A Stereophonic Acoustic Front-End for Distant-Talking Interfaces based on Blind Source Separation

K. Reindl, R. Maas, A. Schwarz, Y. Zheng, S. Meier, A. Sehr, and W. Kellermann
Chair of Multimedia Communications and Signal Processing, Cauerstr. 7, 91058 Erlangen, Germany
Email: {reindl, maas, schwarz, zheng, smeier, sehr, wk}@LNT.de

Abstract

In this contribution, an acoustic front-end for distant-talking interfaces that only requires two microphone signals is presented. It comprises a directional blind source separation (BSS)-based noise and interference estimation scheme and Wiener-type filters for noise and interference suppression. The proposed front-end and its integration into a speech recognition system is analyzed and evaluated in noisy living-room-like environments representing a generalization of the PASCAL CHiME challenge, implying that the potential of the method is verified for a wide range of distances.

Introduction

For human-machine interfaces easing our daily life close-talking microphones are usually unacceptable. However, using distant-talking microphones implies a degradation of the desired speech signal quality caused by additive undesired components resulting from competing speakers and diffuse and generally nonstationary background noise, and reverberation. For automatic speech recognition (ASR) these impairments become a major challenge if they are unpredictable and nonstationary. As in reality, a large variety of (highly nonstationary and unpredictable) interferences are to be expected, we aim here at designing an acoustic front-end that robustly recovers the target speech components from the acquired noisy microphone signals independently of the underlying scenario and the corresponding signal-to-interference-and-noise ratio (SINR).

The acoustic front-end

The proposed two-channel acoustic front-end is shown in Fig. 1, and it comprises a directional blind source separation (BSS)-based blocking matrix (BM) for noise and interference estimation and Wiener-type filters for noise and interference suppression. The acquired microphone signals $x_p,\ p \in \{1,2\}$ usually contain the signals of Q simultaneously active point sources s_q (speech), of which only one signal (here: s_1) is considered as desired signal. Due to reverberation in the acoustic environment also reflections of the source signals are acquired. The acoustic paths between the q-th source and the p-th microphone (h_{qp}) are modelled by finite impulse response (FIR) filters with typical orders of several thousands. Background noise signals n_p constitute additional microphone signal components. In order to reliably cope with underdetermined and unpredictable nonstationary scenarios when

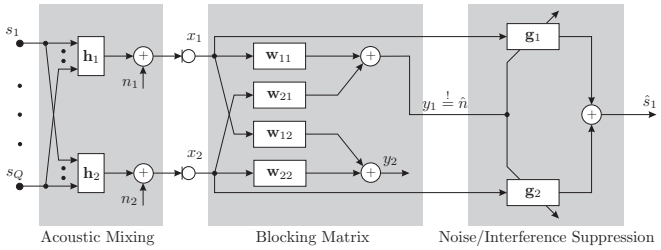

Figure 1: Signal model of the proposed noise and interference suppression scheme

only the minimum number of two microphone signals is available for array-based signal processing, it is beneficial to firstly separate all noise and interference components from the target components, and secondly, apply spectral enhancement filters to suppress the undesired components contained in the microphone signals [1]. For separating all undesired components from the desired signal components, we proposed in [2] a directional BSS approach which only requires a coarse estimate of the target source position. This coarse estimate of the target position provides a geometric constraint for BSS leading to the following cost function [2]

$$\mathcal{J}_{\mathrm{DirBSS}} = \mathcal{J}_{\mathrm{BSS}} + \eta_{\mathrm{c}}\mathcal{J}_{\mathrm{c}}, \tag{1}$$

where $\mathcal{J}_{\mathrm{BSS}}$ and $\mathcal{J}_{\mathrm{c}}$ denote the cost functions of BSS and the geometric constraint, respectively. For our approach, we exploit the generic TRINICON (TRIple-N-Independent component analysis for CONvolutive mixtures) optimization criterion introduced in [3]. The BSS cost function is complemented by a geometric constraint [2], to force a spatial null towards the desired source location. The weight η_{c} indicates the relative importance of the constraint [2]. The overall estimate of all undesired signal components $\hat{n}$ is then used to realize Wiener-type spectral enhancement filters applied to the microphone signals to suppress all unwanted components. In order to realize optimum filters for noise and interference suppression, estimates of the channel-specific noise components are required. However, directional BSS only provides an overall estimate of all noise and interference components. Assuming a spherically isotropic noise field [4], the optimum (mean square error) channel-specific spectral enhancement filters $g_p,\ p \in \{1,2\}$ are obtained in the frequency domain as [1]

$$g_p = 1 - \frac{\mu \hat{S}_{\hat{n}\hat{n}}}{\left(|w_{11}|^2 + |w_{21}|^2 + 2\Re\{w_{11}w_{21}^*\Gamma_{n_1n_2}\}\right)\hat{S}_{x_px_p}}, \tag{2}$$

where $\Gamma_{n_1 n_2}$ denotes the coherence function of the isotropic noise field. $\hat{S}_{\hat{n}\hat{n}}$ and $\hat{S}_{x_p x_p}$ represent estimates of the power spectral densities of the noise reference and the microphone signals, respectively. The real-valued gain factor μ is used to achieve a trade-off between noise reduction and desired speech signal distortion.

Evaluation

The proposed acoustic front-end is evaluated according to a generalization of [5, 6]. For our setup, we consider three different target source distances. Therefore, in contrast to [1, 5, 6], binaural room impulse responses (BRIRs) were measured for distances 1m, 2m, and 3m in a living-room-like environment with a reverberation time of $T_{60} \approx 300$ms. The signal-to-reverberation ratios (SRRs) for the individual distances decrease from 3.5dB to −1.6dB. Analogously to [5], the task was the recognition of commands being uttered in a noisy and reverberant environment. For more information on the recognition task, see [5].

The signal extraction scheme is implemented using a polyphase filter bank. The filter length of the prototype lowpass filter is 1024, the number of subbands (complex-valued) equals 512, and the downsampling factor is set to 128. The sampling rate is 16kHz. The estimated target signal $\hat{s}_1$ is further processed by a speech recognizer based on the ASR toolkit Sphinx-4 [7]. 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 processed input signals, features consisting of 13 mel-frequency cepstral coefficients (MFCCs), 13 delta and 13 acceleration coefficients are derived. Cepstral mean subtraction is applied to compensate for short convolutive distortion. For HMM training, all utterances of the training set are fed into the proposed acoustic front-end. Afterwards, the entire set of preprocessed "noisy" training data is used to perform Baum-Welch training 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. Solely the training data of the concerned speaker over all SINR levels and all distances are exploited resulting in one SINR-multi-style HMM per speaker.

In Table 1 the word accuracies are listed for the proposed acoustic front-end combined with the back-end based on adaptive training. Besides, the baseline results are shown, which are obtained when neither a preprocessing nor noise adaptation are performed. Comparing the baseline results (last row) with the noise-adapted back-end alone (no preprocessing) for all distances and denoted as 'Sphinx', a significant and consistent absolute improvement of the recognition performance ranging from 8% to 46% over all conditions is achieved. Although the applied multi-style training is a very powerful and competitive approach, applying the proposed front-end prior to further processing the signals with an ASR system, still leads to remarkable improvements. With the proposed approach it is possible to reduce the word error rate (WER) by more than 25% for high SINR conditions and up to almost 50% for low SINR conditions compared to a back-end processing alone. It turns out that the results are usually best for a distance of 2m, which is explained by the fact that only a single model is trained for all distances. These results clearly underline the potential of the introduced front-end for a wide range of distances. Besides, it is demonstrated that noise-adaptive training is a promising way of combining a powerful front-end with an ASR system.

dist./ (SRR)	ASR system	SINR in dB −6	−3	0	3	6	9
1m	Prop.	**85.5**	**87.8**	**91.0**	**92.0**	**92.9**	**92.4**
(3.5dB)	Sphinx	76.3	79.0	87.7	87.9	90.5	90.8
2m	Prop.	**86.0**	**89.3**	**92.4**	**92.5**	**94.1**	**93.8**
(1.9dB)	Sphinx	76.3	79.2	88.5	87.9	90.4	90.8
3m	Prop.	**84.5**	**87.1**	**91.5**	**92.6**	**93.3**	**93.0**
(-1.6dB)	Sphinx	72.6	75.1	87.2	88.1	90.4	91.2
Baseline		30.3	35.4	49.5	62.9	75.0	82.4

Table 1: Comparison of word accuracies in % for the proposed acoustic front-end combined with the ASR system, and the ASR system 'Sphinx' alone, for different target source distances and various input SINR conditions

Acknowledgement

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

References

[1] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *International Workshop on Machine Listening in Multisource Environments (satellite event of Interspeech 2011)*, Florence, Italy, Sep. 2011, pp. 41–46.

[2] 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, pp. 253–256.

[3] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 3, Montreal, Canada, May 2004, pp. 889–892.

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

[5] 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

[6] 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 Groove, CA, USA, Nov. 2010.

[7] 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*, no. TR-2004-139, 2004.