NOISE ROBUST DISTANT AUTOMATIC SPEECH RECOGNITION UTILIZING NMF BASED SOURCE SEPARATION AND AUDITORY FEATURE EXTRACTION

Niko Moritz[1], *Marc René Schädler*[2], *Kamil Adiloglu*[3], *Bernd T. Meyer*[2], *Tim Jürgens*[2], *Timo Gerkmann*[2], *Birger Kollmeier*[1,2,3], *Simon Doclo*[1,2], *Stefan Goetze*[1]

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[3]Hörtech gGmbH, Oldenburg, Germany

ABSTRACT

This paper describes our contribution to the 2nd CHiME challenge and focuses on the small vocabulary task, i.e. track one. We present a robust system combination that involves source separation, auditory feature extraction and a modified automatic speech recognition back-end. The source separation code is based on a non-negative matrix factorization approach and the presented auditory feature extraction method uses 2D Gabor filter functions to extract spectral, temporal and spectro-temporal information of the speech signals. In addition we describe the modifications to our classification back-end and discuss the achieved results. On the final CHiME test set the proposed system achieves a maximum keyword recognition rate improvement of 50.25 % for the -6 dB SNR condition, for instance.

Index Terms— CHiME challenge, non-negative matrix factorization, Gabor feature extraction, source separation, automatic speech recognition

1. INTRODUCTION

Automatic speech recognition (ASR) has made the leap to be used in many commercial systems, such as smartphones, personal computers and more rarely in game consoles or TVs. Some of these systems, especially those using distant microphones, have to cope with different room acoustics and an unknown number of real-world, i.e. possibly highly non-stationary, noise sources. It is well known that especially these requirements pose a major challenge to ASR and signal enhancement algorithms. Development of these algorithms so far has not reached the goal of matching or even outperforming human speech recognition (HSR) [1][2]. The expectation that ASR should equal human performance in reverberant or non-stationary noise conditions still prevents an even larger spread of ASR technologies.

The CHiME Speech Separation and Recognition Challenge is designed to represent a platform for developing, testing and combining a variety of algorithms that are able to cope with highly non-stationary real-world interferences. In the 2nd CHiME challenge the task is to recognize voice commands in a noisy home environment. The audio signals were recorded using two distant in-ear microphones of a manikin. The target speaker's position is known with a small degree of uncertainty as head movements of the speaker are simulated. This property is the essential difference to the 1st CHiME challenge (with fixed location).

This contribution focuses on the small vocabulary task (track one) of the 2nd CHiME challenge that is based on the GRID corpus [3]. We present a system combination of speech enhancement, robust feature extraction and robust ASR back-end training. The speech enhancement is based on a source separation algorithm using a multi-level non-negative matrix factorization (NMF) [4] that incorporates a variational Bayesian inference for learning the model parameters including those of NMF and the sources. The feature extraction method we propose for CHiME extracts auditorily motivated features by applying two-dimensional (2D) Gabor filters to a Mel-warped and log-compressed spectro-temporal representation [5][6]. The ASR back-end is based on a triphone HMM architecture utilizing maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) techniques for the speaker adaption step [7].

The paper is structured as follows: The front-end processing parts are explained in Section 2 and 3. Modifications of the ASR back-end are explained and evaluated in Section 4. Final experimental results combining the proposed front-end and back-end modules are presented in Section 5, and Section 6 concludes the paper.

2. SPEECH ENHANCEMENT

For the reduction of interfering noise, we apply a source separation step. With this step, we aim at separating the speech signal from the background sources.

2.1. Algorithm

For the separation of the target speech source, a general, fully Bayesian source separation algorithm based on the

variational inference method is used [4]. The proposed model operates in the short time Fourier transform (STFT) domain. For J source signals in I channels, the mixing equation is written as

$$x_{fn} = A_f s_{fn} + \varepsilon_{fn} \tag{1}$$

where x_{fn} represents the $I \times 1$ vector containing the mixture STFT coefficients, s_{fn} represents the $J \times 1$ vector consisting of the sources, A_f represents the $I \times J$ complex valued mixing matrix and ε_{fn} represents the noise. In this formulation, f is the frequency index, n the time frame index, i the channel index and j the source index.

We assume that each source signal $s_{j,fn}$ follows a zero-mean complex-valued Gaussian distribution with variance $v_{j,fn}$ given as

$$s_{j,fn} = N\left(0, v_{j,fn}\right) \tag{2}$$

The source variances $v_{j,fn}$, which encode the spectral power, are decomposed via an excitation-filter model using a multi-level NMF model [8]. This framework makes it possible to incorporate a wide range of constraints about the sources. For more details about how to constrain spectral and temporal structures, see [8].

In a fully Bayesian treatment, we need to define the prior distributions of the model parameters. As we do not have reliable prior knowledge about any of the model parameters, we define the multilevel NMF parameters of the source variances as well as the mixing system to follow the non-informative priors.

Using this scheme, we aim at obtaining the posterior distributions of the model parameters, particularly those of the sources. As an exact Bayesian inference is intractable, we used a variational Bayesian (VB) approximation. In this approximation, we employ the generalized inverse Gaussian distribution for the multilevel NMF parameters of the source variances that allows for obtaining closed form update equations for the variational parameters [4][9].

2.2. Initialization and Separation

We follow a very similar scheme as Ozerov *et al.* (2011) [10] for the initialization. In order to extract the spectral characteristics of each speaker properly, we pre-train one NMF model per speaker on the reverberated training samples. Furthermore, we also pre-train one mixing matrix for the development and test set using four randomly selected utterances from each speaker in the reverberated training set (i.e. overall 136 utterances).

We assume that two background sources and one target speech source is active. A two-step separation is performed. In the first step, the background sources are separated. Using the annotation files, 10 seconds of data before and after the target speech source are extracted from the background samples. The mixing matrices and the NMF components are initialized to random values and a maximum of 500 VB iterations are performed. In the second step, the trained mixing matrices and NMF components of the background sources from the first step are kept. The mixing matrix of the target speech source is initialized with the pre-trained mixing matrix. Similarly, the pre-trained speaker model is used for initializing the NMF components of the target speech source. In this step a maximum of 1000 VB iterations are performed. The mean source estimates are saved to be used in the following steps.

3. AUDITORY FEATURE EXTRACTION

In [5], a feature extraction scheme that encodes spectro-temporal modulation patterns in high-dimensional feature vectors was proposed. It was shown that this type of features, called GBFB (Gabor filter bank) features, can improve the robustness of ASR systems. A MATLAB reference implementation of the GFBF is available online [11].

GBFB features are extracted by applying a set of 2D Gabor filters to a spectro-temporal representation of a signal. A log Mel-spectrogram is employed for the spectro-temporal representation because it incorporates several properties of the auditory system (i.e., non-linear frequency scaling and compression of amplitude values) and is widely used in ASR. In contrast to the GBFB reference implementation [11] that uses a log Mel-spectrogram with 23 Mel-bands between 64 Hz and 4 kHz, 10 ms window shift, and 25 ms window length, we extend the frequency range to 8 kHz and increase the number of Mel-bands to 31. This results in the lower 23 Mel-bands to cover the frequency range from 64 Hz to ≈4 kHz.

$$\begin{aligned} h_b(x) &= \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2} \\ 0 & \text{else} \end{cases} \\ s_\omega(x) &= \exp(i\omega x) \\ g_{\omega_k,\omega_n,v_k,v_n,\phi}(k,n) &= \underbrace{s_{\omega_k}(k)s_{\omega_n}(n)}_{\text{sinusoidal function } s(k,n)} \cdot \underbrace{h_{\frac{v_k}{2\omega_k}}(k)h_{\frac{v_n}{2\omega_n}}}_{\text{hann function } h(k,n)} \end{aligned} \tag{3}$$

Equation (3) defines the 2D Gabor filter function that is applied to a spectro-temporal representation by calculating the 2D convolution. The filter consists of the product of a 2D sinusoidal $s(k,n)$ function and a 2D envelope function $h(k,n)$, where k and n denote the discrete frequency and time index, and v_k and v_n correspond to the number of semi-cycles under the envelope in spectral and temporal dimension. The spectral and temporal modulation frequencies ω_k and ω_n allow the 2D Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal patterns. The usage of this kind of filter is motivated by their similarity to spectro-temporal patterns of neurons in the auditory cortex of mammals [12].

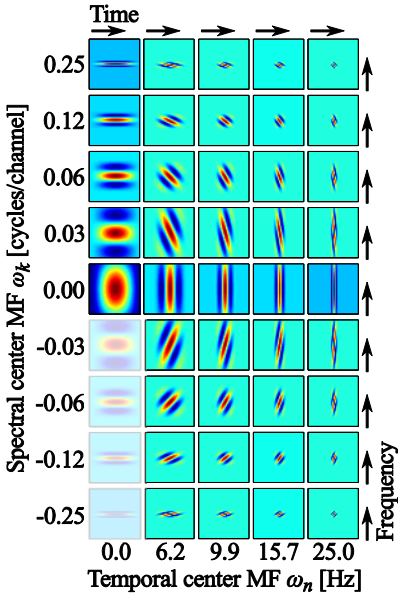

Figure 1: Real components of Gabor filters used for the filter bank, arranged by temporal modulation frequencies (MFs) ω_n and spectral MFs ω_k. Within each tile the horizontal and vertical axis represent time and frequency, respectively. Duplicate filters that occur due to symmetries in the GBFB are greyed out and not used.

The shapes of the different 2D GBFB filters are shown in Figure 1. While purely spectral filters ($\omega_n = 0$) are sensitive to spectral patterns like vowels, purely temporal filters ($\omega_k = 0$) are sensitive to broad-band onsets such as consonants. The spectro-temporal modulations are sensitive to frequency changes over time, like formant transitions. Duplicate filters that occur due to symmetry of the GBFB are removed from the filter set.

In order to reduce correlations and the number of features at the same time, the GBFB filtered log Mel-spectrogram is critically sub-sampled in spectral dimension [6].

To use 2D Gabor filters for feature extraction a set of filters with different parameters have to be identified to capture the different dynamics caused by consonants, vowels, and format transitions. If zero-phase-filters are used as done in this work, four parameters need to be determined for each Gabor filter (ω_k, ω_n, v_k, v_n). Hence, with a set of 30 Gabor filters at least 120 Parameters would have to be estimated.

In [6] this problem is solved by building a filter bank of 2D Gabor filters where each filter covers a range of spectro-temporal modulation frequencies. In this way many parameters are shared and depend on other parameters, reducing the total number of parameters to eight. These parameters were then optimized on the Aurora 2 task, a digit in noise recognition task [13], and remain unchanged in this work. The resulting center modulation frequencies and 41 filter shapes are depicted in Table 1 and Figure 1.

Table 1: Gabor filer bank center modulation frequencies.

Spectral MFs ω_k $\left[\frac{\text{cyc}}{\text{ch}}\right]$
0.000, 0.0293, 0.0599, 0.1223, 0.2500
Temporal MFs ω_n [Hz]
0.00, 6.19, 9.86, 15.70, 25.00

A feature vector with 455 dimensions is obtained by filtering a log Mel-spectrogram with 31 channels using each of the 41 2D Gabor filters and application of the critical sub-sampling. Despite the high dimensionality, GBFB features mesh very well with standard GMM-HMM back-ends and do not need to be mapped to a lower dimensional space. This is a result of limiting correlation between feature dimensions by the critical sub-sampling and restricting the overlap of the Gabor filters in the temporal modulation frequency domain.

In this work, mean and variance normalization (MVN) are applied to the GBFB feature vector on an utterance basis as it was shown to improve the robustness towards additive noise [5]. The resulting feature set is referred to as GBFB-MVN in the following.

4. ASR BACK-END

A properly trained ASR back-end is at least as important as a robust front-end processing to achieve noise robust ASR rates. We employ a speech recognizer based on HTK [7] that uses context dependent triphone HMMs with 3 states per model and 7 Gaussian mixture components per state. Besides the phone models we also trained a silence and a short pause model. After the training of speaker-independent models the MLLR and MAP based speaker adaptation techniques [7] are employed to obtain speaker-dependent HMM models for all 34 speakers. The MLLR adaptation involves two passes. On the first pass a global adaption is performed that is used as an input for the second pass using a regression class tree with up to 32 leaf nodes. After the MLLR adaptation one iteration of MAP adaptation is applied to the means and mixture weights of the GMM-HMM models.

Multi-condition training is a well-known and effective technique to improve the noise robustness of a recognizer. A reverberated (REV), a noisy isolated (ISO) and a clean training set is provided with the challenge data. However, ASR systems can benefit from even more and different noise examples during the training procedure. This is the reason why we created a new multi-condition training set (MCT) by mixing the reverberated training data with the noise backgrounds that are provided in the CHiME challenge data. We generate six SNR conditions ranging from -6 to 9 dB that have been made without level adjustment but by searching for a random suitable noise section. In addition the noise-free reverberated utterances are maintained in the new multi-condition set resulting in seven times the original

Table 2: ASR back-end evaluation. Results are given in keyword recognition rates. STD denotes the standard CHiME speaker adaptation, REV denotes the noise-free reverberated training set, ISO denotes the noisy isolated set and MCT denotes the extended multi-condition training set.

HMM	Speaker	Training	Development Set [%]						Test Set [%]					
Architecture	Adaptation	Condition	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
word	STD	REV	32.08	36.33	50.33	64.00	75.08	83.50	32.17	38.33	52.08	62.67	76.08	83.83
word	STD	ISO	49.75	57.92	67.83	73.67	80.92	82.67	49.08	58.83	67.33	75.08	79.17	82.83
word	STD	MCT	62.50	67.67	77.00	84.08	89.42	90.33	64.33	70.17	78.92	85.42	89.58	91.33
word	MLLR+MAP	REV	46.50	48.75	60.92	73.92	82.33	89.83	44.17	49.17	62.92	72.58	82.50	89.42
word	MLLR+MAP	ISO	62.92	68.58	77.08	83.67	86.67	89.83	61.42	69.67	77.33	82.58	88.33	89.92
word	MLLR+MAP	MCT	65.25	71.17	78.58	86.75	91.42	92.83	66.33	72.25	**82.08**	86.58	**91.33**	92.50
triphone	STD	REV	38.42	41.33	53.92	67.58	74.83	82.83	38.00	38.83	52.75	63.75	75.67	85.92
triphone	STD	ISO	49.92	57.58	67.00	74.08	78.92	82.33	50.50	57.75	68.17	74.83	78.08	82.83
triphone	STD	MCT	62.42	69.33	77.67	85.25	90.92	91.92	64.33	70.33	78.75	87.25	90.00	92.75
triphone	MLLR+MAP	REV	52.00	56.58	68.00	77.83	83.92	90.75	48.83	55.67	68.75	78.83	85.58	91.50
triphone	MLLR+MAP	ISO	67.25	70.67	79.83	84.67	89.17	91.42	65.17	73.00	80.50	85.92	88.17	90.00
triphone	MLLR+MAP	MCT	**68.17**	**73.25**	**80.50**	**87.75**	**92.17**	**93.25**	**67.42**	**73.25**	81.17	**88.00**	91.17	**93.42**

amount of training data. The influence exerted by the amount of noisy training data and by different back-end modifications is illustrated in Table 2 with respect to the keyword recognition rates (KRRs). For the presented experiments we used the baseline MFCC features proposed by the CHiME challenge that use 26 Mel channels, 12 cepstral and one energy coefficient, delta and acceleration coefficients and cepstral mean normalization (CMN).

First, it can be seen that triphone HMMs promise a benefit over the whole-word HMM architecture. This benefit seems to decline with increasing the extrinsic variance in the training data, which can be explained by the fact that phoneme combinations are smaller units of speech than words and thus are more frequent events providing inherently a larger extrinsic and intrinsic variance. Moreover, the benefit due to adding more extrinsic variability to the training data by adding different noise conditions does not result in an equally increasing improvement and drops conspicuously fast as can be seen in Table 2 by comparing the different training conditions. This particularly applies if a more advanced speaker adaptation method is used such as MLLR and MAP adaptation for example. These adaptation methods adapt only the means of the GMM-HMM models to the speaker specific characteristics, whereas the variances and transition probabilities remain unaffected. This is crucial, since only a limited number of training files for each speaker is available, and thus the intrinsic and extrinsic variances become very limited for a small subset of the whole training set.

In the further reading we only publish the ASR improvements of the best back-end settings we evaluated in this section, i.e. the triphone HMM recognizer with the MLLR plus MAP speaker adaption. However, the gain in recognition performance by adding a better front-end processing to the back-end should remain comprehensible, even if the results of every single processing block combination are not shown here.

5. EXPERIMENTS

Experiments and results shown in this paper are all carried out according to the guidelines of the CHiME challenge [14]. Therefore we did not exploit any information about the signal-to-noise ratio (SNR), the fact that the same utterances are used at each SNR and each available training and test set, and the fact that the same noise backgrounds are used in the development and the test set. The final keyword recognition rates are obtained using the scoring scripts provided by the CHiME challenge website [14].

5.1. Robust Feature Extraction

The GBFB-MVN features are evaluated and compared with MFCC features on the CHiME speech recognition task. For the MFCC optimization we increase the number of Mel channels to 30 and use a band limitation for the Mel filter bank with a lower cut-off frequency of 100 Hz and an upper cut-off frequency of 5500 Hz. This domain is chosen, since it approximates the frequency range where speech is most prominent. The new MFCC features consist of 13 cepstral coefficient including the 0^{th} coefficient plus the delta and double-deltas. In addition MVN is performed on the MFCC features for each utterance. The CHiME baseline MFCCs (MFCCs), the proposed MFCC changes (MFCC-MVN) and the GBFB-MVN features are compared in terms of noise robustness in Table 3.

It can be seen that the proposed changes of the MFCC setup already cause a significant improvement over the baseline MFCCs. The average improvement amounts 5.5 % for the REV, 2.84 % for the ISO and approximately 2 % for the MCT set, whereby a maximum gain of up to 10 % is achieved for the lowest SNR conditions, i.e. -6 and -3 dB.

If we compare the GBFB-MVN with MFCC features it becomes obvious that the GFBF-MVN considerably

Table 3: ASR front-end evaluation. Results are given in keyword recognition rates. The total average relates to the development and test set. The ASR back-end here consists of triphone HMMs with MLLR+MAP for the speaker adaptation.

ASR Features	Source Separation	Training Condition	Development Set [%]						Test Set [%]						Total average
			-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	
MFCCs	no	REV	52.00	56.58	68.00	77.83	83.92	90.75	48.83	55.67	68.75	78.83	85.58	91.50	71.52
MFCC-MVN	no	REV	57.17	66.58	74.58	83.33	88.75	92.50	58.92	64.00	74.75	83.58	88.00	92.25	77.03
GBFB-MVN	no	REV	61.25	68.92	78.00	85.58	91.33	94.58	62.42	68.83	78.25	86.58	92.33	94.83	80.24
MFCCs	no	ISO	67.25	70.67	79.83	84.67	89.17	91.42	65.17	73.00	80.50	85.92	88.17	90.00	80.48
MFCC-MVN	no	ISO	70.75	77.33	82.58	85.75	89.75	91.92	71.25	76.92	84.17	87.25	90.25	91.92	83.32
GBFB-MVN	no	ISO	75.17	81.50	85.75	91.08	92.92	94.17	75.42	80.58	88.17	91.92	93.08	95.00	87.06
MFCCs	no	MCT	68.17	73.25	80.50	87.75	92.17	93.25	67.42	73.25	81.17	88.00	91.17	93.42	82.46
MFCC-MVN	no	MCT	71.17	78.00	82.67	88.25	92.17	93.25	71.75	78.00	83.83	89.08	91.83	93.00	84.42
GBFB-MVN	no	MCT	77.92	83.00	88.25	92.42	**95.33**	**96.92**	76.08	83.67	88.58	92.58	**95.42**	**96.00**	88.85
MFCCs	yes	REV	66.25	71.67	78.25	85.33	89.08	89.17	64.58	71.50	79.17	84.33	87.00	90.75	79.76
MFCC-MVN	yes	REV	69.50	76.00	82.00	86.92	89.58	91.67	69.00	74.83	82.25	86.42	90.08	91.92	82.51
GBFB-MVN	yes	REV	70.83	75.33	84.50	88.92	92.42	94.83	70.50	76.67	85.67	88.92	92.17	94.42	84.60
MFCCs	yes	ISO	77.75	80.75	85.67	89.08	90.75	91.50	75.17	79.75	85.83	88.00	90.50	92.00	85.56
MFCC-MVN	yes	ISO	76.33	81.58	86.75	89.17	91.17	91.42	77.08	80.25	87.00	89.75	91.08	91.67	86.10
GBFB-MVN	yes	ISO	78.42	83.50	87.83	91.92	92.92	93.17	79.17	84.50	89.08	93.17	93.92	94.25	88.49
MFCCs	yes	MCT	77.33	81.33	85.75	90.33	91.67	93.08	74.42	79.42	87.17	89.25	91.08	92.92	86.15
MFCC-MVN	yes	MCT	76.50	84.50	88.08	91.58	93.33	93.67	78.42	83.00	87.92	90.33	92.83	93.67	87.82
GBFB-MVN	yes	MCT	**80.50**	**85.75**	**89.50**	**94.25**	94.25	96.58	**82.42**	**86.00**	**90.58**	**93.75**	94.83	95.92	**90.36**

outperform even the enhanced MFCC setup. For the REV training condition an average gain in keyword recognition rates of 3.21 % can be observed compared to the MFCC-MVN setup. It is worth noting that the gain is distributed relatively evenly over all SNR conditions and does not decrease at higher SNRs as it is the case for the MFCC-MVN features compared to the baseline MFCCs.

For the noisy training sets the robustness of the GBFB-MVN features becomes even more significant. An average increase in KRRs of 3.74 % and 4.43 % compared to MFCC-MVN features and 6.58 % and 6.38 % compared to the baseline MFCCs can be assessed on the ISO and MCT set, respectively.

Besides the presented ASR features we investigated also other popular feature extraction methods on the CHiME task, such as the power normalized cepstral coefficients (PNCCs) [15], RASTA-MFCCs [16] and amplitude modulation filter sets [17][18]. Without being able to present all results here in detail due to space constraints, no other method that was tested could outperform the GBFB-MVN results on the CHiME track one task. We believe that a major difference between the tested feature extraction methods is the fact that the GBFB-MVN features have a frequency band selective output, whereby the other methods rely on the Cepstrum and thus do not have this characteristic. This GBFB property may help if just single frequency bands are corrupted while others remain largely undisturbed.

5.2. Source Separation

The speech enhancement is based on the source separation approach presented in Section 2. Table 3 shows the ASR results that are obtained by using the proposed source separation method. It can be seen that the NMF based source separation substantially contributes to improve the ASR robustness in noisy environments with an unknown number of interfering sources. It turns out that the presented source separation can work together with different feature extraction methods and that these modules can take advantage of the each other. By regarding the results of the

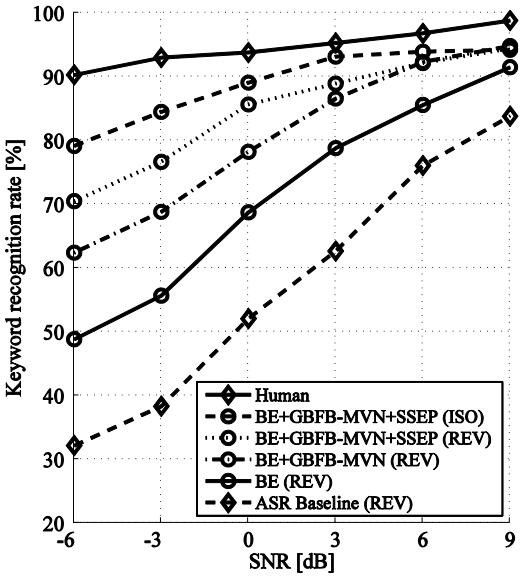

Figure 2: Final ASR results compared to human speech recognition performance and to the CHiME baseline results. The results relate to the test set. BE denotes the ASR back-end modifications and SSEP denotes the source separation.

REV training condition it can be noted that the source separation slightly decreases the ASR results for the 9 dB SNR condition, whereby the results of the lower SNR conditions are increased. This is due to the fact that speech enhancement algorithms in practice can cause distortions to the desired signal. To reduce the influence of this property, we also apply the speech enhancement to the noisy training data in order to adapt the acoustic models to the characteristics of the noise reduction.

For the baseline MFCCs the source separation affects an average gain in KRRs of 8.24 % for the REV, 5.08 % for the ISO and 3.69 % for the MCT training condition. For the GBFB-MVN features the average gain is slightly less with 4.36 % (REV), 1.43 % (ISO) and 1.51 % (MCT).

Figure 2 illustrates the ASR results of our different system modules for the REV and ISO training sets and compares the KRRs to HSR performance and to the CHiME ASR baseline.

6. CONCLUSION

The proposed ASR system could substantially improve the keyword recognition rates on the small vocabulary task of the 2nd CHiME challenge. We obtained an average recognition score of 90.14 % on the development set and 90.58 % on the final test set, which is 33.06 % above the ASR baseline results. Each of the major system changes, which include source separation, robust feature extraction and back-end modifications, could themselves and in combination contribute in great extent to the mentioned improvements. This indicates that each of the methods may have different strengths and properties that can be combined.

7. ACKNOWLEDGEMENTS

This work was partially funded by the DFG Cluster of Excellence 1077 "Hearing4all", the Federal Ministry of Education and Research, BMBF, (project AALADIN, V4PFL013) and the SFB/TRR 31 (the active auditory system).

8. REFERENCES

[1] B.T. Meyer, T. Brand, and B. Kollmeier, "Effects of speech-intrinsic variations on human and automatic speech recognition of spoken phonemes," *J. Acoustic. Soc. Am.* 129, pp. 388-403, 2011.

[2] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *J. Speech Communication* 49, pp. 336-347. 2007.

[3] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, M. Matassoni, "The Second `CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines", *in Proc. ICASSP*, Vancouver, Canada, 2013.

[4] K. Adiloglu and E. Vincent, "Variational Bayesian Inference for Source Separation and Robust Feature Extraction," Technical Report, 2012.

[5] M.R. Schädler and B. Kollmeier. "Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems," *in Proc. of Interspeech*, Portland, USA, 2012.

[6] M.R. Schädler, B.T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", *The Journal of the Acoustical Society of America* 131, pp. 4134-4151, 2012.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woddland, *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2009.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech and Language Processing* 20(4), pp. 1118– 1133, 2012.

[9] M.D. Hoffman, D.M. Blei and P.R. Cook, "Bayesian nonparametric matrix factorization for recorded music," *in Proc. of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

[10] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," *in Proc. of the CHiME 2011 Workshop on Machine Listening in Multisource Environments*, Florence, Italy, 2011.

[11] M.R. Schädler, "GBFB feature extraction reference implementation," http://medi.uni-oldenburg.de/GBFB, 2012, (last visited 01/29/2013).

[12] A. Qiu, C. Schreiner, and M. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *J. Neurophysiol.* 90, pp. 456-476, 2003.

[13] H.G. Hirsch, and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *in Proc. ISCA ITRW ASR*, 2000.

[14] "The 2nd 'CHiME' Speech Seperation and Recognition Challenge": http://spandh.dcs.shef.ac.uk/chime_challenge/ (last visited 01/29/2013).

[15] C. Kim, R.M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias substraction," *in Proc. Interspeech*, pp. 28-31, Brighton, UK, 2009.

[16] H. Hermansky, and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing* 2(4), pp. 578-589, 1994.

[17] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," *in Proc. ICASSP*, Prague, Czech, pp. 5492-5495, 2011.

[18] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation filters as feature sets for robust ASR: constant absolute or relative bandwidth?," *in Proc. Interspeech*, Portland, USA, 2012.