Soft Decision Combining for Dual Channel Noise Reduction

Timo Gerkmann and Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum, 44780 Bochum, Germany
timo.gerkmann@rub.de, rainer.martin@rub.de

10.21437/Interspeech.2006-555

Abstract

We present a flexible dual channel noise reduction algorithm that combines the information of two microphone channels in the discrete Fourier transform domain. This algorithm can cope with substantially different signal-to-noise ratios at different time-frequency bins in both channels. The output is obtained by combining the outputs of two single channel filters and a dual channel filter weighted by the probability of speech presence. The algorithm has a low latency and does not require any additional information such as the microphone positions.

Index Terms: multichannel speech enhancement, speech presence probability.

1. Introduction

In recent years, the usage of digital mobile communication devices such as cell phones and hearing aids has increased rapidly. In mobile scenarios we often have to deal with noise, e.g. babble or vehicle noise. Consequently, there is an increased need for noise reduction (NR) algorithms. Many NR algorithms apply a gain function in the discrete Fourier transform (DFT) domain to estimate the clean speech coefficients [1]. The gain function is commonly derived by assuming a certain distribution of the clean speech DFT coefficients. This distribution is parameterized by the signal-to-noise ratio (SNR), which is computed for instance by using the *decision-directed* approach [1]. Furthermore, an additional weighting function which indicates the probability of speech presence is usually applied, and allows distinct treatment for the cases of speech presence or absence [1, 2, 3].

In adverse conditions like automotive environments, improved NR is desirable. Here, multiple microphones may be used, e.g. two microphones in the sunvisor, or one at the A-pillar and one at the rear view mirror. Here, the relative positioning of the microphones may not be known. Furthermore, one microphone may be severely disturbed at times, e.g. because it is close to a fan or other noise sources. Therefore, the benefit of the additional microphone is *a priori* unknown.

This paper presents a dual-channel algorithm that addresses these problems. The contribution of each microphone channel to the output is controlled by the probability of speech presence at each channel in each time-frequency (TF) point. At TF points where one channel is superior, that channel dominates the output. In case the speech presence probabilities of the two channels are similar, the information of both channels is combined to gain optimal noise reduction. While the proposed scheme will only be described for two microphones the algorithm can easily be extended for more than two microphones.

2. The Signal Model

The NR system uses two microphone channels and the DFT to compute spectral coefficients on short signal segments. The short-time Fourier transforms (STFT) of the input channel signals $y_1(t)$ and $y_2(t)$ are denoted as $Y_1(k,l)$ and $Y_2(k,l)$, where k and l indicate the frequency and time indices, respectively. Each channel m contains a certain amount of noise, $N_m(k,l)$. Speech, $S(k,l)$, may be present in each channel. When speech is present, the signal at each microphone will differ in phase, $\Delta\phi(k,l)$, and amplitude, $A(k,l)$. Speech presence at channel m and each TF-point (k,l) is indicated by $H_m^1(k,l)$. Since the whole system works in STFT-domain, the argument (k,l) may be dropped. Thus, our model has the following form:

$$\begin{array}{lllll} \text{speech presence:} & H_1^1: & Y_1 & = & S + N_1 \\ & H_2^1: & Y_2 & = & A S e^{-j\Delta\phi} + N_2 \\ \text{speech absence:} & H_1^0: & Y_1 & = & N_1 \\ & H_2^0: & Y_2 & = & N_2\,. \end{array} \tag{1}$$

3. Soft Decision Combining

Based on (1), four joint hypotheses can be derived:

- (H_1^1, H_2^0) Speech is only present in channel one
- (H_1^0, H_2^1) Speech is only present in channel two
- (H_1^1, H_2^1) Speech is present in both channels
- (H_1^0, H_2^0) Speech is present in neither channel

Different NR filters are used for each joint hypothesis. The outputs of all four filters are combined using a soft decision approach. Bayes' theorem is used to compute the conditional probabilities of the joint hypotheses. For instance, for (H_1^1, H_2^0) we obtain:

$$P(H_1^1, H_2^0|Y_1, Y_2) = \frac{p(Y_1, Y_2|H_1^1, H_2^0) P(H_1^1, H_2^0)}{p(Y_1, Y_2)}\,. \tag{2}$$

In most acoustic scenarios the two microphone signals will be similar. However, we explicitly aim at exploiting potential differences, to achieve an enhanced output signal at TF-points where one channel is superior. Thus, we assume independence of the two microphone channels, as well as independence of the real and imaginary parts of the speech and noise DFT coefficients. Under these assumptions, the joint probabilities as in equation (2) can be factored, and the soft weights for the four joint hypotheses be written as:

$$\begin{aligned} W_{10} = P(H_1^1, H_2^0|Y_1, Y_2) &\approx \frac{\Lambda_1}{1+\Lambda_1}\,\frac{1}{1+\Lambda_2} \\ W_{01} = P(H_1^0, H_2^1|Y_1, Y_2) &\approx \frac{1}{1+\Lambda_1}\,\frac{\Lambda_2}{1+\Lambda_2} \end{aligned}$$

$$W_{11} = P(H_1^1, H_2^1|Y_1, Y_2) \approx \frac{\Lambda_1}{1+\Lambda_1}\frac{\Lambda_2}{1+\Lambda_2}$$
$$W_{00} = P(H_1^0, H_2^0|Y_1, Y_2) \approx \frac{1}{1+\Lambda_1}\frac{1}{1+\Lambda_2}, \quad (3)$$

where Λ_m are generalized likelihood ratios, defined as:

$$\Lambda_m = \frac{P(H_m^1)}{P(H_m^0)}\frac{p(Y_m|H_m^1)}{p(Y_m|H_m^0)}. \quad (4)$$

For this we require to model the priors $q_m = P(H_m^0) = 1 - P(H_m^1)$, and the distributions $p(Y_m|H_m^1)$ and $p(Y_m|H_m^0)$.

3.1. Modelling the pdf of the input channels

The distribution of the DFT coefficients of speech and noise is assumed to have a Gaussian probability density function (pdf). This does not hold for small frame sizes as usually used in mobile communications, where the pdf becomes more heavy-tailed [4]. However, the Gaussian assumption is still used because it is computationally less expensive. Also the experiments presented in [4] showed that the statistical model used in speech presence uncertainty weighting has only a minor impact on the speech quality. Thus, the pdfs are modelled as follows:

$$p(Y_m|H_m^0) = \frac{1}{\pi\sigma_{N_m}^2}\exp\{-\gamma_m\} \quad (5)$$
$$p(Y_m|H_m^1) = \frac{1}{\pi\sigma_{N_m}^2}\frac{1}{1+\xi_m}\exp\left\{-\gamma_m\frac{1}{1+\xi_m}\right\}, \quad (6)$$

where σ_N^2 is the noise variance estimated using *minimum statistics* [5], $\gamma_m = |Y_m|^2/\sigma_{N_m}^2$ is the *a posteriori* SNR and $\xi_m = \sigma_{S_m}^2/\sigma_{N_m}^2$ is the *a priori* SNR, as gained by the *decision directed* approach [1]. Thus, from equations (4), (5) and (6) we obtain:

$$\Lambda_m = \frac{1-q_m}{q_m}\frac{1}{1+\xi_m}\exp\left\{\frac{\gamma_m\xi_m}{1+\xi_m}\right\}. \quad (7)$$

3.2. Modelling the prior probabilities

The prior probabilities, $q_m = P(H_m^0) = 1 - P(H_m^1)$, should be chosen independent of an observation. They provide a bias in favor of either speech presence or speech absence. The prior speech absence probability q should not be set higher than 0.5, because that would favor the case of speech absence. Thus, at low SNRs, speech may be suppressed. To account for this, in [1] q is set as low as 0.2. This means however, that in cases where speech is absent and the *a priori* SNR is zero, the resulting conditional probability of speech absence would not be larger than $P(H_m^0|Y_m) = 0.2$. However, for our algorithm a proper detection of speech absence is crucial, i.e. this probability should be close to 1 when speech is absent. This may be achieved by tracking *a priori* SNRs based on the observation [2, 3]. For our purposes, however, we find that setting the priors to $q_m = 0.5$ and limiting the SNR in eq. (7) to be higher than $\xi_{\min} = 9$ dB, leads to the desired result.

4. Noise Reduction

For all four joint hypotheses we consider different gain functions that yield four different outputs. To allow a correct superposition of the outputs we need to adjust for the phase difference, $\Delta\phi$, of the two channels.

4.1. Phase Estimation

As in [6], we estimate a dual channel phase compensation term using the instantaneous frequency dependent phase estimate when speech is present, and the relative time delay τ when speech is absent. The relative time delay τ is estimated by searching for the maximum value of the cross-correlation of the two input channels during speech activity, and stored for pauses. The phase compensation factor at each TF-point results in:

$$e^{\jmath\Delta\phi} = (1 - W_{11})e^{\jmath k\omega_0\tau} + W_{11}\frac{Y_1Y_2^*}{|Y_1|\,|Y_2|}, \quad (8)$$

where W_{11} is the probability that speech is present in both channels, as defined in (3). In [6] this weight is only based on the *a priori* SNRs, ξ_m. The *a priori* SNRs are usually gained using the decision-directed approach [1], and adapt rather slowly. Therefore, speech pauses are not instantly detected, and phase estimates may still be based on the instantaneous phase difference. This is unreasonable during speech absence and yields a distorted output signal. However, our weights are based on both the *a priori* and the *a posteriori* SNR. They thus adapt faster and avoid this problem. Note, that it is crucial that the speech presence weights, W_{11}, actually approach zero during speech absence, to avoid speech distortions due to unreliable instantaneous phase estimates (section 3.2).

4.2. Both channels contain only noise

When speech is absent in both channels we leave some residual noise for a natural sounding result. For this, we combine both input channels and attenuate the resulting coefficients, Y_{LN}, by a constant factor α. The combination is done according to the noise power of the two channels, as:

$$Y_{\text{LN}} = \zeta Y_1 + [1-\zeta]Y_2e^{\jmath\Delta\phi}, \quad (9)$$

where the weighting factor ζ is gained as:

$$\zeta = \frac{\sigma_{n2}^2}{\sigma_{n1}^2 + \sigma_{n2}^2}. \quad (10)$$

The factor ζ is smoothed over time and frequency to avoid perceivable switching effects. The resulting coefficients, Y_{LN}, are dominated by the channel with the lower noise power at each TF-point.

4.3. Only one channel contains speech

In the two cases where we detect speech in only one channel, we use single channel noise reduction gain functions. E.g. the spectral amplitude MAP estimator as derived in [7] can be used:

$$G_m^{\text{sgle}} = \frac{\xi_m}{2(\xi_m+1)}\left[1 + \sqrt{1 + \frac{1+\xi_m}{\xi_m\gamma_m}}\right]. \quad (11)$$

4.4. Both channels contain speech

If both channels contain speech, we want to use the information of both channels for an optimal noise reduction. A summation of the two channels, after a correct phase adjustment, will increase the output SNR by up to 3 dB. This signal combination may be done either before or after noise reduction. Using the former method a single channel gain function (section 4.3) is computed on a signal with an increased SNR, yielding less musical tones during speech presence. However, we decided for the latter, because the multi-channel filters adapt faster, and thus yield lower speech distortion.

A multichannel MAP estimator is derived in [8]. For the two channel case we have:

$$G_1^{\text{dual}} = \frac{\xi_1}{1+\xi_1+\xi_2}\left[1+\sqrt{\frac{\gamma_2\xi_2}{\gamma_1\xi_1}}\right], \quad (12)$$

$$G_2^{\text{dual}} = \frac{\xi_2}{1+\xi_1+\xi_2}\left[1+\sqrt{\frac{\gamma_1\xi_1}{\gamma_2\xi_2}}\right]. \quad (13)$$

5. Overall Algorithm and Example

The resulting algorithm is depicted in Figure 1. The input signals are transformed into the DFT domain using a Kaiser window of length 256 at a sampling rate of 8 kHz, and an overlap of 50%. The shaping parameter is 5.5. For synthesis we use a Hann/Kaiser window.

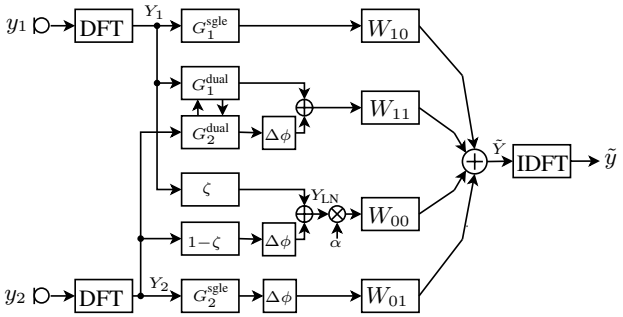
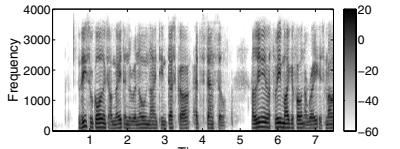
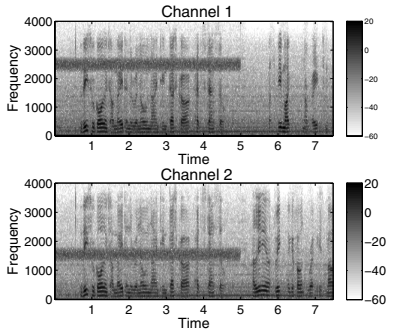


Figure 1: The two channel soft decision combining algorithm

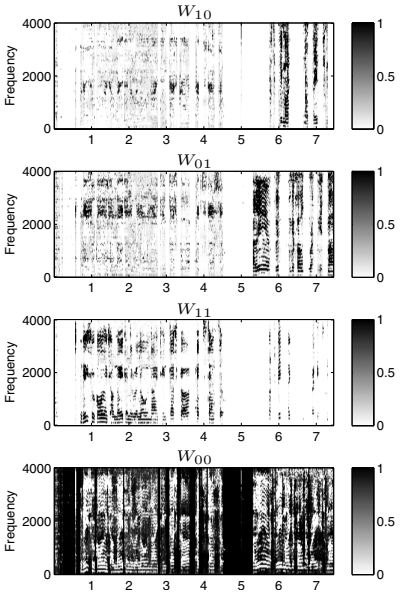
Figure 2 demonstrates the potential of our algorithm. A speech signal (Figure 2(a)) is recorded inside a car with an inter-microphone spacing of 20 cm (this information is not used by our algorithm). Vehicle noise, recorded with the same setup, is added afterwards so that the global input SNR is 10 dB. This test signal is now further modified, to demonstrate the performance in some extreme cases. The resulting spectrograms of the two microphone channels may be seen in Figure 2(b). During the first 5 seconds of this test signal, noise is added at different frequency bands for the two channels. In the remaining 2.5 seconds speech is missing at different time frames of the two channels. Figure 2(c) shows the weights used for soft decision combining and Figure 2(e) the resulting single channel output signal. Except for the frequency bins where narrow-band noise is added, the SNR is similar in both channels during the first 5 seconds. Consequently, the soft-weight W_{11} is close to one where speech is present, allowing the information in both channels to be used (section 4.4). At the frequency band around 2.5 kHz, channel 1 is disturbed by additional noise. Here, mostly channel 2 will contribute to the output signal. This may be seen in the soft-weights: W_{01} is close to one during speech presence, which means that single channel noise reduction will be performed upon channel 2. The equivalent behavior may be observed at the frequency band around 1.5 kHz, where channel 2 is corrupted. Here, single channel noise reduction is performed on channel 1. For the residual noise, the channel with the lower noise power shall be used (section 4.2). Indeed, ζ is close to zero where channel 1 is corrupted and close to one where channel 2 is corrupted (Figure 2(d)). In the last 2.5 seconds of the example, speech is missing at different time frames in the two microphone channels. It may be seen in Figure 2(c), that the soft-weights accurately choose the channel where speech is present. Speech is well reconstructed in the output signal, as may be seen in Figure 2(e).



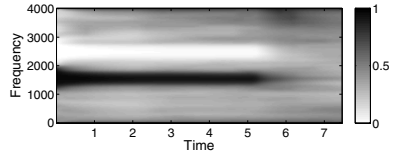
(a) Clean speech (channel 1)



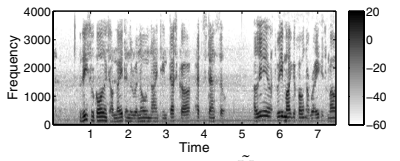
(b) Dual channel inputs Y_1 and Y_2



(c) Soft decision weights



(d) ζ according to (10)



(e) Output $\tilde{Y}$

Figure 2: Demonstration of the algorithm in the TF-plane

6. Evaluation

For the evaluation, we draw on the intelligibility weighted segmental SNR (iwsSNR). We use the *shadow filtering* approach, i.e. the same filters applied to the noisy signals Y_m are applied on the speech and noise signal separately. The filtered signals are indicated by a tilde. The iwsSNR is computed as follows:

$$\text{iwsSNR} \quad = \quad \frac{1}{L}\sum_{l=1}^{L}\sum_{k=k_l}^{k_u} w_{SII}(k)\left[10\log\frac{\tilde{S}_l(k)}{\tilde{N}_l(k)}\right]_{-20}^{+35} \quad (14)$$

The intelligibility weighting factor w_{SII} is defined such that equal weight is given for each auditory critical band between $k_l = 300\,\text{Hz}$ and $k_u = 6400\,\text{Hz}$ [9]. Since our example has a sampling frequency of 8 kHz we set w_{SII} to give equal weights to the bands between $k_l = 300\,\text{Hz}$ and $k_u = 4000\,\text{Hz}$. The segmental SNRs are restricted to be between -20 and +35 dB.

For the evaluation, we use 40 seconds of test data from the TIMIT database and white noise, to achieve results that are easy to reproduce. The phase difference between the channels is zero, i.e. the speakers are located at the broadside of the array (this information is not used by our algorithm). Uncorrelated white gaussian noise is added to both microphone channels. At two different critical bands of the two channels we increase the noise level additionally. This procedure ensures, that the overall iwsSNR is approximately equal in both microphone channels, while the local SNR in the critical bands differ. We corrupt the 8^{th} critical band in channel 1 (1270-1480 Hz), and the 10^{th} critical band in channel 2 (1720-2000 Hz). White noise is added according to an iwsSNR of 21 – 2.5 dB. Independent of this background noise, the noise in the subbands is increased according to a fixed *equivalent broadband SNR*[1] of 0 dB and -20 dB. The results are shown in Table 1. We list the iwsSNR before and after NR for the cases that only white noise is present and for two cases where noise is increased in auditory critical subbands. We compare three different NR algorithms. First we state the results of single channel NR using speech presence uncertainty. Due to the symmetry of the noisy input signals, the output iwsSNRs are similar for both channels, and their mean is stated. The second algorithm we refer to as "dual ch [8]", is based on the multichannel MAP estimator according to [8]. Here, it is assumed that speech is always present in both channels, i.e. $W_{11} = 1$ while all other soft weights are zero. Finally, we state the results of the novel soft decision combining algorithm. If both channels are similar, the dual channel algorithms increase the iwsSNR by 1-2 dB as compared to the single channel filters[2]. When noise is added at different subbands of the two channels, the dual channel algorithm according to [8] performs worse than a single channel NR algorithm, because both corrupted subbands contribute to the output. This case, when the two channels differ at different TF-bins, is when our novel soft decision combining algorithm is most powerful. It will pick the best of both channels, and outperform the other considered algorithms by several dBs, depending on how much the two channels differ.

[1]The *equivalent broadband SNR* is defined as follows: If we would extend the narrowband noise to a broadband noise using the same spectral power in all subbands, we would obtain the "equivalent broadband SNR" values of 0 dB and -20 dB. Thus, the equivalent broadband SNR is computed before narrowband filtering of the noise.

[2]When both channels are similar, W_{01} and W_{10} are very small, and the difference between the two dual channel algorithms is mainly due to the attenuation during speech absence.

iwsSNR broadband noise	NR algorithm	only white noise	*equivalent broadband SNR*[1] of subband noise 0 dB	-20 dB
		overall iwsSNR [dB]		
21.0 dB	no NR	21.0	16.6	13.8
	single ch	21.8	18.2	15.3
	dual ch [8]	22.7	16.3	11.9
	novel dual ch	22.9	20.0	16.7
9.3 dB	no NR	9.3	7.2	5.5
	single ch	10.7	8.9	7.2
	dual ch [8]	12.4	9.3	5.9
	novel dual ch	12.3	10.9	8.9
2.5 dB	no NR	2.5	1.3	0.1
	single ch	3.9	2.9	1.7
	dual ch [8]	5.9	4.3	1.6
	novel dual ch	5.6	4.8	3.2

Table 1: Evaluation results.

7. Conclusion

We have presented a flexible noise reduction algorithm that combines two channels without any knowledge about the geometrical setup and can cope with input signals which differ significantly in their SNR. If one channel is disturbed at time-frequency points where the other is not, the better channel is chosen. If both channels contain speech with similar SNRs, the information of both channels is used for optimal noise reduction.

8. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *ICASSP*, vol. 2, pp. 789–792, 1999.

[3] I. Cohen, "On speech enhancement under signal presence uncertainty," *ICASSP*, vol. 1, pp. 661–664, 2001.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 845–856, 2005.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[6] J. Freudenberger and K. Linhard, "A two-microphone diversity system and its application for hands-free car kits," *Interspeech*, pp. 2329–2332, 2005.

[7] P. Wolfe and S. Godsill, "Simple alternatives to the ephraim and malah supression rule for speech enhancement," *Proc. 11th IEEE Workshop on Stat. Signal Proc.*, pp. 496–499, 2001.

[8] T. Lotter, C. Benien, and P. Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," *EURASIP JASP*, pp. 1147–1156, 2003.

[9] ANSI-S3.5, "American national standard methods for the calculation of the speech intelligibility index," *American National Standards Institute, New York*, 1997.