SPECTRAL COMBINING FOR MICROPHONE DIVERSITY SYSTEMS

Jürgen Freudenberger, Sebastian Stenzel, Benjamin Venditti

Department of Computer Science
University of Applied Sciences Constance, Germany
phone: + (49) 7531 206 647, fax: + (49) 7531 206 559, email: juergen.freudenberger@htwg.konstanz.de
web: www.edc.in.htwg-konstanz.de

ABSTRACT

In telecommunications, diversity combining for multiple receiving antennas is a commonly used technique to achieve robustness for fading channels. This paper proposes a frequency domain diversity approach for two or more microphone signals, e.g. for in-car applications. The microphones should be positioned separately to insure diverse signal conditions. This enables a better compromise for the microphone position with respect to different speaker sizes and noise sources. The microphone signals are weighted with respect to their signal-to-noise ratio and then summed similar to maximum-ratio-combining. The output SNR is significantly improved compared to single microphone noise reduction systems, even if one microphone is heavily corrupted by noise.

1. INTRODUCTION

For safety and comfort reasons, hands-free telephone systems should provide the same quality of speech as conventional fixed telephones. In practice however, the speech quality of a hands-free car kit heavily depends on the particular position of the microphone. Speech has to be picked up as directly as possible to reduce reverberation and to provide a sufficient signal-to-noise ratio. The important question, where to place the microphone inside the car, is, however, difficult to answer. The position is apparently a compromise for different speaker sizes, because the distance between microphone and speaker depends significantly on the position of the driver and therefore on the size of the driver. Furthermore, noise sources like airflow from electric fans or car windows have to be considered. Good noise robustness of single microphone systems requires the use of single channel noise suppression techniques, most of them derived from spectral subtraction [1]. Such noise reduction algorithms improve the signal-to-noise ratio, but they usually introduce undesired speech distortion. Microphone arrays can improve the performance compared to single microphone systems. Nevertheless, the signal quality does still depend on the speaker position. Moreover, the microphones are located in close proximity. Therefore, microphone arrays are often vulnerable to airflow that might disturb all microphone signals.

Alternatively, multi microphone setups have been proposed that combine the processed signals of two or more separate microphones. The microphones are positioned separately (e.g. 80cm apart) in order to ensure incoherent recording of noise [2, 3, 4, 5]. Similar multi channel signal processing systems have been suggested to reduce signal distortion due to reverberation [6, 7]. For hands-free car kits, this diversity technique also enables a better compromise for the microphone position with respect to different speaker sizes, noise sources, and sources of airflow.

In this paper we consider a diversity technique that combines the processed signals of several separate microphones. As we focus on in-car applications our aim is noise robustness. A major issue of multi microphone setups with spread microphones is the coherent addition of the signals [4, 5]. This requires a reliable estimate of the phase differences. In situations where one of the microphone sources is heavily corrupted by noise, this phase estimation is particularly difficult. Therefore, we propose an incoherent combination of the microphone signals. The input signals are used to estimate the power spectrum of the speech signal. The phase of the output signal is the noisy phase of one of the input signals as with most single channel noise suppression techniques.

In section 2, we revise the basic concept of maximum-ratio-combining as required for the following discussion. Some measurement results for the car environment are discussed in section 3. These results motivate spectral diversity combining for multi microphone systems. In the subsequent sections 4 to 6, we describe the signal processing components required for the proposed diversity combining. That is, we consider the design of appropriate signal weights and noise suppression filters based on the noisy observations. Finally, we present some simulation results in section 7.

2. MAXIMUM-RATIO-COMBINING

In telecommunications, the method of diversity combining for multiple receiving antennas is well known. The signals from each channel are added together, using different weights for each channel. The gain of each channel is proportional to the signal level and inversely proportional to the noise level in that channel. From communication theory we know that maximal-ratio-combining (MRC) is the optimum combiner for independent AWGN channels [8].

It is worthwhile to consider this communication situation for a moment. We assume a scenario with M sensors. Let x be the transmitted symbol. The received symbol y_i from the ith sensor is

$$y_i = h_i x + n_i,$$

where h_i is the complex channel coefficient modeling the channel from the transmitter to the i^{th} sensor. n_i is the noise at the i^{th} sensor. Usually, it is assumed that the channel is randomly varying in time, but h_i is known at the receiver. On each receiving antenna, the noise is a Gaussian random variable with zero mean and variance σ_n^2. With maximal-ratio-combining the estimated (equalized) symbol $\hat{x}$ is obtained by the weighted sum

$$\hat{x} = \sum_{i=1}^{M} g_i y_i.$$

SNR IN	100km/h	140km/h	defrost
SNR small speaker	3.9/3	2.0/0	4.4/1.8
SNR tall speaker	2.9/9.0	0.9/7.6	3.4/9.5

Table 1: Input SNR values [dB] from mic. 1/mic. 2 for typical background noise conditions in a car.

The weights g_i depend on the channel coefficients h_i

$$g_i = \frac{h_i^*}{\sum_{j=1}^{M} \mid h_j \mid^2}$$

where h_i^* is the complex conjugate of the channel coefficient h_i. We therefore have

$$\begin{aligned} \hat{x} &= \frac{h_1^*}{\sum_{i=1}^{M} \mid h_i \mid^2} y_1 + \frac{h_2^*}{\sum_{i=1}^{M} \mid h_i \mid^2} y_2 + \dots \\ &= \frac{h_1^*}{\sum_{i=1}^{M} \mid h_i \mid^2} (h_1 x + n_1) + \frac{h_2^*}{\sum_{i=1}^{M} \mid h_i \mid^2} (h_2 x + n_2) + \dots \\ &= x + \frac{h_1^*}{\sum_{i=1}^{M} \mid h_i \mid^2} n_1 + \frac{h_2^*}{\sum_{i=1}^{M} \mid h_i \mid^2} n_2 + \dots . \end{aligned}$$

The estimated symbol $\hat{x}$ is therefore equal to the actually transmitted symbol x plus some weighted noise term.

The overall signal-to-noise ratio of the combined signal is simply the sum of the signal-to-noise ratios of the M received signals [8].

3. MEASUREMENT RESULTS

The basic idea of our spectral combining approach is to apply MRC to speech signals. This is motivated by the measurement results presented in table 1. This table contains the average SNR values for different noise conditions typical in a car. For these measurements we used two cardioid microphones with positions suited for car integration: One microphone (denoted by *mic. 1*) was installed close to the inside mirror in the head unit. The second microphone (*mic. 2*) was mounted at the A-pillar. With an artificial head we recorded speech samples in two different seat positions. Therefore we considered two speaker sizes: a tall speaker of about 194cm height and a small speaker of 164cm. With respect to three different background noise situations, we recorded driving noise at 100km/h and 140km/h. As third noise situation we considered the noise which arises from an electric fan (defroster). For all recordings we used a sampling rate of 11025 Hz. Apparently, for tall speakers the microphone position 2 is the preferred position. For small speakers the position 1 provides the better results.

The SNR differences are even more pronounced when we consider the SNR versus frequency as for example depicted in Fig. 1. From this figure we observe that the SNR values are quite distinct for these two microphone positions with differences of up to 10dB depending on the particular frequency. We also note that the better microphone position is not obvious in this case, because the SNR curves cross several times.

Theoretically, a MRC combining of the two input signals would result in an output SNR equal to the sum of the input SNR values. With two inputs, MRC achieves a maximum gain of 3dB for equal input SNR values. In case of the input SNR values being rather different, the sum is dominated by the maximum value. Hence, for the curves in Fig. 1 the output SNR would essentially be the envelope of the two curves.

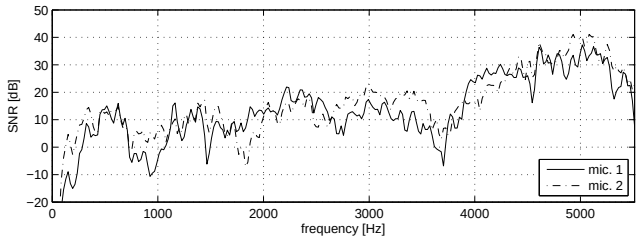

Fig. 1: Input SNR values for a driving situation at a car speed of 100km/h.

4. SPECTRAL COMBINING

The results of the previous section motivate a diversity technique that combines the processed signals of the different microphones in the frequency domain. Ideally we would apply maximum-ratio-combining for each frequency. The problem at hand is that different to the situation in radio communication we have no means to explicitly estimate the room transfer characteristic for our microphone system. In this section, we show that MRC combining can be achieved without explicit knowledge of the acoustic channels. The weights for the different microphones can be calculated based on an estimate of the signal-to-noise ratio for each microphone.

We consider the spectrum of the microphone signal

$$Y_i(f) = H_i(f)X(f) + N_i(f),$$

where $X(f)$ and $N_i(f)$ are the spectra of the speech signal and the noise at the i^{th} microphone, respectively. $H_i(f)$ is the channel transfer function, i.e. the Fourier transform of the impulse response from the speaker to the i^{th} microphone.

With MRC we would weight each microphone input signal $Y_i(f)$ with the gain factor

$$G_i(f) = \frac{H_i^*(f)}{\sum_{j=1}^{M} \mid H_j(f) \mid^2}. \quad (1)$$

In general, the MRC weighting and the coherent addition require exact knowledge of the absolute value and the phase of $H_i(f)$. With speech signals we have no means to explicitly estimate the transfer functions $H_i(f)$ from the speaker to the i^{th} microphone. We therefore estimate the gain factor (1) based on the signal-to-noise ratios of the microphone signals.

Let

$$\gamma_i(f) = \frac{\mathbb{E}\left\{ \mid H_i(f)X(f) \mid^2 \right\}}{\mathbb{E}\left\{ \mid N_i(f) \mid^2 \right\}}$$

denote the signal-to-noise-ratio for the i^{th} microphone at frequency f. For the sake of simplicity, we assume that the speech signal and the noise signals are stationary processes with zero mean, where the noise power is the same for all microphones. Furthermore, we assume that the speech signal and the channel coefficients are uncorrelated. Thus, we obtain

$$\gamma_i(f) = \frac{\mathbb{E}\left\{ \mid H_i(f)X(f) \mid^2 \right\}}{\mathbb{E}\left\{ \mid N_i(f) \mid^2 \right\}} = \frac{\mid H_i(f) \mid^2 P_X(f)}{P_N(f)}, \quad (2)$$

where $P_X(f)$ and $P_N(f)$ are the power spectral densities of the speech and noise process at frequency f, respectively.

We consider the weights

$$\tilde{G}_i(f) = \sqrt{\frac{\gamma_i(f)}{\sum_{j=1}^{M} \gamma_j(f)}} \tag{3}$$

Substituting $\gamma_i(f)$ by equation (2) leads to

$$\tilde{G}_i(f) = \sqrt{\frac{|\,H_i(f)\,|^2}{\sum_{j=1}^{M} |\,H_j(f)\,|^2}} = \frac{|\,H_i(f)\,|}{\sqrt{\sum_{j=1}^{M} |\,H_j(f)\,|^2}}.$$

Hence, with equation (1) we have

$$|\,G_i(f)\,| = \tilde{G}_i(f) \frac{1}{\sqrt{\sum_{j=1}^{M} |\,H_j(f)\,|^2}}.$$

We observe that the magnitude of the weight $\tilde{G}_i(f)$ is proportional to the absolute value of the MRC weights according to equation (1), because the factor

$$\frac{1}{\sqrt{\sum_{j=1}^{M} |\,H_j(f)\,|^2}}$$

is the same for all M microphone signals. Consequently, coherent addition of the sensor signals weighted with the gain factors $\tilde{G}_i(f)$ still leads to a combining, where the signal-to-noise ratio at the combiner output is the sum of the input SNR values. Thus, we obtain the overall SNR

$$\gamma(f) = \sum_{i=1}^{M} \gamma_i(f). \tag{4}$$

5. NOISE SUPPRESSION

Maximum-ratio-combining provides an optimum weighting of the M sensor signals. However, it does not necessarily suppress the noisy signal components. We therefore combine the spectral combining with an additional noise suppression filter. Of the numerous proposed noise reduction techniques in the literature, we consider only spectral subtraction which supplements the spectral combining quite naturally. With the overall SNR $\gamma(f)$ the spectral subtraction filter for the combined signal can be written as

$$G_{NS}(f) = \sqrt{\frac{\gamma(f)}{1+\gamma(f)}} \quad . \tag{5}$$

Multiplying this filter transfer function with equation (3) leads to

$$\begin{aligned} \hat{G}_i(f) &= \tilde{G}_i(f) G_{NS}(f) \\ &= \sqrt{\frac{\gamma_i(f)}{\gamma(f)}} \sqrt{\frac{\gamma(f)}{1+\gamma(f)}} \\ &= \sqrt{\frac{\gamma_i(f)}{1+\gamma(f)}} \quad . \end{aligned} \tag{6}$$

This formula shows that noise suppression can be introduced by simply adding a constant to the denominator term in equation (3).

Most, if not all, implementations of spectral subtraction are based on an over-subtraction approach, where an overestimate of the noise power is subtracted from the power spectrum of the input signal [9]. Over-subtraction can be included in equation (6)) by using a constant ρ larger than one. This leads to the final gain factor

$$\hat{G}_i(f) = \sqrt{\frac{\gamma_i(f)}{\rho+\gamma(f)}} \quad . \tag{7}$$

The parameter ρ does hardly affect the gain factors for high signal-to-noise ratios retaining optimum weighting. For low signal-to-noise ratios this term leads to an additional attenuation. The over-subtraction factor is usually a function of the SNR, sometimes it is also chosen differently for different frequency bands [10]. Nevertheless, a constant value in the range $\rho \in [5, 10]$ leads to reasonable results.

Real world speech and noise signals are non stationary processes. For an implementation of the spectral weighting we have to estimate the short-time power spectral densities (PSD) of the speech signal and the noise components. However, only the noisy speech signals are available. We therefore have to estimate the current signal-to-noise ratio based on the noisy microphone input signals. This is commonly done by using voice activity detection (VAD, see e.g. [11]) or minimum statistics [12] to estimate the noise power spectral density $\mathbb{E}\{|\,N_i(f)\,|^2\}$. The current signal-to-noise ratio is then obtained by

$$\gamma_i(f) = \frac{\mathbb{E}\{|\,Y_i(f)\,|^2\} - \mathbb{E}\{|\,N_i(f)\,|^2\}}{\mathbb{E}\{|\,N_i(f)\,|^2\}},$$

assuming that the noise and speech signals are uncorrelated.

6. MAGNITUDE COMBINING

Similar to the problem of SNR estimation the coherent signal combination requires an estimate of the phase differences of the input signals. Let $\phi_i(f)$ denote the phase of $H_i(f)X(f)$ at frequency f. For signal processing of speech signals our aim is usually not to restore the absolute phase of the speech signal. Hence, it is sufficient to take care of the phase differences between the microphone input signals. However, the phase differences can only be reliably estimated during speech activity. Estimating the phase difference $\phi_{\Delta,i}(f) = \phi_1(f) - \phi_i(f), \forall i = 2, \ldots, M$

$$e^{j\phi_{\Delta,i}(f)} = E\left\{\frac{Y_1(f)Y_i(f)^*}{|\,Y_1(f)\,|\;|\,Y_i(f)\,|}\right\}$$

leads to unreliable phase values for all time-frequency points without speech activity. Diversity combining using this estimate leads to additional signal distortions. A coarse estimate of the phase difference can also be obtained from the time-shift τ_i between the direct path components in both room impulse responses. This time-shift can for example be found by searching for the maximum value of the cross-correlation of the two input signals. The estimate is then $\phi_{\Delta,i}(f) \approx 2\pi f \tau_i$. Note that a combiner using these phase values would in a certain manner be equivalent to a delay-and-sum beamformer. However, for distributed microphone arrays this phase compensation leads to a poor estimate of the actual phase difference.

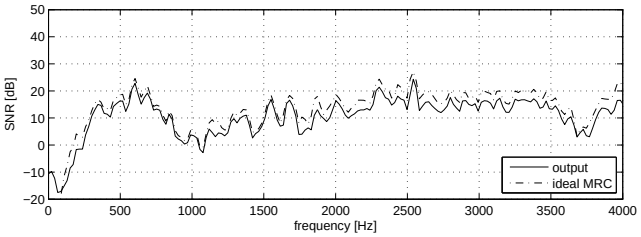

Fig. 3: Output SNR values for spectral combining without additional noise suppression (car speed of 100km/h, $\rho = 0$).

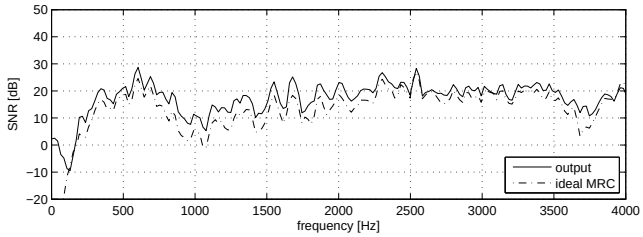

Fig. 4: Output SNR values for spectral combining with additional noise suppression (car speed of 100km/h, $\rho = 10$).

	100km/h	140km/h	defrost
SNR small speaker	3.4	0.6	2.7
SNR tall speaker	8.8	6.1	8.3

Table 2: Output SNR values [dB] for diversity combining without additional noise reduction ($\rho = 0$).

We therefore use a simple magnitude combining approach, where the phase of the output signal is the noisy phase of one of the input signals. Let $\phi_1(f)$ be the phase of the first input signal. With magnitude combining we calculate the combined signal as follows

$$\begin{aligned} \hat{X}(f) &= \hat{G}_1(f)Y_1(f) + \hat{G}_2(f) \mid Y_2(f) \mid e^{j\phi_1(f)} + \ldots \\ &\quad + \hat{G}_M(f) \mid Y_M(f) \mid e^{j\phi_1(f)} \quad . \end{aligned} \tag{8}$$

A corresponding processing system for two inputs is depicted in Fig. 2.

7. SIMULATION RESULTS

For our simulations we consider the same microphone setup as described in section 3. For the spectral combining we used an FFT length of 256 and a Hamming window for time windowing.

Figure 3 presents the output SNR values for a driving situation with a car speed of 100km/h. For this simulation we used $\rho = 0$, i.e. spectral combining without noise suppression. In addition to the output SNR, the curve for ideal maximum-ratio-combining is depicted. This curve is simply the sum of the input SNR values for the two microphones which we calculated based on the actual noise and speech signals. We observe that the output SNR curve closely follows the ideal curve but with a loss of 1-3dB. This loss is essentially caused by the estimation of the power spectral densities which is based on the noisy microphone signals. Table 2 provides the average SNR values for all considered driving situations. Comparing these results with the input SNR values in table 1 we note that the output SNR is close to the maximum of the input SNR values.

In the following we consider the spectral combining with additional noise suppression ($\rho = 10$). Figure 4 presents the corresponding results for a driving situation with a car speed of 100km/h. The output SNR curve still follows the ideal MRC curve but now with a gain of up to 5dB. More simulation results are presented in Tab. 3 and Tab. 4. As an objective measure of speech distortion we calculated the cosh spectral distance (a symmetrical version of the Itakura-Saito distance) between the power spectra of the clean input signal (without reverberation and noise) and the output speech signal (filter coefficients were obtained from noisy data).

	100km/h	140km/h	defrost
SNR small speaker	13.1/10.6	10.7/6.8	11.6/8.7
SNR tall speaker	12.3/17.0	9.7/13.3	10.8/15.6
dist. small speaker	1.8/2.0	1.8/1.9	1.2/1.3
dist. tall speaker	2.7/1.3	2.5/1.3	1.8/1.2

Table 3: Output SNR values [dB] and cosh spectral distances for single channel noise reduction with input signal from mic. 1/mic. 2, respectively.

Table 3 provides results for single channel noise reduction, where we used spectral subtraction as proposed in [11]. We observe from these results that the position of microphone 1 would be a suitable compromise for both speaker sizes, whereas position 2 would result in up to 5dB better SNR for a tall speaker. The results for the diversity combining scheme are given in Table 4. The SNR values are slightly better than the best value of the single channel noise suppression, where the signal distortion is similar to the single channel case. The speech is free of musical tones and sounds more natural compared to ordinary spectral subtraction. The lack of musical noise can also be seen in Fig. 5, which shows the spectrograms of enhanced speech and the input signals. The improved signal quality can be explained by the dereverberation effect of the diversity combining. The spectral combining equalizes frequency dips that occur only in one microphone input (compare Fig. 1 and Fig. 3).

8. CONCLUSIONS

In this paper we have presented a diversity technique that combines the processed signals of several separate microphones. The aim of our approach was noise robustness for in-car hands-free applications, because single channel noise suppression methods are sensitive to the microphone location and in particular to the distance between speaker and microphone.

	100km/h	140km/h	defrost
SNR small speaker	13.2	11.3	12.4
SNR tall speaker	17.6	15.8	17.5
dist. small speaker	1.3	1.4	1.6
dist. tall speaker	1.2	1.4	1.6

Table 4: Output SNR values [dB] and cosh spectral distances for diversity combining ($\rho = 10$).

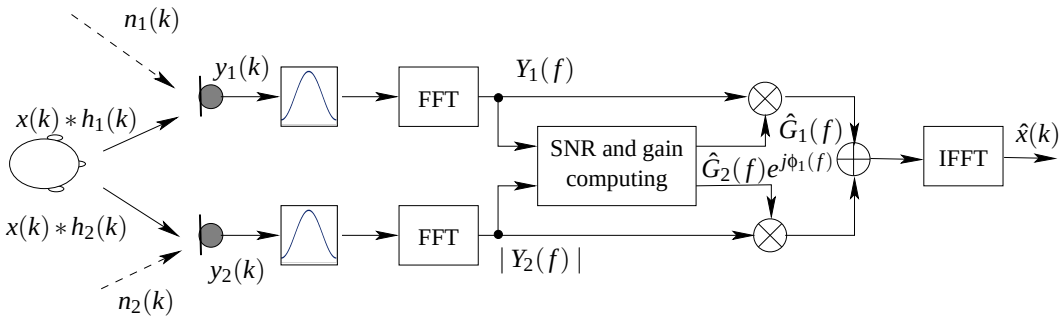

Fig. 2: Basic system structure of the two-channel diversity system.

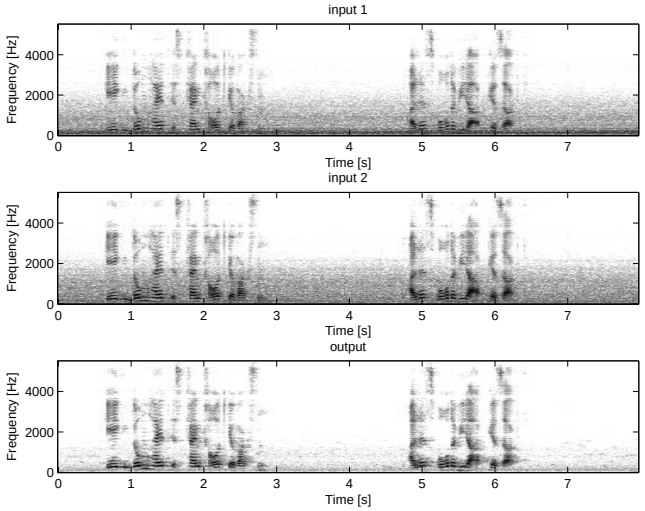

Fig. 5: Spectrograms of the input and output signals (car speed of 100km/h, $\alpha = 10$).

We have shown theoretically that the proposed signal weighting is equivalent to maximum-ratio-combining. Here we have assumed that the noise power spectral densities are equal for all microphone inputs. This assumption is of course unrealistic. However, the simulation results for a two microphone system demonstrate that a performance close to that of MRC can be achieved with real world noise situations. These results were obtained with an SNR estimate based on voice activity detection and with magnitude combing, i.e. without a phase estimation.

The proposed diversity scheme achieves better SNR values than the better of the two single channel systems and is therefore less sensitive to varying speaker positions.

REFERENCES

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

[2] A. Akbari Azirani, R. Le Bouquin-Jeannes, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech, and Audio Processing*, vol. 5, no. 5, pp. 484–487, 1997.

[3] A. Guerin, R. Le Bouquin-Jeannes, and G. Faucon, "A two-sensor noise reduction system: applications for hands-free car kit," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1125–1134, 2003.

[4] J. Freudenberger and K. Linhard, "A two-microphone diversity system and its application for hands-free car kits," in *European Conference on Speech Communication and Technology (INTERSPEECH), Lisbon*, 2005.

[5] T. Gerkmann and R. Martin, "Soft decision combining for dual channel noise reduction," in *The Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP), Pittsburgh*, 2006, p. 21342137.

[6] J. L. Flanagan and R. C. Lummis, "Signal processing to reduce multipath distortion in small rooms," *The Journal of the Acoustical Society of America*, vol. 47, no. 6, pp. 1475–1481, June 1970.

[7] J. B. Allen, D. A. Berlkey, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[8] B. Sklar, *Digital Communications: Fundamentals and Applications*, Prentice Hall, 2001.

[9] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 208–211, 1979.

[10] A. Juneja, O. Deshmukh, and C. Espy-Wilson, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, pp. 4164, 2002.

[11] H. Puder, "Single channel noise reduction using time-frequency dependent voice activity detection," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), Pocono Manor*, 1999, pp. 68–71.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, Jul. 2001.