DUAL-CHANNEL NOISE REDUCTION FOR A VERY SMALL MICROPHONE ARRANGEMENT

F. Cheikh-Rouhou[1] *S. Euler* *U. Gärtner* *U. Haiber*

Darmstadt University Robert-Bosch Robert-Bosch DaimlerChrysler AG
Theory of Signals Dept. FV/FLI Dept. CM/EFS32 Dept. FT3/AV
Merckstr. 25 P.O. Box 106050 Robert-Bosch Str. 200 P.O. Box 2360
D-64283 Darmstadt D–70049 Stuttgart D–31139 Hildesheim D-89013 Ulm

ABSTRACT

In this paper we describe our setup and some first experiments to evaluate the potential of the use of two microphones with a very small distance for noise reduction in the car. We examine the improvement with respect both to speech quality and the error rate of a speech recognition system.

1. INTRODUCTION

As there is an ever-increasing demand for safety in cars, hands-free telephone systems and speech-controlled equipment are nowadays often used. Optimal functionality of such systems requires a sufficient signal-to-noise-ratio (SNR). However, in a car the SNR is frequently very low due to the various types of noise being present. For this reason the enhancement of the SNR by means of a noise reduction unit is often required.

In this contribution the potential of dual channel approaches to noise reduction with very small microphone arrangements is investigated. The work is part of the European research project SENECA (**S**peech control modules for **E**ntertainment, **N**avigation and communication **E**quipment in **CA**rs) [1] in the ESPRIT IV framework. It is the objective of the SENECA project to develop a robust and inexpensive multi lingual dialogue system for the automotive environment.

One major problem is the quality of speech signals recorded through a far talk microphone in a car. Besides the speech signal from the driver the microphone signal in general contains background noise and in a telephone conversation echos from the remote speaker. Within the SENECA project algorithms for noise reduction and echo compensation have been investigated and implemented on the target DSP platform. In the first phase of the project single channel approaches have been considered. Furthermore, in the second phase of the project, the extension to two microphones has been examined.

[1] Now with Deutsche Bank

The automotive market is very cost sensitive. Thereby the tradeoff between improved speech quality and additional costs has to be taken into account. One considerable cost factor is the installation of microphones at different places in the car. In our work we examined the possibility of providing a cost efficient solution by mounting two microphones in one housing. This requires a small distance between the microphones (maximal distance of 5 cm).

In SENECA we have developed a framework to investigate the potential of this approach. The goal of the noise reduction is twofold: Enhanced speech quality for hands free telephony and improved recognition performance in the presence of noise. In the following we first will discuss the selection of the microphone positions and the collection of a multi-channel data base. We then will describe our methods for evaluation of subjective quality and recognition performance. Next, an approach for two-channel noise reduction is presented. Finally, we report first results for this new approach.

1.1. Microphone position

In a first test we recorded speech and noise signals with two microphones in different geometries in order to find an optimal arrangement. In all cases the microphones were mounted near the mirror in the car. Mainly based on the measured coherence function between the noise signals we decided to use a arrangement in which one microphone is pointing to the speaker and the other into the opposite direction (back-toback). The distance between these microphones is around 4 cm.

Additionally, we used two more microphones that are also positioned very close to the first microphone but both point in the direction of the speaker. This arrangement is intended to examine alternative noise reduction approaches.

2. DATABASE

In order to evaluate the system with regard to the two criteria mentioned above we collected a database with

the special microphone arrangement. The database consists of 20 sessions from 20 different speakers.

Each session contains 5 phonetically balanced sentences and 55 spelled city names. The multichannel platform developed in the SpeechDat-Car project [2] was used. This platform allows to record four channels simultaneously. The data is sampled at 16kHz and quantized with 16 bits. The prompts for the different items are shown on a display. In all cases the driver is the speaker. An experimenter guides the speaker through the recording session. The collected speech material is then transcribed according to the SpeechDat-Car specifications.

In general, each session was recorded in a specific driving condition such as city or highway traffic. For the subjective listening tests, however, we wanted to use examples from a given speaker in different driving conditions. Therefore, we collected in four sessions (2 male and 2 female speakers) two test sentences in four different driving conditions: car stopped (motor running), city traffic, acceleration and constant speed 120 km/h. We choose the two sentences [1]

1. August Macke schrieb diese Zeilen neunzehn hundert zehn an seinen Freund
2. Die Burg Altenahr stammt aus dem elften Jahrhundert

3. SUBJECTIVE EVALUATION

The speech quality is examined in subjective listening tests with a number of different listeners. The signal quality is rated on an absolute scale and the original – noisy – recordings are included as reference in the test. The test modalities are designed as follows:

- absolute scaling on a rating scale with 5 steps
- random mixture of test stimuli in the listening test with 2 seconds evaluation pause after each sample (judgment time)
- repeated presentation of the samples during listening test (Trial 2-fold)
- acoustical playback of the stimuli using a calibrated playback equipment
- evaluation of sound samples due to the criteria *Overall Acceptance* and *Speech Quality*.

In order to give the listeners a first impression, each presentation was started with three selected sound samples (subjective worst sound, medium range sound and subjective best sound). These samples were selected based on the results of subjective evaluation of three listeners.

[1]The recordings of the test sentences can be found at the SENECA web page.

This test program was completed in a first step by eight expert listeners and in the main test by 30 consumers (users). The listeners for the user evaluation were selected by means of telephone interviews. As a result of these interviews 50% female and 50% male listeners were chosen, who often call persons with car phones. Criterion for the participation in the test was the experience of the consumer with the quality of hands–free mobile phone systems.

4. EVALUATION OF RECOGNITION PERFORMANCE

As a test for the effect of the noise reduction on the speech recognition we measured the error rates on the spelling sequences. The recognition performance both for the original and the processed spelling sequences is compared. The recognition of spelled names provides a challenging recognition task thereby giving significant results even with a fairly small database. In these tests a unrestricted recognition of letters was performed, i.e. no language model or list of entries was used.

5. NOISE REDUCTION METHOD

Our first simulation experiments focus on an approach for the back-to-back configuration. The processing occurs in three stages. At first, the uncorrelated noise parts are reduced by means of a correlation filter. Residual noise components are than reduced using a spectral subtraction filter. Finally, a post processing filter reduces and masks most of the artifacts arising during the processing. In the following the main elements of this approach (in the following called "combined approach") will be described in some detail.

5.1. Correlation filter

Let x_1 be the signal from the microphone pointing away from the speaker and x_2 the signal from the microphone pointing toward the speaker. Due to the microphone geometry described above the two microphone signals contain a significant amount of uncorrelated noise components, whereas, the speech components are strongly correlated. The correlation filter is defined by the transfer function

$$W_1(b,n) = \min\left\{ \left| \frac{\Phi_{x_2x_1}(b,n)}{\Phi_{x_2x_2}(b,n)} \right|^r, 1 \right\} \qquad (1)$$

b denoting the block index and n the discrete frequency index. $\Phi_{x_2x_1}(b,n)$ denotes the estimated time-dependent cross power spectral density of the two microphone signals. $\Phi_{x_2x_2}(b,n)$ represents the estimated power spectral density of the signal of the microphone which is directed to the speaker. In the following, we will omit the index b and n. Spectral components of X_2 which are uncorrelated to X_1 will be suppressed by

the filter W_1. For these components, which we assume to consist of noise only, W_1 will be sufficiently close to zero. Measurements have shown that the introduced noise reduction by this filter is between 5 and 10dB.

5.2. Voice activity detection in the frequency domain

For the spectral subtraction the information about the presence of voice activity is needed. Voice activity can be estimated from the weighting function of the correlation filter. Due to the different correlation properties of speech and noise in the microphone signals, the weighting function takes higher values during speech activity than during pauses. The sum of the computed weighting function coefficients is calculated as

$$p_a(b) = \sum_{n=0}^{n=N-1} W_1 \tag{2}$$

where N is the length of the used fast Fourier transform (FFT). The calculated sum $p_a(b)$ is smoothed using a forgetting factor α_{vadf}

$$p_t(b) = \alpha_{vadf} \cdot p_t(b-1) + (1-\alpha_{vadf}) \cdot p_a(b) \tag{3}$$

Then, the value

$$p(b) = \min\left\{\frac{p_t(b)}{p_a(b)}, 1\right\} \tag{4}$$

is interpreted as probability of speech absence. For $p(b) \geq 0.75$ we decide for a speech pause.

5.3. Reduction of the residual noise

Residual noise components that can not be eliminated using the correlation properties of the two microphone signals are suppressed in this stage. The suppression of these components is based on the well known spectral subtraction. Using the spectral floor H_{spfl} the weighting function W_2 of the spectral subtraction is given by

$$W_2 = \max\left[1 - \frac{\Phi_{n_2n_2}}{\Phi_{x_2x_2}}, H_{spfl}\right] \tag{5}$$

In our approach we estimate the noise spectrum from the microphone pointing away from the speaker. The underlying assumption is that this channel contains less speech components and therefore a better estimation of the noise is possible. The time-frequency spectrum $\hat{N}_1$ of the extracted background noise is calculated as

$$\hat{N}_1 = \frac{\Phi_{n_1n_1}}{\Phi_{x_1x_1}} \cdot X_1 \tag{6}$$

$\Phi_{n_1n_1}$ and $\Phi_{x_1x_1}$, respectively, denoting the estimated power spectral densities of the noise and the noisy signal from the microphone pointing away from the speaker. Replacing $\Phi_{n_2n_2}$ in (5) by the estimation from $\hat{N}_1$ results in

$$W_2 = \max\left[1 - \frac{c^2 \cdot \Phi^2_{n_1n_1}}{\Phi_{x_1x_1} \cdot \Phi_{x_2x_2}}, H_{spfl}\right] \tag{7}$$

c is a power adjustment factor for the noise signals of the two microphones. Replacing $x_i(k)$ and with $s_i(k) + n_i(k), i+1,2$ yields

$$W_2 = \max\left[1 - \frac{c^2 \cdot \Phi^2_{n_1n_1}}{|\Phi_{s_1s_2}|^2 + |\Phi_{n_1n_2}|^2}, H_{spfl}\right] \tag{8}$$

When speech is predominant the cross power spectral density of the speech signals is high. Then, the quotient in (8) is small and the filter W_2 is close to one. On the other hand, when no speech is present the quotient is close to one and the filter blocks the noise. Finally, the transfer function for the whole system is given as product $W = W_1 \cdot W_2$ of the transfer functions of the two units.

5.4. Post processing

Using the filtering described above, some undesirable residual noise components remain in the output signal, especially, during speech pauses. Fluctuations in the transfer function can be reduced by decreasing its frequency resolution in intervals without speech activities. However, during speech activities, a high resolution is required, so that noise components between the harmonics can be attenuated. This can be achieved by introducing a post processing unit with adaptive frequency resolution. We use a modified version of the tree-structure method proposed by Gülzow [3]. According to this method adjoining sub bands which do not belong to speech activity intervals are combined. At first, the frequency band consisting of M channels is divided into $K \ll M/2$ equidistant sub bands. If in a sub band the signal power is lower than the weighted estimated noise power, all the spectral coefficients of the weighting function in this sub band are replaced by their mean value. Otherwise this sub band is divided into two new equidistant sub bands and the comparison procedure is repeated.

Musical tones are randomly distributed components. The amplitude of such a component can be very high. Consequently, the calculated mean value can be adversely affected by these high amplitudes even if all other spectral coefficients are close to zero. In order to avoid this phenomenon the algorithm has been modified to use the median instead of the mean value.

6. RESULTS

6.1. Subjective evaluation

The results from the subjective evaluation are summarized in Table 1. In this test we included as an alternative a noise reduction system implementing a graphic

Table 1: Rating of the subjective evaluation in expert and user tests

		orig.	equal.	comb.
Expert	Acceptance	3.2	2.5	2.3
Expert	Quality	3.2	3.0	2.4
User	Acceptance	3.4	2.7	2.8
User	Quality	3.8	3.2	2.6

equalizer [4]. Although the noise reduction systems remove a significant amount of background noise they are rated lower both from experts and so-called users. This fact results in our opinion from the non-natural sound of the systems. Some experts remarked a "non-natural, electronic, distorted voice". The system with the graphical equalizer removes less noise and thereby affects the speech less. This results in higher scores in the perceived *Quality*.

A detailed analysis of the scores showed that for the male speaker the score in *Acceptance* is in average about 0.4 points better than for the female speaker. The criterion *Quality* yields a difference of 0.25 points in average. Again the male speaker scores better than the female one. This fact agrees with spontaneous statements from a lot of test listeners (experts and users), who mentioned after the tests that the female voice is less "clearly to understand" than the male voice. Although the graphical equalizer in general scores better than the combined approach, the *Acceptance* for the female speaker is lower.

6.2. Recognition experiment

In the recognition experiment we used the spelled city names from 19 different speakers. In summary 990 utterances with an average of 9 letters per utterance were processed. The 8kHz speech signals resulting from the noise reduction were up-sampled to 12kHz. Recognition rates both for the original and the processed speech signal were measured. Furthermore, we ran one test with the build-in single-channel noise reduction of the recognizer. In the other two tests, this noise reduction was switched off.

The results are summarized in Table 2. Using the new approach results in a small increase of substitutions and a large increase of insertions. The integrated single-channel noise reduction, on the other hand, yields a significantly improved performance. It should be noted, however, that the results can not be compared directly. The integrated method has been applied both in training and recognition. Training a new recognizer with output from the dual channel method would require the collection of a large data base with the special microphone arrangement.

Table 2: Error rates for letters (Substitutions, deletions, insertions, total error rate)

Method	sub	del	ins	err
Original	45.1%	1.1%	0.8%	47.0%
Combined	47.6%	0.6%	5.5%	53.8%
Integrated NR	34.7%	0.2%	2.0%	37.0%

7. CONCLUSIONS

Within the SENECA activities on dual channel noise reduction we have developed a frame work for testing new approaches on microphone arrangements with very small distances. First experiments with one arrangement were performed. Although we did not achieve an improvement the results provide valuable insight and first benchmarks. Further work could include refinement of the proposed approach as well as alternative approaches using some combinations of the four microphone channels.

8. ACKNOWLEDGMENT

The SENECA project is supported by the European Commission within the 4th framework of ESPRIT in the so called program "System Integration and Applications" and under Human Language Technology under the Contract number ESPRIT 26 981. The authors want to thank all the partners involved in the project.

9. REFERENCES

[1] http://www.seneca-project.de/

[2] C. Draxler, R. Grudszus, S. Euler and K. Bengler: *First experiences of the German SpeechDat-Car database collection in mobile Environments.* EUROSPEECH 99, Budapest 1999

[3] T. Gülzow: *Spektrale Subtraktion in Teilbändern mit adaptiven Bandbreiten.* ITG-Fachbericht Sprachkommunikation, Vol. 152, Dresden 1998

[4] A. Korthauer: *Untersuchungen zur zweikanaligen Störgeräuschreduktion für die automatische Spracherkennung im PKW.* Diplomarbeit RWTH Aachen 1997