MULTI-CHANNEL SPEECH PROCESSING ARCHITECTURES FOR NOISE ROBUST SPEECH RECOGNITION: 3RD CHIME CHALLENGE RESULTS

Lukas Pfeifenberger, Tobias Schrank, Matthias Zöhrer, Martin Hagmüller, Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

lukas.pfeifenberger@alumni.tugraz.at,
{tobias.schrank,matthias.zoehrer,hagmueller,pernkopf}@tugraz.at

ABSTRACT

Recognizing speech under noisy condition is an ill-posed problem. The CHiME 3 challenge targets robust speech recognition in realistic environments such as street, bus, caffee and pedestrian areas. We study variants of beamformers used for pre-processing multi-channel speech recordings. In particular, we investigate three variants of generalized sidelobe canceller (GSC) beamformers, i.e. GSC with sparse blocking matrix (BM), GSC with adaptive BM (ABM), and GSC with minimum variance distortionless response (MVDR) and ABM. Furthermore, we apply several postfilters to further enhance the speech signal. We introduce MaxPower postfilters and deep neural postfilters (DPFs). DPFs outperformed our baseline systems significantly when measuring the overall perceptual score (OPS) and the perceptual evaluation of speech quality (PESQ). In particular DPFs achieved an average relative improvement of 17.54% OPS points and 18.28% in PESQ, when compared to the CHiME 3 baseline. DPFs also achieved the best WER when combined with an ASR engine on simulated development and evaluation data, i.e. 8.98% and 10.82% WER. The proposed MaxPower beamformer achieved the best overall WER on CHiME 3 real development and evaluation data, i.e. 14.23% and 22.12%, respectively.

***Index Terms*—** multi-channel speech processing, deep postfilter, automatic speech recognition

1. INTRODUCTION

Background noise is the primary source of performance degradation in speech recognition systems. While the capabilities of single-channel speech pre-processing are limited, multi-channel systems exploit the spatial information of the sound field and usually achieve better speech recognition results. Adaptive beamforming is a widely used technique for multi-channel pre-processing of speech as alternative to blind source separation approaches. For a sufficient amount of noise reduction, beamformers are generally used in conjunction with a postfilter.

The aim of the 3rd CHiME challenge is to develop a multi-channel speech recognition system [1], where we encounter multi-channel recordings of a speaker located in the near-field, embedded in mostly far-field noise. The setup covers different speakers, noise environments, and real-world problems like microphone failure, clipping, and other recording glitches.

In this paper, we present a multi-channel speech enhancement system which tries to cope with these conditions: First, we detect recording glitches using the prediction error of an auto-regressive model. Then, we estimate the position of the speaker relative to the microphone array using our *direction-dependent signal-to-noise ratio* (DD-SNR) algorithm [2], which also provides a sufficiently accurate *voice activity detection* (VAD). The speaker position is used to obtain a steering vector for a *generalized sidelobe canceller* (GSC) beamformer, which we implemented in three different variants. We also present two novelties here: Firstly we introduce a *MaxPower* postfilter (PF), leading to the best speech recognition result on CHiME 3 real data. Secondly we present deep neural PFs – deep neural networks attached to beamformers, improving the overall perceptual quality (OPS) of the target speech significantly and also outperforming baseline systems on simulated data. This front-end, i.e. the three beamformer variants and different PFs, are empirically evaluated using the PESQ and the OPS measures [3].

In the back-end, we use two speech recognition systems based on the Kaldi toolkit [4]. The first is a GMM system which makes extensive use of feature transformations as this was shown to provide good results for distant talk speech recognition [5]. The second is a DNN system that employs pre-training with restricted Boltzmann machines, cross entropy training and state-level minimum Bayes risk training [1]. Our best model, i.e. MaxPower PF with a GMM backend, reduces word error rate (WER) from 37.61% for the baseline enhancement system to 22.12% (41% relative improvement) on the real evaluation set.

The outline of the paper is as follows: In Section 2 we introduce the architecture of the proposed system. Section 3 details the multi-channel speech processing approaches including the proposed beamformers. PFs are introduced in Section 4 while the PESQ and PEASS scores of the front-end are summarized in Section 6.1. The ASR system is presented in Section 5 and the results are discussed in Section 6.2. Section 7 concludes the paper.

2. SYSTEM OVERVIEW

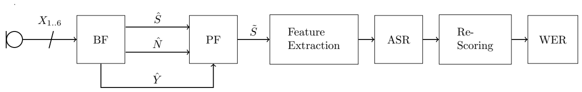

Fig. 1. System overview.

Figure 1 shows the setup of the components of the proposed ASR system. Speech estimate $\hat{S}$, the noise estimate $\hat{N}$ and the beamformer output $\hat{Y}$ are fed into a postfilter predicting an enhanced speech estimate $\tilde{S}$. After feature extraction the signal is fed into the ASR. Next, Language model re-scoring is applied and then the final word error rate (WER) is calculated.

3. MULTI-CHANNEL SPEECH PROCESSING

The input signal vector $\boldsymbol{X}$ of the 6 microphone channels is written as

$$\boldsymbol{X}(k,l) = \boldsymbol{A}(k,l)S(k,l) + \boldsymbol{N}(k,l), \tag{1}$$

where S is the speech signal, $\boldsymbol{N}$ is the noise part of the 6-channel input signal in frequency-domain, k and l denote the frequency bin and time frame, respectively, and $\boldsymbol{A}(k,l)$ denotes the *acoustic transfer function* (ATF) from the true speaker position to each microphone. In this challenge, additional information is supplied by the *noise context*, a short section of noise-only signal before each utterance. The noise context for each utterance is referenced in annotations provided by the challenge organizers. This allows to estimate the spatial noise correlation matrix $\boldsymbol{\Phi}_{NN}$, which is given as

$$\boldsymbol{\Phi}_{NN}(k,l) \triangleq E\{\boldsymbol{N}^H(k,l)\boldsymbol{N}(k,l)\}, \tag{2}$$

where $E\{\cdot\}$ denotes the expectation operation and $\{\cdot\}^H$ the Hermitian transpose.

We found that the noise context contains speech in some utterances, which would cause speech cancellation in a beamformer. We therefore decided to adaptively estimate $\boldsymbol{\Phi}_{NN}$ by using VAD.

3.1. Failed Channel Detection

The above signal model requires signals which strictly adhere to the linear time-invariant theory. Clearly, errors such as recording glitches, amplitude variations, time shifts or total signal loss must be detected before multi-channel speech enhancement such as beamforming. In particular, we noticed that especially channel 4 and 5 exhibit rather complex recording glitches in about 15% of all isolated recordings. To address these problems, a mere energy threshold may not suffice. We therefore employed an auto-regressive linear predictive coding (LPC) on each channel c in time-domain [6, 7], and used the predictor error $e(t)$ as criterion whether a channel is considered as failed, i.e.

$$e(t) = x_c(t) - \sum_{m=1}^{M} x_c(t-m)a(m), \tag{3}$$

where $a(m)$ are LPC coefficients and $M = 100$. A channel $x_c(t)$ is considered as failed if the power of its predictor error $e(t)$ lies outside the $\pm 10dB$ corridor around the median of the energy of the predictor errors of all channels. If a failed channel is detected this channel is not used for further processing.

3.2. Direction Of Arrival Estimation

For successful beamforming an accurate *direction of arrival* (DOA) estimation is required. Therefore, the *steered response power phase transform* (SRP-PHAT) [8] algorithm has been already provided for this purpose. But it lacks a proper VAD estimate, which might also be useful for estimating the spatial noise correlation matrix $\boldsymbol{\Phi}_{NN}$ during speech pauses. For this purpose, we used our DD-SNR algorithm [2], which provides a direction-dependent a-priori SNR $\xi_\tau(k,l)$ under the assumption of an ideal, spherical noise sound field, i.e.

$$\begin{aligned}\xi_\tau(k,l) = \mathrm{Tr}([&\boldsymbol{\Gamma}_{\boldsymbol{XX}}(k,l) - \boldsymbol{A}_\tau(k,l)\boldsymbol{A}_\tau^H(k,l)]^{-1} \\ &\cdot [\boldsymbol{\Gamma}_{\boldsymbol{NN}}(k) - \boldsymbol{\Gamma}_{\boldsymbol{XX}}(k,l)]),\end{aligned} \tag{4}$$

where the DD-SNR ξ is also used as VAD, τ is the relative *time difference of arrival* (TDOA) between all microphone pairs, $\boldsymbol{A}_\tau$ the corresponding ATFs, $\boldsymbol{\Gamma}_{\boldsymbol{XX}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{NN}}$ are the spatial coherence matrices [2] for the multi-channel signals $\boldsymbol{X}$ and noise-only components $\boldsymbol{N}$. The interested reader is referred to [2] for more details.

The optimal TDOA τ also maximizes ξ_τ. It can be detected for each time frame l by searching over a small set of possible delays using $\tau_{OPT}(l) = \arg\max_\tau \frac{1}{K}\sum_{k=0}^{K} \xi_\tau(k,l)$. We quantize τ into 13 equally spaced segments which is sufficient for each microphone pair and the given aperture.

3.3. Beamforming

After evaluating a wide variety of beamforming and multi-channel speech enhancement algorithms [9–13], we decided to use the *general sidelobe canceller* (GSC) [14]. The main

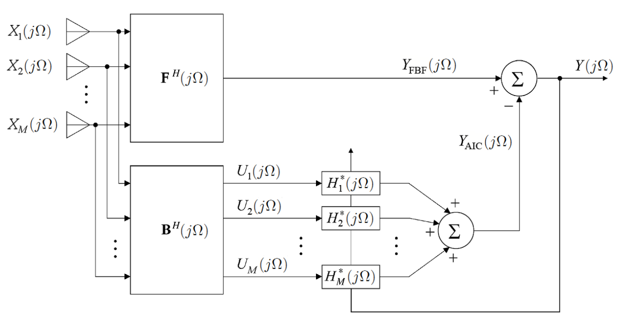

Fig. 2. Block diagram of the generalized sidelobe canceller.

reasons are its observed empirical performance and robustness for the given problem.

The entire beamformer can be expressed as

$$\boldsymbol{W}(k,l) = \boldsymbol{F}(k,l) - \boldsymbol{H}(k,l)\boldsymbol{B}(\boldsymbol{k},\boldsymbol{l}) \tag{5}$$

using the *fixed beamformer* (FBF) $\boldsymbol{F}$, the *adaptive interference canceler* (AIC) $\boldsymbol{H}$, and the *blocking matrix* (BM) $\boldsymbol{B}$. In particular, we implemented the following three GSC variants detailed in the following sub-sections. Details can be found in [2, 15].

3.3.1. GSC with sparse BM

This variant is the standard GSC, as depicted in Figure 2. The FBF is given as $\boldsymbol{F}(k,l) = \frac{\boldsymbol{A}(k,l)}{\boldsymbol{A}^H(k,l)\boldsymbol{A}(k,l)}$. The BM is defined as [16]

$$\boldsymbol{B}(k,l) = \begin{bmatrix} -\frac{A_2^*(k,l)}{A_1^*(k,l)} & -\frac{A_3^*(k,l)}{A_1^*(k,l)} & \cdots & -\frac{A_M^*(k,l)}{A_1^*(k,l)} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{6}$$

with $M = 6$ channels, and channel 1 as *reference microphone*. The asterix in (6) denotes the conjugate complex coefficient. We used the channel with the highest signal energy as reference in our implementations. The AIC $\boldsymbol{H}$ is a non-causal adaptive filter.

3.3.2. GSC with adaptive Blocking Matrix (ABM)

This variant features an adaptive BM presented in Figure 3. The columns of the ABM are designed as non-causal adaptive filters and the coefficients are determined via the *normalized least mean squares* (NLMS) approach [17].

3.3.3. GSC with MVDR and ABM

It is possible to estimate the spatial noise correlation matrix $\boldsymbol{\Phi}_{NN}$ during speech pauses using the DD-SNR from Section 3.2 as VAD. Hence, the GSC may be replaced with the

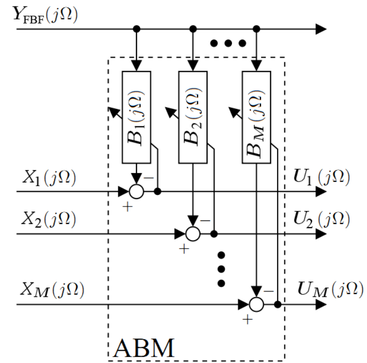

Fig. 3. Block diagram of the adaptive blocking matrix.

minimum variance distortionless response (MVDR) solution [18, 19] given as:

$$\boldsymbol{F}(k,l) = \frac{\boldsymbol{\Phi}_{NN}^{-1}(k,l)\boldsymbol{A}(k,l)}{\boldsymbol{A}^H(k,l)\boldsymbol{\Phi}_{NN}^{-1}(k,l)\boldsymbol{A}(k,l)}. \tag{7}$$

This has already been provided in the baseline enhancement system, however, the estimate $\boldsymbol{\Phi}_{NN}$ may be inaccurate, therefore we only replaced the FBF in Figure 2 with the MVDR solution. This allows for additional noise removal by the ABM and AIC.

4. POSTFILTERING

4.1. MaxPower postfilter

Our first postfilter is based on the GSC with MVDR and ABM. Similar to [15], the beamformer output $Y(k,l)$ is back-projected to the microphones using the ATFs $\boldsymbol{A}(k,l)$. This way, the microphone inputs $\boldsymbol{X}$ can be split into their speech and noise components $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{N}}$:

$$\begin{aligned} \hat{\boldsymbol{S}}(k,l) &= \boldsymbol{A}(k,l)Y(k,l) \\ \hat{\boldsymbol{N}}(k,l) &= \boldsymbol{X}(k,l) - \boldsymbol{A}(k,l)Y(k,l) \end{aligned} \tag{8}$$

The final output of this method is chosen to be the maximum energy of $|\hat{\boldsymbol{S}}(k,l)|^2$ for each frequency bin k and time frame l. As the phases of $\hat{\boldsymbol{S}}(k,l)$ do not match, there would be no reconstruction back into time domain. To circumvent this limitation, each channel in $\hat{\boldsymbol{S}}(k,l)$ has been aligned to the geometric origin of the setup.

4.2. Multi-Channel postfilter

As second postfilter we used our parametric multi-channel Wiener filter (PMWF) proposed in [2]. With the noise PSD matrix $\boldsymbol{\Phi}_{NN}$ being already available, estimating the residual noise power in the beamformer becomes straightforward.

With the beamforming filter $\boldsymbol{W}$, the residual noise power in the beamformer output is given as

$$\Phi_{Y_N Y_N}(k,l) \triangleq E\{\boldsymbol{W}^H(k,l)\boldsymbol{\Phi}_{NN}(k,l)\boldsymbol{W}(k,l)\}. \quad (9)$$

Together with the overall output power of the beamformer

$$\Phi_{YY}(k,l) \triangleq E\{\boldsymbol{W}^H(k,l)\boldsymbol{\Phi}_{XX}(k,l)\boldsymbol{W}(k,l)\} \quad (10)$$

the real-valued gain mask G is obtained as

$$G(k,l) = \frac{\zeta(k,l)}{1+\zeta(k,l)}, \quad (11)$$

where $\zeta(k,l) = \frac{\Phi_{YY}(k,l)}{\Phi_{Y_N Y_N}(k,l)} - 1$ can be identified as the output SNR. Further smoothing over time may be achieved using a spectral subtraction algorithm like the mean-square error log-spectral amplitude estimator [20].

4.3. Deep neural postfilter

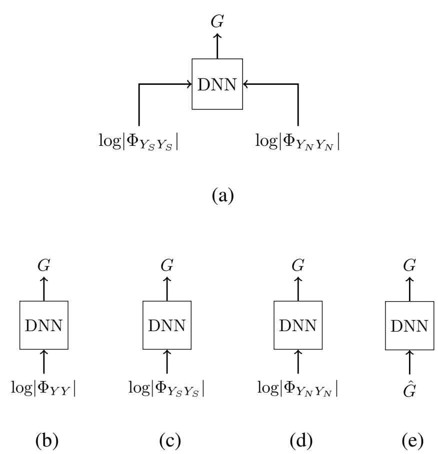

Fig. 4. Variants of deep postfilter models. A neural network maps the beamformed speech $\Phi_{Y_S Y_S}$, noise $\Phi_{Y_N Y_N}$ or estimated gain mask $\hat{G}$ to the optimal gain mask G. The first column shows the different combinations of various beamformer components (a-d), respectively.

In [21–24] deep neural networks (DNNs) were applied to single channel source separation, improving the overall quality of speech in terms of PESQ and OPS scores. In order to analyze the enhancement capabilities of DNNs for multichannel inputs, we introduce deep postfilter models: In particular, we use DNNs to map beamformed log-spectrogram outputs to the optimal gain mask G estimated from the close talking microphone (channel 0). Figure 4 shows variants of these postfilters using different beamformer components. In particular, model (a) uses concatenated beamformed speech log-spectrograms $\Phi_{Y_S Y_S}$ and noise log-spectrograms $\Phi_{Y_N Y_N}$

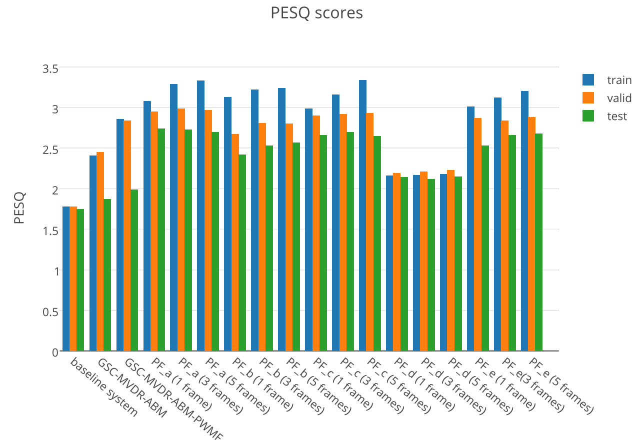

Fig. 5. PESQ scores of deep postfilter models (a-f).

as input. $\Phi_{Y_N Y_N}$ is estimated as in (9). $\Phi_{Y_S Y_S}$ can be caculated directly as $\Phi_{Y_S Y_S}(k,l) = \Phi_{YY}(k,l) - \Phi_{Y_N Y_N}(k,l)$. In case of the models (b-e) Φ_{YY}, $\Phi_{Y_S Y_S}$, $\Phi_{Y_N Y_N}$, or the estimated gain mask $\hat{G}$ were fed into the network. After training, mask estimates are applied to the output signal of the beamformer obtaining enhanced speech $\tilde{S}$ and noise estimates.

We trained 3 layer multi-layer perceptrons [25] with rectifier activation functions using a context window of $1, 3, 5$ frames and a MSE criteria on a subset of the CHiME 3 database. In particular we selected 400 utterances, 50 validation utterances and 50 test utterances from the simulated training corpus. Figure 5 and Figure 6 show the PESQ and OPS scores [3] of the postfilter (PF) models (a-e), respectively. For objective evaluation the estimated speech was compared to the output of the GSC with MVDR and ABM (with/without PMWF postfilter) and the baseline system. The best deep postfilter, i.e. PF variant a (PF$_a$), achieved an OPS score of 71.97, a validation score of 54.03 and a test OPS of 50.83. It outperforms the beamformed signal GSC-MVDR-ABM (with/without PMWF postfilter) as well as the provided CHiME 3 baseline system. Therefore, we further investigate this approach when applied to ASR.

5. ASR

Both ASR systems employed in this paper are based on the baseline system provided by the 3rd CHiME challenge [1]. The GMM system uses mel frequency cepstral coefficients (MFCC) as features which are input to a series of feature-space transformations. The features are in this order transformed by applying linear discriminant analysis, maximimum likelihood linear transformation and feature-space maximum likelihood linear regression. In addition, inter-speaker differences are compensated for by doing speaker-adaptive training. This pipeline proved to be highly competitive in

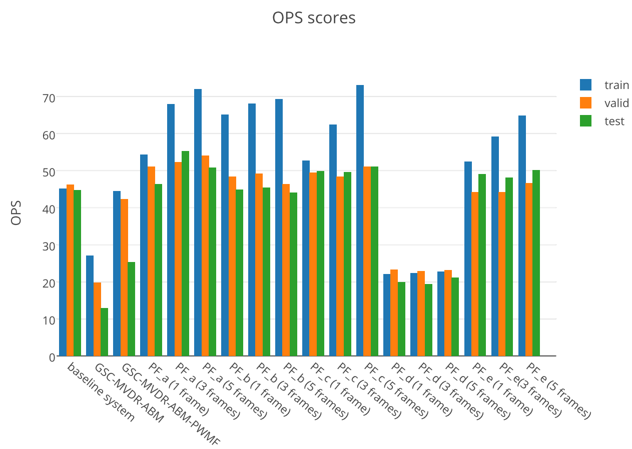

Fig. 6. OPS scores of deep postfilter models (a-f).

the CHiME 2 challenge [5]. The DNN system employs 40-dimensional filterbank features and is pre-trained using restricted Boltzmann machines with 6 hidden layers. The actual training stage of the DNN uses 4 hidden layers and also does cross entropy training. Finally, sequence discriminative training is performed using a state-level minimum Bayes risk criterion.

In the following sections, we describe the changes we made to the baseline system. These are to be found in the frontend and in the postprocessing stage.

5.1. Feature extraction

In contrast to the baseline which uses MFCC features, we additionally employ power-normalised cepstral coefficients (PNCC) [26]. For these features, we use a Hamming window with a window duration of 25 ms and a step size of 10 ms. Parallel to MFCCs, we extract 13 features and collect deltas and delta-deltas of these.

5.2. Rescoring

The postprocessing step features n-best list language model rescoring. For this, we collect the 36 best hypotheses for each utterance and reweight them with a class-based recurrent neural network language model (RNN-LM) [27]. The RNN-LM is trained on the official training data only and is configured to use a class size of 50.

6. RESULTS AND DISCUSSION

The data of the challenge and the recording setup is described in detail in [1]. The data is a collection of two sets of recordings: real data and simulated data. The first are speech recordings made in noisy environments. The second are clean recordings mixed with noise that has been recorded in the same noisy environments. The real recordings were made using 6 microphones custom-fitted to a tablet hand-held device. The recordings with this device were conducted in four different environments: on a bus (BUS), in a café (CAF), in a pedestrian area (PED), and at a street junction (STR). For real data, there is an additional channel recorded with a head-mounted close-talking microphone. This channel, however, may not be used directly for obtaining ASR results but is only to be used in training.

6.1. Preprocessing results

To evaluate our three beamformers, we used PESQ and OPS scores. Evaluation is performed against the close-talking microphone channel for the real data set, and against the WSJ corpus for the simulated data set. Tables 1 and 2 show the scores for our four beamformers, and the baseline enhancement system for comparison. Again the GSC-MVDR with ABM and deep postfilter (PF_a) outperforms the other beamformers in terms of OPS and PESQ scores. In particular the proposed system achieved an average relative improvement of 17.54% in OPS and 18.28% in PESQ compared to the baseline enhancement system.

	set	train	dev	eval
Baseline enhancement	simu	2.00	1.64	1.72
system	real	1.59	1.42	1.50
GSC with sparse BM,	simu	2.15	1.73	1.81
and PMWF	real	1.51	1.37	1.35
GSC with ABM,	simu	1.53	1.49	1.52
and PMWF	real	1.36	1.30	1.36
GSC with MVDR	simu	2.05	1.60	1.73
and ABM	real	1.60	1.45	**1.73**
GSC with MVDR	simu	**2.55**	**2.17**	**2.28**
and ABM, and PF_a	real	**1.73**	**1.56**	1.56
GSC with MVDR	simu	1.98	1.69	1.63
and ABM, and MaxPower PF	real	1.51	1.39	1.44

Table 1. PESQ scores for our beamformers with PFs and the baseline.

	set	train	dev	eval
Baseline enhancement	simu	54.80	44.22	47.31
system	real	44.66	40.98	31.48
GSC with sparse BM,	simu	59.64	46.99	46.77
and PMWF	real	38.69	33.05	29.04
GSC with ABM,	simu	48.61	43.84	43.71
and PMWF	real	43.16	42.81	**38.02**
GSC with MVDR	simu	52.4	45.87	47.18
and ABM	real	48.26	45.87	37.93
GSC with MVDR	simu	**63.94**	**53.83**	**54.53**
and ABM, and PF_a	real	**48.69**	**46.54**	37.72
GSC with MVDR	simu	56.08	44.82	44.48
and ABM, and MaxPower PF	real	47.18	44.90	36.96

Table 2. OPS scores for our beamformers with PFs and the baseline.

6.2. ASR results

Table 3 shows ASR results for the preprocessing methods presented in this paper. MaxPower outperforms all other proposed methods on the real development data and the real evaluation data (14.53% WER and 22.14% WER, respectively), whereas PF_a achieved the best ASR scores on simulated data, i.e. 8.98% and 10.82% on development and evaluation, respectively. When comparing MFCCs and PNCCs , on average, PNCCs lead to an improvement of 6.04% WER on the real evaluation set. Improvements vary, however, depending on noise environment and preprocessing. After language model rescoring, the scores for the real development set and the real evaluation set descrease slightly to 14.23% WER and 22.12% WER, respectively (see Table 4).

Due to time constraints, our results for the DNN-based ASR system are limited to MaxPower which achieves best results among GMM-based systems. While considerable improvements are gained for the system using MFCCs (-3.02% WER on real evaluation set), DNNs lead to increased WER for the system using PNCCs ($+2.03\%$ WER on real evaluation set).

		development		evaluation	
	features	real	simu	real	simu
Baseline	MFCC	20.38	9.72	37.61	11.10
GSC sparse BM	MFCC	26.14	10.39	44.01	12.75
GSC ABM	MFCC	15.66	20.15	36.39	79.05
+ MVDR	MFCC	16.78	10.16	27.45	11.47
+ PF_a	MFCC	17.93	**8.98**	27.72	**10.82**
+ MaxPower	MFCC	15.70	10.77	25.22	14.86
+ DNN	FBANK	14.54	9.52	22.20	15.67
Baseline	PNCC	18.99	11.14	31.57	12.15
GSC sparse BM	PNCC	22.32	11.17	36.98	13.87
GSC ABM	PNCC	15.60	21.96	34.02	77.47
+ MVDR	PNCC	16.34	11.01	24.55	12.69
+ PF_a	PNCC	16.77	10.64	25.58	12.37
+ MaxPower	PNCC	**14.53**	12.05	**22.14**	15.08
+ DNN	FBANK	15.79	10.42	24.17	16.72

Table 3. ASR results for our beamformers and the baseline enhancement system.

	development		evaluation	
environment	real	simulated	real	simulated
BUS	16.17	10.52	29.00	12.46
CAF	13.78	13.97	24.04	15.61
PED	11.73	9.53	19.75	14.81
STR	15.26	13.38	15.69	16.64
AVG	14.23	11.85	22.12	14.88

Table 4. Detailed results for single best system, MaxPower using PNCC features and RNN language model rescoring.

7. CONCLUSION

We presented a comparison of different beamformers and postfilters applied to the CHiME 3 speech database. We studied three variants of GSC beamformers, i.e. GSC with sparse blocking matrix (BM), GSC with adaptive BM (ABM), and GSC with minimum variance distortionless response (MVDR) and ABM. In addition we investigated three postfilters (PF), a MaxPower PF, a parametric multi-channel Wiener filter, and a deep neural PF. The proposed ASR systens use either MFCC or PNCC features calculated from the the preprocessed signals which are fed into GMM or DNN-based systems. Finally n-best list re-scoring, using a recurrent neural network (RNN) language model, was applied.

We evaluated the overall perceptual score (OPS), and perceptual evaluation of speech quality (PESQ) of the proposed beamformers and postfilters. Deep neural postfilters using an GSC-MVDR-ABM beamformer outperformed other BF systems significantly, achieving an average relative improvement of 17.54% in OPS and 18.28% in PESQ compared to the baseline system. However, improvements in OPS were not reflected in the ASR performance on the real data set, although the best scores were achieved on the simulated data. The GSC-MVDR-ABM beamformer followed by the MaxPower postfilter and GMM ASR achieved the best WER on real data. This configuration obtained a 22.14% WER and a 22.12% WER on the real evaluation set, with or without rescoring, respectively.

8. ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF) under the project number P27803-N15, and the K-Project ASD. The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), the Province of Styria – Government of Styria and the Technology Agency of the City of Vienna (ZIT). The program COMET is conducted by Austrian Research Promotion Agency (FFG). Furthermore, we acknowledge NVIDIA for providing GPU computing resources.

9. REFERENCES

[1] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ’CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, submitted.

[2] L. Pfeifenberger and F. Pernkopf, “Blind source extraction based on a direction-dependent a-priori SNR,” in *Interspeech*, 2014.

[3] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing Society.

[5] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A chime challenge benchmark," in *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME*, 2013, pp. 19–24.

[6] T. D. Rossing, *Springer Handbook of Acoustics*, Springer, Berlin–Heidelberg–New York, 2007.

[7] P. Vary and R. Martin, *Digital Speech Transmission*, Wiley, West Sussex, 2006.

[8] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[9] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, 2007.

[10] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, 2009.

[11] W. Herbordt and W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for nonstationary broadband signals," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002.

[12] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 73–76, 2008.

[13] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, 2011.

[14] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, 1999.

[15] L. Pfeifenberger and F. Pernkopf, "A multi-channel postfilter based on the diffuse noise sound field," in *European Association for Signal Processing Conference*, 2014.

[16] M. G. Shmulik, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints GSC beamformer," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[17] J. Li, Q. Fu, and Y. Yan, "An approach of adaptive blocking matrix based on frequency domain independent component analysis in generalized sidelobe canceller," *IEEE 10th International Conference on Signal Processing*, pp. 231–234, 2010.

[18] K. Lae-Hoon, M. Hasegawa-Johnson, and S. Koeng-Mo, "Generalized optimal multi-microphone speech enhancement using sequential minimum variance distortionless response(MVDR) beamforming and postfiltering," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2006.

[19] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, 1985.

[21] M. Zöhrer and F. Pernkopf, "Representation models in single channel source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015.

[22] M. Zöhrer and F. Pernkopf, "Single channel source separation with general stochastic networks," in *Interspeech*, 2014.

[23] M. Zöhrer, R. Peharz, and F Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015, accepted.

[24] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Neurocomputing: Foundations of Research*, James A. Anderson and Edward Rosenfeld, Eds., pp. 673–695. MIT Press, Cambridge, MA, USA, 1988.

[26] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4101–4104.

[27] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, 2010.