IET Signal Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Block-online multi-channel speech enhancement using deep neural network-supported relative transfer function estimates

ISSN 1751-9675
Received on 26th June 2019
Revised 30th September 2019
Accepted on 11th December 2019
E-First on 30th January 2020
doi: 10.1049/iet-spr.2019.0304
www.ietdl.org

Jiri Malek[1] ✉, *Zbyněk Koldovský*[1], *Marek Bohac*[1]

[1]*Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, Liberec, Czech Republic*
✉ *E-mail: jiri.malek@tul.cz*

Abstract: This work addresses the problem of block-online processing for multi-channel speech enhancement. Such processing is vital in scenarios with moving speakers and/or when short utterances are processed, e.g. in voice assistant applications. We consider several variants of a system that performs beamforming supported by deep neural network-based voice activity detection followed by post-filtering. The speaker is targeted through estimating relative transfer functions between microphones. Each block of the input signals is processed independently to make the method applicable in highly dynamic environments. Due to short processed blocks, the statistics required by the beamformer are estimated less precisely. The influence of this inaccuracy is studied and compared to batch processing regime, when recordings are treated as one block. The experimental evaluation is performed on large datasets of CHiME-4 and another dataset featuring moving target speaker. The experiments are evaluated in terms of objective and perceptual criteria. Moreover, word error rate (WER) of a speech recognition system is evaluated, for which the method serves as a front-end. The results indicate that the proposed method is robust for short length of the processed block. Significant improvements in terms of the criteria and WER are observed even for the block length of 250 ms.

1 Introduction

Speech enhancement in real-world conditions is an open audio signal processing problem with many applications ranging from distant communication and automatic speech recognition (ASR) to hearing aids. The enhancement aims at the suppression of distortions present in target speech recordings such as background noise, reverberation, or cross-talk. It is a challenging problem as real-world environments involve various situations, rooms, interferences, transient sounds, and non-stationary noises. In particular, it is difficult to develop systems that can efficiently operate in general conditions with low processing delays [1].

One of the basic distinctions among enhancement techniques is the number of input signals they use. Single-channel techniques rely only on spectral filtering, whereas multi-channel methods also exploit the spatial information. Single-channel enhancement is constituted by classical techniques such as spectral subtraction [2, 3] or minimum mean square estimator [4], as well as modern techniques based either on statistical principles [5, 6] or supported by machine learning (e.g. [7] and overview in [8]). Owing to more information available, multi-channel filters usually achieve better enhancement compared with single-channel ones [9]. Moreover, spatial information is important to determine the target speaker, which can be done, e.g. based on localisation. Multi-channel approaches can be based on classical array processors, commonly referred to as 'beamformers' or on blind source separation (BSS). BSS techniques try to separate individual source signals that are simultaneously active and mixed [10]. Beamformers exploit known parameters of the signal mixture, which, however, must also be estimated when only mixed signals are observed. For example, under the free-field propagation model, the direction of arrival and array geometry are used to select the optimum filter coefficients. In reverberant environments, acoustic transfer functions and/or signals' spatial covariance matrices must be estimated to compute an optimum beamformer [1].

A conventional technique is the delay-and-sum beamformer (DSB) [11], which is based on the free-field assumption and performs the enhancement through summing signals from delay-compensated channels. More advanced beamformers take into account the reverberation. To name the most popular approaches: the speech distortion weighted multi-channel Wiener filter (SDW-MWF) minimises the square distance between filtered inputs and the unknown target [12]. It involves a free parameter, which provides a trade-off between noise suppression and speech distortion. A particular choice of this parameter leads to the classical minimum mean square error (MMSE) beamformer [13], which is sometimes referred to as the MWF [9]. For another choice of the parameter, the minimum variance distortionless response (MVDR) beamformer is obtained [14]. The maximum signal-to-noise ratio (MSNR) beamformer attempts to maximise the SNR in the output signal [15]. Its application in the enhancement of speech was presented in [16], where it was referred to as generalised eigenvalue beamformer (GEV). For a more comprehensive overview of the beamforming techniques, see [9].

The beamformers rely on robust estimation of their inner parameters such as covariance matrices of speech/noise or relative transfer functions (RTFs). RTF is a transfer function between two channels in response to the target source, as defined in [17]. Since the speech/noise covariance values are not known in noisy situations, the methods usually exploit voice activity detectors to localise signal frames or bins in the time–frequency domain, where only noise or speech is active. The desired statistics are then estimated using these frames/bins.

Parametric voice activity detection (VAD) methods rely on statistical models for speech/noisy signals [18]. Often, the drawback resides in their limited ability to model highly non-stationary noise and transient interferences [19]. To overcome this drawback, several approaches have been proposed. One is to focus solely on harmonic properties of voiced speech instead of assuming any specific properties of diverse noise [20]. Another approach is to exploit machine-learning principles [21], especially deep neural networks (DNNs) [22].

Concerning the RTF estimation, various approaches exist in the literature. The covariance subtraction [23] and covariance whitening [24] methods require the noise-speech and noise covariance matrices. The method from [25] alleviates this requirement by assuming that the noise is stationary and that the noise statistics can be estimated along with the RTF. The RTF can also be estimated using BSS; an analysis of the techniques based

on independent component analysis can be found, e.g. in [26, 27]; see also [28–30] for approaches exploiting sparsity.

Multi-channel enhancement techniques can be used as pre-processors for systems performing robust ASR as described in overviews [31, 32].

1.1 Community-based campaign: CHiME

Although the theoretical properties of the enhancement techniques discussed in Section 1 are well known, their application in real scenarios is difficult due to the need to robustly estimate their inner parameters. This motivates the researchers working in the fields of speech enhancement and robust speech recognition to organise evaluation campaigns, where the most recent technologies are compared in practical tasks. The campaigns thus reveal the current state-of-the-art methods. The most recognised is the CHiME speech separation and recognition challenge (CHiME). This work focuses on the data from the CHiME-4 [33] challenge, i.e. real-world multi-channel recordings of a single speaker originating in four distinct noisy environments.

The solutions proposed for CHiME-4 are mainly focused on the robust speech recognition of multi-channel data since the evaluation proceeds in terms of word error rate (WER). Nevertheless, the front-end processing before the feature computation is one of the key components, because it can improve the final WER by as much as several per cents. Most of CHiME-4 front-end processors combine beamforming along with information obtained through machine learning. For example, a Gaussian-mixture-model-based VAD is utilised in [34] to assist the estimation of speech/noise covariance matrix for MVDR. For a similar purpose, Heymann *et al.* [35] utilise a DNN-based VAD along with MVDR and GEV beamformers; see also [36]. In [37], the enhancement is performed in two passes. First, MVDR is used to obtain enhanced speech, which is then forwarded to an external speech recogniser. In the second pass, a fine-tuned beamforming is performed using a VAD that exploits signal segmentation provided by the speech recogniser.

An important aspect of CHiME-4 is that though the acoustic conditions are varying across different recordings, the position of the speaker is rather static within the single recording. Many state-of-the-art methods proposed for CHiME-4 take advantage of this fact and use the whole recording as one batch of data to estimate the necessary speech/noise spatial covariance matrices as precise as possible. Some methods perform several passes [37, 38] to optimise the enhancement of the given recording. This approach is not suitable for recordings that are more dynamic. For example, when the target speaker is moving, short blocks of the signals need to be processed, to update the necessary statistics for the changing position continuously. The design of a technique intended for short data segments and its comparison with techniques used to solve CHiME-4 in the batch manner are the objectives of this work.

Very recently, CHiME-5 challenge [39] was released, featuring a dinner party of four speakers in a domestic scenario. This dataset contains recordings with more speaker movement. However, the experiments in this work are focused on the CHiME-4 data rather than CHiME-5. The recordings of the latter dataset contain a high amount of cross-talk (up to four speakers are active simultaneously), which is beyond the scope of the current research.

1.2 Contribution

We propose an online multi-channel enhancement system that processes data block-by-block. To maximise the adaptation speed of the system, the estimation of inner statistics is *not* recursive, i.e. it does not utilise estimates from previous blocks. Such processing is vital in scenarios with moving speakers and/or when very short utterances or mere keywords are processed. The proposed system is implemented in two variants, respectively, with the approximate minimum mean-squared error beamformer [40] and with a robust inverse RTF (IRTF) beamformer.

The beamformer steering is based on an explicit estimation of RTFs, as compared with the recently published supervised techniques that mainly rely on the estimation of covariance matrices. For example, some methods (e.g. [35, 36, 41–43]) estimate the steering vectors as principal components of speech covariance matrices. The methods in [44, 45] rely on alternative MVDR computations that solely depend on the covariance matrices of noise and of noisy data. In our experiments, the methods based on eigenvalue decomposition of covariance matrices are represented by GEV from [35]. Our experiments indicate that the performance of GEV deteriorates when short processing blocks are analysed.

In contrast, the proposed method, based on direct RTF estimation, yields stable performance. The RTF estimator used in this work is endowed with a DNN-based VAD, which improves the estimation accuracy and lowers the WER of the ASR back-end. Nevertheless, the VAD inclusion is optional and may be excluded, if, e.g. computational demands due to the VAD are of concern. Recently, a similar solution using the RTF estimation for steering the MVDR beamformer have appeared in [46]. However, this paper focused on the batch processing and the RTF estimator utilised in work cannot be used without VAD.

The system's capabilities are validated on the CHiME-4 datasets and a multi-channel dataset featuring moving speaker. Our study is focused on how enhancement performance depends on the length of the processing blocks. The results for CHiME-4 are evaluated in terms of WER achieved by the baseline CHiME-4 speech recognition system. Also, the speech enhancement capabilities of the systems are quantified using objective criteria reflecting the output signal quality.

Recently, researchers started investigating the (block)-online behaviour of DNN-endowed beamformers. The works in [47, 48, 49, 50] present block-online/frame-by-frame VAD-based MVDR beamformers, which are all based on covariance matrix operations. In contrast to the current work, these papers neither analyse the dependence of the input length on the performance, nor the perceptual quality of the output.

This paper is organised as follows. In Section 2, the multi-channel signal enhancement problem is formulated, and basic beamforming techniques are introduced. Section 3 is devoted to a detailed description of the proposed multi-channel enhancement system. In Section 4, the results of extensive tests and comparisons are presented. Section 5 provides conclusions.

2 Problem description

2.1 Model

Considering the target speaker as a directional source, a noisy recording can be described in the short-time Fourier domain as

$$\boldsymbol{x}(k, \ell) = \underbrace{\boldsymbol{g}(k, \ell)s(k, \ell)}_{\boldsymbol{s}(k, \ell)} + \boldsymbol{y}(k, \ell), \tag{1}$$

where $\boldsymbol{x}(k, \ell)$ is the $M \times 1$ vector of noisy microphone signals, M is the number of microphones, and $s(k, \ell)$ is the target speech as observed on a reference microphone. Next, $\boldsymbol{y}(k, \ell)$ is a vector involving all noise components, which are assumed to be uncorrelated with $s(k, \ell)$; k represents the frequency index; and ℓ stands for the time frame index.

The elements of $\boldsymbol{g}(k, \ell)$ contain the RTFs of all microphones with respect to the reference one as in [17]. The speaker can move during the utterances; therefore, the RTFs can vary in time. It is assumed that the changes are slow, that is, $\boldsymbol{g}(k, \ell)$ remains approximately constant during each block of frames. The spatial image of the target speech as measured on the microphones is given by $\boldsymbol{s}(k, \ell) = \boldsymbol{g}(k, \ell)s(k, \ell)$. In the following text, the indexes k and ℓ are omitted for the sake of brevity, where no confusion arises.

2.2 Multi-channel processing

This section describes the basic forms of the beamforming techniques discussed in this work. Let the output of a beamformer be denoted as

$$u(k, \ell) \triangleq \boldsymbol{w}(k, \ell)^{\mathrm{H}}\boldsymbol{x}(k, \ell), \tag{2}$$

where $\boldsymbol{w}$ is the steering vector of the beamformer, which takes the following forms.

The MVDR beamformer is a popular multi-channel processor for noise suppression but it may also be used for dereverberation, as suggested in [51]. It is defined as the constrained minimiser

$$\boldsymbol{w}_{\mathrm{MVDR}} = \arg\min_{\boldsymbol{w}} \left\{ \boldsymbol{w}^{\mathrm{H}} \boldsymbol{C}_{yy} \boldsymbol{w} \ \mathrm{s.t.} \ \boldsymbol{w}^{\mathrm{H}} \boldsymbol{g} = 1 \right\}, \tag{3}$$

which gives

$$\boldsymbol{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{C}_{yy}^{-1} \boldsymbol{g}}{\boldsymbol{g}^{\mathrm{H}} \boldsymbol{C}_{yy}^{-1} \boldsymbol{g}}, \tag{4}$$

where $\boldsymbol{C}_{yy} = E\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}\right]$ denotes the covariance matrix of noise $\boldsymbol{y}$; $E[\cdot]$ stands for the expectation operator, and $(\cdot)^H$ denotes the conjugate transpose.

The MMSE beamformer is defined through

$$\boldsymbol{w}_{\mathrm{MMSE}} = \arg\min_{\boldsymbol{w}} \ E\left\{ \left| \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x} - s \right|^2 \right\}. \tag{5}$$

MMSE can be implemented as a cascade of MVDR and a single-channel Wiener filter [9]. The single-channel filter suppresses the residual noise $r_{\mathrm{MVDR}} = \boldsymbol{w}_{\mathrm{MVDR}}^{\mathrm{H}} \boldsymbol{y}$ in the beamforming output u. Specifically

$$\boldsymbol{w}_{\mathrm{MMSE}} = \boldsymbol{w}_{\mathrm{MVDR}} \underbrace{\frac{E[|u|^2] - E[|r_{\mathrm{MVDR}}|^2]}{E[|u|^2]}}_{\mathrm{Wiener\ post-filter}}. \tag{6}$$

The IRTF beamformer, as defined in this work, utilises the RTFs represented by $\boldsymbol{g}(k, \ell)$ to synchronise the spatial images of the target source on all microphones with the reference one, and sums them together. Therefore

$$\boldsymbol{w}_{\mathrm{IRTF}} = \frac{1}{m}\left(\boldsymbol{g}^{-1}\right), \tag{7}$$

where $\boldsymbol{g}^{-1}$ contains the reciprocal values of the elements of $\boldsymbol{g}$. Compared to MVDR, processing through IRTF alleviates the need to estimate $\boldsymbol{C}_{yy}$. This renders the beamforming operation robust in respect to estimation errors. The IRTF could be seen as a generalisation of the delay-and-sum beamforming for reverberant environments, because the DSB and IRTF coincide in free-field conditions.

The IRTF can also be followed by a single-channel post-filter such as the Wiener filter described by (6). The residual noise estimate is obtained $r_{\mathrm{IRTF}} = \boldsymbol{w}_{\mathrm{IRTF}}^{\mathrm{H}} \boldsymbol{y}$.

The GEV beamformer is designed to maximise the SNR in each frequency bin using

$$\boldsymbol{w}_{\mathrm{GEV}} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{C}_{ss} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{C}_{yy} \boldsymbol{w}}, \tag{8}$$

where $\boldsymbol{C}_{ss} = E\left[\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\right]$ is the covariance matrix of target speech. This optimisation problem leads to finding the generalised eigenvector [36]

$$\boldsymbol{C}_{ss} \boldsymbol{w}_{\mathrm{GEV}} = \lambda \boldsymbol{C}_{yy} \boldsymbol{w}_{\mathrm{GEV}}, \tag{9}$$

where λ is the maximal generalised eigenvalue. Since the norm of $\boldsymbol{w}_{\mathrm{GEV}}$ can be arbitrary, the output of the GEV beamformer has an ambiguous spectrum. This problem can be solved by applying a single-channel post-filter

$$G_{\mathrm{BAN}} = \frac{\sqrt{\boldsymbol{w}_{\mathrm{GEV}}^{\mathrm{H}} \boldsymbol{C}_{yy} \boldsymbol{C}_{yy}^{\mathrm{H}} \boldsymbol{w}_{\mathrm{GEV}} / M}}{\boldsymbol{w}_{\mathrm{GEV}}^{\mathrm{H}} \boldsymbol{C}_{yy} \boldsymbol{w}_{\mathrm{GEV}}}, \tag{10}$$

which is called blind analytic normalisation (BAN). In theory, the cascade of the GEV beamformer and the BAN post-filter results in the MVDR beamforming; see, e.g. [16].

3 Proposed method

The proposed enhancement algorithm, visualised in Fig. 1, consists of the following operations that are applied independently to each block of input signals:

(i) Detection of microphone failures.
(ii) Transformation into a time–frequency domain using short-time Fourier transform (STFT).
(iii) Estimation of time–frequency masks based on voice activity as detected by pre-trained DNN.
(iv) Estimation of RTFs.
(v) Application of selected beamforming technique.
(vi) Post-filtering applied to the output of the beamformer.
(vii) Synthesis of the speech waveform using the inverse discrete Fourier transform and overlap-add.

Two processing regimes are considered: (i) in block-online mode, the signals are processed block-by-block, where the length of each block is fixed, whereas (ii) in batch mode, the entire input recording is processed as a single block.

The signal processing part is implemented in MATLAB. The training of the VAD network is performed using the Torch framework [52] and the Lua language. Source codes are available on the Internet [53].

3.1 Detection of microphone failures

Various hardware malfunctions or measurement errors caused by the user can occur in real-world recordings. These effects violate the validity of the mixing model (1), which can result in a distorted output of the multi-microphone enhancement algorithm. It is, therefore, beneficial to detect channels, which contain such errors and to skip them for the time interval of the malfunction.

The proposed failure detection proceeds as follows. For the ith microphone signal, $i = 1 \ldots M$, time-domain correlation coefficients with the other channels are computed. Let the maximum correlation in absolute value be denoted by μ_i. If μ_i is smaller than a predefined threshold t_μ, the ith channel is discarded from further processing.

3.2 Time–frequency domain transformation

The STFT is applied to signals sampled at 16 kHz with a frame length of 512 samples and a shift of 128 samples. The frames are weighted by the Hamming window. In the block-online regime, each block consists of 30–250 STFT frames, which corresponds to 0.25–2 s.

3.3 Voice activity detection

The VAD is implemented as a DNN. This type of VAD can adapt to complex, noisy conditions, provided that enough training data is available. The network aims to classify which STFT bins are dominated by target speech.

The training data consists of noisy input speech and corresponding target labels, which reflect the true amount of speech in the input signal. The labelling of targets is performed automatically, using a known decomposition of the input signal into the speech component and the background noise. Our data for VAD training are taken from the simulated part of the CHiME-4 [33] training dataset, i.e. 7138 utterances are utilised. A major part of the dataset (80%) is used to learn the network parameters; the remaining part is used to validate the obtained VADs.

The VAD is proposed for a single channel, that is, the spatial information provided by the multi-channel data is not used. Henceforth, our detector is denoted as multi-output VAD (mVAD). It performs the detection in each frequency bin separately, so it

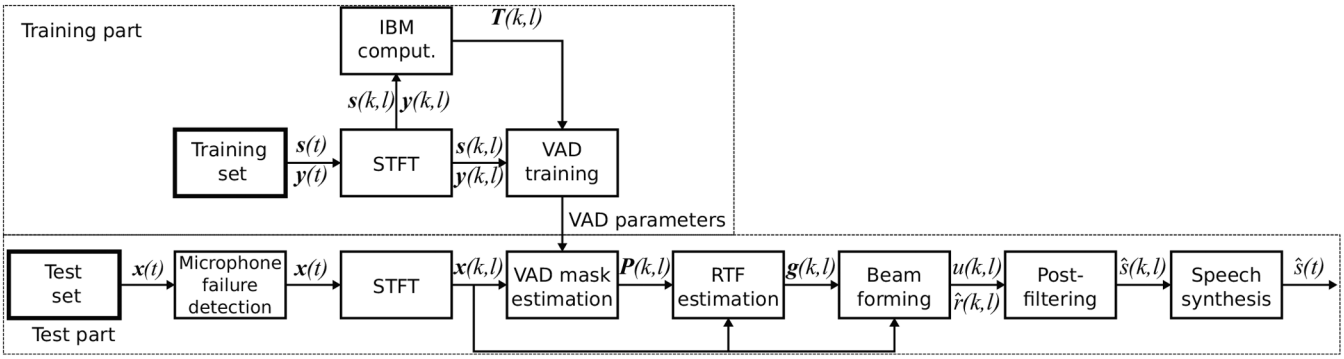

Fig. 1 *Block diagram of the proposed method. Variables* $\boldsymbol{s}(t)$, $\boldsymbol{y}(t)$ and $\boldsymbol{x}(t)$ *denote time-domain representations of one block of target speech, noise, and noisy speech signals, respectively*

outputs a vector value whose dimension corresponds with the frequency resolution.

For practical reasons stemming from the block-online utilisation, the mVAD is designed to have feed-forward fully connected topology. Although, e.g. the bi-directional long–short-term memory network can achieve slightly better detection (see, e.g. [35]), it requires the long context of frames to operate. The mVAD is designed to function without any context frames, which allows fully independent processing of subsequent data blocks.

The input layer of mVAD consists of 257 neurones; the input vector contains spectral magnitudes of the current STFT frame. Next, there are two hidden layers, each containing 1024 neurones. The activation function between hidden layers is rectified linear unit; the activation preceding the output layer is sigmoid non-linearity. The output layer consists of 257 units, which present the VAD decision for each of the input frequency bins.

The mVAD is trained to estimate ideal binary masks defined through

$$T_i^{\mathrm{mVAD}}(k,\ell)=\begin{cases}1, & 10\log\left(\dfrac{|s_i(k,\ell)|^2}{|y_i(k,\ell)|^2}\right)>t_{\mathrm{SNR}},\\ 0, & \text{otherwise}.\end{cases} \tag{11}$$

where i is the microphone index and t_{SNR} denotes a threshold parameter. During the training, which is finished after 40 epochs, the mean-square error criterion is optimised.

The mVAD is trained without pre-training, using the gradient descent method. Before the training, the input vectors are normalised to zero mean and unit variance. The initialisation of both weights and biases are random, drawn from the normal distribution $\mathcal{N}\left(0,\frac{6}{a+b}\right)$, where a and b are the numbers of network's input and output neurones, respectively. The output of the networks are masks $P_i(k,\ell)$, $i=1\ldots M$. Although the targets T_i^{mVAD} are binary, the networks output values from [0, 1].

As an optional step, the masks $P_i(k,\ell)$ can be condensed into a single mask that applies to all channels, which is an operation referred to as pooling. This post-processing step improves the robustness of VAD. We utilise median pooling, that is, the condensed mask is computed as

$$P(k,\ell)=\mathrm{median}_{i=1\ldots M}P_i(k,\ell). \tag{12}$$

3.4 RTF estimation

The proposed RTF estimation is based on a modification of the method from [25]. This estimator exploits the non-stationarity of speech while assuming that the noise is stationary. The latter assumption is somewhat restrictive, since the noise is often non-stationary in real-world situations. Our idea for the improvement of this estimator is as follows. The estimates of covariance between microphones are weighted by the output of VAD, and thus emphasise bins, which exhibit high SNR. In this way, even if the assumption of noise stationarity is not completely true, the contribution of the noisy bins to the RTF estimate is limited and the estimates are computed mainly from bins dominated by speech.

Specifically, let x_i, y_i, and g_i^{-1} denote the ith elements of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{g}^{-1}$, respectively. Using (1), the relationship between the reference and the ith channel can be expressed as

$$x_{\mathrm{ref}}(k,\ell)=g_i^{-1}(k)\,x_i(k,\ell)+v_i(k,\ell) \tag{13}$$

where $v_i(k,\ell)=y_{\mathrm{ref}}(k,\ell)-g_i^{-1}(k)\,y_i(k,\ell)$. Now, let each processing block be divided into N equally long sub-blocks. On the basis of (13), the (cross-) power spectral densities (PSDs) of the channels within the nth sub-block satisfy

$$\phi_{x_{\mathrm{ref}},\,x_i}(k,n)=g_i^{-1}(k)\phi_{x_i,\,x_i}(k,n)+\phi_{v_i,\,x_i}(k,n), \tag{14}$$

where $\phi_{v_i,\,x_i}(k,n)$ is independent of n when the noise is stationary within the block [25]. By substituting the PSDs with their sample-based estimates (denoted by $\hat{\phi}$), N linear equations are obtained (one for each sub-block)

$$\hat{\phi}_{x_{\mathrm{ref}},\,x_i}(k,n)=g_i^{-1}(k)\hat{\phi}_{x_i,\,x_i}(k,n)+\phi_{v_i,\,x_i}(k,n)+\epsilon(k,n), \tag{15}$$

where $\epsilon(k,n)$ is the estimation error

$$\epsilon(k,n)=\hat{\phi}_{v_i,\,x_i}(k,n)-\phi_{v_i,\,x_i}(k,n). \tag{16}$$

Let the sub-block length be denoted by L_0. The weighted (cross-)channel PSD estimates within the nth sub-block are

$$\begin{aligned}\hat{\phi}_{x_{\mathrm{ref}},\,x_i}(k,n)&=\sum_{\ell=nL_0}^{(n+1)L_0-1}P_i(k,\ell)x_{\mathrm{ref}}(k,\ell)x_i^*(k,\ell)\\ \hat{\phi}_{x_i,\,x_i}(k,n)&=\sum_{\ell=nL_0}^{(n+1)L_0-1}P_i(k,\ell)x_i(k,\ell)x_i^*(k,\ell),\end{aligned} \tag{17}$$

where $\cdot^*$ denotes conjugation. By setting $P_i(k,\ell)=1$, the weighting of the frames by VAD output is disabled, and the RTF estimation coincides with the original method from [25].

Now, formula (15), where $n=0,\ldots,N-1$ gives an over-determined system of linear equations with two unknown variables $g_i^{-1}(k)$ and $\phi_{v_i,\,x_i}(k,n)$. As suggested in [25], the least-squares solution gives us the final estimates of these quantities. They can be computed using the closed-form expression (the arguments k and n are omitted)

$$\hat{g}_i^{-1}=\frac{\left\langle\hat{\phi}_{x_{\mathrm{ref}},\,x_i}\hat{\phi}_{x_i,\,x_i}\right\rangle-\left\langle\hat{\phi}_{x_{\mathrm{ref}},\,x_i}\right\rangle\left\langle\hat{\phi}_{x_i,\,x_i}\right\rangle}{\left\langle\hat{\phi}^2_{x_i,\,x_i}\right\rangle-\left\langle\hat{\phi}_{x_i,\,x_i}\right\rangle^2}, \tag{18}$$

where $\langle\cdot\rangle$ denotes the averaging operator over the sub-block index n.

Input: Multi-channel speech signal
Discard channels with microphone failures;
Apply STFT to input speech to compute $\boldsymbol{x}(k,\ell)$;
for $i \leftarrow 2$ **to** M **do** {excluding reference channel 1}
Compute VAD mask $P_i(k,\ell)$ using $x_i(k,\ell)$, mVAD; {Section 3.3}
Estimate RTF $g_i(k)$ using $P_i(k,\ell)$ and $x_i(k,\ell)$; {Equations (17) and (18)}
end for
Compute beamforming output $u(k,\ell)$ using either IRTF (7) or MVDR (19)-(23);
Apply post-filter (25)-(26) to gain target $\hat{s}(k,\ell)$;
Reconstruct the enhanced speech in time-domain;

Fig. 2 *Algorithm 1: enhancement of a processing block by the proposed method*

Table 1 Values of the free parameters used in the experiments

Parameter	Value	Meaning
M	5 or 6	number of channels: CHiME-4
	4	number of channels: dynamic dataset
N		sub-block number, RTF estimation (see Section 3.4)
	3,5,10,25	for block lengths 0.25, 0.4, 0.8, 2 s
t_{SNR}	5 dB	threshold, mVAD [see (11)]
$f_{\max}$	3125 Hz	threshold, Wiener filter (see Section 3.6)
$f_{\min}$	100 Hz	threshold, Wiener filter (see Section 3.6)
t_{VAD}	0.3	threshold, Wiener filter (see Section 3.6)
t_{μ}		threshold, mic. failure detection (see Section 3.1)
	0.05	CHiME-4 simulated, dynamic
	0.40	CHiME-4 real-world datasets

Values were selected based on preliminary experiments with the CHiME-4 data. The sub-block number N is selected such that the sub-block always contains ten frames (80 ms of speech).

3.5 Beamforming: implementation details

To implement the MVDR beamformer (4) in practise, an estimate of the noise covariance $\boldsymbol{C}_{\boldsymbol{yy}}$ must be available. Since the RTF estimate has already been computed, we proceed by constructing a blocking matrix that is orthogonal to $\widehat{\boldsymbol{g}}$; it blocks the target signal and outputs a noise reference

$$\boldsymbol{v} = \boldsymbol{B}\boldsymbol{x}\,. \tag{19}$$

The blocking matrix $\boldsymbol{B}$ can have the structure

$$\boldsymbol{B} = \begin{bmatrix} -1 & g_2^{-1} & 0 & \dots & 0 \\ -1 & 0 & g_3^{-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \dots & g_M^{-1} \end{bmatrix}, \tag{20}$$

where it is assumed, without any loss of generality, that the reference channel $\boldsymbol{x}_{\mathrm{ref}}$ corresponds to the first microphone.

To obtain the estimate of noise signals $\boldsymbol{y}$ as they appear in mixture (1), the least-square estimate [40] is applied

$$\hat{\boldsymbol{y}} = \boldsymbol{C}_{\boldsymbol{xx}}\boldsymbol{B}^{\mathrm{H}}\left(\boldsymbol{B}\boldsymbol{C}_{\boldsymbol{xx}}\boldsymbol{B}^{\mathrm{H}}\right)^{-1}\boldsymbol{B}\boldsymbol{v}, \tag{21}$$

where $\boldsymbol{C}_{\boldsymbol{xx}} = E\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}\right]$ is replaced by its sample-based estimate. The covariance of $\widehat{\boldsymbol{y}}$ is equal to

$$\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}} = \mathrm{E}\left[\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^{\mathrm{H}}\right] = \boldsymbol{C}_{\boldsymbol{xx}}\boldsymbol{B}^{\mathrm{H}}\left(\boldsymbol{B}\boldsymbol{C}_{\boldsymbol{xx}}\boldsymbol{B}^{\mathrm{H}}\right)^{-1}\boldsymbol{B}\boldsymbol{C}_{\boldsymbol{xx}}, \tag{22}$$

which is substituted into (4). However, $\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}$ is rank deficient having the rank $\leq M-1$; therefore, its inverse matrix does not exist. As proposed in [40], the matrix $\boldsymbol{C}_{\boldsymbol{yy}}^{-1}$ is replaced by the Moore–Penrose pseudoinverse of $\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}$, denoted as $\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}^{\dagger}$. The implementation of MVDR is thus, finally, given by

$$\hat{\boldsymbol{w}}_{\mathrm{MVDR}} = \frac{\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}^{\dagger}\hat{\boldsymbol{g}}}{\hat{\boldsymbol{g}}^{\mathrm{H}}\boldsymbol{C}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}^{\dagger}\hat{\boldsymbol{g}}}\,. \tag{23}$$

In the case of the IRTF beamformer, the implementation is straightforward through formula (7), where the true RTF is replaced by its estimated value.

3.6 Single-channel post-filtering

The single-channel post-filtering is applied to minimise the residual noise in the beamforming output. In our work, the residual noise is estimated as

$$\begin{aligned} \hat{r}_{\mathrm{MVDR}} &= \hat{\boldsymbol{w}}_{\mathrm{MVDR}}^{\mathrm{H}}\hat{\boldsymbol{y}}, \\ \hat{r}_{\mathrm{IRTF}} &= \hat{\boldsymbol{w}}_{\mathrm{IRTF}}^{\mathrm{H}}\hat{\boldsymbol{y}}, \end{aligned} \tag{24}$$

where $\hat{\boldsymbol{y}}$ is obtained using (20) and (21). Assuming that the beamforming output $u(k, l)$ consists of the target speech signal and the residual noise, the approximate Wiener mask is applied defined as

$$G(k,\ell) = \frac{\max\left\{\left|u(k,\ell)\right|^2 - \left|\hat{r}(k,\ell)\right|^2, \delta\right\}}{\left|u(k,\ell)\right|^2 + \delta}, \tag{25}$$

where δ is a small positive constant to prevent from division by zero. Finally, the estimate of the target speech is given by

$$\hat{s}(k,\ell) = G(k,\ell)u(k,\ell)\,. \tag{26}$$

We add the following three heuristic modifications of (25) based on practical experience to improve the performance of this post-filtering step:

(i) For frequencies lower than $f_{\min}$Hz, $G(k,\ell) = 0.01$. As observed on CHiME-4 data, the very low frequencies usually contain only noise, so it is better to attenuate them.
(ii) For higher frequencies than $f_{\max}$Hz, $G(k,\ell) = 1$. This is because a speech signal is more difficult to block in the higher-frequency band, as there is typically a lower SNR. The RTF estimate has a higher variance, which causes leakage of the target signal into the residual noise estimate, and finally, a distortion due to the post-filter.
(iii) We set $G(k,\ell) = 1$ when $P(k,\ell) > t_{\mathrm{VAD}}$. The goal is to preserve time–frequency regions, where VAD detects speech with sufficiently high probability.

To summarise, the proposed method is described in Algorithm 1 (Fig. 2). Specific choices of the parameters that are used in experiments are listed in Table 1.

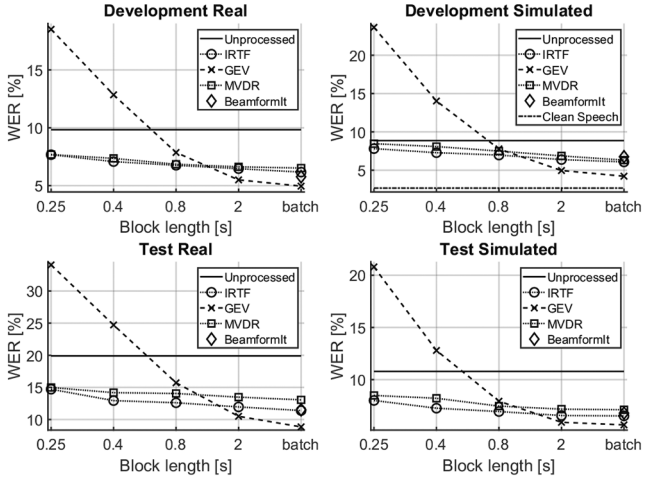

Fig. 3 *CHiME-4: WER (%) as a function of processing block length. The proposed beamformers use mVAD detector and mask pooling. WER of clean speech is available only for the development simulated set due to the rules of CHiME-4*

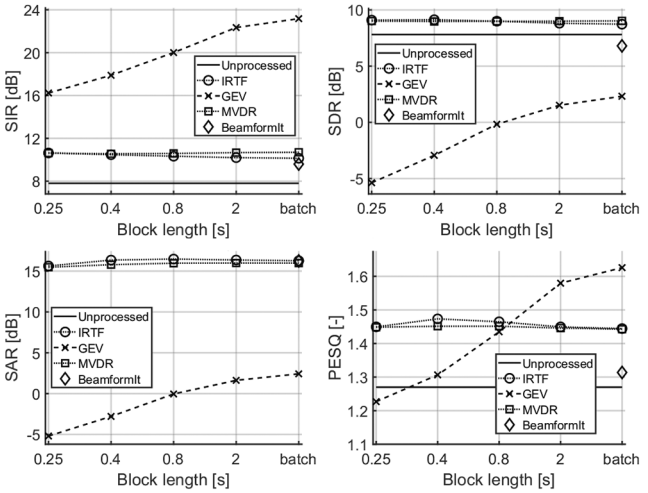

Fig. 4 *CHiME-4, development SIMU: objective criteria and PESQ as functions of the processing block length. The proposed beamformers use mVAD detector and mask pooling. BeamformIt is presented only for the batch processing mode. SAR of the unprocessed signals is infinite; therefore, it is not shown*

4 Experimental results and discussion

This section presents experiments conducted on two datasets: the CHiME-4 dataset [33] and another dataset featuring moving speaker, which is further denoted as 'Dynamic'. On both datasets, we perform objective evaluation using the criteria defined in the BSS_Eval toolbox [54] and the evaluation of perceptual quality of enhanced signals using perceptual evaluation of speech quality (PESQ) [55]; the wideband (16 kHz) implementation from [56] is used. The PESQ measure yields high correlation coefficient with overall quality and enhanced signal distortion obtained using subjective listening tests [57]. For CHiME-4, the availability of a trained ASR system and of labelled test utterances is taken into account. The multi-channel enhancement is used to improve the performance of ASR in terms of WER. In the experiments, the influence of various processing steps and short duration of the input signal on the performance is analysed.

CHiME-4 defines three tracks where utterances of speaking persons are recorded in noisy conditions; our experiments focus on the multi-channel track with six-channel recordings. These were either simulated (SIMU) or acquired by a tablet device in real-world situations (REAL). With the REAL recordings, we do not use microphone 2, because it is oriented in a direction away from the speaker. For the SIMU test set, we use another experimentally determined subset of channels, namely 2–6. The dataset contains data from four different noisy environments (bus, cafeteria, street, and pedestrian area). The utterances are taken from the wall street journal corpus [58] and are provided along with the respective transcription. There are 5920 test files overall, which are divided into development and test datasets.

In the dynamic dataset, the signal part corresponds to a 59 s long utterance recorded on four microphones in a highly reverberated room ($T_{60} \simeq 700$ ms). The speaker was sitting about 0.5 m in front of the microphone array and was leaning its upper body to sides by about 25 cm to the left or right from the central position. As the noise part of the data, the noise types from the CHiME-4 campaign were used (bus, cafeteria, street and pedestrian area), because the utilised VADs were trained for these environments. For the detailed analysis, the speech was mixed with noise at global SNR of 5 dB. We present also additional experiments, where the influence of the input SNR is studied; here, the SNR varies from 0 to 8 dB. Channels 1, 3, 4, 6 of the six-channel track are used. There are thus four noisy instances of the recorded speech, so the total length of the dynamic dataset is about 4 min.

To perform the objective evaluation, the ground-truth clean speech signals are needed. Therefore, we present the objective evaluation only for the SIMU development dataset of the CHiME-4 and of the dynamic dataset, where the reference data is available. BSS_Eval is applied by using the `bss_decomp_tvfilt` function for signal decompositions. This function utilises a time-varying filter (we set its length to 32 taps) as an allowed distortion of the estimated target source. The signal-to-interference ratio (SIR), signal-to-distortion (SDR), and signal-to-artefact (SAR) ratios are subsequently computed as defined in [54].

Speech recognition is performed for the CHiME-4 datasets, where the text reference is available, by the baseline transcription system provided by CHiME-4 organisers [59]. The recogniser features a hybrid DNN-hidden Markov model for acoustic modelling and a 5-gram language model. Subsequently, rescoring of the lattice is performed using a recurrent neural network language model.

The performance of the proposed system has been compared with two state-of-the-art techniques that were successfully used to solve the CHiME-4 challenge, namely (i) the baseline method BeamformIt [38] and (ii) the GEV beamformer that is the front-end part of the system described in [35, 36]. These methods represent two frequently used approaches for beamformer steering in the reverberant environment: the compensation of time difference of arrival and the eigenvector decomposition of the speech covariance matrix. While BeamformIt operates only in the batch regime (thus, we apply it only to more static CHiME-4 data) as it passes twice through each recording to optimise its inner parameters, GEV is also modified for the block-online processing.

The training procedure for the VAD in GEV was kindly provided to us by Heymann *et al.* [35]. This VAD was retrained using the CHiME-4 training datasets. The fully connected variant of the VAD (denoted as hVAD) is utilised because it possesses comparable topology with that of mVAD.

For the sake of clarification, the compared systems will be denoted by *A*:*B*, where *A* denotes the beamforming method and *B* denotes the particular modification focused on by the experiment. For example, when VAD selection is discussed, the labels could be IRTF:mVAD, IRTF:hVAD etc. The presented results are averaged over the four noisy environments; the resulting WER reductions are absolute. Also, see Table 1 for other values of parameters utilised in the experiments.

4.1 Block-online versus batch processing

This experiment is focused on the performances of the compared methods as they depend on the length of the processing block. The proposed systems utilise mVAD, pooling of VAD masks, and the Wiener post-filtering. The GEV beamformer is applied with the BAN post-filtering.

4.1.1 CHiME-4 datasets: The results are shown in Figs. 3 and 4. They clearly show that all the methods improve their ability to

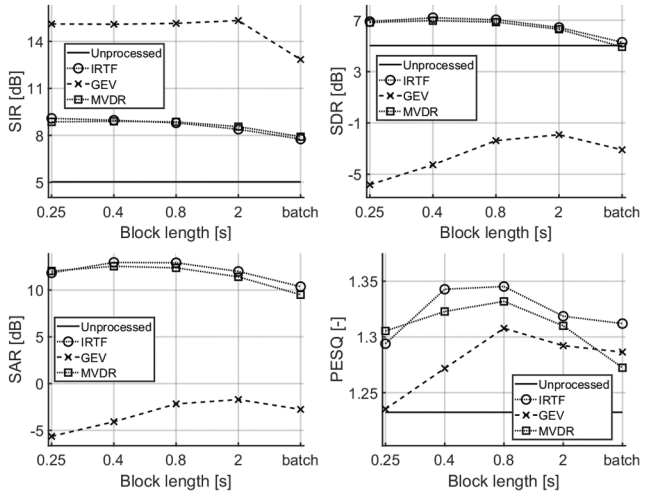

Fig. 5 *Dynamic: objective criteria and PESQ as functions of the processing block length. The proposed beamformers use mVAD detector and mask pooling. SAR of the unprocessed signals is infinite; therefore, it is not shown*

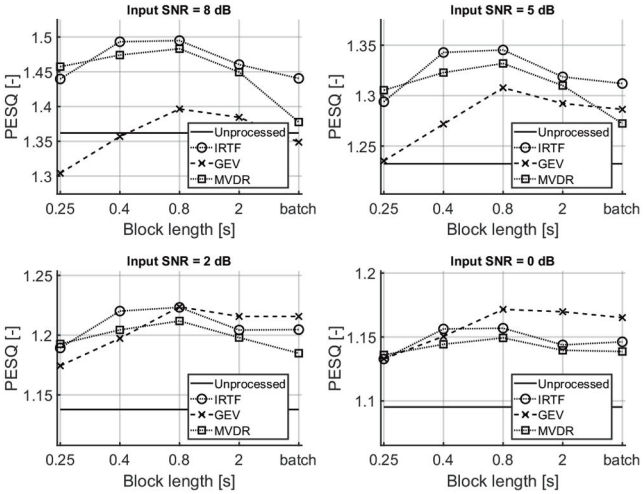

Fig. 6 *Dynamic: perceptual quality measured by PESQ as a function of the input SNR and length of the processed block. The proposed beamformers use mVAD detector and mask pooling*

suppress the background noise with the growing length of the processing blocks, and achieve their optimum performance in the batch processing mode. The latter observation confirms the fact that the target speaker has, in CHiME-4 recordings, an almost fixed position, because, otherwise, the batch processing would have failed.

The proposed methods yield stable performance for the length of the processing block. IRTF achieves comparable or slightly lower WER (by up to 1.5%) compared with MVDR. GEV:batch achieves the best WER and surpasses IRTF:batch by 0.8–2.2% WER. BeamformIt yields slightly better results compared with IRTF:batch on real datasets and slightly worst results by 0.3–0.7% on the simulated ones.

The performance of GEV drops down with the decreasing length of the processing block, especially when this goes below 2 s. This sensitivity seems to be mainly caused by the increased amount of artefacts in the enhanced speech, as shown by SDR and SAR in Fig. 4. We conjecture that this is partly due to less precise estimates of speech/noise covariance matrices and partly due to the limitations of BAN, as stated in [16]. BAN relies on the precise estimation of $\boldsymbol{w}_{\mathrm{GEV}}$ and assumes that the norm of $\boldsymbol{g}(k, \ell)$ is approximately constant across frequencies. Using short blocks of signals, these quantities are estimated with higher estimation errors, and BAN seems to be ineffective. Compared to unprocessed speech, PESQ (ranging from 5 for the best quality to 1 for the worst) achieved by GEV:batch is increased by about 0.4, whereas GEV:0.25 s lowers the PESQ by about 0.05. This performance decrease for short processing blocks could be probably alleviated through the recursive implementation of GEV. This would, however, decrease the adaptation speed of the technique.

Compared to the proposed solutions, GEV achieves strong noise suppression measured by SIR. For example, GEV:0.8 s outperforms IRTF:0.8 s by about 9 dB in terms of SIR. However, SDR and SAR obtained by GEV:0.8 s are lower by about 9 and 16 dB, respectively, compared with IRTF:0.8 s.

Concerning the functionality for noise type: All the compared methods are successful in suppressing the cafeteria, street and pedestrian area noises to a certain extent. For the batch processing, the difference in WER for these environments is up to 3% for the IRTF beamformer and up to 2% for GEV. A lower performance is achieved for the bus environment (especially in the REAL task); WER is higher by about 6–9% compared with the other environments for GEV and IRTF. We conjecture that the deterioration arises because the bus noise has a narrowband character concentrated in low frequencies (up to 1 kHz). These frequencies (above 250 Hz) are important to ASR because they contain important vowel formants. However, the low-frequency noise is difficult to suppress for the beamformers, because of the long wavelengths of the sound.

4.1.2 Dynamic dataset: The results in Fig. 5 suggest that the most suitable block length for the dynamic dataset is 0.8 s. With increasing block duration up to 0.8 s, the compared methods exhibit improvements in most of the observed metrics. In contrast to CHiME-4 datasets, most of the metrics deteriorate when the block size exceeds 2 s. This indicates that the signal and noise spatial statistics change in time due to the movements of the speaker.

For processing block lengths 0.8 s and lower, the obtained results are consistent with our findings reported on CHiME-4 datasets. The proposed methods yield stable performance in terms of SIR, SDR, and SAR. In contrast, the performance of the GEV beamformer deteriorates due to the presence of distortions and artefacts in the processed speech, as is indicated by the SDR and SAR metrics. However, GEV retains its strong noise suppression ability indicated by high SIR values.

The influence of input SNR on the enhancement is studied in Fig. 6 using PESQ, which evaluates both noise suppression and speech distortion in a single measure. The overall character of the results does not change. The proposed method is generally superior using the short blocks. The best PESQ is achieved for the block length 0.8 s, and it deteriorates for longer blocks due to the movements of the speaker. For input SNR = 0 dB, noise attenuation is more crucial to the perceptual quality than low distortion. Here, GEV achieves the highest PESQ, which points to its strong noise suppression ability. With increasing input SNR, the proposed method yields higher PESQ in most cases, because it does not distort the target speech much.

4.2 Influence of post-filtering

In this experiment, the compared methods are tested with or without their post-filtering parts. The block-online regime with a block length of 0.8 s is considered together with mVAD and pooling of VAD masks. The results achieved with this setting on CHiME-4 datasets are shown in Table 2 and in Fig. 7; an example of a spectrogram demonstrating the effects of post-filtering is shown in Fig. 8. The overall results show that the post-filtering improves both the perceptual quality as well as the accuracy of recognition. When post-filtering is omitted, the WER is increased by about 0.5% for IRTF as well as for MVDR. The influence of BAN on the performance of GEV is more significant: without BAN, the WER of GEV is increased by 1.7–3.8%.

For MVDR and IRTF, the omission of the Wiener post-filtering results in a decrease of SIR by about 2 dB and of SDR by about 1 dB. Although the SAR is slightly increased (0.8 dB), the PESQ value is decreased by about 0.1. The performance of GEV:BAN improves in terms of SIR, SDR as well as SAR, by about 2 dB compared with GEV:none, and PESQ is improved by 0.03.

Table 2 CHiME-4: influence of post-filtering in terms of the WER (%)

System: post-filter	Dev		Test	
	Real	SIMU	Real	SIMU
unprocessed	9.83	8.86	19.90	10.79
IRTF:none	7.02	**6.98**	13.34	7.36
IRTF:Wiener	**6.76**	**6.98**	**12.61**	**6.95**
MVDR:none	7.03	7.95	14.72	7.97
MVDR:Wiener	6.85	7.50	14.05	7.47
GEV:none	10.02	9.49	19.52	10.27
GEV:BAN	7.87	7.78	15.71	7.94

Proposed beamformers use mVAD detector and mask pooling. The best-achieved results are written in bold.

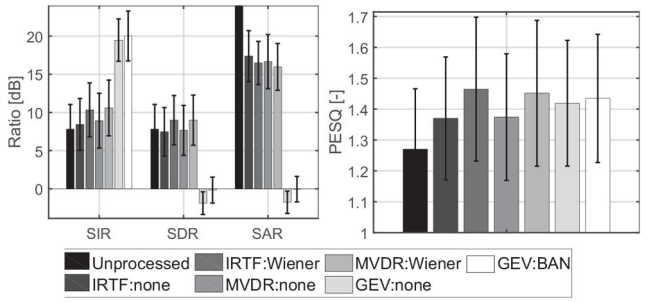

Fig. 7 *CHiME-4: influence of post-filtering in the terms of mean objective criteria and PESQ along with respective standard deviations. The proposed beamformers use mVAD detector and mask pooling. The SAR values for unprocessed data are theoretically infinity; thus, are truncated in the graph*

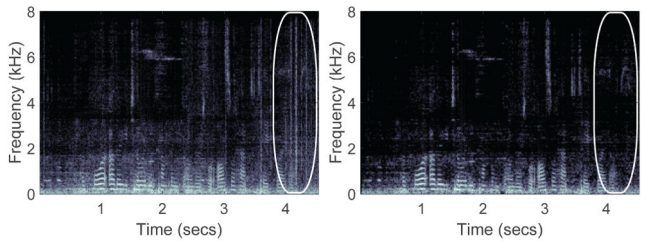

Fig. 8 *Demonstration of post-filter effects. On the left, there is a signal without post-filtering, on the right is signal with applied post-filter ($f_{\max}$ was selected 8 kHz to emphasise the effects of the post-filter). The residual noise and the highlighted artefact around time 4 s are suppressed*

Table 3 CHiME-4: influence of VAD-mask pooling in terms of WER (%)

System: pooling	Dev		Test	
	Real	SIMU	Real	SIMU
unprocessed	9.83	8.86	19.90	10.79
IRTF:no	7.08	7.02	**13.16**	**7.32**
IRTF:yes	**7.02**	**6.98**	13.34	7.36
MVDR:no	7.82	8.86	16.16	9.06
MVDR:yes	7.03	7.95	14.72	7.97

Beamformers use mVAD detector and do not use post-filtering. The best-achieved results are written in bold.

4.3 Influence of pooling and VAD inclusion

This section investigates the effects of VAD output pooling on beamforming. Moreover, the contribution of VAD itself to the enhancement performance is measured. To this end, the results of beamforming with included/disabled VAD-mask weighting are compared, as discussed in Section 3.4 below (17).

The results were achieved on CHiME-4 datasets in block-online regime with the block length 0.8 s without post-filtering.

The results shown in Table 3 suggest that the pooling is beneficial for the MVDR beamformer. Here, the omission of pooling increases the WER by 0.8–1.5%. Pooling does not improve (but neither deteriorate) the WER of IRTF. The objective measures did not show any significant benefits of pooling; that is why they are not presented.

Table 4 indicates that the utilisation of mVAD is beneficial for the beamforming. IRTF with mVAD performs better by 0.1–0.4% compared with IRTF:none, whereas MVDR is improved by 0.4–0.9% by utilisation of VAD. Again, the objective criteria are not presented, since these do not reflect the inclusion of VAD significantly.

4.4 Computational aspects

To evaluate the computational burden due to the compared methods, we measure the time necessary to enhance the simulated development dataset in the bus environment. The experiment was performed on a PC with Intel processor i7–2600 K, 3.4 GHz, and 16 GB RAM. To maintain comparable conditions, only techniques implemented in MATLAB are compared. Moreover, only a single computational thread is enabled during the experiments. All competing techniques are tested with mVAD (including GEV) or without any VAD (excluding GEV, which requires VAD to work properly).

Fig. 9 shows that the computational burden is decreasing with the growing length of the block. This is, obviously, caused by the savings as the same processing is applied to longer segments of signals. Most of the burden is due to VAD. In the case of batch processing, mVAD-mask computation consumes about 90% of the processing time for all techniques.

The results in Section 4.3 indicate that the utilisation of VAD improves the performance of the presented technique. However, the method can be utilised without VAD if moderate performance loss is allowable. Thus, if the computational burden is of concern, the computational cost can be significantly reduced by the omission of mVAD (or utilisation of a network with a lesser number of parameters).

5 Conclusion

A block-online multi-channel speech enhancement method based on beamforming has been proposed. In this method, the steering vector is constructed using estimates of RTFs between microphones. Performed experiments indicate that the performance of the proposed method is robust for the short length of the processed data.

An alternative approach to the computation of the steering vector consists of eigenvalue decomposition of the speech covariance matrix. In our experiments, the covariance-based techniques are represented by the state-of-the-art GEV beamformer. Considering the CHiME-4 datasets, the proposed methods achieve lower WER and higher PESQ values when blocks of length lower than 0.8 s are used, whereas the GEV beamforming is superior for blocks of length 2 s and higher. The proposed methods introduce a significantly smaller amount of distortions, as shown through the achieved SDR and SAR.

The proposed IRTF beamformer appears to be computationally simpler than its MVDR counterpart, and it achieves lower WER and higher PESQ. We find it beneficial to perform single-channel post-filtering after beamforming; it improves the perceptual quality of the enhanced speech as well as it lowers the WER of the recogniser.

Concerning the future work, the proposed system (even without VAD) can serve as a strong initialisation for BSS methods, which can refine the resulting estimates of clean speech.

6 Acknowledgments

This work was supported by The Czech Science Foundation through Project No. 17–00902S and by California Community Foundation through Project no. DA-15-114599.

Table 4 CHiME-4: influence of the VAD on beamforming in terms of the WER (%)

System: VAD type	Dev		Test	
	Real	SIMU	Real	SIMU
unprocessed	9.83	8.86	19.90	10.79
IRTF:none	7.21	7.34	13.56	7.52
IRTF:mVAD	**7.08**	**7.02**	**13.16**	**7.32**
MVDR:none	8.19	9.26	17.06	9.67
MVDR:mVAD	7.82	8.86	16.16	9.06

Techniques use neither pooling of VAD masks nor post-filtering. The best-achieved results are written in bold.

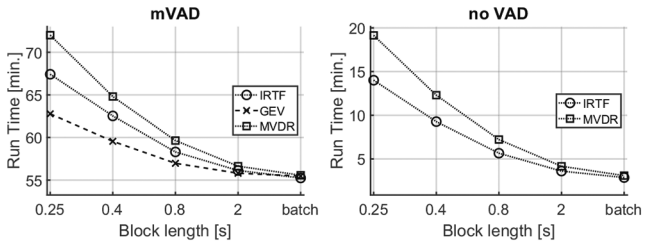

Fig. 9 *CHiME-4: time required to enhance the simulated development 'bus' dataset; the total length of the recordings is 43.2 min*

7 References

[1] Cohen, I., Benesty, J., Gannot, S.: '*Speech processing in modern communication: challenges and perspectives*', vol. **3** (Springer Science & Business Media, Switzerland, 2009)

[2] Boll, S.: 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Trans. Acoust. Speech Signal Process.*, 1979, **27**, (2), pp. 113–120

[3] Samui, S., Chakrabarti, I., Ghosh, S.K.: 'Improved single-channel phase-aware speech enhancement technique for low signal-to-noise ratio signal', *IET Signal Process.*, 2016, **10**, (6), pp. 641–650

[4] Ephraim, Y., Malah, D.: 'Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator', *IEEE Trans. Acoust. Speech Signal Process.*, 1984, **32**, (6), pp. 1109–1121

[5] Mahmmod, B.M., Ramli, A.R., Abdulhussian, S.H., *et al.*: 'Low-distortion MMSE speech enhancement estimator based on Laplacian prior', *IEEE Access*, 2017, **5**, pp. 9866–9881

[6] Mahmmod, B.M., Ramli, A.R., Baker, T., *et al.*: 'Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials', *IEEE Access*, 2019, **7**, pp. 103485–103504

[7] Li, Y., Kang, S.: 'Deep neural network-based linear predictive parameter estimations for speech enhancement', *IET Signal Process.*, 2016, **11**, (4), pp. 469–476

[8] Wang, D., Chen, J.: 'Supervised speech separation based on deep learning: an overview', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2018, **26**, (10), pp. 1702–1726

[9] Gannot, S., Vincent, E., Markovich Golan, S., *et al.*: 'A consolidated perspective on multimicrophone speech enhancement and source separation', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2017, **25**, (4), pp. 692–730

[10] Makino, S., Lee, T.W., Sawada, H.: '*Blind speech separation*', vol. **615** (Springer, Switzerland, 2007)

[11] Van Veen, B.D., Buckley, K.M.: 'Beamforming: a versatile approach to spatial filtering', *IEEE ASSP Mag.*, 1988, **5**, (2), pp. 4–24

[12] Doclo, S., Spriet, A., Wouters, J., *et al.*: 'Speech distortion weighted multichannel Wiener filtering techniques for noise reduction', in Benesty, J., Makino, S., Chen, J. (Eds.): '*Speech enhancement. Signals and communication technology*' (Springer, Berlin, Heidelberg, 2005), pp. 199–228

[13] Van Trees, H.L.: '*Optimum array processing: part IV of detection, estimation, and modulation theory*' (John Wiley & Sons, USA, 2004)

[14] Doclo, S., Gannot, S., Moonen, M., *et al.*: 'Acoustic beamforming for hearing aid applications', in Haykin, S., Ray Liu, K.J. (Eds.): *Handbook on array processing and sensor networks,* (John Wiley & Sons, USA, 2010), pp. 269–302

[15] Cox, H., Zeskind, R., Owen, M.: 'Robust adaptive beamforming', *IEEE Trans. Acoust. Speech Signal Process.*, 1987, **35**, (10), pp. 1365–1376

[16] Warsitz, E., Haeb Umbach, R.: 'Blind acoustic beamforming based on generalized eigenvalue decomposition', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (5), pp. 1529–1539

[17] Gannot, S., Burshtein, D., Weinstein, E.: 'Signal enhancement using beamforming and non-stationarity with applications to speech', *IEEE Trans. Signal Process.*, 2001, **49**, (8), pp. 1614–1626

[18] Chang, J.H., Kim, N.S., Mitra, S.K.: 'Voice activity detection based on multiple statistical models', *IEEE Trans. Signal Process.*, 2006, **54**, (6), pp. 1965–1976

[19] Dov, D., Talmon, R., Cohen, I.: 'Kernel method for voice activity detection in the presence of transients', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2016, **24**, (12), pp. 2313–2326

[20] Kim, J.T., Jung, S.H., Cho, K.H.: 'Efficient harmonic peak detection of vowel sounds for enhanced voice activity detection', *IET Signal Process.*, 2018, **12**, (8), pp. 975–982

[21] Zhang, X.L., Wu, J.: 'Deep belief networks based voice activity detection', *IEEE Trans. Audio Speech Lang. Process.*, 2013, **21**, (4), pp. 697–710

[22] Goodfellow, I., Bengio, Y., Courville, A.: '*Deep learning*' (MIT Press, USA, 2016). Available at http://www.deeplearningbook.org

[23] Cohen, I.: 'Relative transfer function identification using speech signals', *IEEE Trans. Speech Audio Process.*, 2004, **12**, (5), pp. 451–459

[24] Markovich, S., Gannot, S., Cohen, I.: 'Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals', *IEEE Trans. Audio Speech Lang. Process.*, 2009, **17**, (6), pp. 1071–1086

[25] Shalvi, O., Weinstein, E.: 'System identification using non-stationary signals', *IEEE Trans. Signal Process.*, 1996, **44**, (8), pp. 2055–2063

[26] Meier, S., Kellermann, W.: 'Analysis of the performance and limitations of ICA-based relative impulse response identification'. 2015 23rd European Signal Processing Conf. (EUSIPCO), France, 2015, pp. 414–418

[27] Khan, A.H., Taseska, M., Habets, E.A.P.: 'A geometrically constrained independent vector analysis algorithm for online source extraction'. 12th Int. Conf. Latent Variable Analysis and Signal Separation: LVA/ICA 2015, Springer, 2015, pp. 396–403

[28] Koldovsky, Z., Malek, J., Gannot, S.: 'Spatial source subtraction based on incomplete measurements of relative transfer function', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015, **23**, (8), pp. 1335–1347

[29] Giri, R., Rao, B.D., Mustiere, F., *et al.*: 'Dynamic relative impulse response estimation using structured sparse Bayesian learning'. 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), China, 2016, pp. 514–518

[30] Katzberg, F., Mazur, R., Maass, M., *et al.*: 'Spatial interpolation of room impulse responses using compressed sensing'. 2018 16th Int. Workshop on Acoustic Signal Enhancement (IWAENC), Japan, 2018, pp. 426–430

[31] Cutajar, M., Gatt, E., Grech, I., *et al.*: 'Comparative study of automatic speech recognition techniques', *IET Signal Process.*, 2013, **7**, (1), pp. 25–46

[32] Zhang, Z., Geiger, J., Pohjalainen, J., *et al.*: 'Deep learning for environmentally robust speech recognition: an overview of recent developments', *ACM Trans. Intell. Syst. Technol. (TIST)*, 2018, **9**, (5), p. 49

[33] Vincent, E., Watanabe, S., Nugraha, A.A., *et al.*: 'The 4th CHiME speech separation and recognition challenge', 2019. Available at http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/, last submission on 27 November 2019

[34] Hongyu, W.B., Ou, Z.: 'The THU-SPMI CHiME-4 system: lightweight design with advanced multi-channel processing, feature enhancement, and language modeling'. Proc. the Fourth Int. Workshop on Speech Processing in Everyday Environments CHiME-4, USA, 2016

[35] Heymann, J., Drude, L., Haeb Umbach, R.: 'Neural network-based spectral mask estimation for acoustic beamforming'. 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), China, 2016, pp. 196–200

[36] Heymann, J., Drude, L., Haeb Umbach, R.: 'Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition'. Proc. Fourth Intl. Workshop on Speech Processing in Everyday Environments CHiME-4, USA, 2016

[37] Xiang, H., Wang, B., Ou, Z.: 'The USTC-iFlytek system for CHiME-4 challenge'. Proc. Fourth Int. Workshop on Speech Processing in Everyday Environments CHiME-4, USA, 2016

[38] Anguera, X., Wooters, C., Hernando, J.: 'Acoustic beamforming for speaker diarization of meetings', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (7), pp. 2011–2022

[39] Barker, J., Watanabe, S., Vincent, E.: 'The 5th CHiME speech separation and recognition challenge', 2019. Available at http://spandh.dcs.shef.ac.uk/chime_challenge/index.html, last submission on 27 November 2019

[40] Koldovsky, Z., Nesta, F.: 'Approximate MVDR and MMSE beamformers exploiting scale-invariant reconstruction of signals on microphones'. 2016 15th Int. Workshop on Acoustic Signal Enhancement (IWAENC), China, 2016

[41] Araki, S., Okada, M., Higuchi, T., *et al.*: 'Spatial correlation model-based observation vector clustering and MVDR beamforming for meeting recognition'. 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), China, 2016, pp. 385–389

[42] Nakatani, T., Ito, N., Higuchi, T., *et al.*: 'Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming'. 2017 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, New Orleans, LA, USA, 2017, pp. 286–290

[43] Pfeifenberger, L., Zohrer, M., Pernkopf, F.: 'DNN-based speech mask estimation for eigenvector beamforming'. 2017 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, New Orleans, LA, USA, 2017, pp. 66–70

[44] Ochiai, T., Watanabe, S., Hori, T., *et al.*: 'Unified architecture for multichannel end-to-end speech recognition with neural beamforming', *IEEE. J. Sel. Top. Signal Process.*, 2017, **11**, (8), pp. 1274–1288

[45] Erdogan, H., Hershey, J.R., Watanabe, S., *et al.*: 'Improved MVDR beamforming using single-channel mask prediction networks'. INTERSPEECH, USA, 2016, pp. 1981–1985

[46] Wang, Z.Q., Wang, D.: 'Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR'. 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, Calgary, Canada, 2018, pp. 1–5

[47] Boeddeker, C., Erdogan, H., Yoshioka, T., *et al.*: 'Exploring practical aspects of neural mask-based beamforming for far-field speech recognition'. 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, Calgary, Canada, 2018, pp. 1–5

[48] Higuchi, T., Kinoshita, K., Ito, N., *et al.*: 'Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming'. 2018 IEEE Int. Conf.

Acoustics, Speech and Signal Processing (ICASSP) IEEE, Calgary, Canada, 2018, pp. 1–5

[49] Higuchi, T., Ito, N., Yoshioka, T., *et al.*: 'Robust MVDR beamforming using time–frequency masks for online/offline ASR in noise'. 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, Shanghai, People's Republic of China, 2016, pp. 5210–5214

[50] Togami, M.: 'Simultaneous optimization of forgetting factor and time-frequency mask for block-online multi-channel speech enhancement'. ICASSP 2019–2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) IEEE, Brighton, UK, 2019, pp. 2702–2706

[51] Schwartz, O., Gannot, S., Habets, E.A.P.: 'Multi-microphone speech dereverberation and noise reduction using relative early transfer functions', *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015, **23**, (2), pp. 240–251

[52] Collobert, R., Farabet, C., Kavukcuoglu, K., *et al.*: 'Torch – a scientific computing framework for LuaJIT', 2019. Available at http://torch.ch/, last submission on 27 November 2019

[53] Malek, J., Koldovsky, Z.: 'Multichannel-enhancement', 2019. Available at https://asap.ite.tul.cz/downloads/enhancer/, last submission on 27 November 2019

[54] Vincent, E., Gribonval, R., Fevotte, C.: 'Performance measurement in blind audio source separation', *IEEE Trans. Audio Speech Lang. Process.*, 2006, **14**, (4), pp. 1462–1469

[55] Rix, A.W., Beerends, J.G., Hollier, M.P., *et al.*: 'Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs'. 2001 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2001 Proc. (ICASSP'01) IEEE, Salt Lake City, UT, USA, 2001, vol. **2**, pp. 749–752

[56] Loizou, P.C.: '*Speech enhancement, theory and practice*' (CRC Press, USA, 2013, 2nd edn.)

[57] Hu, Y., Loizou, P.C.: 'Evaluation of objective quality measures for speech enhancement', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **16**, (1), pp. 229–238

[58] Garofolo, J., Graff, D., Paul, D., *et al.*: 'CSR-I (WSJ0) complete LDC93S6A', Philadelphia: Linguistic Data Consortium, 1993. Available at http://spandh.dcs.shef.ac.uk/chime_challenge/index.html, last submission on 27 November 2019

[59] Vincent, E., Watanabe, S., Nugraha, A.A., *et al.*: 'An analysis of environment, microphone and data simulation mismatches in robust speech recognition', *Comput. Speech Lang.*, 2016, **46**, pp. 535–557