3 Beamforming and BSS

Combining the beamforming method and source separation techniques is a topic discussed previously in the literature. In Ref. 10, an algorithm was proposed for multiple blind source localizations using source separation and beamforming

a)Author to whom correspondence should be addressed.

techniques. Although the method can locate more sources than the number of microphones in the array, it requires a specific microphone layout. It has only been tested under simulated static conditions. Another approach, presented in Ref. 11, proposes to combine convolutional source separation with geometric beamforming to address ambiguities in source separation and degrees of freedom provided by additional sensors. This method has also not been tested with moving sources or microphones. In Ref. 12, a method was proposed based on the parallel combination of BSS and beamforming. The target source and corresponding direction of arrival (DOA) are first obtained using subband ICA-based BSS, where the average DOA is computed across all the frequency bins. Subsequently, the target source signal is estimated with a null beamforming scheme by employing the average DOA obtained in the previous step. Finally, the target source signal is chosen by either the ICA-based BSS estimate or the null beamforming estimate, depending on the reliability of the frequency-bin-based estimation of the DOA.

SNR and word error rate (WER) results in simulated static conditions using real room impulse responses (RIRs) are reported in Ref. 12. A cascade method that employs a beamforming scheme as a preprocessor of BSS is described in Ref. 13. By assuming that the sound source locations are known, the mixed signals are enhanced with the D&S beamforming scheme; then, the enhanced signals are used as input for Infomax-based BSS. This method considers the beamforming and BSS techniques as independent stages that can be applied one after the other. This implies that the BSS method receives the beamformed outputs without any information on how they were obtained. SNR results in simulated static conditions with reverberation times (referred to as RT or RT60) between 0.3 and 0.7 s using real RIRs are reported in Ref. 13.

Historically, from the signal processing point view, beamforming and BSS have been thought of as representing two different families of techniques. This scenario has changed with deep learning-based beamforming architectures that also perform BSS.[14–18] Those methods are usually treated as "black-boxes" that require training. This paper describes a beamforming-based BSS scheme that can be easily understood and fully integrates the two types of techniques. This method does not require a specific microphone array configuration, applies to scenarios with moving microphones or sources, does not require any prior information about path attenuation or reverberation, and does not require training. In our efforts to integrate beamforming and BSS techniques, we consider a system that develops multiple beams simultaneously, one for each source. If each such output is considered to be a mixture, estimating the acoustic sources could be considered to be a BSS problem. We note that the estimation of TDOA is more feasible in HRI scenarios with computer vision. This paper presents a solution to the problem of BSS by assuming that generated beamformed outputs can be considered mixtures without the need for statistics, such as in ICA or NMF. The resulting solution can be applied frame-by-frame, coping well with dynamic environments. In our approach, we assume that we know the direction of arrival of the sources as well as the physical configuration of the microphones, and we estimate the transfer functions that incorporate microphone frequency response mismatches and reverberation effects to represent a given source recorded in a reference channel given the same source recorded in a different channel. We validate the method by experimental results using real static and dynamic HRI scenarios, comparing our methods to other approaches that integrate beamforming schemes with BSS.

2. BF and source separation

2.1 Formulation of the classical beamforming method

In the short-time Fourier transform (STFT) domain, the weighted D&S beamformer can be expressed as

$$B_l(t,\omega) = \sum_{c=1}^{C} S_c(t,\omega) w_{l,c}(\omega) e^{-i\omega\tau_{l,c}}, \tag{1}$$

where $S_c(t,\omega)$ denotes the observed signal recorded by microphone, c, at frequency, ω, and frame, t; C is the number of microphones; $B_l(t,\omega)$ is the beamformed output pointing to source l; and $w_{l,c}(\omega)$ and $\tau_{l,c}$ correspond to the weight and time delay, respectively, applied to the microphone, c, for steering the beam toward source, l. Specifically, the ordinary D&S beamformer is obtained when $w_{l,c}(\omega) = 1$. The observed signal in microphone c, $S_c(t,\omega)$, corresponds to the summation of the signals received directly from all sources plus reverberation. If the latter is considered negligible, $S_c(t,\omega)$ can be expressed as

$$S_c(t,\omega) = \sum_{j=1}^{J} S_{j,c}(t,\omega), \tag{2}$$

where $S_{j,c}(t,\omega)$ denotes the signal from source signal, j, received by microphone, c, at frequency, ω, and frame, t, and J is the number of sources. The beamformed output can be written as

$$B_l(t,\omega) = \sum_{c=1}^{C} \sum_{j=1}^{J} S_{j,c}(t,\omega) w_{l,c}(\omega) e^{-i\omega\tau_{l,c}}. \tag{3}$$

In Eq. (3), time delays, $\tau_{l,c}$, are given by the angular positions of the sources, which are considered known, and the estimation of weights, $w_{l,c}(\omega)$, and the resulting signal, $B_l(t,\omega)$, depends on the beamforming algorithm considered. It is worth noting that in Eq. (3), given a beamforming method, the only unknown terms are those corresponding to sources $S_{j,c}(t,\omega)$.

2.2 The ideal case

If we can assume that reverberation is negligible, the microphone array-source distance is much greater than the inter-microphone separation, and the microphones have identical responses, each source signal in microphone c, $S_{j,c}$, can be written in terms of the same source signal input to an arbitrary reference microphone, $S_{j,1}$, as follows:

$$S_{j,c}(t,\omega) = e^{-i\omega(\tau_{j,1}-\tau_{j,c})}S_{j,1}(t,\omega). \tag{4}$$

Equation (4) states that source j at microphone c can be considered to be a delayed version of the source received at microphone 1, which is used as a reference. Then, Eq. (3) can be written as

$$B_l(t,\omega) = \sum_c \sum_j S_{j,1}(t,\omega) w_{l,c}(\omega) e^{-i\omega(\tau_{j,1}-\tau_{j,c}+\tau_{l,c})}. \tag{5}$$

By steering the beamformer to each source, a system of linear equations is obtained for each time frame, t, and frequency bin, ω, such that

$$\boldsymbol{b}_{t,\omega} = \boldsymbol{A}_\omega \boldsymbol{s}_{t,\omega}, \tag{6}$$

where

$$\boldsymbol{b}_{t,\omega} = \begin{bmatrix} B_1(t,\omega) \\ \vdots \\ B_J(t,\omega) \end{bmatrix}, \quad \boldsymbol{s}_{t,\omega} = \begin{bmatrix} S_{1,1}(t,\omega) \\ \vdots \\ S_{J,1}(t,\omega) \end{bmatrix}, \quad \boldsymbol{A}_\omega = \begin{bmatrix} a_{11} & \cdots & a_{1J} \\ \vdots & \ddots & \vdots \\ a_{J1} & \cdots & a_{JJ} \end{bmatrix}, \quad \text{and} \quad a_{lj} = \sum_{c=1,\ldots,C} w_{l,c}(\omega) e^{-i\omega(\tau_{j,1}-\tau_{j,c}+\tau_{l,c})}.$$

The sources, $S_{j,1}(t,\omega)$, in microphone 1 can be separated by applying the equation

$$\boldsymbol{s}_{t,\omega} = \boldsymbol{A}_\omega^{-1} \boldsymbol{b}_{t,\omega}, \tag{7}$$

where $\boldsymbol{A}_\omega^{-1}$ was set equal to the zero matrix if $\det(\boldsymbol{A}_\omega)$ was smaller than a given threshold. We note that the matrix $\boldsymbol{A}_\omega$ in Eqs. (6) and (7) depends on the values of the weights, w, used by the beamforming formulation as described in Eq. (1). Consequently, the source separation process and the beamforming algorithm are fully integrated. We also note that the number of speech or noise sources is arbitrary in Eqs. (6) and (7).

2.3 Including corrections for real environments

Equation (4) assumes identical microphones and no reverberation. These assumptions can be relaxed by generalizing our formulation as follows:

$$S_{j,c}(t,\omega) = e^{-i\omega(\tau_{j,1}-\tau_{j,c})}S_{j,1}(t,\omega)H^1_{j,c}(t,\omega), \tag{8}$$

where $H^1_{j,c}(t,\omega)$ is a transfer function that represents the transformation from signal $S_{j,1}(t,\omega)$ to signal $S_{j,c}(t,\omega)$. The transfer function, $H^1_{j,c}(t,\omega)$, presumably incorporates the effects of microphone mismatches, reverberation effects, and other attributes of the room acoustics to represent source, j, recorded in microphone, c, given the same source, j, recorded in microphone 1. The ideal scenario described above is obtained by assuming that $H^1_{j,c}(t,\omega) = 1$ for all t and ω. The beamforming in real conditions can be written as

$$B_l(t,\omega) = \sum_c \sum_j S_{j,1}(t,\omega) w_{l,c}(\omega) H^1_{j,c}(t,\omega) e^{-i\omega(\tau_{j,1}-\tau_{j,c}+\tau_{l,c})}, \tag{9}$$

and sources $S_{j,1}(t,\omega)$ in microphone 1 can now be separated using Eq. (7) and expressing a_{lj} in $\boldsymbol{A}_\omega$ as

$$a_{lj} = \sum_{c=1,\ldots,C} w_{l,c}(\omega) H^1_{j,c}(\omega) e^{-i\omega(\tau_{j,1}-\tau_{j,c}+\tau_{l,c})}. \tag{10}$$

If the source positions are known, the problem of source separation is reduced to the estimation of $H^1_{j,c}(\omega)$ in Eqs. (9) and (10) such that

$$H^1_{j,c}(t,\omega) = \frac{S_{j,c}(t,\omega)}{e^{-i\omega(\tau_{j,1}-\tau_{j,c})}S_{j,1}(t,\omega)}. \tag{11}$$

2.4 Estimation for the real case correction

Let $\hat{S}_{j,c}(t,\omega)$ represent an estimate of $S_{j,c}(t,\omega)$. An estimate of $H^1_{j,c}(t,\omega)$, $\hat{H}^1_{j,c}(t,\omega)$, can be obtained as

$$\hat{H}^1_{j,c}(t,\omega) = \frac{\hat{S}_{j,c}(t,\omega)}{e^{-i\omega(\tau_{j,1}-\tau_{j,c})}\hat{S}_{j,1}(t,\omega)}. \tag{12}$$

Using Eq. (2), an estimate for the l th source recorded in microphone c, $\hat{S}_{l,c}(t,\omega)$, can be given as

$$\hat{S}_{l,c}(t,\omega) = S_c(t,\omega) - \sum_{\substack{j=1 \\ j\neq l}}^{J} S_{j,c}(t,\omega). \tag{13}$$

The beamformed output, $B_j(t,\omega)$, pointing to the jth source can be used as an estimate of the source j in microphone 1, $S_{j,1}(t,\omega)$. The jth source in microphone c, $S_{j,c}(t,\omega)$, could be approximated as

$$\hat{S}_{j,c}(t,\omega) = B_j(t,\omega)\, e^{-i\omega(\tau_{j,1}-\tau_{j,c})}. \tag{14}$$

Then, an estimate for $\hat{S}_{l,c}(t,\omega)$ can be obtained as

$$\hat{S}_{l,c}(t,\omega) = S_c(t,\omega) - \sum_{j\neq l} \lambda_{j,c}\, B_j(t,\omega) e^{-i\omega(\tau_{l,1}-\tau_{l,c})}, \tag{15}$$

where $\lambda_{j,c}$ is a coefficient that needs to be tuned or estimated. For the two-source case, Eq. (15) can be written as

$$\hat{S}_{l,c}(t,\omega) = S_c(t,\omega) - \lambda_{j,c}\, B_j(t,\omega) e^{-i\omega(\tau_{l,1}-\tau_{l,c})}, \tag{16}$$

where $l=1$, $j=2$ or $l=2$, $j=1$; the weight, $\lambda_{j,c}$, was set equal to a constant for all l and j. For simplicity, $B_j(t,\omega)$ in Eq. (14) corresponds to the D&S beam.

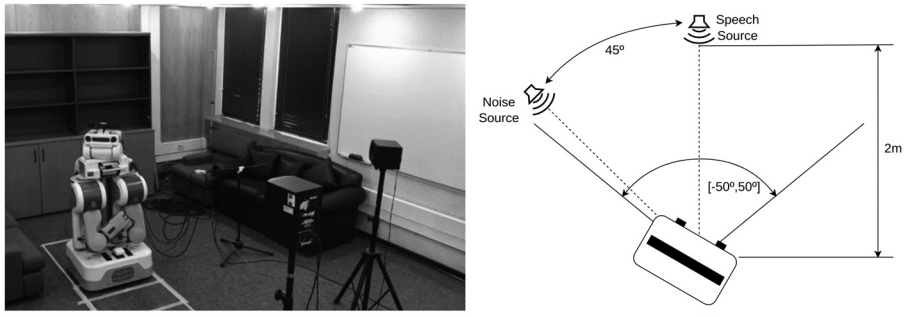
3. Experiments

3.1 Database

The experiments were performed using the 330 test utterances from the Aurora-4 database,[19] which were re-recorded in Ref. 20 using the Personal Robot 2 (PR2), with a Microsoft Xbox 360 Kinect mounted on its head. The setup consists of a speech source 2 mfrom the microphone array and a noise source 2 m from the array and 45° to the left of the speech source [see Fig. 1(b)]. In the static condition, the microphone array points to the speech source (0°). In the dynamic scenario, which simulates a relative robot-source movement, the robot head and microphone array rotate between 50° and –50° at a speed of 24.1°/s. Speech and noise were also recorded in different sessions to generate a simulated version of the static condition. Figure 1(a) shows the PR2 robot in the experimental setup used to record the databases. The estimated RT60 in the room was 0.5 s.

3.2 Procedures

The proposed scheme was tested using D&S (Ref. 21) and MVDR.[22] In both cases, the angle of the robot's head was employed to determine the TDOAs on a frame-by-frame basis applied in the frequency domain. The implementation of MVDR considered a free-field model. The noise covariance matrices were computed using the nonspeech segments determined by a voice activity detector (VAD).[23] The noise sources correspond to the mechanics of the robot or ambient noise that is generated by the loudspeaker. Both types of noise are added in the microphone signals for the purpose of speech/



(a) (b)

Fig. 1. (a) Experimental setup for database recording with PR2 robot and (b) setup diagram for the HRI database recording showing the rotation of the robot head are displayed.

nonspeech detection. During each nonspeech frame, a noise matrix was calculated. The noise matrix was determined during the speech intervals by interpolating between the previous and following nonspeech noise matrices. The breaks between spoken words and the breaks between pauses in speech are, on average, 253and 3.291 ms, respectively.

The proposed BSS-based enhancement procedure was performed in the frequency domain using Eq. (7). Each microphone time-domain signal was segmented into 25-ms time frames with a 15-ms overlap. A fast Fourier transform (FFT) was applied to compute the complex spectrum for each time frame. The sampling frequency was 16 kHz, and a 512-sample FFT was employed, resulting in 257 frequency bins. D&S and MVDR were used to generate $B_i(t,\omega)$ in Eq. (5). $\lambda_{j,c}$ in Eq. (16) was set equal to 0.5. The resulting enhanced speech source in the time domain was obtained by using the overlap-add procedure with $S_{1,1}(t,\omega)$ obtained using Eq. (7). The following metrics were employed to compare the methods tested here with simulated conditions (Table 1): output SNR, perceptual evaluation of speech quality (PESQ),[24] and short-time objective intelligibility (STOI).[25] However, only the SNR metric was employed with the real static and dynamic datasets because the reference signals needed to evaluate PESQ and STOI were unavailable. Table 1 also shows WER results obtained with a DNN-HMM automatic speech recognition (ASR) system run in Kaldi and trained as in Ref. 20 with the Aurora-4 database. The training audios were generated by convolving the speech and noise signals with RIRs measured in the same room where the tests were recorded. Then, the speech and noise signals were added with SNR between 0 and 10 dB and processed using the same techniques as employed for the test utterances according to Table 1 to obtain matched training-testing conditions.

Results obtained using three different types of models are compared: (1) a simulated "anechoic" environment in which it is assumed that $H^1_{j,c}(t,\omega) = 1$;(2) an "Oracle H" environment in which the room transfer function, $H^1_{j,c}(t,\omega)$, was calculated with Eq. (11) using speech samples captured in the absence of noise; and (3) the "Hest" environment in which the transfer function, $H^1_{j,c}(t,\omega)$, is estimated in realistic conditions using Eq. (12). Results in Tables 1 and 2 obtained using these three models of the environment are designated D&S-BSS-1 and MVDR-BSS-1, D&S-BSS-Oracle, and MVDR-BSS-Oracle, and D&S-BSS-Hest and MVDR-BSS-Hest.

We also provide performance comparisons to two standard strategies of combining beamforming techniques with BSS in parallel[12] and cascade[13] configurations, which are discussed in Sec. 1.3. The approaches are referred to in Tables 1 and 2 as ICA//Null BF or ICA//MVDR and D&S → INFOMAX and MVDR → INFOMAX, respectively.

The methods published in Refs. 12 and 13 were chosen for comparison purposes because they present two standard strategies to combine beamforming techniques and BSS in parallel and cascade configurations, respectively. Additionally, they do not require a special microphone array, and they consider the number of sources to be known. Because the parallel scheme in Ref. 12 requires the number of sources to be equal to the number of microphones, two of the four Kinect channels were employed for BSS and beamforming. The two selected channels were those at the left and right ends of the microphone array. Also, in addition to the parallel combination of ICA and null beamforming proposed in Ref. 12, MVDR was combined with ICA, leading to two different configurations of the same strategy, i.e., ICA//Null BF and ICA//MVDR, respectively. The null beamforming method was replaced by MVDR because the latter provides a SNR gain that is significantly larger than the former. The source DOAs are considered to be known in both cases. The magnification parameter of the threshold defined in Ref. 12 was tuned by maximizing the average SNR of the resulting enhanced speech signals. It is worth observing that ICA//Null BF or ICA//MVDR require a single DOA per frequency bin, therefore, they do not apply to the dynamic condition where the robot head rotates. Regarding the cascade method described in Ref. 13, two frequency-domain beamforming techniques, D&S and MVDR, were evaluated and applied using the four Kinect channels. Two beamformed signals were generated by pointing to both sources and input to the Infomax-based BSS system on a bin-by-bin basis. Then, permutation and scaling were corrected as in Ref. 13. Two cascade configurations were evaluated depending on the beamforming scheme employed before Infomax: D&S → INFOMAX and MVDR → INFOMAX, respectively.

Table 1. Average SNR, PESQ, and STOI with simulated static HRI condition.

Enhancement	SNR	WER	PESQ	STOI
Clean speech at channel 1	27.9 dB	5.16%	4.50	1.00
Noisy speech at channel 1	7.6 dB	41.27%	2.31	0.67
Standard D&S	9.1 dB	19.93%	2.37	0.70
D&S-BSS-Oracle	28.0 dB	4.71%	4.27	0.97
D&S-BSS-1	10.8 dB	15.73%	2.50	0.74
D&S-BSS-Hest	12.5 dB	16.46%	2.41	0.71
Standard MVDR	13.3 dB	12.78%	2.47	0.69
MVDR-BSS-Oracle	27.9 dB	5.19%	4.18	0.98
MVDR-BSS-1	13.1 dB	12.65%	2.50	0.68
MVDR-BSS-Hest	15.9 dB	11.69%	2.42	0.70

Table 2. Average SNR with real static and dynamic HRI conditions.

Enhancement	Static	Dynamic
Noisy speech at channel 1	7.3 dB	6.9 dB
Standard D&S	9.2 dB	8.2 dB
Standard MVDR	12.5 dB	10.6 dB
INFOMAX	9.7 dB	7.8 dB
ICA	9.6 dB	7.8 dB
D&S-BSS-Hest	11.8 dB	10.2 dB
MVDR-BSS-Hest	14.3 dB	12.5 dB
ICA//Null BF	8.7 dB	N/A
ICA//MVDR	11.7 dB	N/A
D&S → INFOMAX	9.4 dB	9.5 dB
MVDR → INFOMAX	11.9 dB	9.7 dB

3.3 Results and discussion

Table 1 shows results obtained using simulated static scenarios in which the speech and noise sources were recorded separately and combined offline. Unsurprisingly, the best performance in terms of SNR is obtained for clean speech, and the worst performance is observed in response to uncompensated noisy speech. D&S and MVDR beamforming methods provide improvement in all the metrics. With Oracle processing, {D&S,MVDR}-BSS-Oracle, the speech and noise waveforms were recorded separately, and the transfer functions were estimated directly using perfect knowledge of these waveforms. With {D&S,MVDR}-BSS-1, the imagined room was assumed to be anechoic and the microphones were considered identical, leading to diagonal spatial covariance matrices. The {D&S,MVDR}-BSS-Hest data represent real results without any unrealistic assumptions based on Eqs. (12)–(16). We note, first, that the D&S-BSS-Oracle and MVDR-BSS-Oracle methods recovered the original SNR observed with clean speech, which validates our overall model of the environment and confirms the existence of an optimal $H^1_{j,c}(t,\omega)$ as defined in Eq. (8). In general, our proposed combination algorithms, D&S-BSS-Hest and MVDR-BSS-Hest, provide substantial improvements in SNR compared to beamforming without BSS. These results are confirmed in the spectrograms of Fig. 2, which depict clean speech, speech with broadband noise limited to 4 kHz, and results obtained using some of the compensation schemes described in this paper. It can be observed that the results obtained using MVDR-BSS-Hest provide the best noise elimination. Nevertheless, the use of D&S-BSS-Hest provided small improvements in the PESQ and STOI metrics, whereas no improvement at all was observed through the use of MVDR-BSS-Hest compared to MVDR alone. These results may be a consequence of possible speech-distortion artifacts

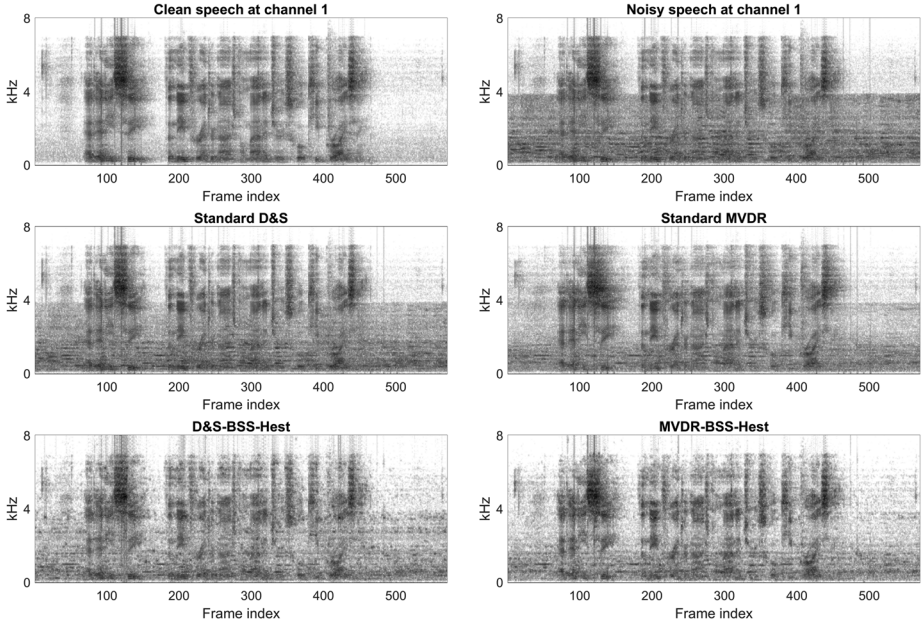


Fig. 2. Spectrograms in static HRI condition as in Table 1.

introduced by the {D&S,MVDR}-BSS-Hest processing. In any case, the impact of the speech-distortion artifact in the ASR results appears to be small.

Table 2 summarizes our results obtained using data that were recorded in real static and dynamic scenarios, as described in Sec. 3.1. As expected, the results indicate consistently that the dynamic scenarios are more challenging to the system than the static scenarios. We also note that for every possible comparison of results, the method that we propose for combining BSS and beamforming techniques is more effective than either the BSS or beamforming techniques used in isolation or the traditional cascade and parallel methods to which we compared our data. This was true using the MVDR and D&S beamformers. We observe that MVDR beamforming is more effective than D&S beamforming for every comparable condition considered. We also note that the parallel integration of ICA with beamforming schemes, as described in Ref. 12, requires defining a single DOA per each frequency bin, thus, it cannot be applied to the dynamic HRI scenario where the robot head rotates. Compared to D&S or MVDR beamforming alone, parallel and cascade combination methods lead to little or no improvements in SNR in real HRI conditions. These results suggest that the real scenarios addressed here are challenging to model. It is also worth highlighting that no assumption about path attenuation or reverberation is needed to achieve the performance that we observed.

4. Conclusions

In this paper, we have presented a unified model that encompasses beamforming and BSS in experimental laboratory conditions that are very similar to natural acoustical environments. The scheme is based on describing each beamformed output as an explicit combination of the audio sources and representing the output of each microphone as a function of a reference microphone. Expressing each microphone output as a delayed version of the reference microphone signal multiplied by a relative transfer function allows us to apply the traditional matrix inversion solution to undo the effects of the mixing. This transfer function models the interchannel mismatch produced by the array's microphones and reverberation. It can be approximated with the difference between the mixed sources in the channel and a linear combination of the delayed versions of beamformed outputs pointing to the other sources. Oracle experiments show that the target speech sources can be accurately recovered, which validates our assumptions and model. Experiments with real static and dynamic HRI datasets recorded with a PR2 robot suggest that the proposed BSS scheme with MVDR can lead to a substantial increase in SNR when compared to standard MVDR only and provide a performance that exceeds that of previous parallel and cascade combinations of BSS and beamforming schemes. The proposed approach with MVDR provided a SNR that was 2.6 and 2.4 dB greater than the parallel and cascade systems with MVDR, respectively, with real static HRI data. Compared to the cascade architecture, a SNR gain equal to 2.8 dB was observed for the real dynamic HRI condition. The parallel integration scheme is not applicable to the dynamic scenario where the robot head rotates. Moreover, the results of ASR experiments suggest that the impact of the speech-distortion artifact is small. The objective of future research will be to develop more precise estimates of the inter-microphone relative transfer functions.

Acknowledgments

The research reported here was funded by Agencia Nacional de Investigación y Desarrollo (ANID), Chile, under Grant Fondecyt No. 1211946.

Author Declarations

Conflict of Interest

The authors have no conflicts to disclose.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1]Q. Hu, Z. Hou, K. Chen, and J. Lu, "Learnable spectral dimension compression mapping for full-band speech enhancement," JASA Express Lett. **3**(2), 025204 (2023).
[2]T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," JASA Express Lett. **2**(10), 104802 (2022).
[3]D. Lee, B. Jo, and J.-W. Choi, "Direction-of-arrival estimation with blind surface impedance compensation for spherical microphone array," JASA Express Lett. **1**(7), 074801 (2021).
[4]B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE ASSP Mag. **5**(2), 4–24 (1988).
[5]Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004).
[6]P. Comon, "Independent component analysis, a new concept?," Signal Process. **36**(3), 287–314 (1094).
[7]A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," IEEE Trans. Neural Networks **13**(4), 888–893 (2002).
[8]M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," Comput. Stat. Data Anal. **52**(1), 155–173 (2007).

[9]H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," IEEE Trans. Audio. Speech Lang. Process. **14**(2), 666–678 (2006).
[10]H. Pu, C. Cai, M. Hu, T. Deng, R. Zheng, and J. Luo, "Towards robust multiple blind source localization using source separation and beamforming," Sensors **21**(2), 532 (2021).
[11]L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," IEEE Trans. Speech Audio Process. **10**(6), 352–361 (2002).
[12]H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," EURASIP J. Adv. Signal Process. **2003**(11), 569270.
[13]L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," EURASIP J. Audio, Speech, Music Process. **2010**(1), 1–13.
[14]J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "NICE-beam: Neural integrated covariance estimators for time-varying beamformers," arXiv:2112.04613 (2021).
[15]Z. Zhang, Y. Xu, M. Yu, S. X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada (June 6–11, 2021) (IEEE, New York, 2021), pp. 6089–6093.
[16]X. Li, Y. Xu, M. Yu, S.-X. Zhang, J. Xu, B. Xu, and D. Yu, "MIMO Self-Attentive RNN Beamformer for Multi-Speaker Speech Separation," in *Proceedings of the Interspeech 2021*, Brno, Czech Republic (August 30–September 1, 2021 (ISCA, Baixas, France, 2021), pp. 1119–1123.
[17]Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal RNN beamformer for target speech separation," Proc. Interspeech **2021**, 3076–3080.
[18]T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," IEEE/ACM Trans. Audio. Speech. Lang. Process. **31**, 835–845 (2023).
[19]G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," European Telecommunications Standards Institute (ETSI) Speech Transmission Quality (STQ) Aurora Digital Speech Recognition Working Group (2002).
[20]J. Novoa, R. Mahu, J. Wuth, J. P. Escudero, J. Fredes, and N. B. Yoma, "Automatic speech recognition for indoor HRI scenarios," ACM Trans. Hum-Robot. Interact. **10**(2), 1–30 (2021).
[21]M. Omologo, M. Matassoni, and P. Svaizer, "Speech recognition with microphone arrays," in *Microphone Arrays, Signal Processing Techniques and Applications*, edited by M. Brandstein and D. Ward (Springer, Berlin, 2001), pp. 331–353.
[22]J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing* (Wiley, New York, 2017).
[23]Team Silero, "Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier" (2021), available at https://github.com/snakers4/silero-vad (Last viewed October 27, 2022).
[24]ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs" (International Telecommunication Union, Geneva, Switzerland, 2001).
[25]C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio. Speech. Lang. Process. **19**(7), 2125–2136 (2011).