A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Reverberant Environments

Roland Maas, Andreas Schwarz, Yuanhang Zheng, Klaus Reindl,
Stefan Meier, Armin Sehr, Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Erlangen, Germany
{maas,schwarz,zheng,reindl,smeier,sehr,wk}@LNT.de

Abstract

An acoustic front-end for robust automatic speech recognition in noisy and reverberant environments is proposed in this contribution. It comprises a blind source separation-based signal extraction scheme and only requires two microphone signals. The proposed front-end and its integration into the recognition system is analyzed and evaluated in noisy living room-like environments according to the PASCAL CHiME challenge. The results show that the introduced system significantly improves the recognition performance compared to the challenge baseline.

Index Terms: PASCAL CHiME challenge, robust automatic speech recognition, blind source extraction, speech enhancement

1. Introduction

Automatic speech recognition (ASR) with distant-talking microphones constitutes a major challenge and, at the same time, a major chance of our days. We can imagine many possible applications easing our daily life, such as voice interaction with television sets, humanoid robots, or smart homes.

In such scenarios however, an ASR system has to deal with unwanted additive interference and reverberation picked up by the microphones besides the desired signal. Therefore, the system's robustness to such distortions has to be increased, which can either be achieved by applying signal or feature enhancement techniques or by adapting the acoustic models of the ASR system to capture the distortions.

In real-world scenarios, a large variety of interferences must be expected, and there can be highly nonstationary and unpredictable noise and interference components leading to very low signal-to-noise ratios (SNRs). A reliable recognition of spoken commands is therefore hardly possible without proper preprocessing of the noisy signals. Correspondingly, it is highly desirable to design an acoustic front-end that reliably extracts the target speech components from the acquired noisy microphone signals independently of the underlying scenario and the corresponding SNR level. In terms of speech recognition, this means that the target speech components extracted from the noisy mixtures should lead to speech features that are largely independent of the SNR level and the underlying scenario. The goal is here to design such a robust acoustic front-end for the PASCAL CHiME challenge [1].

The PASCAL CHiME challenge calls for recognizing speech commands uttered in noisy living room-like environments and captured by two distant-talking in-ear microphones placed in a manikin. Hereby, the main problem arises from the partly nonstationary additive distortions in low SNR levels. To cope with this problem, we propose a speech enhancement technique based on Blind Source Extraction (BSE) for interference estimation and Wiener filtering for interference suppression. Adaptive training techniques are used to integrate the preprocessing system into the recognizer. Recognition experiments carried out on the CHiME corpus [2] show that a reduction in word error rate (WER) of up to 71% can be achieved.

This paper is structured as follows: The proposed acoustic front-end and its integration into the ASR system is explained in Sec. 2. Experimental results are discussed in Sec. 3, and Sec. 4 concludes the paper.

2. Acoustic Front-End

The signal model for robust ASR in adverse environments is depicted in Fig. 1. It is based on a two-channel audio capture and a BSE scheme followed by the ASR system. The acquired microphone signals $x_p,\ p \in \{1, 2\}$, contain the signals of Q simultaneously active point sources, where only one signal (here: s_1 without loss of generality) is considered as desired signal to be extracted, and the remaining $Q - 1$ source signals are regarded as interfering signals. Moreover, background noise denoted by $n_{\mathrm{b},p},\ p \in \{1,2\}$, is present in the observed microphone signals. The mixing of the original sources is modeled by finite impulse response (FIR) filters of length M (denoted by $\mathbf{h}_{qp} = [h_{qp}(0), \ldots, h_{qp}(M-1)]^T$ in Fig. 1) which capture reverberation in real environments leading to the sensor signals

$$x_p(k) = \sum_{q=1}^{Q} \sum_{\kappa=0}^{M-1} h_{qp}(\kappa) s_q(k-\kappa) + n_{\mathrm{b},p}(k),\ p \in \{1,2\}, \tag{1}$$

where $h_{qp}(k),\ k = 0, \ldots, M-1$ denote the coefficients of the FIR filter model from the q-th source $s_q,\ q = 1, \ldots, Q$ to the p-th sensor $x_p,\ p \in \{1,2\}$. As mentioned above, for a speech recognizer it is important that the target speech components (here: s_1) are properly extracted from the acquired noisy microphone signals. Therefore, the microphone signals are fed into a two-channel BSE unit. This concept extracts the desired speech signal components by suppressing all noise and interference components. The output signal of the BSE scheme denoted by $\hat{s}_1$, which represents an estimate of the spoken command, is then fed into the ASR system where the command should be robustly recognized.

The applied signal extraction scheme is illustrated in Fig. 2. It consists of two building blocks: a blocking matrix that yields a reference of all noise and interference components (denoted by $\hat{n}$) and a noise suppression unit providing an estimate of the desired signal (here: $\hat{s}_1$). These two building blocks are described in more detail in the following. The approach of separating all noise and interference components from the target speech components and subsequently suppressing the estimated noise signals contained in the noisy mixtures seems to be an adequate strategy for the given case, where only two microphone signals are available and where the scenarios may vary widely. This has already been shown to be a promising strategy in other contexts, e.g., in [3, 4].

2.1. Noise and Interference Estimation

For separating all noise and interference components from the desired signal, a blocking matrix based on the TRINICON (TRIple-N-Independent component analysis for CONvolutive mixtures) framework (see, e.g., [5, 6]) is designed. Before the blocking matrix is introduced, the generic broadband algorithm derived from the TRINICON framework for separating convolutive mixtures is briefly reviewed.

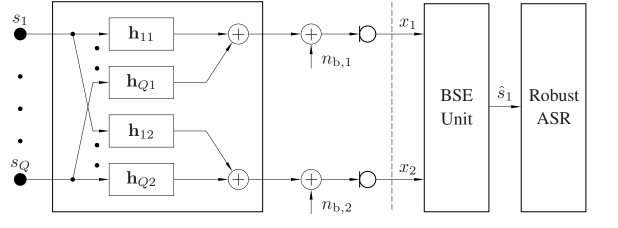

Figure 1: Signal model for robust speech recognition

Source separation algorithms aim at finding a demixing system (denoted by $\mathbf{w}_{pq} = [w_{pq}(0), \ldots, w_{pq}(L-1)]^T$ in Fig. 2), whose output signals $y_q(k)$ are described by (here, the determined case of two active sources and two microphone signals is considered)

$$y_q(k) = \sum_{p=1}^{2} \sum_{\kappa=0}^{L-1} w_{pq}(\kappa) x_p(k-\kappa), \quad q \in \{1,2\}, \tag{2}$$

where $w_{pq}(k),\ k = 0, \ldots, L-1$ denote the weights of the adaptive demixing filter from the p-th sensor channel to the q-th output channel within the MIMO (mulitple input/ multiple output) demixing system. The generic TRINICON cost function for blind source separation (BSS) is given by the Kullback-Leibler divergence (KLD) between the estimated PD-variate joint probability density function (PDF) $\hat{f}_{y,PD}(\mathbf{y}_1, \ldots, \mathbf{y}_P)$ of the output signals of the demixing system and the product $\prod_{p=1}^{P} \hat{f}_{y_p,D}(\mathbf{y}_p)$ of the estimated D-variate marginal output PDFs [6]:

$$\mathcal{J}_{\mathrm{BSS}}(n) = \sum_{i=0}^{\infty} \beta(i,n) \cdot \tilde{\mathcal{J}}_{\mathrm{BSS}}(i), \tag{3}$$

$$\tilde{\mathcal{J}}_{\mathrm{BSS}}(i) = \frac{1}{N} \sum_{j=iL}^{iL+N-1} \left\{ \log \left(\frac{\hat{f}_{y,PD}(\mathbf{y}_1, \ldots, \mathbf{y}_P)}{\prod_{p=1}^{P} \hat{f}_{y_p,D}(\mathbf{y}_p)} \right) \right\}, \tag{4}$$

where i and n denote block indices and the vectors $\mathbf{y}_p$ contain D consecutive output samples each. $\beta(i,n)$ denotes a window function that allows for offline, online, and block-online algorithms. In general, the KLD involves the expectation operator which is here replaced by a short-time average $\tilde{\mathcal{J}}_{\mathrm{BSS}}$ over N blocks of length D. If and only if the BSS outputs are statistically independent, i.e., for perfect separation of mutually independent source signals, (4) becomes zero as the joint output PDF can then be written as the product of the marginal output PDFs. A natural-gradient-descent approach is applied for iterative optimization of the BSS filter coefficients [7].

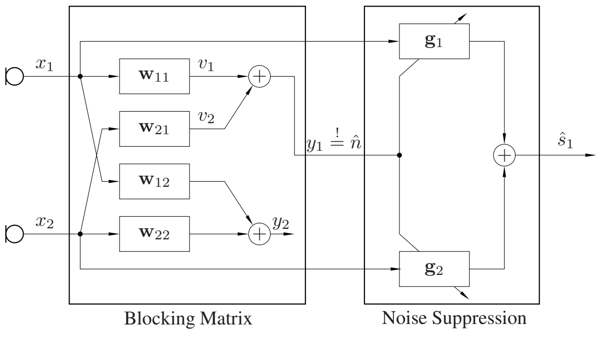

Figure 2: Realization of the BSE unit

For the noise and interference estimation approach an efficient second-order-statistics (SOS) realization of the TRINICON update rule was derived based on multivariate Gaussian probability density functions [8].

In a real scenario, where only two microphone signals are available, it is unrealistic that always only two competing speakers are active. As for the intended application more than two point sources and also diffuse sources may be active, determined BSS algorithms are not directly applicable without modifications. As there is no determined solution for a demixing matrix to separate the individual sources in an underdetermined case (more active sources than available microphone signals) or a noisy scenario, the generic TRINICON cost function as given by (3) and (4) is modified so that all noise and interference components can be separated from the target signal when only two microphone signals are available. The cost function of this 'directional BSS' concept [9] is given by

$$\mathcal{J}_{\mathrm{DirBSS}} = \mathcal{J}_{\mathrm{BSS}} + \eta_{\mathrm{C}} \mathcal{J}_{\mathrm{C}}, \tag{5}$$

where $\mathcal{J}_{\mathrm{C}}$ represents a geometrical constraint and is given by

$$\mathcal{J}_{\mathrm{C}} = \parallel w_{11}(k) + w_{21}(k - \tau_\phi) \parallel^2 . \tag{6}$$

The weight η_{C}, typically in the range $0.4 < \eta_{\mathrm{C}} < 0.6$, indicates the relative importance of the geometrical constraint [9]. The directional constraint as given in (6) forces a spatial null towards the desired source location, which has to be estimated or is known a priori in real applications. τ_ϕ describes the time difference of arrival (TDOA) of the target source between the two sensors. It has to be noted that in real applications, this can be any fractional delay. If a-priori information about the target angular position is missing, the localization concept as discussed, e.g., in [10, 11] can be applied. Owing to the property of BSS to produce independent output signals, directional BSS also suppresses correlated components arriving from other directions, i.e., reflections and reverberation will also be suppressed to the greatest extent possible so that the BSS output signal can be used as an interference estimate $\hat{n}(k)$.

The ability to suppress also reflections of the desired signal makes directional BSS superior to conventional beamforming techniques, e.g., null-beamformers, in suppressing the target signal, especially in reverberant environments (see [9]). Moreover, in contrast to many beamformer techniques, e.g., [12, 13], no voice-activity detector is needed and no prior knowledge on the microphone positions is required. The output signal of directional BSS can be approximated by

$$\begin{aligned} \hat{n}(k) &= w_{11}(k) * x_1(k) + w_{21}(k) * x_2(k) \\ &\approx \sum_{p=1}^{2} \left(\sum_{q=2}^{Q} h_{qp}(k) * s_q(k) + n_{\mathrm{b},p}(k) \right) * w_{p1}(k), \end{aligned} \tag{7}$$

where $*$ denotes convolution and $w_{p1},\ p \in \{1,2\}$ denote the demixing coefficients obtained by directional BSS.

2.2. Noise and Interference Suppression

In order to extract the desired speech signal components from noisy mixtures, either single-channel or multichannel noise reduction techniques can be applied. However, multichannel techniques require reliable estimates of the noise and interference components in all available microphones. Since, in practice, it is extremely challenging to obtain these separate noise and interference estimates in highly nonstationary scenarios, the combination of BSS methods for noise and interference estimation with single-channel Wiener filtering techniques for noise and interference suppression to obtain an estimate of the desired speech signal components $\hat{s}_1$ is investigated. To this end, the single-channel noise and interference reference $\hat{n}(k)$ (7) obtained by directional BSS is used to control spectral enhancement filters $\mathbf{g}_p,\ p \in \{1,2\}$, as shown in Fig. 2. Optimum spectral weights for a Wiener filtering strategy are given by

$$g_{\mathrm{opt},p}(\nu) = 1 - \frac{\hat{S}_{n_p n_p}(\nu)}{\hat{S}_{x_p x_p}(\nu)}, \quad p \in \{1,2\}, \tag{8}$$

where ν represents the frequency index. $\hat{S}_{n_p n_p}(\nu)$ and $\hat{S}_{x_p x_p}(\nu),\ p \in \{1,2\}$, represent power spectral density (PSD) estimates of the true noise and interference components contained in channel p and the microphone signals, respectively. However, from the directional BSS algorithm discussed above, only a single noise reference is obtained that rather describes all noise and interference components than the channel-specific components. The corresponding PSD estimate of (7) reads:

$$\hat{S}_{\hat{n}\hat{n}}(\nu) = \hat{S}_{\tilde{n}_1\tilde{n}_1}(\nu) + \hat{S}_{\tilde{n}_2\tilde{n}_2}(\nu) + 2\Re\{\hat{S}_{\tilde{n}_1\tilde{n}_2}(\nu)\}, \tag{9}$$

$$\hat{S}_{\tilde{n}_p\tilde{n}_p}(\nu) = |W_{p1}(\nu)|^2 \hat{S}_{n_p n_p}(\nu), \quad p \in \{1,2\}, \tag{10}$$

$$\hat{S}_{\tilde{n}_1\tilde{n}_2}(\nu) = W_{11} W_{21}^*(\nu) \hat{S}_{n_1 n_2}(\nu), \tag{11}$$

where $\Re\{\cdot\}$ represents the real part. In order to approximate the optimum spectral weights (optimum in the Wiener sense) given by (8), a method is derived to obtain a PSD estimate of the channel-specific noise and interference components from $\hat{S}_{\hat{n}\hat{n}}(\nu)$ (9). Therefore, let us define the noise power ratio $R(\nu)$ and the coherence as follows: The noise power ratio between the two channels $R(\nu)$ is defined as

$$R(\nu) = \frac{\hat{S}_{\tilde{n}_1\tilde{n}_1}(\nu)}{\hat{S}_{\tilde{n}_2\tilde{n}_2}(\nu)}, \tag{12}$$

and the coherence of all noise and interference components between the two channels is given by

$$\Gamma_{\tilde{n}_1\tilde{n}_2}(\nu) = \frac{\hat{S}_{\tilde{n}_1\tilde{n}_2}(\nu)}{\sqrt{\hat{S}_{\tilde{n}_1\tilde{n}_1}(\nu)\hat{S}_{\tilde{n}_2\tilde{n}_2}(\nu)}}. \tag{13}$$

Using (12) and (13), the PSD estimate (9) can be written as

$$\hat{S}_{\tilde{n}\tilde{n}}(\nu) = \hat{S}_{\tilde{n}_1\tilde{n}_1}(\nu) \cdot \frac{F(\nu)}{R(\nu)} \qquad (14)$$

$$= \hat{S}_{\tilde{n}_2\tilde{n}_2}(\nu) \cdot F(\nu), \qquad (15)$$

where $F(\nu)$ is given by

$$F(\nu) = 1 + R(\nu) + 2\sqrt{R(\nu)}\Re\{\hat{\Gamma}_{\tilde{n}_1\tilde{n}_2}(\nu)\}. \qquad (16)$$

Assuming a spherically isotropic diffuse noise field, which is usually given in reverberant environments with large distances between the sources and the microphones, (12) simplifies to $R(\nu) = 1$, and the coherence (13) is given by [14]

$$\Gamma_{\text{diff}}(\nu) = \frac{\sin\left(2\pi f_s \cdot c^{-1} \cdot d \cdot \nu \cdot M^{-1}\right)}{2\pi f_s \cdot c^{-1} \cdot d \cdot \nu \cdot M^{-1}},$$

$$\nu = 0, ..., M-1, \qquad (17)$$

where c and d represent the speed of sound and the distance between two omnidirectional microphones, respectively. f_s denotes the sampling frequency and M the total number of frequency bins. $F(\nu)$ (16) thus simplifies to

$$F_{\text{diff}}(\nu) = 2\left(1 + \Gamma_{\text{diff}}(\nu)\right) \qquad (18)$$

and estimates of the PSDs of the channel-specific noise components can be obtained by

$$\hat{S}_{\tilde{n}_p\tilde{n}_p}(\nu) = \frac{\hat{S}_{\tilde{n}\tilde{n}}(\nu)}{2\left(1 + \Gamma_{\text{diff}}(\nu)\right)}, \quad p \in \{1, 2\}. \qquad (19)$$

Finally, the spectral weights for noise and interference suppression $g_p,\ p \in \{1, 2\}$ are calculated as

$$g_p(\nu) = \max\left[1 - \mu\frac{\hat{S}_{\tilde{n}_p\tilde{n}_p}(\nu)}{\hat{S}_{v_pv_p}(\nu)}, g_{\min}\right], \quad p \in \{1, 2\}, \qquad (20)$$

where μ and $g_{\min}$ denote a gain factor and the spectral floor, respectively. These parameters are real-valued constants and are used to achieve a trade-off between noise and interference suppression and speech distortion.

Summarizing the above, the main advantages of the acoustic preprocessing as illustrated in Fig. 2 are as follows: Applying BSS algorithms in unknown noisy and underdetermined scenarios for noise and interference estimation is very powerful as a reference with very low target speech components can be obtained as already shown, e.g., in [4, 9]. Moreover, no voice activity detection algorithm is necessary, which is usually unreliable in nonstationary and unpredictable scenarios at low SNRs. Besides, the scheme requires only two sensor channels, which makes it very attractive for practical applications.

2.3. Integration into the ASR System

For optimum performance, the ASR system has to be tuned to the preprocessing algorithm. Since perfect speech enhancement cannot be achieved in practice, the output of the front-end will always contain some residual interference and some distortion of the desired signal. Therefore, it is beneficial to carefully adjust the hidden Markov models (HMMs) of the ASR system to the front-end. To this end, we perform speech enhancement both on the test and training data. For capturing different interference characteristics by the acoustic model, the set of training utterances comprises different noise conditions as in multi-style training. As the speech enhancement reduces the variability of the training data due to different interferences significantly, a more compact acoustic model can be obtained from the enhanced data compared to the noisy data. Since the acoustic model trained with the enhanced data does not have to spend many degrees of freedom on capturing different noise characteristics, it can represent the variability due to different words more accurately. This type of training efficiently combines speech enhancement and multi-style training and can be considered as noise-adaptive training [15] or model-independent adaptive training [16].

3. Experiments

The experiments are carried out according to the PASCAL CHiME challenge conditions [1]. In the following, we therefore describe the most important challenge settings, the configuration of the employed acoustic front-end and ASR back-end, and present the achieved recognition results.

3.1. PASCAL CHiME Setup

The PASCAL CHiME challenge addresses the problem of recognizing commands being uttered in a noisy living room environment. All utterances are taken from the Grid corpus [17] and are artificially convolved with binaural room impulse responses (BRIRs) measured with a binaural manikin at a distance of 2 meters in broadside direction [2]. For testing, the utterances are mixed with binaural background noise recordings from the CHiME domestic audio corpus and controlled such that different SNR levels are obtained [2]. The measurements of both the BRIRs and the background noise are performed in a lounge and a kitchen with a reverberation time T_{60} of 300 ms each. The noise sources are typical for a family home, e.g., TV, kitchen and laundry appliance sounds, footsteps, electronic gadget sounds, or playing children. All recordings are sampled at a rate of 16 kHz. The ASR task itself is speaker-dependent and imposes a simple grammar superseding any kind of language model. Each utterance is of the form <command–color–preposition–letter–number–adverb>, where only the so-called *keyword accuracy* is of interest, which is based on the number of correctly recognized <letter> and <number> tokens. Two disjoint test sets are considered, namely the *development test set* and the *final test set*, where solely the former one is allowed to be used for parameter tuning.

Note that for all performed tests, we did furthermore neither exploit the available continuous audio streams [1], nor the SNR labels, nor the fact that the test sets contain the same utterances at each SNR level.

3.2. Acoustic Front-End

For the directional BSS, we use filters of the length $L_{\mathrm{DirBSS}} = 1024$. The relative importance of the directional constraint η_{C} is set to 0.5. For the processing of the training data and the development and test set, we use fixed filter coefficients for each speaker that are obtained by adapting the filter over all utterances of the speaker from the training, development or test set, respectively.

In order to achieve a trade-off between noise and interference suppression and speech distortion of the Wiener filtering concept, the parameters μ and $g_{\min}$ are set to 1.2 and 0.15, respectively. The Wiener filter is implemented using a polyphase filterbank with a filter length of 1024, 512 complex-valued subbands, and a downsampling rate of 128.

3.3. ASR Back-End

In order to evaluate the performance of our proposed acoustic front-end, we employed a speech recognizer based on the ASR toolkit Sphinx-4 [18].

The recognizer uses triphone HMMs with 3 states per model, 8 Gaussian output densities per state, and a total number of 600 tied states. From the input signals, features consisting of 13 mel-frequency cepstral coefficients (MFCCs) as well as 13 delta and 13 acceleration coefficients are derived. Furthermore, cepstral mean subtraction is applied to compensate for short convolutive distortions. We created our own training data by mixing each utterance of the provided training set with two isolated background noise sequences for each SNR level. The set of noise sequences corresponds to the one underlying the development test set [1]. All utterances are then fed into the proposed acoustic front-end (except for the case “w/o front-end” in Table 1). Afterwards, the entire set of preprocessed “noisy” training data is used to perform Baum-Welch training [19] leading to a speaker-independent HMM. To obtain speaker-dependent HMMs, the adaptation techniques MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum A Posteriori) are applied to the means of the HMM’s output densities [20]. Solely the training data of the concerned speaker over all SNR levels are exploited resulting in one SNR-multi-style HMM per speaker.

ASR system	SNR in dB					
	-6	-3	0	3	6	9
prop. front-end + Sphinx	78.9	84.5	88.4	90.8	94.3	94.3
Sphinx (w/o front-end)	70.8	75.3	84.1	87.4	91.4	94.1
CHiME baseline	31.1	36.8	49.1	64.0	73.8	83.1

ASR system	SNR in dB					
	-6	-3	0	3	6	9
prop. front-end + Sphinx	79.8	83.3	88.3	92.8	92.6	95.1
Sphinx (w/o front-end)	70.1	74.5	82.8	89.8	90.5	93.8
CHiME baseline	30.3	35.4	49.5	62.9	75.0	82.4

Table 1: Comparison of keyword accuracies in % for the development (top) and final test set (bottom).

We would like to underline that the development and the final test set consist of different utterances, temporally positioned at different points in the CHiME background noise recordings [1]. For the training of the recognizer, only the noise recordings from the development test set have been exploited. Hence, no data from the final test set was used for training.

The CHiME challenge baseline recognition system is based on HTK [21] with word-level HMMs and 7 Gaussian output densities per state. The same type of features as for Sphinx is used. To create both speaker-independent and speaker-dependent HMMs, the Baum-Welch [21] method is performed on the provided noise-free training data [1]. The CHiME challenge baseline results are then obtained when no preprocessing algorithm is applied to the test data.

3.4. Experimental Results

Table 1 compares the keyword accuracies for the development and the final test set. Besides the results of the Sphinx back-end with and without employing the proposed acoustic front-end, the CHiME challenge baseline results are listed.

For all SNRs and both test sets, the noise-adapted back-end itself without front-end leads to a consistent improvement of the recognition performance. Applying the proposed acoustic front-end achieves another remarkable gain especially for low SNRs. In the case of the final test set at an SNR of -6 dB, the reduction of the WER due to the back-end adaptation is 57% relative to the challenge baseline results. The additional improvement by the front-end is 32% resulting in an overall relative WER reduction of 71%.

These results clearly underline the potential of the introduced speech enhancement algorithm based on BSS and Wiener filtering. Furthermore, they show that noise-adaptive training is a promising way of combining a powerful front-end with an ASR system.

4. Summary and Conclusions

A two-channel acoustic front-end for robust automatic speech recognition was presented. The concept is based on blind source separation and Wiener filtering strategies. Its integration into the ASR system was realized via noise-adaptive training of the recognizer's acoustic model. Experiments under noisy and reverberant conditions showed a remarkable reduction in word error rate of up to 71%. These results indicate that the application of multichannel speech enhancement techniques along with the adaptation of the recognizer represents a very powerful combination significantly increasing the reliability of distant-talking ASR systems.

5. Acknowledgements

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work (contract number KE 890/4-1).

6. References

[1] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent. The PASCAL CHiME speech separation and recognition challenge 2011. [Online]. Available: http://www.dcs.shef.ac.uk/spandh/chime/challenge.html

[2] H. Christensen, J. Barker, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," *Proc. Interspeech*, 2010.

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

[4] K. Reindl, Y. Zheng, A. Lombard, A. Schwarz, and W. Kellermann, "An acoustic front-end for interactive TV incorporating multichannel acoustic echo cancellation and blind signal extraction," in *Proc. 44th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, November 2010.

[5] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," in *Int. Symp. Independent Component Analysis and Blind Separation (ICA)*, Nara, Japan, April 2003, pp. 945–950.

[6] ——, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio signal processing for next-generation multimedia communication systems*, Y. Huang and J. Benesty, Eds. Boston: Kluwer Academic Publishers, 2004, pp. 255–293.

[7] S. I. Amari, "Natural gradient works efficiently in learning," in *Neural Computation*, vol. 10, 1998, pp. 251–276.

[8] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," in *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, Jan. 2005, pp. 120–134.

[9] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Int. Workshop on Comp. Advances in Multi-Sensor Adapt. Proc. (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, August 1976.

[11] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 233–236.

[12] O. Hoshuyama, B. Begasse, A. Hirano, and A. Sugiyama, "A realtime robust adaptive microphone array controlled by an SNR estimate," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 1998.

[13] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, "Application of a double-talk resilient DFT- domain adaptive filter for bin-wise stepsize controls to adaptive beamforming," in *Int. Workshop on Nonlinear Signal and Image Processing (NSIP)*, Sapporo, Japan, May 2005.

[14] H. Kuttruff, *Room Acoustics*. London: Taylor & Francis, 2000.

[15] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," *Proc. ICSLP*, vol. 3, pp. 806–809, 2000.

[16] M. Gales, "Adaptive training for robust ASR," *Proc. ASRU*, pp. 15–20, 2001.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition," *Sun Microsystems Technical Report*, 2004.

[19] CMUSphinx Wiki. Training acoustic model for CMUSphinx. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorialam

[20] ——. Adapting the default acoustic model. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorialam

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. University of Cambridge, 2009.