MULTI-CHANNEL SPEECH ENHANCEMENT IN A CAR ENVIRONMENT USING WIENER FILTERING AND SPECTRAL SUBTRACTION

Joerg Meyer and Klaus Uwe Simmer

University of Bremen, FB–1, Dept. of Telecommunications
P.O. Box 33 04 40, D-28334 Bremen, Germany, Fax: +(49)-421/218-3341, e-mail: meyer@comm.uni-bremen.de

ABSTRACT

This paper presents a multichannel-algorithm for speech enhancement for hands–free telephone systems in cars. This new algorithm takes advantage of the special noise characteristics in fast driving cars. The incoherence of the noise allows to use adaptive Wiener filtering in the frequencies above a theoretically determined frequency. Below this frequency a smoothed spectral subtraction (SSS) is used to get an improved noise suppression. The algorithm yields better results in noise reduction with significantly less distortions and artificial noise than spectral subtraction or Wiener filtering alone.

1. INTRODUCTION

The handset equipment for telephones in cars is a restriction and a potential risk for the driver. Only hands–free devices can overcome this problem. Two different approaches for hands–free devices can be pursued. The first one uses only one microphone [1, 2], whereas the second one is a multichannel approach [3, 4]. The most often used single-sensor method is spectral subtraction.

However, this method introduces various other problems, as eg. musical tones. Using smoothed versions of the spectral subtraction in order to avoid these tones leads to signal distortions. All multi-channel approaches, in contrast, use many microphones and cause high computational costs. With less microphones they do not work well on the special noise field in cars due to the dominating noise power at low frequencies.

In the first part, this contribution presents the special noise conditions for multi-channel recordings in cars. Then, the new algorithm is introduced. The final part shows the experimental results compared to other algorithms.

2. NOISE FIELD

The noise field in a car varies in many aspects and depends on the position and type of the microphones, the road, the speed of the car, and the car type. In this study, four omnidirectional microphones are fixed at the left side of the windscreen in distances of 10cm. (In the car industry, this is one of the favoured potential ways to arrange microphones in cars) The car used was a mid-sized Volkswagen (Diesel).

In order to describe the special noise conditions in our experiment, we have measured the power spectral density and the coherence function. The main energy is concentrated at low frequencies (see figure 1), and it depends on the speed of the car. At higher speed, more high frequency components appear, caused by wind noise. The SNR varies from -5dB to +10dB according to the noise situation.

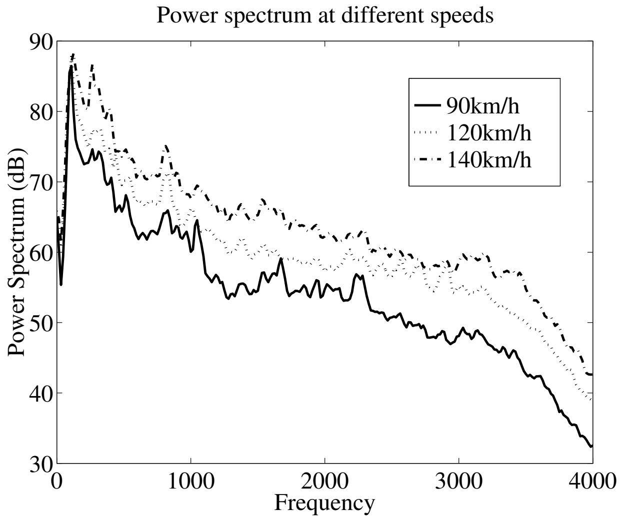

Figure 1: Power spectrum measured in a car

To show whether or not the noise field is diffuse we use the complex coherence function , defined as

$$\Gamma_{ij}(\omega) = \frac{P_{X_{ij}}(\omega)}{\sqrt{P_{X_{ii}}(\omega)\, P_{X_{jj}}(\omega}} \tag{1}$$

where $P_{X_{ij}}$ denotes the crosspower spectral density of the signals x_i and x_j. $P_{X_{ii}}$ and $P_{X_{jj}}$ are the autopower spectral densities. If the signal is spatially diffuse, then

$$\Gamma_{ij}(\omega) = \mathrm{si}\left(\omega \frac{d_{ij}}{c}\right) \tag{2}$$

where d_{ij} denotes the distance between the sensors and c the speed of sound [5]. Another importent tool to describe the acoustic environment is the well known magnitude squared coherence function (MSC)

$$C(\omega) = \frac{|P_{X_{ij}}(\omega)|^2}{P_{X_{ii}}(\omega) P_{X_{jj}}(\omega)} \quad . \tag{3}$$

Compressed postscript files of our publications are readily available from our WWW server `http://www.comm.uni-bremen.de`.

The MSC for the spatially diffuse noise field is simply the square of eq. 2. In cars the noise field is mainly diffuse (see figure 2). This characteristic does not change with the speed.

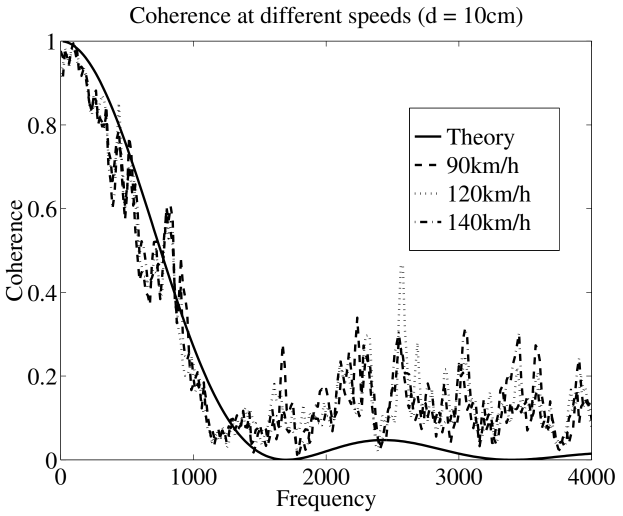

Figure 2: MSC measured in a car (microphone distance = 10cm)

3. ALGORITHM

Because of the noise condition in cars an efficient algorithm has to work well for diffuse noise fields, and it has to have high capabilities to suppress low frequencies. The proposed algorithm consists of three parts (see block diagram 3): Delay & sum beamformer (DS), a spectral subtraction algorithm for low frequencies and a Wiener filter for high frequencies.

After a supposed ideal time delay compensation and a FFT analysis (Hamming window, 50% overlap), our algorithm devides the spectrum in two parts. The lower region with high coherence and the upper part with less coherence. The transient frequency we use, is the first minimum of the MSC function. This minimum is given by $f = c/(2d)$. where c denotes the speed of sound and d the distance of the microphones. The high frequencies are processed by an adaptive Wiener filter, because of the small coherence in this region. It can be shown that the adaptive Wiener filter works well in this case.

The DS beamformer and the Wiener filter have similar properties because their noise reduction factors are functions of the average complex coherence function of the noise field for all sensor pairs $i \neq j$.

$$\bar{\Gamma}(\omega) = \frac{2}{N^2 - N} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} Re\left\{\Gamma_{ij}(\omega)\right\} \tag{4}$$

In a diffuse noise field $\bar{\Gamma}$ is completely determined by the distance d_{ij} between the sensors. The first stage of the system is a DS beamformer. The power spectral density (psd)

$$Y_b(\omega) = \frac{1}{N} \sum_{i=0}^{N-1} X_i(\omega) \tag{5}$$

at the output of the beamformer can be computed as follows

$$\begin{aligned} P_{Y_b Y_b}(\omega) &= Y_b(\omega) Y_b^*(\omega) \\ &= \left(\frac{1}{N}\sum_{i=0}^{N-1} X_i(\omega)\right)\left(\frac{1}{N}\sum_{i=0}^{N-1} X_i^*(\omega)\right) \\ &= \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} X_i(\omega) X_j^*(\omega) \\ &= \frac{1}{N^2}\sum_{i=0}^{N-1} X_i(\omega) X_i^*(\omega) \\ &\quad + \frac{1}{N^2}\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1}\left(X_i(\omega)X_j^*(\omega) + X_j(\omega)X_i^*(\omega)\right) \\ &= \frac{1}{N^2}\sum_{i=0}^{N-1} P_{X_{ii}} + \frac{2}{N^2}\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} Re\left\{P_{X_{ij}}\right\} \end{aligned} \tag{6}$$

If we assume that signal s and noise n are uncorrelated, $P_{X_{ii}} = P_{ss}$ at all sensors and $\bar{\Gamma}(\omega) = 1$ for the signal s then the psd at the output of the beamformer is

$$P_{Y_b Y_b}(\omega) = P_{ss}(\omega) + P_{nn}(\omega)\left(\frac{1}{N} + \left(1 - \frac{1}{N}\right)\bar{\Gamma}(\omega)\right) \tag{7}$$

The improvement of the SNR or noise reduction factor $NR(\omega)$ of the DS beamformer is given by

$$NR_b(\omega) = \frac{1}{\frac{1}{N} + \left(1 - \frac{1}{N}\right)\bar{\Gamma}(\omega)} \tag{8}$$

NR_b is close to N for frequencies where $\bar{\Gamma}$ is small and close to one if $\bar{\Gamma}$ approaches 1.

The noise reduction factor $NR_w(\omega)$ of the post filter can be computed from its transfer function.

$$NR_w(\omega) = \frac{1}{W^2(\omega)} \tag{9}$$

We have chosen the following estimate for the post filter transfer function

$$\begin{aligned} \widehat{W}(\omega) &= \frac{\frac{2}{N^2 - N}\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} Re\left\{X_i(\omega)X_j^*(\omega)\right\}}{\left|Y_b(\omega)\right|^2} \\ &= \frac{\frac{2}{N^2 - N}\sum_{i=0}^{N-2}\sum_{j=i+1}^{N-1} Re\left\{P_{X_{ij}}\right\}}{P_{Y_b Y_b}} \end{aligned} \tag{10}$$

Using the same assumptions as in the case of the DS beamformer the transfer function of post filter can be written

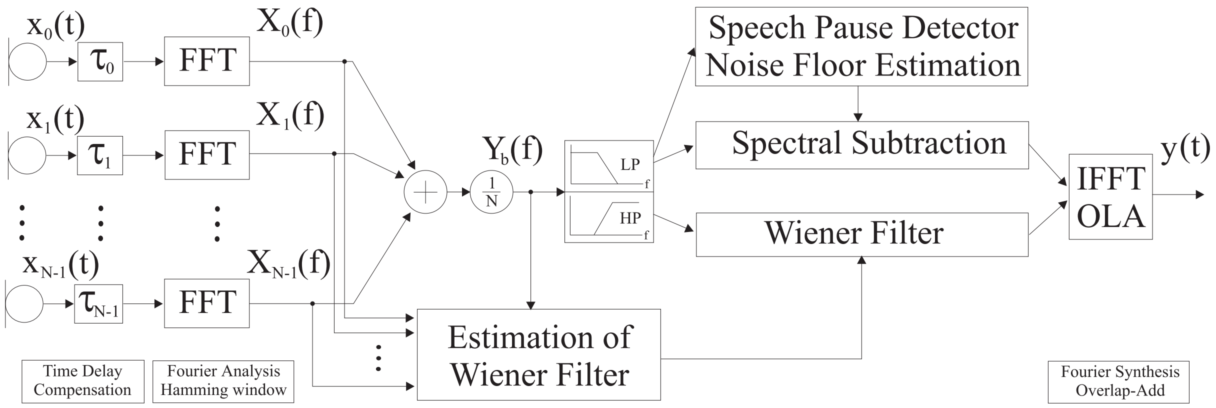

Figure 3: Microphone array with adaptive postfilter or spectral subtraction.

as

$$\begin{aligned}\widehat{W}(\omega) &= \frac{P_{ss}(\omega)+P_{nn}(\omega)\bar{\Gamma}(\omega)}{P_{ss}(\omega)+P_{nn}(\omega)\left(\frac{1}{N}+\left(1-\frac{1}{N}\right)\bar{\Gamma}(\omega)\right)} \\ &= \frac{P_{ss}(\omega)}{P_{ss}(\omega)+P_{nn}(\omega)NR_b(\omega)} \\ &\quad+\frac{P_{nn}(\omega)}{P_{ss}(\omega)+P_{nn}(\omega)NR_b(\omega)}\bar{\Gamma}(\omega) \qquad (11) \\ &= \frac{SNR_b(\omega)}{1+SNR_b(\omega)}+\frac{NR_b(\omega)}{1+SNR_b(\omega)}\bar{\Gamma}(\omega) \qquad (12)\end{aligned}$$

where $SNR_b(\omega)$ is the signal to noise ratio at the output of the beamformer. A similar result, including array shading and steering delay effect, can be found in [6]. If $\bar{\Gamma}(\omega)=0$ then $\widehat{W}(\omega)$ is a Wiener filter

$$W(\omega)=\frac{SNR_b(\omega)}{1+SNR_b(\omega)} \qquad (13)$$

that can have an arbitrarily high noise reduction factor for frequencies with low SNR. However, equation 12 also shows that the post filter has a poor performance at low frequencies where $\bar{\Gamma}(\omega)$ is close to 1. To overcome this problem we use spectral subtraction for low frequencies. If the noise $n(t)$ is stationary and uncorrelated with the speech signal $s(t)$, the power spectrum of the noisy speech $x(t)$ is the sum of the power spectra of the speech and of the noise. Therefore, a clean speech signal power spectrum can be estimated by subtracting the current noise power spectrum. To unterstand how to get a clean speech signal we introduce the interpretation of the spectral subtraction as a time-varying filter.

$$\begin{aligned}\hat{Y}(\omega) &= H(\omega)\,Y_b(\omega) \qquad (14) \\ H(\omega) &= \sqrt{\frac{P_{Y_bY_b}-P_{nn}}{P_{Y_bY_b}}} \qquad (15) \\ \hat{y}(t) &= \mathrm{IFFT}\{\hat{Y}(\omega)\} \qquad (16)\end{aligned}$$

However, the noise power spectrum cannot be estimated during the speech activity. But it is possible to get an estimation of the noise power spectrum by averaging the noise spectrum during non speech intervalls. Due to the fluctuation in the current noise spectrum, randomly spaced spectral peaks appear, called the musical tones. There are different techniques to reduce this effect eg. [7]. We use:

- Overestimating the noise ($\alpha > 1$)
- Soft rectification ($0 < \beta \ll 1$)
- Averaging the PSD of the current signal

The final estimation of the filter is:

$$H(\omega)=\max\left\{\beta\,\bar{P}_{nn}\left|\sqrt{\frac{\bar{P}_{Y_bY_b}-\alpha\bar{P}_{nn}}{\bar{P}_{Y_bY_b}}}\right.\right\} \qquad (17)$$

The final step is the combination of the two output spectra of the algorithms and the following inverse Fourier transform and the overlap add method.

4. EXPERIMENTS AND RESULTS

We generate noisy signals by artificially adding clean speech signals to real multi-channel car noise data with different SNRs. We assume ideal equalisation and an ideal gain normalisation. The results present a comparison between the noise reduction behaviours of the SSS and the new algorithm. For objective measurement of speech distortion we use the log area ratio distance (LAR) which has been shown in [8] to have the highest correlation between subjective and objective speech quality measurement .

The results for the noise reduction are shown in figure 4. The new algorithm yields a better noise reduction at all SNRs. The objective speech quality measurement results are more complex (see figure 5). For a good SNR the speech quality for the Wiener filter is little better than for the new algorithm. The new algorithm outperforms SSS in all cases. Informal listening tests confirm these results. The processed speech sounds more natural and less muffled as with the SSS.

In a second test we used a speaker–independent isolated word recognition system to test the performance of the new algorithm for dialing and commands. The system was trained with 16 persons, and it was tested with four

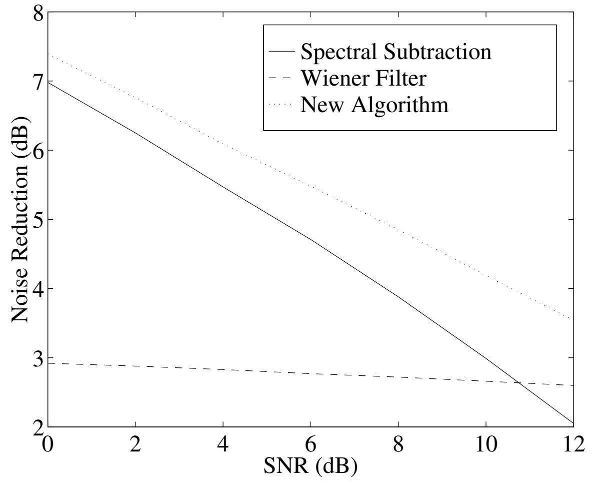

Figure 4: Noise reduction vs. input SNR

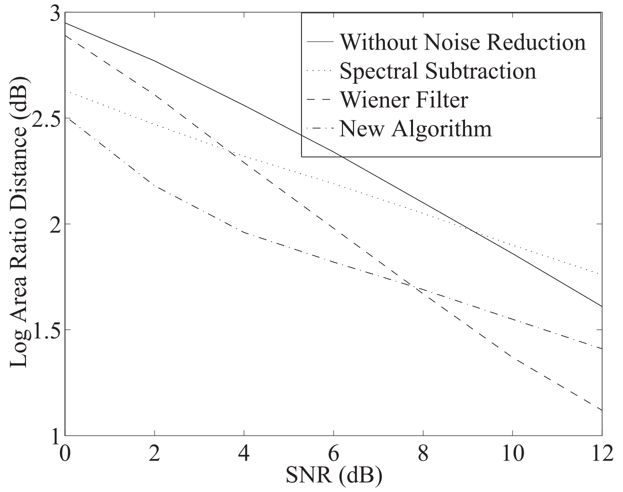

Figure 5: Log area ratio distance vs. input SNR

other persons. The vocabulary consists of 40 words, including the 10 digits and words to control computers. For the results see figure 6. The new algorithm significantly increases the recognition rate.

5. CONCLUSION

This paper has presented a new algorithm for multichannel speech enhancement in cars. The proposed algorithm is a combination of a DS beamformer, a Wiener filter and spectral subtraction. This combination is justified by the theoretical limits of the performance of delay & sum beamformer and post filter. The new combination overcomes the problem of too many microphones and reduces the annoying musical tones. In simulated experiments with only four microphones and real–world noise signals the new algorithm outperforms related algorithms like spectral subtraction or adaptive Wiener filtering.

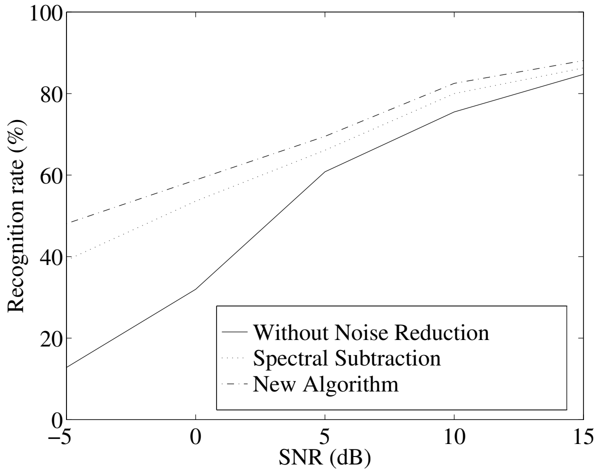

Figure 6: Speaker–independent word recognition rate

6. REFERENCES

[1] L. Arslan, A. McCree, and V. Viswanathan, "New methods for adaptive noise suppression," in *Proc. IEEE Int. Conference Acoustic, Speech and Signal Processing, ICASSP–95*, (Detroit, Michigan), pp. 812–815, Mai 1995.

[2] R. Le Bouquin, "Enhancement of noisy speech signals: Application to mobile radio communication," *Speech Communication*, vol. 18, pp. 3–19, 1996.

[3] A. Affes and Y. Grenier, "Test of adaptive beamformers for speech acqusition," in *Int. Conference on Signal Processing Applications and Technology, ICSPAT 94*, (Dallas, Texas), pp. 154–159, October 1994.

[4] S. Nordholm, I. Claesson, and I. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. on Vehicular Technology*, vol. 42, no. 4, pp. 514–518, November 1993.

[5] J. S. Bendat and A.G. Piersol, *Engineering applications of correlation and spectral analysis.* Wiley Interscience, New York, 1980.

[6] C. Marro, Y. Mahieux, and K.U. Simmer, "Performance of adaptive dereverberation techniques using directivity controlled arrays," in *European Signal Processing Conference (EUSIPCO–96)*, (Trieste, Italy), pp. 1127–1130, September 1996.

[7] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conference Acoustic, Speech and Signal Processing, ICASSP–79*, (Washington DC), pp. 208–211, April 1979.

[8] S. R. Quakenbusch, T. P. Barnwell, and B.A. Clemens, *Objective Measures of Speech Quality.* Prentice-Hall, Englewood Cliffs, NJ, 1988.