CHiME4: Multichannel Enhancement Using Beamforming Driven by DNN-based Voice Activity Detection

Zbyněk Koldovský, Jiri Malek, Marek Boháč, and Jakub Janský

Faculty of Mechatronics, Informatics, and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic.
zbynek.koldovsky@tul.cz

Abstract

In this work, we focus on methods for enhancing the six-channel CHiME4 data using beamforming that is driven by voice activity detectors (VAD). We propose two beamformers and two VADs that are based on trained deep neural networks (DNN). Their combinations are compared when used as front-ends whose outputs are forwarded to the baseline automatic speech recognition system. Results in term of Word-Error-Rate (WER) achieved when the acoustic model of the baseline is or is not adapted for the given front-end (re-trained on enhanced training sets) are reported.

1. Introduction

Many multichannel speech enhancement systems apply beamforming methods such as the conventional Delay-and-Sum Beamformer (DSB), various implementations of minimum variance distortionless (MVDR) beamformer, or a generalization of the latter one, the linearly constrained minimum variance (LCMV) beamformer [1]. In order to achieve optimum performance, parameters have to be estimated and tracked with a sufficient accuracy. If not, the target signal in the system output can be distorted, which often deteriorates the final performance achieved by back-end processors (e.g., automatic speech recognition systems) even if the Signal-to-Noise Ratio (SNR) in a beamformer's output is improved.

In the conventional beamforming, the free-field sound propagation is assumed, and the DSB relies purely on the Time-Difference-Of-Arrival (TDOA) estimation. By contrast, the MVDR and LCMV can regard reverberation and multiple sources when using relative transfer functions (RTFs); see [2]. Such systems tend to be less robust as compared to the conventional approach. In particular, they are more sensitive to possible nonlinearities in the signal path as well as to various measurement (sensor) failures. On the other hand, their performance is potentially higher than that of the DSB, especially, in multi-source and reverberant conditions. The goal of this work is to compare the methods within CHiME4.

The baseline system of CHiME4 utilizes a state-of-the-art DSB technique named BeamformIt, proposed in [3]. The method estimates TDOAs using generalized cross-correlations (GCC-PHAT) and performs a robust multichannel TDOA tracking, which significantly helps to avoid sudden changes and estimation errors in TDOA. This and other straightforward modifications such as a mechanism that helps to avoid microphone failures make BeamformIt robust and useful for CHiME4. A practical drawback is that BeamformIt is passing through the signals several times before the output is computed, which hampers its direct applicability in continuous (on-line) processing.

Multichannel enhancement systems applying MVDR or LCMV with the aid of Deep Neural Networks (DNN) were applied to CHiME3 data; see, e.g., [4, 5]. The beamformers rely on the estimation of the noise covariance and of the source steering vector from masked signals, where the masks are obtained as outputs of DNNs.

In this work, approximate Minimum Mean-Squared Error beamformer (MMSE), recently proposed in [6], is modified in order to be applied within CHiME4. Similarly to [4, 5], the beamformer exploits DNNs, however, the DNNs are used to control the estimation of RTFs, not the estimation of noise/speech covariances. This is done through applying the RTF estimator from [7] where speech presence probabilities are obtained as the outputs of Voice-Activity Detectors (VAD) that are realized using the DNNs.

The performance of the MMSE depends purely on the accuracy of the estimated RTFs. As such, the beamformer strongly relies on the linearity of the observed signals. However, this appears to be often violated in the CHiME4 data, e.g., because of microphone failures and nonlinear gain fluctuations. The results of this work thus provide a comparison of the advanced beamforming with BeamformIt. We compare also a Filter-and-Sum Beamformer (FSB) based on the estimated RTFs, which could be seen as a solution on the half way between the MMSE and BeamformIt.

The paper is organized as follows. Section 2 describes the problem and basic beamforming approaches. Section 3 provides details of the proposed multichannel enhancement systems. Section 4 defines the back-end solutions that we use for CHiME4. Section 5 reports the results and Section 6 concludes the paper.

2. Problem Description

2.1. Model

A noisy recording of a directional source observed through m microphones can be described, in the short-term frequency domain, as

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

where $\mathbf{x}(k,\ell)$ is the $m \times 1$ vector of the signals on microphones, $s(k,\ell)$ is the target speech as observed on a reference microphone, and $\mathbf{y}(k,\ell)$ involves all other interfering sources and noise components that are uncorrelated with $s(k,\ell)$; k is the frequency index and ℓ is the frame index.

The vector $\mathbf{g}(k,\ell)$ determines the position of the target speaker. Its elements contain relative transfer functions (RTFs) related to the reference microphone [2]. Since the speaker can perform movements during utterances, $\mathbf{g}(k,\ell)$ is varying in

time. Nevertheless, we assume that the changes are slow, so $\mathbf{g}(k, \ell)$ is approximately constant during blocks of frames.

From now on we will omit the arguments k and ℓ from the notation. They will be used only when the more precise notation is needed.

2.2. MVDR and MMSE beamforming

The MVDR beamformer is a popular multichannel processor that extracts s from $\mathbf{x}$, thereby reduces noise, enhances or even dereverberates the target signal [8]. Its output is $u = \mathbf{w}_{\mathrm{MVDR}}^H \mathbf{x}$ where

$$\mathbf{w}_{\mathrm{MVDR}} = \frac{\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{g}}{\mathbf{g}^H \mathbf{C}_{\mathbf{y}}^{-1}\mathbf{g}}. \quad (2)$$

Here, $\mathbf{C}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$ is the covariance matrix of the noise signal $\mathbf{y}$, $\mathrm{E}[\cdot]$ stands for the expectation operator, and $\cdot^*$ and $\cdot^H$ denote the conjugate value and the conjugate transpose, respectively.

The beamformer can be followed by a Wiener postfilter that attenuates the residual noise $y_{\mathrm{res}} = \mathbf{w}_{\mathrm{MVDR}}^H \mathbf{y}$ in the output of MVDR. The whole operation is equivalent to the Minimum Mean Square Error (MMSE) beamforming [1] and is given by

$$\mathbf{w}_{\mathrm{MMSE}} = \mathbf{w}_{\mathrm{MVDR}} \underbrace{\frac{\mathrm{E}[|u|^2] - \mathrm{E}[|y_{\mathrm{res}}|^2]}{\mathrm{E}[|u|^2]}}_{\text{Wiener postfilter}}. \quad (3)$$

To apply MMSE and MVDR efficiently in practice, it is crucial to estimate $\mathbf{C}_{\mathbf{y}}$, $\mathbf{g}$ and y_{res} with a sufficient accuracy.

2.3. Previous MVDR implementations for CHiME3

In [4, 5], $\mathbf{C}_{\mathbf{y}}$ is estimated with the aid of trained DNNs that compute frequency-dependent speech presence probabilities. The probabilities are used to control the noise covariance update so that the update is suspended during the speaker activity and vice versa. Then, the steering vector is estimated as the principal vector of the target covariance, which is estimated as the difference between the covariance of input signals $\mathbf{C} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ and that of noise $\mathbf{C}_{\mathbf{y}}$.

The principal vector can be significantly biased in low SNR conditions. In the frequency bands where the target signal is not active, a vector steered towards another directional (interfering) source can be obtained instead. The above noise covariance estimation is not effective in two aspects. First, the computation of masks requires to pass data through a large DNN with as many outputs as is the number of frequency bins, which is computationally expensive. Second, the noise covariance should be updated continuously, also during the speaker activity, when the noise is nonstationary. The methods we propose here aims to overcome these drawbacks.

2.4. Filter-and-sum beamforming

The computation of the inversion matrix in (2) increases the computational burden and makes the MVDR (MMSE) beamformer sensitive to estimation errors. Once the steering vector $\mathbf{g}$ is estimated, a method that is less sensitive to possible errors and does not require the knowledge (estimation) of $\mathbf{C}_{\mathbf{y}}$ is represented by

$$\mathbf{w}_{\mathrm{FSB}} = \frac{1}{m}(\mathbf{g}^{-1})^*, \quad (4)$$

where $\mathbf{g}^{-1}$ contains the reciprocal values of the elements of $\mathbf{g}$. This method, in fact, performs a filter-and-sum beamforming (FSB) that is a generalization of the DSB for reverberant environments. Indeed, in the free-field conditions, the FSB coincides with the DSB, because the elements of $\mathbf{g}$ correspond to pure delay filters, and $\mathbf{g}^{-1}$ are their respective inverse delays.

The FSB can be followed by the Wiener postfilter similarly to (3) if any estimate of the residual noise in the FSB output (i.e., an estimate of $\mathbf{w}_{\mathrm{FSB}}^H \mathbf{y}$) is available.

3. Front-End

In this section, details of four different systems for multichannel speech enhancement are described. Each system is a combinations of a VAD and of a beamformer.

Two VADs are considered where both are designed through trained DNNs. One VAD performs a detailed speech presence detection, that is, within each frequency bin. The other VAD performs only the per-frame detection. The VADs are used to estimate $\mathbf{g}^{-1}$ using the method from [7].

Then, two beamformers are considered: A variant of the approximate MMSE beamformer described in [6], and the simpler FSB, which was described above.

The processing of signals proceeds in the short-time Fourier (STFT) domain where the window length is 512 samples and the frame shift is 128 samples. The systems operate in a batch-online processing regime. Each batch of 100 STFT frames is processed independently in the following steps.

1. The input channels are selected based on their time domain correlation coefficients. Specifically, for the ith channel, the maximal correlation coefficient with the other channels is computed; let us denote the value μ_i. If this value is smaller than a threshold, the channel is not used. However, at least two channels are kept for further processing (the channels with maximum μ_i).
2. The reference channel is CH5 unless it has been withdrawn in the previous step. If yes, the channel with the maximum μ_i is selected.
3. VAD is applied to the selected channels.
4. The steering vector $\mathbf{g}$ as well as $\mathbf{g}^{-1}$ are assumed to be approximately constant within the batch of frames. The elements of $\mathbf{g}^{-1}$, that is, the respective RTFs related to the reference channel, are estimated using the estimator from [7] where speech presence probabilities are replaced by the outputs of VAD.
5. A given beamformer is applied. Its output is transformed back to the time domain using the inverse Fourier transform and overlap-add.

3.1. VAD using DNN

We consider two VADs: The first detector, referred to as sVAD, yields the speech activity over every frame of the processed signal. The second one detector, referred to as dVAD, estimates the speech activity for every frequency bin and every signal frame. Both VADs are implemented as DNNs trained using the Torch framework[1]. Training as well as testing sets were created from the CHiME4 training data.

sVAD is trained to estimate Wiener gains (values between 0 and 1). Each STFT frame is represented by raw magnitude of the 40 mel filter bank features (which are not decorrelated). The input feature vector concatenates the analyzed frame, 10 frames

[1] http://torch.ch

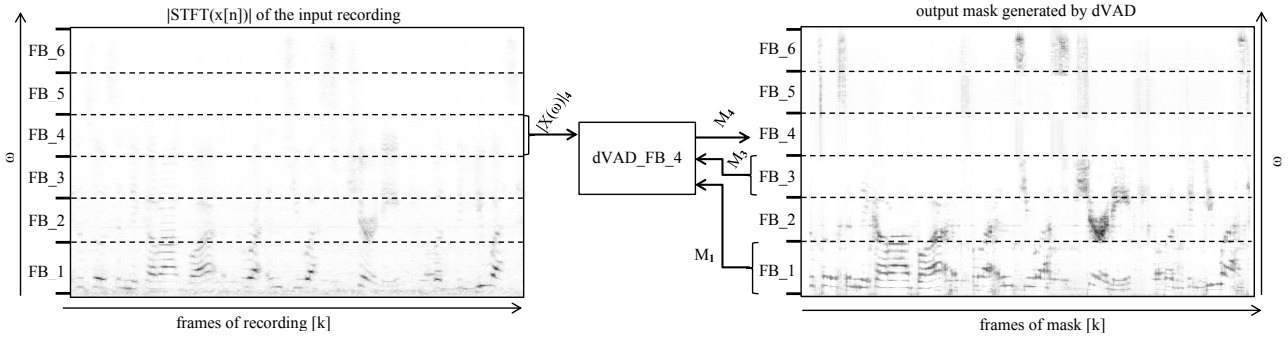

Figure 1: Illustration of the data flow of dVAD_FB_4 (white–black color scale refers to 0–maximum values)

before and 2 after it. Global zero-mean and unit-variance normalization is applied (computed from the training data).

sVAD consists of 5 hidden layers (3x256 and 2x128 neurons, respectively) all with sigmoid activation function. Binary Cross Entropy criterion was optimized using 1024 minibatches and finished within 50 epochs. No pre-training or dropout was used, data order was randomized every epoch.

dVAD consists of 6 smaller DNNs, referred to as dVAD_FB_1,...,dVAD_FB_6. Each DNN has one of six frequency bands (FB_1,...,FB_6) on its input together with reduced outputs of the previous DNNs. For example, the input of dVAD_FB_4 is illustrated in Figure 1.

The output of each DNN is a vector of values from the interval $[0; 1]$ containing the speech presence probabilities for the respective frequency band and frame. The reduced outputs (used on the inputs of the other DNNs) contain averages over 10 neighboring bins. For a given frequency bin, the training output label is zero if the SNR for the frequency is smaller than 5 dB. Otherwise, the label is set to one.

The structure of dVAD is computationally cheaper by about 50% as compared to a VAD that resides in a big DNN that computes the speech probabilities in all frequency bins simultaneously. dVAD_FB_1,...,dVAD_FB_6 were trained subsequently. Therefore, zero mean and unit variance normalization of the input data was applied between the training of each DNN.

Each dVAD_FB_x consists of 5 hidden layers (2x350, 256 and 2x128 neurons, respectively) all with ReLU activation function. For the kth frame, the context of frames $k-8$, $k-6$, $k-4$, $k-2$, $k+2$ and $k+4$ is used. Mean Square Error criterion is optimized within 1024 minibatches. No pre-training was applied; training data order was randomized. The training was finished between epochs 54 and 60.

3.2. Approximate MMSE beamformer

We implement the MMSE beamformer as an approximate MVDR followed by the Wiener post-filter. The MVDR part exploits a blocking matrix to obtain noise reference signals. The blocking matrix is defined as (without any loss on generality, assume that the reference channel is CH1)

$$\mathbf{B} = \begin{pmatrix} -1 & g_2^{-1} & 0 & \dots & 0 \\ -1 & 0 & g_3^{-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \dots & g_m^{-1} \end{pmatrix}, \tag{5}$$

where g_i^{-1} denotes the ith element of $\mathbf{g}^{-1}$. The noise reference signal is obtained by passing the input through the blocking matrix, that is,

$$\mathbf{u} = \mathbf{B}\mathbf{x}, \tag{6}$$

however, this signal is different from the noise term $\mathbf{y}$ in (1). Since the beamformer operates with a batch of frames, the least-square estimate of $\mathbf{y}$ using $\mathbf{u}$ can be computed as

$$\widehat{\mathbf{y}} = \mathbf{C}\mathbf{B}^H(\mathbf{B}\mathbf{C}\mathbf{B}^H)^{-1}\mathbf{B}\mathbf{x}, \tag{7}$$

where $\mathbf{C} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ is replaced by its sample mean estimate. The estimator (7) is scale-invariant in the sense that any scaling substitution $\mathbf{B} \leftarrow \mathbf{\Lambda}\mathbf{B}$ where $\mathbf{\Lambda}$ is regular does not have any influence on $\widehat{\mathbf{y}}$. In particular, this property is useful when $\mathbf{B}$ is derived using blind methods such as Independent Component Analysis (ICA) that can estimate $\mathbf{B}$ only up to the unknown scaling factor $\mathbf{\Lambda}$; see, e.g., [9].

The covariance of $\widehat{\mathbf{y}}$ is equal to

$$\mathbf{C}_{\widehat{\mathbf{y}}} = \mathrm{E}[\widehat{\mathbf{y}}\widehat{\mathbf{y}}^H] = \mathbf{C}\mathbf{B}^H(\mathbf{B}\mathbf{C}\mathbf{B}^H)^{-1}\mathbf{B}\mathbf{C}. \tag{8}$$

In the approximate MVDR, the strategy is to replace $\mathbf{C}_{\mathbf{y}}$ in (2) by $\mathbf{C}_{\widehat{\mathbf{y}}}$. The steering vector $\mathbf{g}$ can be computed directly from $\mathbf{g}^{-1}$; an alternative approach is to compute $\mathbf{g}$ as a vector from the null space of $\mathbf{B}$.

Since the rank of $\mathbf{C}_{\widehat{\mathbf{y}}}$ is $m-1$, its inversion matrix does not exist. We therefore replace $\mathbf{C}_{\widehat{\mathbf{y}}}^{-1}$ by the Moore-Penrose pseudoinverse denoted as $\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}$. Then, the approximate MVDR beamformer is represented by

$$\widehat{\mathbf{w}}_{\mathrm{MVDR}} = \frac{\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}\mathbf{g}}{\mathbf{g}^H\mathbf{C}_{\widehat{\mathbf{y}}}^{\dagger}\mathbf{g}}. \tag{9}$$

In case that the target channel is different from the reference channel, the scale-invariant least-squares can be applied as in (7). Then, all enhanced channels can be obtained as $\widehat{\mathbf{W}}_{\mathrm{MVDR}}\mathbf{x}$ where

$$\widehat{\mathbf{W}}_{\mathrm{MVDR}} = \frac{\mathbf{C}\widehat{\mathbf{w}}_{\mathrm{MVDR}}(\widehat{\mathbf{w}}_{\mathrm{MVDR}})^H}{(\widehat{\mathbf{w}}_{\mathrm{MVDR}})^H\mathbf{C}\widehat{\mathbf{w}}_{\mathrm{MVDR}}}. \tag{10}$$

From now on, let $\widehat{\mathbf{w}}_{\mathrm{MVDR}}$ denote the approximate MVDR for the selected target channel. Let the output be denoted as $v = \widehat{\mathbf{w}}_{\mathrm{MVDR}}^H\mathbf{x}$.

Using (7), the residual noise in the output can be estimated as

$$r = \widehat{\mathbf{w}}_{\mathrm{MVDR}}^H\widehat{\mathbf{y}}. \tag{11}$$

According to (3), the gain of the Wiener postfilter can be approximated as

$$G(k,\ell) = \frac{\max\{|v(k,\ell)|^2 - |r(k,\ell)|^2, \epsilon\}}{|v(k,\ell)|^2 + \epsilon}, \tag{12}$$

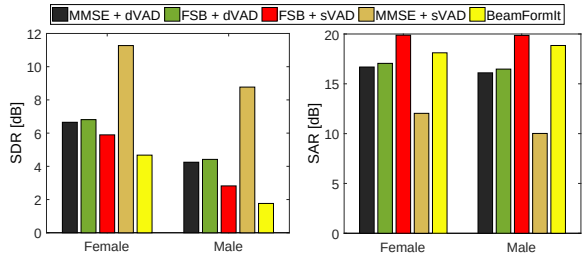

Figure 2: Results of the objective evaluation experiment in terms of SDR and SIR. The results were averaged over the four noisy environments BUS, CAF, STR and PED.

where ϵ is a small positive constant that prevents from division by zero. The final output of the approximate MMSE is

$$\widehat{s}(k,\ell) = G(k,\ell)v(k,\ell). \tag{13}$$

It is worth noting that $G(k,\ell)$ can be modified in various heuristic ways before it is applied in (13). In CHiME-4, we set $G(k,\ell) = 1$ for k corresponding to frequencies higher than 3 kHz. By contrast, for the frequencies below 100 Hz, $G(k,\ell) = 0.01$. The gain could be also modified according to the output of the VAD. For example, if for given k the VAD yields speech probability higher than 0.5, we set $G(k,\ell) = 1$ to avoid the distortion of the speech in the system output.

4. Back-End Solutions

For the experimental evaluation, we consider two automatic speech recognition back-ends:

1. the baseline DNN+RNNLM back-end [10] provided by CHiME4 organizers, and
2. the same back-end with a re-trained acoustic model.

The front-end processing usually introduces additional artifacts into the processed speech signals, which are unknown to the acoustic model trained on the unprocessed signals. This may lead to a deterioration of the performance of the ASR system and motivates us to adapt the acoustic model for the given front-end. This is done as follows. The training set is enhanced by the front-end processor, by which a new training set is obtained. Then, this set is used by the training procedure of the baseline DNN models, which results in an adapted acoustic model.

5. Experiments and Results

5.1. Objective evaluation

Here, we describe an experiment where the proposed multichannel enhancement (front-end) systems are compared with BeamformIt in terms of signal separation criteria from BSS_Eval [11][2]. Two utterances were selected from the development set: `F01_421C0201` (a female speaker) and `M04_052C0112` (a male speaker). Four simulated (SIMU) noisy variants of each utterance (BUS, CAF, STR and PED) were processed by the enhancement systems. The outputs were evaluated in terms of Signal-to-Distortion Ratio (SDR) and Signal-to-Artefact Ratio (SAR). The results in terms of Signal-to-Interference Ratio (SIR) were similar to SDR, but we do not show them to save space. Averaged SDR and SAR over the environments are shown in Figure 2.

The proposed systems outperform BeamformIt in terms of SDR and SIR, which confirms their advanced ability to enhance the signal. The best SDR was achieved by MMSE+sVAD. On the other hand, the results in terms of SAR show that the proposed systems tend to introduce more artifacts into the enhanced signal. Only FSB+sVAD yields higher SAR than BeamformIt. The worst SAR yields MMSE+sVAD, which is the compromise for the high SDR and SIR.

5.2. CHiME4

Now we present the speech recognition results achieved by 10 systems. Each proposed ASR system is denoted by $A(B)$ where A denotes the front-end system, e.g., MMSE:sVAD, and B denotes the acoustic model used within the baseline ASR system, which is either "Base" (original model) or "Adapt" (the model adapted to the front-end). The case when the CHiME4 data are sent directly to the baseline ASR without any processing is denoted as "Unprocessed".

The resulting absolute Word Error Rates (WER) are shown in Table 1. Detailed results of FSB:sVAD(Base) and of the baselines for different noisy environments are presented in Table 2.

Comparing the proposed front-end systems, those using the FSB beamformer yield superior results compared to those with MMSE. The difference in simulated sets is about 2-3% WER. In case of the real-worlds recordings, the difference is up to 9%.

The choice of the VAD does not appear to have much influence on the final WER, especially in the combination with FSB. Considering the MMSE beamforming, the dVAD improves the WER compared to sVAD by 0-6%.

The adaptation of the acoustic models appears to be beneficial for the systems with MMSE, where it improves the performance by 0-2%. On the other hand, the re-training did not bring any significant improvement for the FSB technique.

Table 1: Absolute WER (%) averaged over four environments for the 6-channel track. The best achieved results are written in bold.

System	Dev		Test	
	real	simu	real	simu
Unprocessed (Base)	9.83	8.86	19.90	10.79
BeamformIt (Base)	**5.77**	**6.76**	**11.52**	10.91
MMSE:sVAD (Base)	10.91	9.31	22.39	9.72
MMSE:sVAD (Adapt)	10.56	9.21	20.61	9.11
MMSE:dVAD (Base)	7.78	9.84	16.27	9.68
MMSE:dVAD (Adapt)	7.89	9.28	16.09	9.40
FSB:sVAD (Base)	7.26	7.23	13.48	**7.70**
FSB:sVAD (Adapt)	7.23	7.68	13.46	7.95
FSB:dVAD (Base)	7.09	8.00	13.48	7.85
FSB:dVAD (Adapt)	7.43	8.24	14.40	8.16

6. Conclusions

From the results of our experiments we conclude that, among the proposed systems, FSB:sVAD(Base) appears to be the most effective for CHiME4. It achieves WER between 7%-13%, which improves the WER achieved on unprocessed data by about 1.5%-6.5%. The system is computationally simple, because the FSB does not use the matrix pseudo-inversion in (9), and the sVAD performs the computationally save per-frame de-

[2]We use version 2.3 of BSS_Eval, which contains `bss_decomp_tvfilt.m`, a function that enables us to evaluate time-variant mixtures.

Table 2: Absolute WER (%) per environment. The best achievements are written in bold.

(a) FSB:sVAD (Base)

Envir.	Dev		Test	
	real	simu	real	simu
BUS	10.27	6.21	22.21	**5.68**
CAF	6.55	9.73	12.59	**8.91**
PED	4.57	5.52	10.71	**6.85**
STR	7.54	7.46	**8.40**	**9.34**

(b) BeamformIt (Base)

Envir.	Dev		Test	
	real	simu	real	simu
BUS	**7.43**	**5.97**	**16.88**	7.66
CAF	**5.77**	**8.13**	**10.20**	11.52
PED	**3.73**	**5.47**	**9.87**	10.35
STR	**6.15**	**7.45**	9.13	14.12

(c) Unprocessed (Base)

Envir.	Dev		Test	
	real	simu	real	simu
BUS	16.06	10.07	33.17	9.58
CAF	8.44	10.59	19.22	11.95
PED	5.44	6.34	14.63	9.64
STR	9.38	8.44	12.61	11.97

tection. The method achieves the best WER over the compared systems in the simulated test set.

For the other sets, in particular in the real-world sets, the best WER was achieved with BeamformIt. The experiment of Section 5.1 has demonstrated on typical simulated recordings that BeamformIt achieves lower SDR as well as lower SAR compared to FSB:sVAD. While the simulated recordings are sufficiently linear and do not contain microphones failures, the real-world recordings of CHiME4 do. We therefore attribute the better WER achieved by BeamformIt in real-world sets to its robustness against nonlinear effects rather than to its ability to enhance the target signal.

The improvement of the proposed methods in terms of the robustness against microphone failures and other nonlinearities is the subject of our future progress.

7. Acknowledgments

We thank Francesco Nesta and Trausti Thormundsson for helpful discussions. This work was supported by The Czech Science Foundation through Project No. 14-11898S and by California Community Foundation through Project No. DA-15-114599.

8. References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. Wiley, 2004.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.

[3] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[5] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and mvdr beamforming for meeting recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 385–389.

[6] Z. Koldovský and F. Nesta, "Approximate mvdr and mmse beamformers exploiting scale-invariant reconstruction of signals on microphones," in *Acoustic Signal Enhancement (IWAENC), 2016 15th International Workshop on*, September 2016.

[7] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sept 2004.

[8] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multimicrophone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, Feb 2015.

[9] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, May 2009.

[10] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, to appear.

[11] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.