INTEGRATION OF NOISE REDUCTION AND ECHO ATTENUATION FOR HANDSET-FREE COMMUNICATION

Marc Ihle, Student Member, IEEE, and Kristian Kroschel
Institut fuer Nachrichtentechnik / Automation und Robotik
Universitaet Karlsruhe, Kaiserstr. 12, D-76128 Karlsruhe / Germany
ihle@etec.uni-karlsruhe.de, kroschel@etec.uni-karlsruhe.de

ABSTRACT

In this paper a front-end for handset-free telecommunication is presented, which combines noise reduction and echo attenuation, exploiting phenomena of hearing physiology. Simple and robust techniques for the adaptation to the speaker render a system which can be used in extremely noisy environments around 0 dB, which are typical in small vans. The implementation in hardware is highly economic because some functions are used commonly for both tasks.

1 INTRODUCTION

Handset-free telecommunication facilities for cars suffer from low intelligibility due to the long distance between the mouth of the local speaker and the microphone(s). Thus surrounding noise and the speech signal of the far-end speaker, emitted by a loudspeaker inside the car, interfere with the local speech signal, leading to a low signal-to-noise ratio (SNR) and an irritating echo signal on the far-end side.

To obtain a system especially for small vans, which is capable to handle an SNR around 0 dB, we developed a system using novel speech processing techniques [10],[3]. For the enhancement of the SNR, we propose a noise reduction system, using the spectral subtraction method. The required noise reference is derived from the signals of a microphone array, processed by the sub-band array method [4]. In contrast to other array implementations, e.g. [9], no adaptive delay compensation or equalization is needed.

Instead of the classical echo canceller, a frequency dependent echo attenuation unit is used to reduce the echo of the far-end speaker to a given value. This is done because of three disadvantages of the canceller: first, depending on the geometry of the space in which the echo is generated, the canceller requires a high number of coefficients, in a car typically 64. Second, in environments with high noise power the adaptation of the coefficients becomes critical. Third, the canceller reduces only linear components of the echo whereas the transmission path in a van might be non-linear.

The block diagram of the whole system is shown in Fig. 1. As the system only needs moderate computational power, it was possible to adapt it to run on a single Motorola 96002 signal-processor platform in real-time.

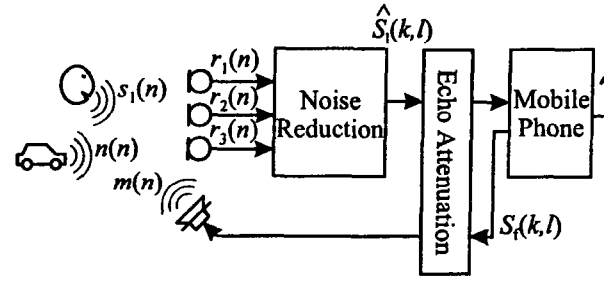

Fig. 1: A handset-free front-end with integrated noise reduction and echo attenuation

2 NOISE REDUCTION UNIT

The front-end of the system is a linear array of three microphones mounted on the A column of a car left of the driver (left driven car assumed). In Fig. 2 the geometry of this array can be seen.

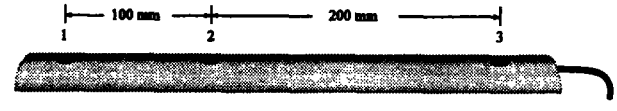

Fig. 2: Geometry of the linear microphone array

The three microphone signals

$$r_i(n) = s_{1i}(n) + n_i(n) + m_i(n),\ i = 1,2,3 \qquad (1)$$

consist of the speech signal of the local speaker $s_1(n)$, the additive noise $n(n)$ and the monitor signal of the far-end speaker $m(n)$. To enhance the coherence of the speech signals $s_{1i}(n)$, the signals $r_i(n)$ are fed through an equalizer (see Fig. 3).

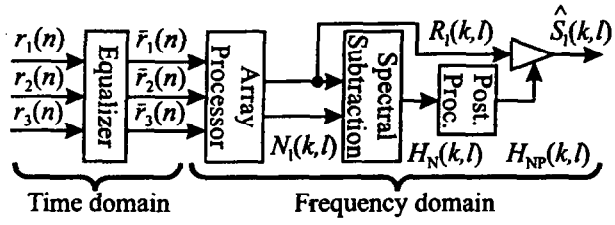

Fig. 3: Block diagram of the noise reduction unit

This equalizer is initialized once before the use of the handset-free system. For this purpose a person must sit inside the car in a normal driving position, while only low ambient noise is present. Then he must place the handset or a separate microphone near his mouth and speak some words. By means of the FLMS algorithm [2], the coefficients of the equalizer's three FIR filters are trimmed to match the signals of the microphone array to the signal of the reference microphone. Thus, the local speech signal sounds much less reverberant and the far-end speaker gets the impression that the local speaker is not so far away.

The filtered array signals

$$\bar{r}_i(n) = \bar{s}_{1i}(n) + \bar{n}_i(n) + \bar{m}_i(n) \qquad (2)$$

consist of approximate coherent speech signals $\bar{s}_{1i}(n)$ of the local speaker, where

$$\bar{s}_{11}(n) \approx \bar{s}_{12}(n) \approx \bar{s}_{13}(n), \qquad (3)$$

as well as of noise signals $\bar{n}_i(n)$ and echo signals $\bar{m}_i(n)$, that can be assumed to be uncorrelated, resulting in

$$\begin{aligned} &\mathrm{E}\{\bar{n}_i(n) \cdot \bar{n}_j(n)\} \approx 0 \\ &\mathrm{E}\{\bar{m}_i(n) \cdot \bar{m}_j(n)\} \approx 0 \end{aligned} \quad , \quad i,j = 1,2,3,\ i \neq j. \qquad (4)$$

Expression (3) is only true for short microphone distances or low frequencies whereas for Eq. (4) large microphone distances or high frequencies are required. A good compromise can be obtained using the sub-band array processing method [4]. Therefore the three signals $\bar{r}_i(n)$ are transformed into the frequency domain using the Short Time Fourier Transform (STFT) [8]. This results in the short time spectra $\bar{R}_i(k,l)$, where k denotes the spectral bin and l the block index. The output signals $R_1(k,l)$ and $N_1(k,l)$ of the array processor are calculated by

$$R_1(k,l) = \begin{cases} \bar{R}_1(k,l) + \bar{R}_3(k,l) & , \ k_1 \le k < k_2 \\ \bar{R}_2(k,l) + \bar{R}_3(k,l) & , \ k_2 \le k < k_3 \\ \bar{R}_1(k,l) + \bar{R}_2(k,l) & , \ k_3 \le k < k_4 \\ 0 & , \ \text{elsewhere} \end{cases} \qquad (5)$$

$$N_1(k,l) = \begin{cases} \bar{R}_1(k,l) - \bar{R}_3(k,l) & , \ k_1 \le k < k_2 \\ \bar{R}_2(k,l) - \bar{R}_3(k,l) & , \ k_2 \le k < k_3 \\ \bar{R}_1(k,l) - \bar{R}_2(k,l) & , \ k_3 \le k < k_4 \\ 0 & , \ \text{elsewhere} \end{cases} \qquad (6)$$

The constants k_1 to k_4 split the frequency range into three relevant band pass regions. The corresponding corner frequencies are 312.5 Hz, 750 Hz, 1.625 kHz and 3.406 kHz. The two output signals of the array processor derive their information from the two furthest microphones 1 and 3 for low frequencies, from the microphones 2 and 3 for middle frequencies and from the closest microphones 1 and 2 for high frequencies.

The signal $N_1(k,l)$ can now be used to estimate the noise level within $R_1(k,l)$, as the coherent speech signals are subtracted from each other, while the uncorrelated noise is enhanced up to 3 dB in both output signals. In addition, $R_1(k,l)$ has an enhanced speech level of 6 dB.

The advantage of the sub-band array processor technique is that an estimate of the noise level is always available, so that unstationary noise processes can even be handled while the local person speaks. In addition, the system is insensitive to movements of the local speaker, with no delay compensation or adaptive filtering of the microphone channels required.

The second component of the noise reduction unit is the spectral subtracter with the transfer function

$$H_N(k,l) = \begin{cases} H'_N(k,l) & \text{if } H'_N(k,l) > b \\ b & \text{elsewhere} \end{cases}, \qquad (7)$$

where

$$H'_N(k,l) = 1 - a(k)\frac{S_-(k,l)}{S_+(k,l)} \qquad (8)$$

represents the filter function for magnitude spectrum subtraction [5]. The estimated levels of the corrupted speech, $S_+(k,l)$, and noise, $S_-(k,l)$, are derived by a low-pass filter with exponential impulse response

$$S_+(k,l) = (1-\alpha) \cdot |R_1(k,l)| + \alpha \cdot S_+(k,l-1) \qquad (9)$$

$$S_-(k,l) = (1-\alpha) \cdot |N_1(k,l)| + \alpha \cdot S_-(k,l-1), \qquad (10)$$

where α determines the forgetting factor. At sampling rates of 125 Hz of the down-sampled spectra, $\alpha = 0.6$ leads to good intelligibility. The *overestimate* weight $a(k)$ has two tasks: first, it compensates for the distortion of the channels in which $R_1(k,l)$ and $N_1(k,l)$ are calculated, and second, it enhances the speech spectrum with respect to the noise [7]. To force the transfer function in Eq. (7) to be positive, the *spectral floor* b in Eq. (8) is set to $b > 0$, resulting in a transmission path that is never interrupted. The transfer function $H_N(k,l)$ is fed through a post-processor, described in section 5, resulting in

$H_{NP}(k,l)$, to reduce the occurrence of transient estimation errors that otherwise can be heard as so called *musical tones*. The output of the spectral subtracter is finally given by $\hat{S}_l(k,l) = H_{NP}(k,l) \cdot R_l(k,l)$ which is fed into the echo attenuation unit.

3 ECHO ATTENUATION UNIT

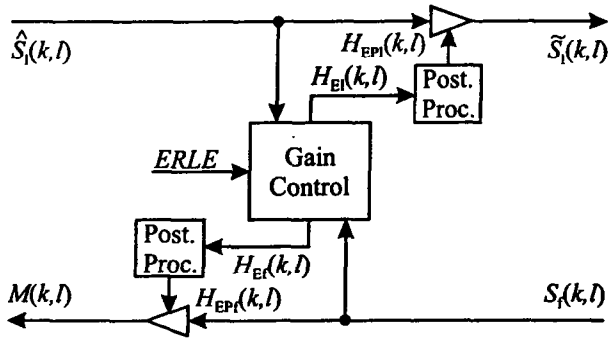

Fig. 4: Block diagram of the gain control subsystems

The echo attenuation unit consists of one gain control subsystem for each frequency bin, as shown in the block diagram in Fig. 4. They act independently from each other, with the exception of the post processing blocks, as described in section 5.
The transfer functions $H_{El}(k,l)$ and $H_{Ef}(k,l)$ for the local and the far-end speech signals are given by

$$H_{El}(k,l) = \begin{cases} c_l \dfrac{|\hat{S}_l(k,l)|}{|S_f(k,l)|} & \text{if } c_l \dfrac{|\hat{S}_l(k,l)|}{|S_f(k,l)|} < 1 \\ 1 & \text{elsewhere} \end{cases} \tag{11}$$

and

$$H_{Ef}(k,l) = \begin{cases} c_f \dfrac{|S_f(k,l)|}{|\hat{S}_l(k,l)|} & \text{if } c_f \dfrac{|S_f(k,l)|}{|\hat{S}_l(k,l)|} < 1 \\ 1 & \text{elsewhere} \end{cases}, \tag{12}$$

respectively. The parameters c_l and c_f are chosen in such a way that a specified value of the echo return loss enhancement (*ERLE*) is met.
The post processing blocks have three tasks. First, they reduce musical tones, occuring when the transfer functions $H_{El}(k,l)$ and $H_{Ef}(k,l)$ change quickly during double talk situations. Second, due to the overlapping frequency bands of the STFT analysis filter banks, $H_{El}(k,l)$ and $H_{Ef}(k,l)$ interfere with adjacent bands $H_{El}(k+i,l)$ and $H_{Ef}(k+i,l)$, $|i| \leq i_{max}$, respectively, where i_{max} is in the range from 1 to 3, depending on the window function used for the STFT. To ensure that a desired *ERLE* is met for all frequencies, the interference between adjacent bands is reduced by an additional attenuation of the transfer functions for $|i| \leq i_{max}$. The third and last task of the post processors is to reduce the interference between successive samples of $H_{El}(k,l+j)$ and $H_{Ef}(k,l+j)$. This is necessary due to the overlapping windowed time slices used by the STFT analysis filter. As the effect is comparable to the one described above for adjacent frequency bands, the same post processing method can be used, when the frequency index k and the time index l are exchanged. For a Short Time Fourier Transform with an FFT length of $N = 256$ bins and an overlapping portion of $L = 32$ samples between succeeding frames, j must be considered up to $\pm j_{max} = N/L = 8$.

4 INFLUENCE OF HEARING PHYSIOLOGY

From the various effects of hearing physiology, the simultaneous masking effect [11] can best be used to enhance the quality of the proposed system. If the far end speaker pronounces a loud vowel, he is still able to hear a consonant from the local speaker, whereas a vowel with spectrum similar to his utterance would be masked. Using this effect, the frequency dependent gain control unit compares the incoming spectra and attenuates mainly those spectral bins which are masked and thereby inaudible. It is possible to achieve even better results when Eqs. (11) and (12) are changed to consider the signal to mask level $S_{SML}(k,l)$ that can be derived from the MPEG audio codec [1]. If, for example, the far-end speaker's signal is below the mask level for the local person $S_{SMLl}(k,l)$, $H_{El}(k,l)$ may be set to zero without information loss, allowing the transfer function $H_{Ef}(k,l)$ for the other direction to be set to one.

5 SYNERGY EFFECTS BETWEEN THE UNITS

Combining noise reduction and echo attenuation enhances the intelligibility of the system and leads to more efficient use of the hardware resources.
Both the noise reduction unit and the echo attenuation unit produce musical tones due to estimation errors. The tones can easily be reduced by means of median filters, which remove temporal spikes within the transfer functions $H_N(k,l)$, $H_{El}(k,l)$ and $H_{Ef}(k,l)$, respectively [6]. Therefore, the two filters for $H_N(k,l)$ and $H_{El}(k,l)$ can be replaced by a single one, when $H_N(k,l)$ and $H_{El}(k,l)$ are first multiplied and their output is fed into a single median filter.
As the gain control block interacts with the noise reduction unit, a nearly constant echo return loss enhancement (*ERLE*) can be achieved. This leads to higher intelligibility for both transmission directions. If, for example, $H_{NP}(k,l)$ is set to the spectral floor $b < ERLE$ due to severe interfering noise,

the echo attenuation unit may pass both signal directions without loss. In that situation, both speakers can hear each other simultaneously. In addition, the transfer function $H_{NP}(k,l)$ of the noise reduction unit, the duration of the reverberation within the room, and effects of hearing physiology can be considered to obtain best results.
Both noise reduction unit and echo attenuation unit are realized in the frequency domain. Thus no additional computation power is needed for the Short Time Fourier Transform when the echo attenuation unit is added to the noise reduction system.

6 RESULTS

In small vans, the signal-to-noise ratio at the output of a microphone used for handset-free communication may lie below 0 dB, so that communication with the far-end customer without any noise reduction technique is almost impossible. This has been demonstrated using a Siemens C5 telephone with integrated handset-free communication facility in a van of the type Volkswagen VW LT28. Driving faster than 80 km/h, no communication was possible, whereas, using the described system, a signal-to-noise ratio enhancement in the range of 10 to 12 dB has been measured, which improved the intelligibility for the far-end customer significantly and made communication possible. Furthermore, even double talk was possible. The echo return loss enhancement $ERLE$ is a free parameter and can easily be set to any desired value. In a field test at the Deutsche Telekom AG, an $ERLE$ value of 33 dB was selected, so that no audible degradation of the speech signal and no echo was heard. With $ERLE \geq 40\,\mathrm{dB}$, the degradation of speech is noticeable even in presence of background noise.

Acknowledgment
The demonstrator system described in this paper has been investigated and financially supported by the Deutsche Telekom AG.

References

[1] ISO/IEC: *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, International Standard ISO/IEC 11172-3; 1993(E)

[2] Ferrara, E. R.: *Fast Implementation of LMS Adaptive Filters*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 4; August 1980, pp. 474 - 475

[3] Ihle, M; Kroschel, K.: *Verfahren und Vorrichtung zur Stoer- und Echounterdrueckung*, German patent application No. 196 50 410.4; Munich, 05.12.1996

[4] Kroschel, K.; Lange, K.: *Subband Array Processing for Speech Enhancement*, Proc. Eurospeech 93; Berlin Sept. 1993, pp. 621-624

[5] Kroschel, K.: *Statistische Nachrichtentheorie*, Springer series in information sciences; Berlin, Heidelberg, 3rd edition 1996

[6] Linhard, K.; Haulick, T.: *Nichtlineare Glaettung und Geraeuschreduktion bei gestoerter Sprache*, 9. Aachener Kolloquium Signaltheorie; RWTH Aachen; 1997, pp. 251-254

[7] Lockwood, P.; Boudy, J.: *Experiments with a Non-linear Spectral Subtractor (NSS), Hidden Markov Models and a Projection, for Robust Speech Recognition in Cars*, Speech Communication vol. 11; June 1992, pp. 215-218

[8] Portnoff, M. R.: *Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 1; Februar 1980, pp. 55 - 69

[9] Zelinski, R.: *A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms*, Proc. Intl. Conf. on ASSP; ICASSP New York 1988, pp. 2578-2581

[10] Zelinski, R.: *Verfahren und Vorrichtung zur Kanalentzerrung fuer ein Mikrofonarray*, German patent No. DE 195 38 880 A1; Munich, 19.10.1995

[11] Zwicker, E.; H. Fastl: *Psychoacoustics* , Springer series in information sciences; Berlin, Heidelberg, 1990