A Teleconference System with Distributed Microphones

Sebastian Stenzel, Eric Böhmler, Jürgen Freudenberger
Institute for System Dynamics,
University of Applied Sciences (HTWG) of Konstanz,
Email: {sstenzel, eboehmle, jfreuden}@htwg-konstanz.de

Introduction

In this work we present an approach which can be used as a frontend for teleconference systems. The problem in such hands-free telecommunication scenarios is the possible large distance between the local speakers and the microphones. Therefore we propose the use of a distributed microphone array, to accomplish a better compromise for different speaker positions. To combine the microphone signals with rather different signal conditions, we use a microphone diversity approach which exploits these different signal and noise conditions. In general teleconference systems require additional signal processing for echo cancellation. To reduce possible background noise from the microphone input signals the system contains a noise suppression filter. In this filter we also include the suppression of the residual echo.

System Overview

In general we consider one loudspeaker, several local speakers and M microphones. The acoustic system is assumed to be linear and time-invariant. Hence, the i^{th} microphone signal $y_i(k)$ can be modeled in the frequency domain as

$$Y_i(\kappa,\nu) = S_i(\kappa,\nu) + H_i(\nu)F(\kappa,\nu) + N_i(\kappa,\nu)\,, \qquad (1)$$

where $S_i(\kappa,\nu)$, $F(\kappa,\nu)$ and $N_i(\kappa,\nu)$ denote the short-time spectra of the speech, the far-end signal and the noise, respectively. $H_i(\nu)$ is the acoustic transfer function, which corresponds to the room impulse response $h_i(k)$ between the loudspeaker and the i^{th} microphone. The subsampled time index and the frequency bin index are denoted by κ and ν, respectively.

In Fig. 1 the block diagram of the proposed system for the two microphone case is depicted. The noisy signals $y_i(k)$ and the signal of the far-end speaker $f(k)$ are transformed into the frequency domain using a Short-Time Fourier Transform (STFT) of length L. Subsequent blocks are overlapping by K samples.

The first processing step is the cancellation of the echoes of the far-end speaker in each microphone signal. Here a frequency domain implementation is used. To handle potentially long room impulse responses, the echo cancellation is done in several partitions. The adaptation of the filter coefficients is done implicitly according to [1], this means there is no need for an explicit double-talk detector.

Then the signals are combined using the Spectral Combiner (SC) to compensate differences in the acoustic transfer functions from the local speaker and to take advantage from the different noise and echo conditions [2]. For the system it is important to place the microphones at several positions, but no explicit assumption of the spatial adjustment is done. Therefore the user can place the microphones arbitrarily. To enable an appropriate noise and residual echo suppression, a spectral subtraction filter is introduced to the combining filter function. In a last processing step the phase differences of the speech signals are compensated.

Signal Combining

With speech signals, recorded under diversity conditions, we would weight each microphone input signal under the constraint of a maximal output SNR of the combined output signal. In [2] a microphone diversity system was proposed, where the input signals are weighted with respect to the SNR ratios at each channel

$$G_{\text{SC}}^{(i)}(\kappa,\nu) = \sqrt{\frac{\gamma_i(\kappa,\nu)}{\sum_{j=1}^{M}\gamma_j(\kappa,\nu)}}\,, \qquad (2)$$

where $\gamma_i(\kappa,\nu)$ denotes the signal-to-noise ratio (SNR) at the i^{th} microphone. It was shown, that a coherent addition of the sensor signals weighted with the gain factors $G_{\text{SC}}^{(i)}(\kappa,\nu)$ leads to a combining, where the signal-to-noise ratio at the combined output is the sum of the input SNR values.

Coherent addition requires additional phase estimation. For the phase compensation it is sufficient to estimate the phase differences $\Delta_i(\kappa,\nu)$ to a reference microphone, e.g. to the first microphone. The estimation of the phase differences is done by an adaptive filter $G_{\text{LMS}}^{(i)}(\kappa,\nu)$, which uses the Frequency domain Least Mean Square (FLMS) algorithm. The adaptation of the filter $G_{\text{LMS}}^{(i)}(\kappa,\nu)$ is chosen to minimize the following expectation value

$$\mathbb{E}\{\mid \tilde{Y}_i^{\text{AEC}}(\kappa,\nu)G_{\text{LMS}}^{(i)}(\kappa,\nu) - \tilde{Y}_1^{\text{AEC}}(\kappa,\nu)\mid^2\}\,, \qquad (3)$$

where $i = 2, 3, \ldots, M$.

Applying the phases of the filter $G_{\text{LMS}}^{(i)}(\kappa,\nu)$ to the corresponding filter functions $G_{\text{SC}}^{(i)}(\kappa,\nu)$ a cophasal addition is achieved

$$\hat{S} = G_{\text{SC}}^{(1)}\tilde{Y}_1^{\text{AEC}} + G_{\text{SC}}^{(2)}e^{j\Delta_2}\tilde{Y}_2^{\text{AEC}} + G_{\text{SC}}^{(3)}e^{j\Delta_3}\tilde{Y}_3^{\text{AEC}}\ldots\,, \qquad (4)$$

where we omitted the dependency on (κ,ν).

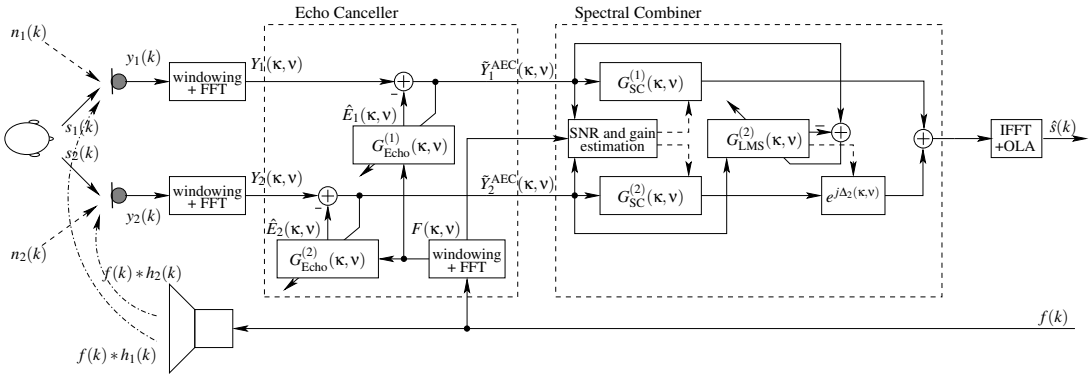

Fig. 1: Basic system structure of the system with two inputs.

Noise and Residual Echo suppression

The Spectral Combining, as described in the previous section, guarantees a signal combining with an optimal SNR ratio, but for an appropriate noise and residual echo reduction further signal processing is needed. Therefore we introduced the well-known spectral subtraction filter function [3] to the SC weighting. As shown in [2] this leads to

$$G^{(i)}(\kappa,\nu) = \sqrt{\frac{\gamma_i(\kappa,\nu)}{\rho+\gamma(\kappa,\nu)}}\,, \tag{5}$$

where ρ is an over-subtraction factor.

For the signal combining the input SNR in each channel has to be estimated. Therefore an estimate of the short-time power spectral densities (PSD) of the speech signal, the noise components $N_i^{\mathrm{AEC}}(\kappa,\nu)$ and the residual echo $\tilde{E}_i^{\mathrm{AEC}}(\kappa,\nu)$ is needed. The noise PSD $P^{(i)}_{N^{\mathrm{AEC}}}(\kappa,\nu)$ is commonly estimated by the use of a voice activity detection or by minimum statistics. The PSD of the residual echo $P^{(i)}_{\tilde{E}^{\mathrm{AEC}}}(\kappa,\nu)$ is calculated using an estimate of the channel specific coupling factor. With these estimates, the current signal-to-noise ratio is then obtained by

$$\gamma_i(\kappa,\nu) = \frac{P^{(i)}_{\tilde{Y}^{\mathrm{AEC}}}(\kappa,\nu) - P^{(i)}_{N^{\mathrm{AEC}}}(\kappa,\nu) - P^{(i)}_{\tilde{E}^{\mathrm{AEC}}}(\kappa,\nu)}{P^{(i)}_{N^{\mathrm{AEC}}}(\kappa,\nu) + P^{(i)}_{\tilde{E}^{\mathrm{AEC}}}(\kappa,\nu)}\,, \tag{6}$$

assuming that the noise, the residual echo and speech signals are uncorrelated.

Evaluation

To evaluate the proposed approach we simulate a microphone diversity system with three microphones. Therefore we measured with an artificial head impulse responses in a conference room for five local conference participants (pos. 1 - pos. 5). We also recorded the background noise from different noise sources typical in a conference room, e.g. from a beamer or a laptop. These recordings were done with omni-directional microphones at a sampling rate of $16000Hz$. The distances between the microphones were chosen in the range of $1m - 1.2m$. This results in a distance for the local speakers to the microphones between $50cm$ to $1.30m$ for that scenario. To simulate the echo signals at the microphones, we also measured the transfer functions between the loudspeaker and the three microphones. The distance between the loudspeaker and the microphones was $0.4m$ for mic. 2 and $0.7m$ for mic. 1 and mic. 3.

	pos. 1	pos. 2	pos. 3	pos. 4	pos. 5
SER ch. 1	-5.7	-5.6	-5.5	-3.0	-0.2
SER ch. 2	-8.8	-7.0	-7.6	-7.4	-9.0
SER ch. 3	-0.1	-3.5	-2.8	-4.8	-4.1
SER out	39.4	39.3	38.1	39.4	40.8
ERLE	35.4	36.6	35.1	36.1	36.7

Tab. 1: In-/Output SER and ERLE of the simulated scenario.

In our simulations the FFT size of the algorithm was chosen to $L = 1024$ and the overlap between two successive frames was $K = 768$ samples. The number of the echo canceler partitions was set to $P = 4$. Tab. 1 contains the input signal-to-echo ratios (SER) at the different microphone positions for each local speaker. Therefore the power of the local speech and echo signal was calculated continuously, also when speech is absent. Tab. 1 presents also the SER at the system output and the total echo return loss enhancement (ERLE) in comparison to mic. 1. It is obvious that the overall SER enhancement is about $40dB$ for all speaker positions.

Conclusion

In this paper we have presented a system for combined echo and noise reduction for distributed microphones. Therefore we introduced the residual echo suppression into the Spectral Combining scheme. The presented system can be used to combine several distributed microphones, e.g. in conference scenario.

References

[1] J. Bourgeois, J. Freudenberger, and G. Lathoud, "Implicit control of noise canceller for speech enhancement," in *European Conference on Speech Communication and Technology , Lisbon*, 2005.

[2] J. Freudenberger, S. Stenzel, and B. Venditti, "Microphone Diversity Combining for In-Car Applications," *EURASIP Journal on Advances in Signal Processing*, 2010, Article ID 509541, 13 pages.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.