COMPARISON OF SIGNAL ENHANCEMENT TECHNIQUES IN COMMUNICATIONS AND SPEECH CONTROL TASKS FOR A SINGLE-DSP IN-CAR APPLICATION

Rico Petrick, Diane Hirschfeld, Christian Gruber, Gregor Kinast

voice INTER connect GmbH, Dresden
{petrick|hirschfeld|gruber|kinast}@voiceinterconnect.de

ABSTRACT

Speech related in-car tasks can be subdivided into hands-free communication and speech control oriented tasks. While the former is characterized by signal processing in the telephone bandwidth, the latter uses multimedia-bandwidth. The accuracy and ergonomics of both types of applications is severely influenced by external noise conditions and the technologies applied for signal enhancement.

Different techniques for signal enhancement are discussed like a four channel microphone array, a single channel noise reduction and an acoustic echo cancellation, implemented on a single DSP chip. A suitable system design is introduced which matches both types of applications by an optimal combination of the different signal enhancement approaches. A number of objective and subjective experiments using real world speech and noise corpora recorded in a car and in a truck environment are accomplished for evaluation of the system quality. Finally, some recommendations for noise reduction techniques in low cost applications are derived.

1. INTRODUCTION

In-car speech processing is getting more and more interesting. On one hand, cars are already equipped like moving offices, on the other hand communication applications are safety-relevant in a running car. Therefore, applications that use the voice as third hand have established in the car over the last years.

In the car, robustness against interfering noises and high ergonomics as well as straight dialogues and short control operations decide over the success of a voice application.

Since most applications demand also a very low price, this paper presents a small but efficient solution to combine speech enhancement methods for speech communication tasks as well as command and control for speech interaction.

The main focus is laid on the issues of noise reduction techniques and system architecture. Communication and speech control differ in their bandwidth requirements. Speech control uses multimedia-bandwidth whereas communication is limited to telephone bandwidth. These considerations influence the selection of appropriate target algorithms. To be cost effective all algorithms are integrated on a single DSP.

Different techniques for signal enhancement are discussed like a four channel microphone array and a single channel noise reduction. All approaches are investigated for their suitability and their impact on the two types of applications by a combined objective and subjective evaluation method.

2. NOISE REDUCTION TECHNIQUES

verbKEY-car is a complete in-car speech based telephone control based on verbKEY, a small footprint speech recognition system developed for embedded applications on the 16 bit, fixed point vicCORE DSP-platform. Figure 1 shows the system architecture.

Compared to earlier approaches using single channel noise reduction [2], a four channel microphone array and an acoustic echo cancellation are added.

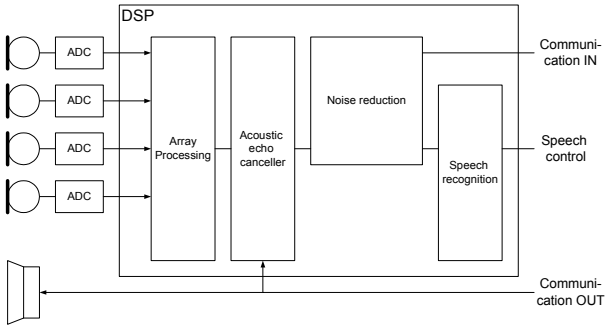

Figure 1: ***Block diagram of the combined solution of a single channel noise reduction, an acoustic echo canceller and a 4-channel microphone array for a command word recognizer as a single DSP solution.***

The following chapters describe the noise reduction techniques more in detail. For all experiments telephone bandwidth is assumed.

2.1. Acoustic echo cancellation

Typical noises arising in a communications situation in a car environment result from the system itself.

For easy and ergonomic hands-free interaction, the system should be robust against the interruption of system prompts by the user and should not transmit the speech signals of the far-end speaker during a telephone conversation. This task is fulfilled by an *acoustic echo cancellation (AEC)* algorithm. The algorithm models the impulse response of the loudspeaker-room-microphone system (LRM-System) of the car-room by estimating a time-variant digital filter.

For the single-chip verbKEY-car solution, an *AEC* algorithm was selected by the following criteria:

- Robust and fast adaptation
- Good echo suppression (ERLE – echo return loss enhancement)
- Low numerical complexity and needed calculation power and
- The potential to combine processing stages with other modules of the speech processor

Therefore, the FBLMS-Algorithm (frequency domain block LMS [6]) was chosen. It combines the advantages of different algorithms. The numerical complexity is reduced compared to time-domain approaches. The calculation complexity is bigger per iteration step, but because of the block wise processing less iterations per time interval are needed. In the frequency domain, a step control can be implemented, that is very fast, frequency selective and therefore robust against near end speech.

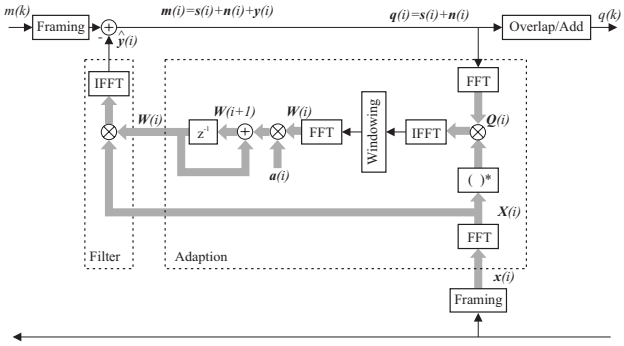

Figure 2: ***Scheme of the FBLMS-algorithm***

2.1.1. Algorithm of the FBLMS

The loudspeaker signal *x(k)* and the microphone signal *m(k)* are framed to the blocks *x(i)* und *m(i)*. The calculation of the estimated signal *ŷ(k)* is done by fast convolution (multiplication in the frequency domain) with the adaptive filter *W(i)*. The parameters like the filter coefficients *W(i)* and the gradient vector *ΔW(i)* for the adaptation of *W(i)* are calculated in the frequency domain.

$$\text{time domain}: \quad \Delta \boldsymbol{w}(i) = \sum_{l} \boldsymbol{x}(iL-l) \cdot q(iL-l)$$

$$\text{frequency domain}: \quad \Delta \boldsymbol{W}(i) = \boldsymbol{X}(i)^{*} \cdot \boldsymbol{Q}(i)$$

L denotes the frame length in the upper equation. The Adaptation of W(i) is done by

$$\boldsymbol{W}(i+1) = \boldsymbol{W}(i) + \boldsymbol{a}(i) \cdot \Delta \boldsymbol{W}(i)$$

Die calculation of the gradient vector *ΔW(i)* for the adaptation of *W(i)* is also done by correlation analysis in the frequency domain. The components of the step control vector ***a(i)*** consist of different step control factors for each frequency band. The step control is adaptive and depends from the block-index *i* and from frequency. The output signal *q(k)* of the *AEC* is reconstructed from the block signal *q(i)* by overlap-and-add.

For long impulse responses the blocks are decomposed to smaller ones in order to avoid unacceptable delays in the processing chain (not displayed in Figure 2).

2.2. Single channel noise reduction

The single channel *Automatic Noise Reduction (ANR)* employs the principle of spectral subtraction. It is carried out block wise on the time signal with overlapping blocks (*k* - block index) of the FFT size 512 (sampling frequency 16000 *Hz*) or 256 (sampling frequency 8000 *Hz*) which corresponds to a frame length of 32 *ms*.

2.2.1 Adaptation of the ANR

For spectral subtraction it is necessary to know the spectral noise characteristics. It is estimated on the basis of a minimum statistics of the spectrum of the microphone signal and is based on the fact that the non-correlated noise and speech signal combine additively to the received microphone signal. Since the noise signal is more stationary than the speech signal, values of the spectrum of the microphone signal $X_{in}(f)$ are bigger than that of the noise spectrum $N(f)$ at all times.

Minimum Statistics Spectra $X_{min}(f)$ are averaged to estimate the noise spectrum $N_{est}(f)$.

$$|N_{est}(f,k)| = \rho \cdot |N_{est}(f,k\text{-}1)| + (1-\rho) \cdot |X_{\min}(f,k)|$$

The adaptation factor ρ describes the speed of adaptation. Fast transient changes such as a door clap do not affect the estimation but slow changes such as the motor sound during car acceleration do.

2.2.1 Spectral subtraction

The estimated noise spectrum $N_{est}(f)$is subtracted from the spectrum of the (disturbed) microphone signal $X_{in}(f)$ according to the following equation:

$$|X_s(f)| = |X_{in}(f)| - \alpha \cdot |N_{est}(f)|$$

The subtraction factor α determines the intensity of the spectral subtraction and is set between 1.2 and 3. It is also possible to have an adaptive behavior of $\alpha(SNR)$ for the entire frequency band or in sub bands $(\alpha(SNR, f))$ [3]. $X_s(f)$ is the mathematical result of the spectral subtraction which will be floored in order to avoid musical tones with the flooring factor β to the cleaned signal $X_{out}(f)$ according the following equation:

$$|X_{out}(f)| = \begin{cases} |X_s(f)| & ; |X_s(f)| > \beta \cdot |N_{est}(f)| \\ \beta \cdot |N_{est}(f)| & ; |X_s(f)| \le \beta \cdot |N_{est}(f)| \end{cases}$$

Flooring is also required to have a sensible level of comfort noise. Experiments to find the optimal value for β resulted in $\beta = 0.1$. β also can be determined in an adaptive way as a function of SNR or by a different setting for sub bands.

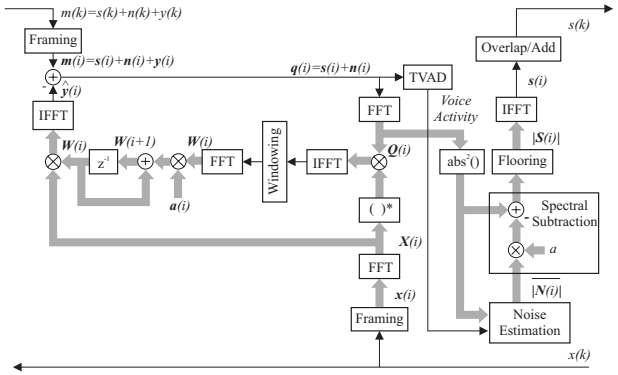

Figure 3: ***Scheme of combined AEC and ANR***

2.3. Combination of *ANR* and *AEC*

If the *AEC* is followed by *ANR* (Figure 1), ANR can reduce the residual echo, acting like a postfilter [7] that filters residual echo and noise from the speech signal.
Because both algorithms, *AEC* and *ANR*, are working in the frequency domain, by combination of the parts a very efficient implementation can be achieved.
In *AEC* for instance, the calculation of the resulting signal *q(k)* by overlap-and-add and framing and FFT before the *ANR* can be dropped (Figure 3).

2.4. Multi-channel noise reduction with microphone arrays

A single microphone is only able to measure the acoustic pressure over the time. A geometric arrangement of (more than one) microphones - *microphone array (MA)* - is able to gain additional information about the location of a signal source. The result is a directional characteristic of the microphone array compared to the omni directional characteristic of a single low cost microphone.

2.4.1 Theoretical aspects of microphone arrays

A one dimensional microphone array is mainly divided in equidistant and harmonic distant arrangements of microphones. In order to create a broadside line array the maximal microphone distance measures $d_{\max} \le \lambda_{\min} / 2$ with $\lambda_{\min}$ characterizing the wave length of the highest frequency component to process.

The basic method of processing the incoming channel signals x_m to the output signal y is the *delay-and-sum* algorithm, where the w_m are 1 for the M Microphones:

$$y(k) = \sum_{m=1}^{M} w_m x_m(k)$$

A more sophisticated approach is the filter-and-sum algorithm which employs FIR filters for the factors w_m [4].

2.4.1 A microphone array for speech recognition

The chosen microphone array is an equidistant line array including 4 electret microphones WM61 of Panasonic, and is designed for telephone bandwidth (*300 - 4000 Hz*).

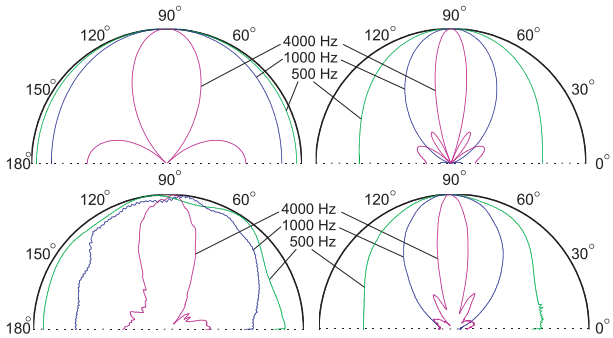

Figure 4: ***Simulation (top) and measurement (bottom) of the directional characteristics of the used microphone arrays for typical frequencies. Left: 2 mic. array. Right: 4 mic. array. A single microphone is omni directional according to the data sheet***

To fulfill the sampling theorem in the given frequency range the distance d between the microphones needs to be less or equal 4.2 cm. The distance between the microphones was chosen experimentally to d = 6.0 cm.

Beam steering is achieved at frequencies higher than 800 Hz. For lower frequencies no significant directivity for the array dimensions are achieved.

Array processing is based on a filter-and-sum algorithm ([4]) combining the directivities of a 4 microphone array (distance d) and a 2 microphone sub array of inter microphone distance 3d. In consequence, there are small side lobes but a gain of a better directivity at low frequencies (Figure 4).

As expected, the 4-channel microphone array forms a smaller beam than the 2-channel microphone array, which in fact only affects the upper frequencies over 1000 *Hz*. The measured directional characteristics correspond to the simulated ones (Figure 4). In the best measured case, the

microphone array results in a damping of noises coming from directions outside the main lobe by 12 dB.

3. SYSTEM EVALUATION

Since the objective of the presented system is to improve communications and speech control, an evaluation method for the system quality is proposed that combines objective and subjective criteria.

3.1. Handsfree-evaluation – combined *AEC* and *ANR*

Combined *AEC* and *ANR* were evaluated in a real world scenario in a truck´s cabin. Noise signals were coming from the motor, running at different speeds, and from the air condition. Two persons were sitting in the truck's cabin, doors closed. Noise and speech were recorded by 3 microphones, mounted in 3 different places:

- Pos1: Microphone with cardioid characteristics, mounted at the wind screen 60 cm above the drivers mouth.
- Pos. 2: Microphone with cardiod characteristics, mounted at the middle console in front 100 cm from the drivers mouth.
- Pos. 3: Microphone with cardioid characteristics, mounted at the upper blending in the centre of the wind screen, about 80 cm from the drivers mouth.

The microphone signals were connected separately to the postprocessors. Mic1 was connected to the *AEC*/*ANR* device, the raw signal and the cleaned signal were simultaneously recorded at a laptop. The raw signals from Pos2 and Pos3 were recorded on an analogue device.

For every prompt, 60 s of the signal were recorded. The driver read aloud a text. The acoustical environment was changed between the following alternatives:

- motor in standby
- medium speed (800 rev/min)
- near maximum speed (2000 rev/min)

As system signal for the test of the *AEC,* a rock song was chosen, played back by a loudspeaker on the front seat, 145 cm away from microphone on Pos1. The playback level was adapted, so that speech and music had the same loudness at the microphone.

Because the clean speech signal S is not available, the SNR was calculated as evaluation parameter as follows from the microphone signal M_S:

$$SNR = 10 \cdot \log_{10} \frac{S}{N} = 10 \cdot \log_{10} \frac{M_S - N}{N}$$

$$= 10 \cdot \log_{10} \frac{m_{s,eff}^{\ 2} - n_{eff}^{\ 2}}{n_{eff}^{\ 2}}$$

The power of the noise signal N is estimated in speech pauses. N contains the sum of all noises (motor and echo). The power of the speech signal is estimated from the microphone signal during speech activity.

Table 1: ***SNR before and after the AEC/ANR unit, damping of noises (NR&ERLE)***

Exp	Acoustic environment	SNR before	SNR after	NR & ERLE
1	No echo / motor 800/min	13.5 dB	26.7 dB	15.3 dB
2	No echo / motor 2000/min	2.8 dB	21.0 dB	20.2 dB
3	Music / motor off	7.7 dB	27.9 dB	21.7 dB
4	Music / motor 800/min	3.9 dB	21.4 dB	22.0 dB
5	Music / motor 2000/min	2.2 dB	17.9 dB	17.6 dB

Experiments 1 and 2 show only *ANR*, because the echo signal is missing. Motor noise is damped by 20 dB. The resulting speech has a good quality.

Experiment 3 shows the function of the *AEC*. The echo suppression is about 22 dB. Speech quality is very good, residual echo is inaudible.

Experiment 4 and 5 use *AEC* as well as *ANR*. Subjective and objective evaluation show that both components work excellent even under difficult conditions.

3.2. Evaluation corpus for *ANR* and *MA*

In order to get reproduceable results from the evaluation experiments, a speech corpus was recorded displaying typical noise conditions, speech signals and the spatial effects exploited by the microphone array.

For the comparison of different microphone configurations, the array was prepared to deliver the following different signals simultaneously (Figure 5):

- a 4-microphone array signal with filter-and-sum algorithm a).
- a 2-microphone signal with delay-and-sum b) and
- a single channel microphone c)

The three audio signals were simultaneously recorded with *mobiLAB* [5] in the anechoic chamber of the University of Technology Dresden under different noise conditions. The microphone array together with the electronics based on *ADSP-BF533* from Analog Devices was placed in the center of the room. Speech and noise sources could be moved in a radius of 1 *m* around the array for directional measurements (Figure 4) and corpus recordings.

Speech signal: A signal source (studio loud speaker) was fixed at the optimal 90° position in front of the microphone array. A corpus of 100 male and 100 female speakers each once speaking the German numbers from 0

to 9 was played over the loudspeaker for every of the following noise conditions. The *equivalent sound pressure level* L_{eq} of the speech signal was adjusted to 70 dB at the front edge of the microphone.

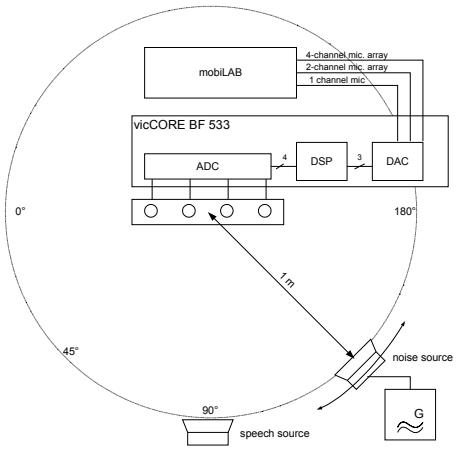

Figure 5: ***Experimental setup for recording of the noisy speech corpus with varying noise source localization.***

Noise condition: During the recordings, the speech signal was superposed by one of the following 3 different noises, typical for the target application:

- *car noise* (recorded at 150 kmh^{-1}): stationary, low frequency characteristics
- *cocktail party noise* (babble noise): non stationary, wide frequency distribution
- *fan noise:* stationary, wide frequency distribution, single harmonics

The L_{eq} of the noise signal was adjusted to 60 dB.

The directional effects were achieved by playing the disturbing noise over a second loud speaker in 3 different positions relative to the microphone: 90°, 45°, 0°. For all 9 conditions and the different microphone localizations, a corpus of 54000 words was recorded.

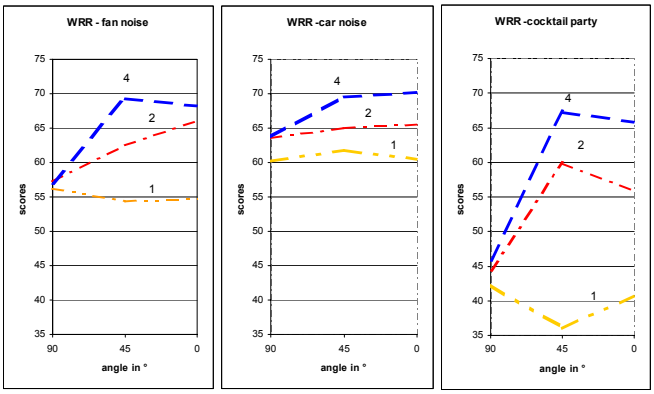

Figure 6: ***Word recognition rate without ANR. Numbers at the curves mark the number of microphones in the array.***

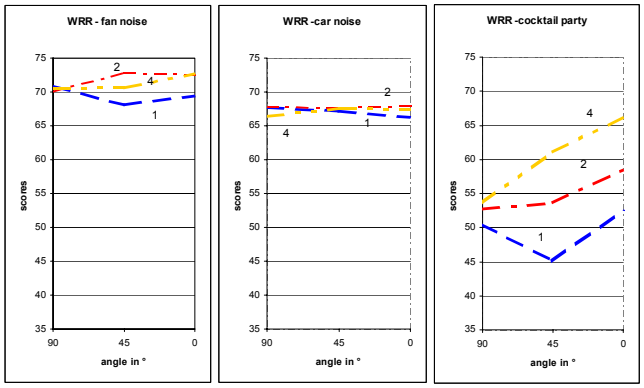

Figure 7: ***Word recognition rate with ANR. Numbers at the curves mark the number of microphones in the array.***

3.3. Objective Evaluation of *MA* and *ANR*

The corpus was taken as the input for a command word recogniser to measure the *Word Recognition Rate WRR* as an objective evaluation criterion. Both noise reduction techniques - the microphone array (*MA*) and the automatic noise reduction (*ANR*) - are connected in series in the signal chain (Figure 1) and can be switched on or off.

In order to see the influence of each component, the *WRR* was measured for each combination of the 3 types of microphones, and for each angle position and noise scenario and with *ANR* switched on or off. The results are graphically displayed in Figure 6 and Figure 7.

AEC was switched off during the evaluation of the *MA* and *ANR* components.

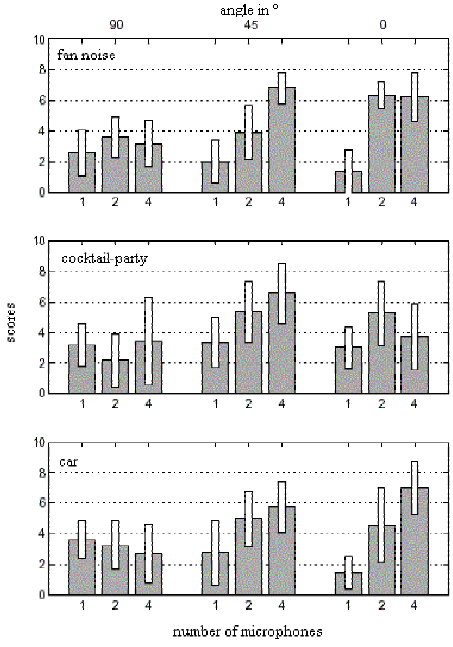

Figure 8: ***Subjective evaluation results***

3.4. Subjective Evaluation of *MA* and *ANR*

For communication tasks, human performance is still the most important evaluation criterion. Therefore, besides the objective evaluation of directional characteristics of the microphone array the resulting speech quality after signal enhancement was evaluated by a subjective test.

Examples were chosen randomly from the corpus, paired according to noise and angle condition and presented to subjects in a pair comparison test. Subjects decided, which sample's quality was better. A score was given to each winner sample and scores were summed up for a given sample over the whole test. The evaluation for a given noise / angle condition was based on the total score – the higher the score, the better was the subjective impression of noise reduction. Figure 8 shows the results (mean score for a given microphone).

4. DISCUSSION OF RESULTS

The baseline speech recogniser worked on the clean data with a *WRR* of 98 %. Low *WRRs* in the experiments are due to the low SNR (especially in the 90° scenario, where speech and noise came from the same direction). Since the focus was on the noise reduction techniques, *WRRs* show qualitative information only and no attempt was made to further improve recognition accuracy.

The cocktail party noise seems to be the worst for the speech recognizer. This can be explained by the wide and non stationary frequency characteristics just in the range of the speech signal. There are four interesting combinations of the evaluated noise reduction techniques:

- Single microphone without *ANR* (naked recogniser, Figure 6) - Recognition accuracy as expected is the poorest.
- Single microphone with *ANR* - *ANR* improves robustness in case of stationary noises (Figure 7)
- *MA* without *ANR* - the MA improves robustness for stationary and instationary wide band noises. The effect is distinct for instationary and broadband noises (Figure 6)
- *MA* with *ANR* – the combination of *ANR* and *MA* shows best performance (Figure 7). The experiment shows that for stationary noises the microphone array does not have a significant influence. This especially holds true at the car environment, because of the low frequency distribution of the noise.

The results of objective tests are supported by the results of the subjective evaluation. Subjects show clear preference for the 4-microphone array in the 45° scenario. For the 0° scenario, judgements are somewhat indifferent, but the 2-microphone array gives clearly improved decisions under all environments compared to the single microphone.

5. SUMMARY

Noise reduction techniques enhance the speech signal while suppressing disturbing noises. In reality, noise and speech signals are mixed so that it is difficult to separate them. Single channel *ANR* is limited and can result in speech distortions.

In this paper we showed that in the case of instationary noise, that was a difficult condition for the single channel *ANR*, significant increase of recognition accuracy can be gained by using the microphone array's capability to separate spatially separated sources. Already a 2-channel microphone array can improve the SNR and thereby speech quality significantly as proven by objective and subjective evaluations.

In a car environment, microphone arrays do not gain much in case of the most prominent car noises, but show their advantage in broad band and babble noises that are clearly spatially separated from the speech source. Therefore, application of microphone arrays in a car is only recommended in special cases.

6. REFERENCES

[1] D. Hirschfeld, J. Bechstein, U. Koloska, T. Richter, R. Petrick. *Development steps of a hardware recognizer with minimal footprint (in German),* Proc. 13th Conf. On Electronic Speech Signal Processing (ESSV), Dresden, 2002.

[2] Petrick, R., Hirschfeld, D., Richter, T., Hoffmann, R.: "verbkey - A Single-Chip Speech Control for the Automobile Environment", In: *DSP in Mobile and Vehicular Systems*, editors: H. Abut, K. Takeda and J. H.L. Hansen, Kluwer Academic publishers, 2004.

[3] W. Hess P. Vary, U. Heute, *Digital Speech Signal Processing (in German)*, Teubner, Stuttgart, 1998.

[4] M. Brandstein, D. Ward, *Microphone Arrays*, Springer Verlag Berlin, Heidelberg [et al.], 2001.

[5] Maase, J., Hirschfeld, D., Koloska, U., Westfeld,T., Helbig, J.: "Towards an Evaluation Standard for Speech Control Concepts in Real-World Scenarios", *Proc. of EUROSPEECH 2003, Geneva.*

[6] Farhang-Boroujeny, B.: "Adaptive Filters – Theory and Applications", John Wiley & Sons, 1998.

[7] Gustafsson, S., Jax, P., Kamphausen, A., Vary, P.: " A Postfilter for Echo and Noise Reduction avoiding the Problem of musical tones", ICASSP, 1999.