APPLICATIONS OF ACOUSTIC ECHO CONTROL – AN OVERVIEW

Gerhard Schmidt

Temic SDS, Research,
Söflinger Str. 100, 89077 Ulm, Germany
E-mail: gerhard.schmidt@temic-sds.com

ABSTRACT

Acoustic echo control has become an important feature in modern speech processing systems, such as hands-free telephones, multi-channel teleconferencing systems, public address systems, car interior communication devices, hearing aids and voice recognition systems.

This paper presents an overview on applications of acoustic echo control, explains the boundary conditions of these applications and shows recent solutions. For some selected applications (hands-free telephones and car-interior communications systems) results achieved with real systems are presented.

1. INTRODUCTION

The reduction of acoustic echoes can be realized by three different approaches:

- The most simple one is *echo suppression* by means of either broadband or frequency selective attenuation. In order to keep the background noise at a constant level this approach is often combined with the injection of so-called *comfort noise*.
- Whenever echo as well as desired speech components are present at the same time (double talk) echo suppression introduces audible distortions to the desired speech signal. In this case acoustic *echo cancellation* by means of an adaptive filter which models the loudspeaker-enclosure-microphone (LEM) system provides much better quality. In case of stereo or multi-channel systems the correlation between the excitation signals needs to be reduced by utilizing either nonlinear or time-varying decorrelation methods.
- In some applications such as hearing aids or car-interior communication systems the adaptation of the echo cancellation filter is rather difficult because of strong correlation between the excitation signal and the local signals. In such applications it is beneficial to exploit the spatial separation of the loudspeaker(s) and the local speaker by means of microphone arrays and *beamforming*.

In recent systems these three approaches are often combined. For the implementation different structures such as low delay broadband or computationally efficient subband or frequency domain processing can be utilized. The latter two structures have the drawback of introducing a delay into the signal path. The most suitable processing structure depends crucially on the boundary conditions of the specific application.

2. APPLICATIONS

In the following we will introduce four applications of acoustic echo control. The aim of this section is not to explain these systems in detail – the interested reader is referred to the cited references – but to highlight specific boundary conditions and problems of each application. Note that only those parts of the algorithms related to the reduction of acoustic echo or feedback, respectively, are mentioned (e.g. everything related to background noise reduction is omitted).

2.1 Hands-Free Telephone Systems

Hands-free telephones [3, 10, 13] started with just a loss control unit that allowed a half-duplex communication only. A full-duplex hands-free system requires an echo cancelling filter $\hat{\boldsymbol{h}}(n)$ in parallel to the LEM system (see Fig. 1). The filter has to be matched to this highly time-variant system. A microphone array can improve the ratio of the local speech to the echo from the loudspeaker and the local noise. Local noise and echoes remaining due to filter mismatch are suppressed by a filter $\boldsymbol{w}_{\mathrm{T}}(n)$ within the outgoing signal path.

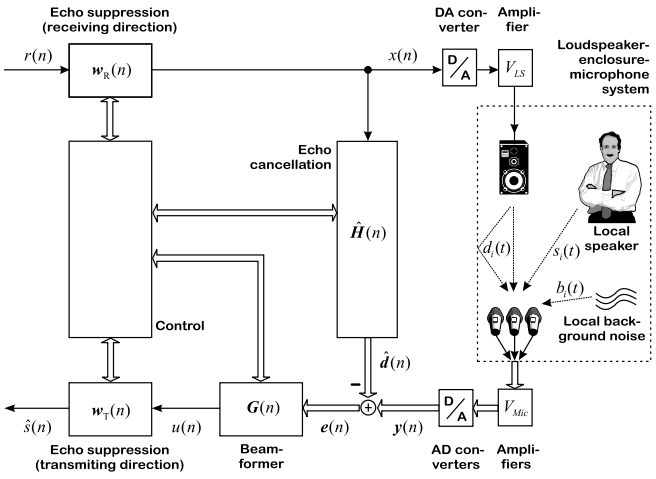

Figure 1: Structure of a hands-free telephone system.

Whenever more than one microphone is involved we will number the microphone channels by a subscript *i*. Lowercase and uppercase boldface letters indicate vectors and matrices, respectively. Each microphone requires one echo cancellation filter $\hat{\boldsymbol{h}}_i(n)$ if the beamformer is adaptive. All filter impulse responses are grouped within the matrix

$$\hat{\boldsymbol{H}}(n) = \left[\hat{\boldsymbol{h}}_0(n), ..., \hat{\boldsymbol{h}}_{L-1}(n)\right], \tag{1}$$

with L being the number of microphones. One of the most important units within an echo reduction system is the control unit. Only if all algorithmic parts are controlled reliably, the entire system is able to achieve high performance in terms of large echo reduction while not affecting the local speech signal. The filter $\boldsymbol{w}_{\mathrm{R}}(n)$ within the receiving path of the system can be utilized for inserting half of the required

attenuation during periods of double talk. Furthermore, the filter can be used to remove large amplitudes by means of a soft limiter. This helps to maintain the linear relationship between the excitation signal $x(n)$ and the echo component $d(n)$ (nonlinear effects – caused, e.g., by clipping of the AD converters – can not be modeled by a linear echo cancellation filter).

Boundary Conditions:

If we assume that the outgoing signal $\hat{s}(n)$ of a hands-free telephone is not (or only marginally) coupled back into the receiving signal $r(n)$ the excitation signal $x(n)$ of the adaptive filter and the local signals $s(n)$ (speech) and $b(n)$ (background noise) are statistically independent. If we further assume a zero-mean incoming signal the following cross correlations will disappear[1]:

$$\mathrm{E}\{x(n)s(n+l)\} = 0, \quad \text{and} \tag{2}$$

$$\mathrm{E}\{x(n)b(n+l)\} = 0, \quad \text{for all } n \text{ and } l. \tag{3}$$

In this case, any adaptive filter which converges to the Wiener solution [18] will perform a true system identification:

$$\hat{H}_{\mathrm{opt}}(\Omega) = \frac{S_{xy}(\Omega)}{S_{xx}(\Omega)} = H(\Omega)\,. \tag{4}$$

This means that even during double talk it is possible – a sufficiently small step size and/or a sufficiently large regularization parameter assumed – to identify the impulse response of the LEM system.

The International Telecommunication Union (ITU) and the European Telecommunications Standards Institute (ETSI) set up requirements for the maximally tolerable front end delay and the minimal acceptable echo suppression of telephones connected to the public telephone network. For ordinary telephones the additional delay introduced by echo and noise control must not exceed 2 ms [23]. For mobile telephones up to 39 ms additional delay is allowed [7]. It is obvious that these are severe restrictions for the types of algorithms usable for any front end processing. Especially in case of ordinary telephones computationally efficient frequency or subband domain procedures can not be applied. As the level of the echo is concerned standards require an attenuation of at least 45 dB in case of single talk. During double talk (or during strong background noise) this attenuation can be lowered to 30 dB. In these situations the residual echo is masked at least partially by the double talk and the noise.

2.2 Car Interior Communication Systems

Car interior communication systems (also called *intercom systems*) are not as intensively studied as hands-free telephones. For this reason, we will introduce those systems in more detail.

In limousines and vans communication between passengers in the front and in the rear may be difficult. Driver and front passengers speak towards the windshield. Thus, they are hardly intelligible for those sitting behind them. In the directions rear-to-front and left-to-right the acoustic loss is smaller. This can be measured by placing a so-called artificial mouth loudspeaker[2] at the driver's seat and torsos with

[1] This is not the case for all of the following applications.

[2] This is a loudspeaker which has (nearly) the same radiation pattern as the human speech apparatus.

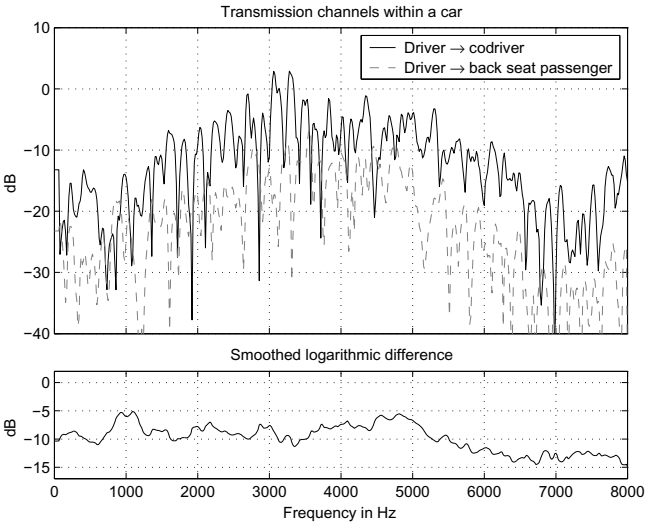

Figure 2: Frequency responses of different (driver to codriver and driver to left rear passenger) communication directions within a car.

earmicrophones [24] at the codriver's seat and at the backseats, respectively. In Fig. 2 the frequency responses measured between the driver's mouth and the left ear of the codriver, respectively the left ear of the rear passenger behind the codriver are depicted. On average the acoustic loss is 5 to 15 dB larger to the backseat passenger (compared to the codriver).

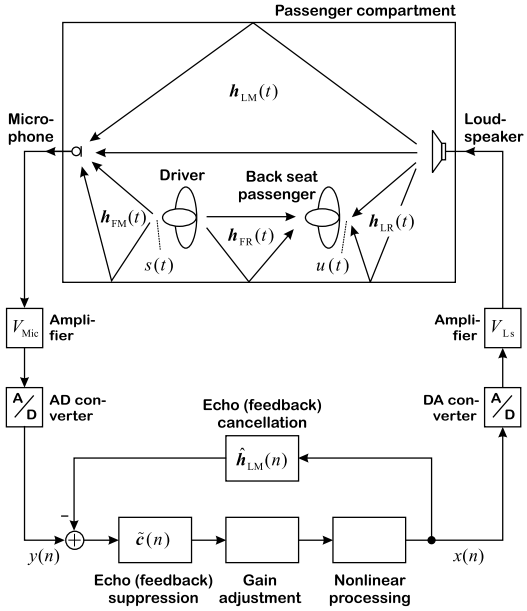

Figure 3: Structure of a car interior communication system.

Fig. 3 sketches the structure of a simple car interior communication system [26, 28] aimed to support only front-to-rear conversations with one microphone and one loudspeaker. Since driver and front passenger are located at well defined positions, fixed microphone arrays (not depicted in Fig. 3) can point towards each of them requiring fixed beamformers only. This allows to start with the echo and feedback cancellation after the beamformer (and to reduce the computational complexity because only one echo cancellation filter

is required). Feedback suppression by means of an adaptive notch filter $\tilde{\boldsymbol{c}}(n)$ can improve the system. Thus, the howling margin is improved. A device with non-linear characteristic attenuates large signal amplitudes. The output gain of a car interior communication system needs to be adjusted continuously according to the current driving condition. While only a moderate gain is required whenever the car does not drive a large gain is required and more artifacts will be tolerated at high speed.

Boundary Conditions:

The feedback cancellation turns out to be extremely difficult since the adaptation of the filter $\hat{\boldsymbol{h}}_{\mathrm{LM}}(n)$ is disturbed by the strong correlation between the excitation signal of the adaptive filter $x(n)$ and the speech signals of the driver and co-driver $s(n)$:

$$\mathrm{E}\{x(n)s(n+l)\} \neq 0\,. \tag{5}$$

Algorithms which are converging towards the Wiener solution [18] will converge towards[3]

$$\hat{H}_{\mathrm{LM,opt}}(\Omega) = \frac{S_{xy}(\Omega)}{S_{xx}(\Omega)} = H_{\mathrm{LM}}(\Omega) + \frac{S_{xs}(\Omega)}{S_{xx}(\Omega)} H_{\mathrm{FM}}(\Omega)\,, \tag{6}$$

which is not the desired solution $(\hat{H}_{\mathrm{LM,opt}}(\Omega) = H_{\mathrm{LM}}(\Omega))$. For this reason, the adaptation is usually performed only at falling signal edges of the excitation signal (whenever the speaking person stops talking for a short moment). During such periods the correlation between the excitation signal and the echo component is much stronger than the correlation between $x(n)$ and $s(n)$. Furthermore, during noise only periods (none of the passengers is speaking) the output signal consists only of background noise. By replacing the output signal with artificially generated noise the undesired cross-correlations can be forced to become zero. However, in this case only a low excitation-to-noise ratio can be expected and the convergence speed is slowed down.

Another approach uses nonlinear or time-variant procedures – known from stereo and multi-channel echo cancellation – applied on the system output signal $x(n)$ in order to reduce the correlation with the signal $s(n)$.

The special challenge in developing systems for this task consists in designing a system that exhibits at most 10 ms delay [17]. Signals from the loudspeakers delayed for more than that will be perceived as echoes and reduce the subjective quality of the system. For this reason, only broadband processing or block processing with very small block sizes can be applied if a high system quality should be achieved.

2.3 Public Address Systems

The problems here are similar to those of car interior communication systems. The enclosure, however, is much larger. Consequently the reverberation time is considerably longer. Loudspeakers are placed at several locations in the auditorium in order to provide a uniform audibility at all seats. Listeners in front rows should hear the speaker directly and via (more than one) loudspeaker at about the same loudness levels and with about the same delays. A microphone array at the lectern should be sensitive towards the speaker and insensitive in the directions of the loudspeakers.

[3] For the definitions of the signals, impulse and frequency responses see Fig. 3.

In addition to feedback cancellation and feedback suppression a frequency shift can be applied to prevent howling [29]. As the name indicates the recorded signals are shifted in frequency by a few cycles before playing them via the loudspeakers. When increasing the amplification of the loudspeakers such systems would start howling at a certain frequency. The frequency shift, however, moves the howling frequency by a few cycles at each loop through the system. As soon as a spectral valley in the frequency responses of LEM systems is reached the howling is attenuated. Adaptive notch filters to attenuate maxima of the closed-loop transfer function are also applied.

Boundary Conditions:

The boundary conditions of public address systems are similar to those of car interior communication systems, except that the size of the enclosures is much larger compared to passenger compartments of cars. Furthermore, the sampling frequencies are also larger (up to 48 kHz). For this reason echo and feedback cancellation filters of much higher order are required. On the other hand, the acceptable prize of public address systems is much higher, hence more powerful hardware can be utilized.

2.4 Hearing Aids

Compared with the applications already mentioned hearing aids are “a world of their own”. They are powered by very small batteries. This fact prohibits – at least at the time of writing – the use of standard signal processing circuits. Devices featuring extremely low power consumption are especially designed for hearing aids. An ear hook connects the loudspeaker to the outer auditory canal. It provides a tight fit and, thus, reduces acoustical feedback to the microphone. Completely-in-the-canal hearing aids are small enough to be worn in the outer ear. Fig. 4 shows the architecture of a completely-in-the-canal system.

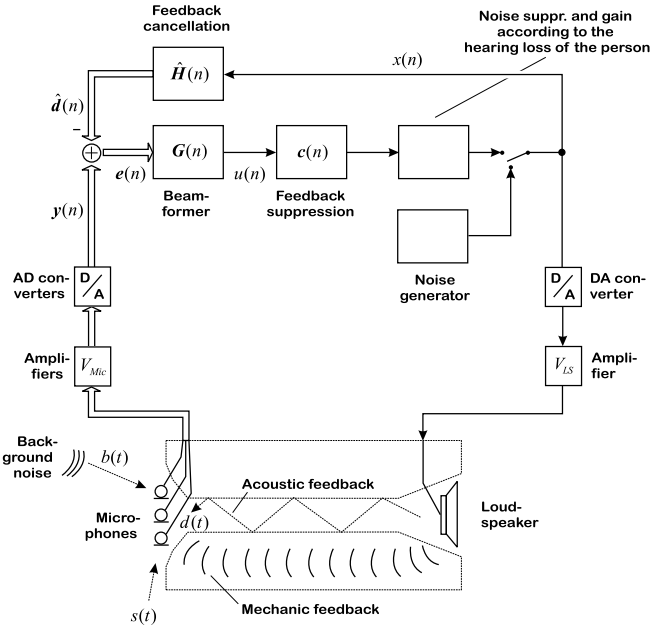

Figure 4: Architecture of a hearing aid.

Hearing aids can provide amplifications of more than 60 dB. Thus, feedback developing through the ventilation open-

ings and the cabinet is extremely critical. In forward direction the signal is split in up to 32 subbands by a filter bank. In order to ensure short delays the filter output signals are not subsampled. Thus, no synthesis filter bank is necessary. The subband signals are amplified according to the needs of the hearing impaired person.

In behind-the-ear systems noise reduction is achieved by two or more microphones that are located close to each other and by differential processing of their outputs [16]. If it is required by the acoustical environment they can be switched to an omnidirectional characteristic. Because there is no free sound field inside the ear canal a different noise reduction method has to be applied in completely-in-the-canal systems. It uses the filter bank outputs and estimates their amplitude modulation depths. If the modulation depth of a subband signal is low, noise is assumed and the band is attenuated. Improved processing capabilities will also allow spectral subtraction techniques in the future. In speech pauses the microphone output is switched off and an electronically generated noise is inserted. Since this is not correlated with the microphone output signals it is very well suited for identifying the acoustical/mechanical feedback loop.

Boundary Conditions:

The boundary conditions of hearing aids are comparable to those of car interior communication systems and public address systems. Additionally, one has to deal with only very limited word length (often less than 16 bits) of the available hardware. Thus, numerical inaccuracies have to be taken into account when designing algorithms for hearing aids.

3. SOLUTIONS

In this section we will introduce solutions to the problems mentioned while describing the applications of acoustic echo control. As in the last section going into the details of each method would go far beyond the scope of this paper. For this reason, the interested reader, once more, is referred to the references cited within the corresponding sections.

3.1 Echo Cancellation

In contrast to loss controls, which will be described in Sec. 3.3.1, echo cancellation by means of an adaptive filter $\hat{\boldsymbol{h}}(n)$ offers the possibility of full-duplex communication. As already explained in Sec. 2.1 an adaptive filter is used as a digital replica of the LEM system (see Fig. 5). If the A/D and D/A converters, as well as the amplifiers are not overloaded the system can be modeled with sufficient accuracy as a linear system with the finite impulse response

$$\boldsymbol{h}(n) = \left[h_0(n), h_1(n), ..., h_{N-1}(n)\right]^{\mathrm{T}}. \tag{7}$$

The locally generated signals (local speech $s(n)$ and local background noise $b(n)$) are summarized within the signal $n(n)$. By subtracting the estimated echo

$$\widehat{d}(n) = \sum_{k=0}^{N-1} \hat{h}_k(n) x(n-k) = \hat{\boldsymbol{h}}^{\mathrm{T}}(n)\, \boldsymbol{x}(n) \tag{8}$$

from the microphone signal $y(n)$ a nearly perfect decoupling of the system is possible (assuming that the echo cancellation filter has converged).

In most applications the adaptive filter requires a few hundred or even more than thousand coefficients in order to model the LEM system with sufficient accuracy. If more than one microphone is used and echo cancellation is performed independently for each microphone the computational load is rather large even for powerful modern signal processors. For this reason, other processing structures than fullband processing are often applied (see next section). Besides the complexity aspect also the delay restrictions of the application have to be taken into account when choosing the processing structure. Furthermore, it is very important to control the adaptation process reliably by weighting the adaptation with a step size or by introducing a regularization parameter [14, 27].

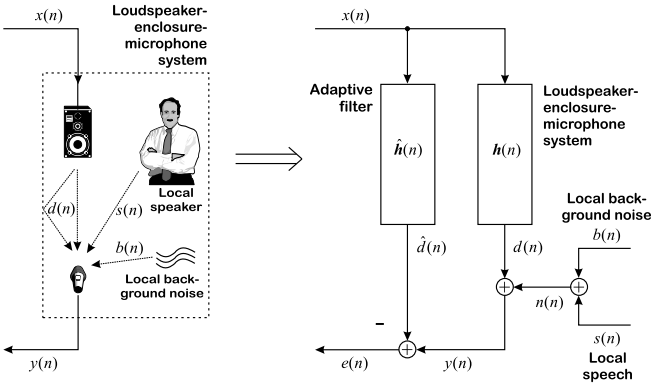

Figure 5: Structure of an acoustic echo canceller.

When changing from monophonic excitation signals to stereophonic or even multi-channel signals, not only the complexity increases but also a non-uniqueness problem [2] within the system identification arises. This can be solved by inserting either a nonlinearity [1] or a time-varying filter [30] into the receiving channel of the system.

3.1.1 Processing Structures

Beside selecting an adaptive algorithm [11] like LMS, affine projection, RLS, etc., the system designer also has the freedom to choose between different processing structures. The most popular ones are

- fullband processing,
- block processing[4], and
- subband processing.

Frequency-domain adaptive filtering (block processing) and subband structures reduce the complexity problem. Furthermore, both structures allow a frequency dependent normalization of the input signal. This method improves the convergence rate of simple adaptive algorithms like NLMS in case of colored excitation signals like speech.

In fullband structures it is possible to adjust the control parameters of an algorithm differently for each sample interval. For this reason we have a very good time resolution for control purposes. On the other hand it is not possible to choose frequency dependent control parameters differently. Subband processing and block processing in the frequency domain offer this possibility. Due to the subsampling and the collection of a block of input signals, respectively, the time-resolution decreases.

[4] By block processing we mean here performing the convolution and/or the adaptation in the frequency domain and using overlap-add or overlap-save techniques.

Block processing with large block sizes offers the highest reduction in computational complexity – but also introduces the largest delay. Subband structures can be seen as a compromise. If the analysis and synthesis filter banks are well designed, it is possible to achieve acceptable time and frequency resolutions, to introduce only a moderate delay, and to keep the computational load at a moderate level. Nevertheless, block (frequency domain) and subband processing are closely related to each other.

3.2 Beamforming

An alternative to the algorithms based on the subtraction of an estimated echo is the usage of a microphone array. Here, the output signals of several microphones are combined in such a way that the signals coming from a predefined source direction ϕ_0 (e.g. the direction pointing to the local speaker) are added in phase. Signals from other directions are combined only diffusely or even destructively, leading to an attenuation of signals from these directions. In case of non-adaptive combination of the microphone signals no time-variant distortions are introduced in the signal path. This is one of the most important advantages of beamformers over conventional single-channel echo suppression systems based on spectral attenuation (described in Sec. 3.3).

Beamforming was first developed in the field of radar, sonar, and electronic warfare [31]. In these applications the ratio of bandwidth and center frequency of the band of interest is very small. Thus, algorithms designed for those subjects can not be ported directly to audio applications.

The output signal $u(n)$ of a beamformer is given by the addition of the (FIR) filtered microphone signals (see Fig. 6):

$$u(n) = \sum_{i=0}^{L-1} \sum_{k=0}^{M-1} y_i(n-k)\, g_{i,k}(n) = \sum_{i=0}^{L-1} \boldsymbol{y}_i^{\mathrm{T}}(n)\, \boldsymbol{g}_i(n)\,, \qquad (9)$$

where L and M denote the number of microphones and the length of the FIR filters, respectively. The simplest type of beamformer is the delay-and-sum beamformer. In this case the filters are time-invariant ($\boldsymbol{g}_i(n) = \boldsymbol{g}_i$) and designed such that

$$G_i(\Omega) \approx \frac{1}{L} e^{-j\Omega\tau_i}\,, \qquad (10)$$

where τ_i represents the delay of the i^{th} microphone channel for time-aligning the signals from a predefined source direction. If this direction is not known a-priori it has to be estimated [22, 25]. The so-called *beampattern* for linear arrays (all microphones are located within one line and are equally spaced), which is defined as the squared magnitude of

$$B(\Omega, \phi) = \sum_{i=0}^{L-1} G_i(\Omega)\, e^{-j\Omega i \frac{d\, f_s}{c}}\,, \qquad (11)$$

is depicted in the lower right of Fig. 6 for an array consisting of $L = 4$ ideal omnidirectional sensors with a distance of $d = 5$ cm between two adjacent microphones. The quantity c is denoting the speed of sound ($c \approx 340$ m/s). If a better directivity at low frequencies should be achieved the delay-and-sum principle can be extended to a filter-and-sum approach. In this case the filters $\boldsymbol{g}_i$ are designed such, that the output power of the beamformer is minimized while keeping

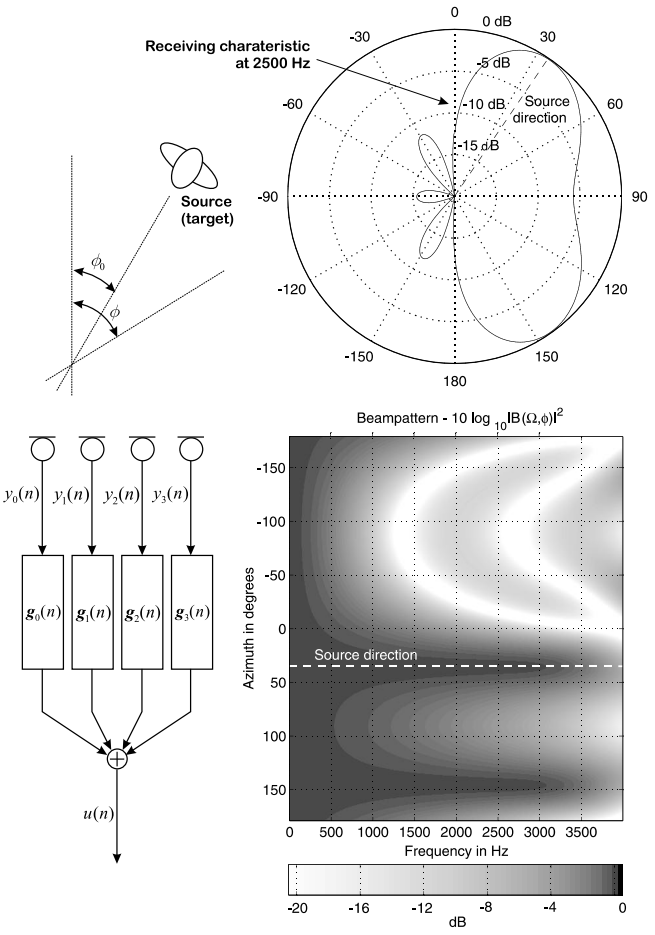

Figure 6: Characteristics of a microphone array.

the receiving characteristic of the desired direction ϕ_0 as a pure delay of K_0 samples:

$$\mathrm{E}\left\{ u^2(n) \right\} \longrightarrow \min\,, \quad \text{with} \quad B(\Omega, \phi_0) = e^{-j\Omega K_0}\,. \qquad (12)$$

Additionally more constraints can be added to the design process, making the result more robust against positioning tolerances and sensor imperfections [19]. When minimizing the output power according to Eqn. 12 the temporal and spatial correlation properties of the noise (e.g. echo or feedback) need to be known. If they are not known a-priori adaptive beamforming can be applied. The most popular approach is generalized sidelobe cancellation according to [12]. In order to make the adaptive approach more robust against several kinds of distortions, a variety of extensions such as an adaptive blocking matrix [21] or adaptive microphone calibration have been proposed. Because of reverberation effects it is very important to adapt a beamformer in generalized sidelobe cancellation structure only during speech pauses of the desired speaker [20] – otherwise signal cancellation (delayed version of the desired signal, caused by reflections on the boundary of the enclosure, cancel the signal corresponding to the direct path) appears.

Finally, an important feature of multi-microphone processing should be considered. By using more than one microphone it becomes possible to estimate the direction of a sound source. With this information it is possible to distinguish between different sound sources (e.g. the local speaker and the loudspeaker which emits the signal from the remote

speaker) which are very similar in their statistical properties. This spatial information can be exploited for enhanced system control [8].

3.3 Echo Suppression

Usually echo cancellation and beamforming are combined with additional residual echo suppression. This could be achieved in a broadband or in a frequency-selective manner.

3.3.1 Controlled Broadband Attenuation

One of the oldest attempts to solve the acoustic feedback problem is the use of speech activated switching or loss controls – a device for hands-free telephone conversation using voice switching was already presented in 1957 [5]. Attenuation units are placed into the receiving path $a_R(n)$ and into the transmitting path $a_T(n)$ of a hands-free telephone (see Fig. 7). In case of single talk of the remote speaker only the transmitting path is attenuated. If only the local speaker is active all attenuation is inserted into the receiving path. During double talk (both partners speak simultaneously) the control circuit decides which direction (transmitting or receiving) will be opened. This approach is the only one that can guarantee the amount of attenuation required by the ITU-T or ETSI recommendations.

The basic principle has been enhanced by adding several features. Before introducing the attenuation of the loss control the incoming signals are weighted with a slowly varying gain or attenuation in order to achieve a determined signal level during speech activity. Low voices will be amplified while loud speech is attenuated (up to a certain level). This mechanism is called *automatic gain control*. In case of hands-free telephones automatic gain controls are also used to adjust the signal level according to the distance of the local speaker from the microphone.

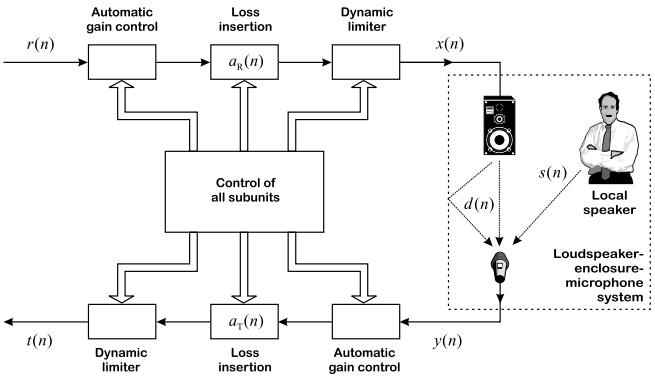

Figure 7: Dynamic processing in a hands-free telephone.

Before the attenuated or amplified signals are converted from digital to analog it must be ensured that the amplitudes do not exceed the range of the amplifiers. If this is the case clipping would occur which reduces the speech quality significantly. For this reason *dynamic limiters* are introduced. Such a device monitors the short-term envelope of a signal. Only large peaks will be attenuated. All other parts of the signal are not attenuated in order to keep the dynamics of the signal. Further algorithms which belong to the group of dynamic processing can be found e.g. in [32].

3.3.2 Spectral Attenuation

The coupling between the loudspeaker and the microphone is reduced in proportion to the degree of the match of the echo cancellation filter $\hat{\boldsymbol{h}}(n)$ to the original system $\boldsymbol{h}(n)$. Since in real applications a perfect match (over all times and all situations) cannot be achieved the remaining signal $e(n)$ (see Fig. 5) still contains echo components. To reduce these further, a Wiener-type filter [18] with a transfer function

$$W_{\text{opt}}(\Omega) = \frac{S_{en}(\Omega)}{S_{ee}(\Omega)} \tag{13}$$

can be utilized[5]. For reasons of better readability the undisturbed error signal $e_u(n) = e(n) - n(n)$ is introduced. In case of orthogonal excitation and local signals the cross power spectral density in Eqn. 13 can be simplified to $S_{en}(\Omega) = S_{ee}(\Omega) - S_{e_u e_u}(\Omega)$. When applying Eqn. 13, the power spectral densities have to be replaced by their short-term estimates. Therefore, the quotient may become smaller than 0. Consequently, the filter exhibits a phase shift of π. To prevent that, Eqn. 13 is (heuristically) modified to

$$\hat{W}_{\text{opt}}(\Omega, n) = \max\left\{1 - \beta\,\frac{\hat{S}_{e_u e_u}(\Omega, n)}{\hat{S}_{ee}(\Omega, n)}, W_{\min}\right\}; \tag{14}$$

where the spectral floor $W_{\min} \geq 0$ determines the maximal attenuation of the filter. The so-called *overestimation parameter* β is a second heuristic modification of the transfer function of the echo suppression filter. Using this parameter the "aggressiveness" of the filter can be adjusted. In order to estimate the short-term power spectral density of the error signal $\hat{S}_{ee}(\Omega, n)$ a short-term frequency analysis – such as a filter bank or a DFT – is required.

According to our model of the LEM system, the undisturbed error $e_u(n)$ can be expressed by a convolution of the excitation signal $x(n)$ and the system mismatch $\Delta h_i(n) = h_i(n) - \hat{h}_i(n)$:

$$e_u(n) = \sum_{i=0}^{N-1} \Delta h_i(n)\, x(n-i) = \boldsymbol{\Delta h}^T(n)\, \boldsymbol{x}(n)\,. \tag{15}$$

Hence, the power spectral density of the undisturbed error signal can be estimated by multiplying the short-term power spectral density of the excitation signal with the squared magnitude spectrum of the estimated system mismatch:

$$\hat{S}_{e_u e_u}(\Omega, n) = \hat{S}_{xx}(\Omega, n)\,\left|\Delta H(\Omega, n)\right|^2. \tag{16}$$

For estimating the system mismatch, a so called double-talk detector [9] and a detector for enclosure dislocations [27] are required. A detailed description of such detectors as well as procedures for estimating $|\Delta H(\Omega, n)|^2$ can be found in [15].

Beside the advantages of residual echo suppression, the disadvantages should be mentioned here as well. In contrast to echo cancellation, echo suppression schemes introduce attenuation into the sending path of the local signal. Therefore, a compromise between attenuation of residual echoes and reduction of the local speech quality has to be made. In case of large background noise, a modulation due to the attenuation of the noise during remote single talk appears. In such scenarios the echo suppression should always be combined with (background) noise suppression and comfort noise injection.

[5] It is assumed that no local background noise is present ($b(n) = 0$). Thus, we have $n(n) = s(n)$.

3.4 Feedback Suppression

Whenever closed-loop acoustic echo control systems – such as public address or car interior communication systems – are operating close to the stability threshold, some sort of "emergency brake" should be implemented. One possibility to realize this, is to implement a feedback suppression filter according to Fig. 8. For $\alpha = 0$ the structure resembles a predictor error filter. The FIR filter $\boldsymbol{c}(n)$ is an adaptive filter which is adjusted such that the power of the output signal $e(n)$ is minimized – e.g. using the NLMS algorithm. If howling occurs at a certain frequency the feedback suppression filter tries to attenuate this frequency. According to the filter structure this is possible as long as the inverse of the howling frequency is larger than N_D and smaller than $N_\mathrm{D} + N_\mathrm{C}$ sample intervals. N_C is denoting the length of the filter $\boldsymbol{c}(n)$.

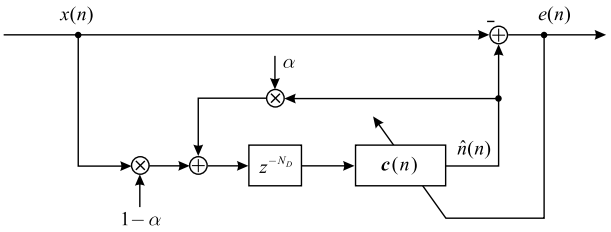

Figure 8: Structure of a basic feedback suppression system.

Delaying the incoming signal before filtering is necessary because otherwise the short-term correlation within the speech signal would be removed. In this case the spectral envelope would be flattened. A delay of about 2 ms is sufficient to avoid this. Due to periodic components of speech signals the memory of the adaptive filter should not exceed a time interval equivalent to the pitch period[6]. For this reason the filter should not contain more than 40 to 60 coefficients (at 8 kHz sampling rate). Otherwise also the periodic components of the speech signal would be suppressed.

Due to short-term correlated components within speech signals the filter also tries to suppress parts of the speech signal. By using a small step size μ this behaviour can be avoided and only periodic distortions which are present for a longer time interval are cancelled. A small step size, on the other hand, leads to a slow convergence. Sudden (periodic) distortions would be attenuated only after a non-negligible period of time. For this reason a compromise for the step size has to be found. Usually, the NLMS algorithm with a fixed, but small step size $\mu \in \{0.01, 0.00001\}$ is utilized. It is needless to say that other adaptation algorithms can be used as well. Besides the ones suitable for echo cancellation algorithms which use higher order statistics have also been applied to this problem [6].

The basic FIR structure of a feedback suppression filter can be extended by a weighted feedback path [4] as depicted in Fig. 8. By varying the feedback gain α it is possible to modify the filter from an FIR structure ($\alpha = 0$) to an adaptive oscillator ($\alpha = 1$). The motivation behind this IIR configuration is to achieve some of the benefits of a noise canceller with a separate pure periodic reference. With the extended structure it is possible to achieve very narrow notches. Nevertheless, due to the IIR structure the filter might become instable if the adaptation process forces the poles to move out of the unit circle within the z-domain. For this reason, the stability of the structure needs to be checked periodically.

[6]The specification is only true for FIR structures. In case of IIR schemes the group delay should not exceed the specified range.

4. REAL SYSTEMS

Some results achieved with real systems are presented here. In Fig. 9 the input and output signal of a hands-free telephone system are depicted. A commercially available system (StarRec® from Temic) was installed in a car. At a speed of about 50 km/h the driver gets a phone call. During the first 4 seconds the remote speaker is active. Afterwards double talk appears for about 4 seconds and finally only the driver speaks. Four microphones located within the rear-view mirror were connected to the system. For playing the signals received from the remote side the standard car speakers were utilized. To remove the echo components echo cancellation, beamforming and frequency selective post filtering in combination with comfort noise injection were applied. Furthermore, a noise suppression unit reduces the background noise up to 8 dB.

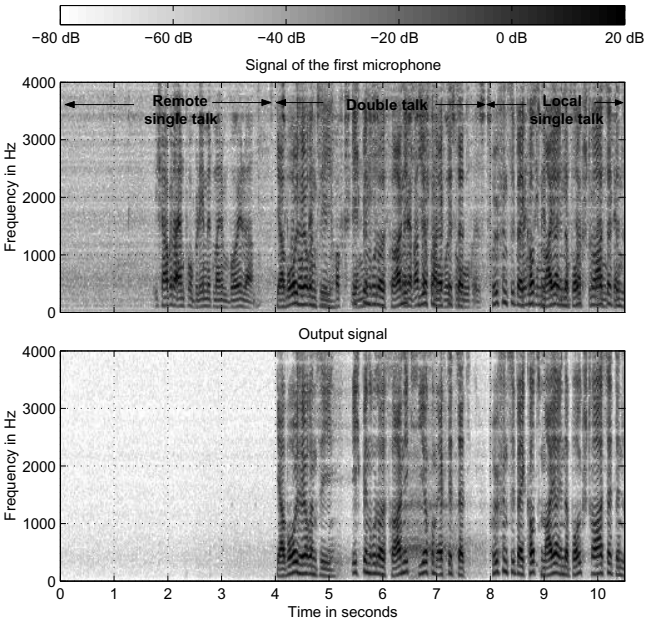

Figure 9: Input (mircophone) and output signal of a hands-free telephone system.

When analyzing the time-frequency analyses depicted in Fig. 9 the removal of the echo components within the first 4 seconds is clearly visible. Double talk without disturbing artifacts is possible as well.

Fig. 10 shows the results of a car interior communication system. The system utilizes 8 microphones (2 per passenger) and 6 loudspeaker channels (standard car loudspeakers). To obtain high speech intelligibility beamforming, feedback cancellation, loss control and dynamic processing were applied. Especially at high speeds (90 km/h or more) a clear improvement of the communication quality could be achieved. To visualize this gain a binaural recording was made with a torso located on the seat behind the driver. Fig. 10 shows a time-frequency analysis of the output signal of the microphone located in the torso's left ear. The car was driving at about 160 km/h. The driver was talking with the same loudness during the entire recording. After 16 seconds the system was deactivated for about 7 seconds to demonstrate

the system performance. Within the time-frequency analysis the speech components of the driver are recognizable whenever the system is activated. During deactivation, however, the driver's speech is mostly masked by the driving noise.

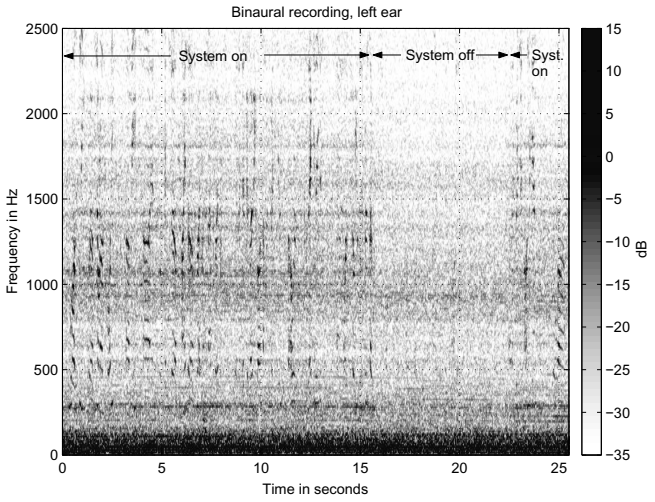

Figure 10: Time-frequency analysis of the right channel of a binaural recording made on the left backseat.

5. CONCLUSIONS AND OUTLOOK

In case of no correlation between the excitation and the local signal single-channel acoustic echo control is well understood and solutions that provide acceptable speech quality are available. However, a variety of problems are still waiting to be solved. Stereo and multichannel echo cancellation systems that do not disturb the playback of high quality audio (music) or adaptation algorithms which can handle correlated distortions in a reliable manner are possible candidates for further research. Due to the continuous progress in designing even more powerful hardware at reduced cost, new algorithms might come up. For this reason, research as well as product development in speech and audio processing continues to be a very interesting profession.

REFERENCES

[1] J. Benesty, D. R. Morgan, M. M. Sondhi: *A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation,* Proc. ICASSP '97, vol. 1, pp. 303-306, 1997.

[2] J. Benesty, D. R. Morgan, M. M. Sondhi: *A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation,* IEEE Trans. Acoust. Speech Signal Process., vol. 6, no. 2, pp. 156-165, 1998.

[3] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, J. Tilp: *Acoustic Echo Control*, IEEE Signal Process. Mag., vol. 16, no. 4, pp. 42-69, 1999.

[4] J. Chang, J.R. Glover: *The Feedback Adaptive Line Enhancer: A Constrained IIR Adaptive Filter*, IEEE Trans. Signal Process., vol. 41, no. 11, pp. 3161-3166, 1993.

[5] W. F. Clemency, F. F. Romanow, A. F. Rose: *The Bell System Speakerphone*, AIEE. Trans., vol. 76(I), pp. 148-153, 1957.

[6] A. G. Constantinides, K.M. Knill, J.A. Chambers: *A Novel Orthogonal Set Adaptive Line Enhancer Tuned with Fourth-Order Cumulants,* Proc. ICASSP '92, vol. 4, pp. 241-244, 1992.

[7] ETS 300 903 (GSM 03.50): *Transmission Planning Aspects of the Speech Service in the GSM Public Land Mobile Network (PLMS) System,* ETSI, France, 1999.

[8] M. Fuchs, T. Haulick, G. Schmidt: *Noise Suppression for Automotive Applications Based on Directional Information,* accepted for publication on ICASSP'04, 2004.

[9] T. Gänsler, M. Hansson, C.-J. Ivarsson, G. Salomonsson: *A Double-Talk Detector Based on Coherence,* IEEE Trans. Commun., vol. COM-44, no. 113, pp. 1421-1427, 1996.

[10] A. Gilloire, E. Moulines, D. Slock and P. Duhamel: *State of the Art in Acoustic Echo Cancellation*, in A.R. Figueiras–Vidal (ed.), Digital Signal Processing in Telecommunications, Springer, London, UK, pp. 45-91, 1996.

[11] G. Glentis, K. Berberidis, S. Theodoridis: *Efficient Least Squares Adaptive Algorithms for FIR Transversal Filtering: A Unified View*, IEEE Signal Process. Mag., vol. 16, no. 4, pp. 13-41, 1999.

[12] L. J. Griffiths, C. W. Jim: *An Alternative Approach to Linearly Constrained Adaptive Beamforming,* IEEE Tans. Antennas and Propagation, vol. AP-30, no. 1, pp. 24-34, 1982.

[13] E. Hänsler: *The Hands–Free Telephone Problem – An Annotated Bibliography*, Signal Process., vol. 27, no. 3, pp. 259-271, 1992.

[14] E. Hänsler, G. Schmidt: *Control of LMS-Type Adaptive Filters,* in S. Haykin, B. Widrow (eds.), Least-Mean-Square Adaptive Filters, pp. 175-240, New York: Wiley, 2003.

[15] E. Hänsler, G. Schmidt: *Single-Channel Acoustic Echo Cancellation,* in J. Benesty, Y. Huang (eds.), Adaptive Signal Processing, pp. 59-93, Berlin: Springer, 2003.

[16] V. Hamacher: *Comparison of Advanced Monaural and Binaural Noise Reduction Algorithms for Hearing Aids,* Proc. ICASSP '02, vol. 4, pp. 4008-4011, 2002.

[17] T. Haulick, G. Schmidt: *Signalverarbeitungskomponenten zur Verbesserung der Kommunikation in Fahrzeuginnenräumen,* Proc. ESSV '03, pp. 130-137, 2003 (in German).

[18] S. Haykin: *Adaptive Filter Theory*, 4. Edition, Englewood Cliffs, NJ: Prentice Hall, 2002.

[19] W. Herbordt, W. Kellermann: *Adaptive Beamforming for Audio Signal Acquisition,* in J. Benesty, Y. Huang (eds.), Adaptive Signal Processing, pp. 155-194, Berlin: Springer, 2003.

[20] O. Hoshuyama, B. Begasse, A. Sugiyama, A. Hirano: *A Realtime Robust Adaptive Microphone Array,* Proc. ICASSP '98, vol. 6, pp. 3605-3608, 1998.

[21] O. Hoshuyama, A. Sugiyama, A. Hirano: *A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters,* IEEE Trans. Signal Process., vol. 47, no. 10, pp. 2677-2684, 1999.

[22] Y. Huang, J. Benesty, G. Elko: *Microphone Arrays for Video Camera Steering,* in S. Gay, J. Benesty, (eds.), Acoustic Signal Processing for Telekommunication, pp. 239-259, Boston, MA: Kluwer, 2000.

[23] ITU-T Recommendation G.167: *General Characteristics of International Telephone Connections and International Telephone Circuits – Acoustic Echo Controllers,* Helsinki, Finland, 1993.

[24] ITU-T Recommendation P.581: *Use of Head And Torso Simulator (HATS) for Hands-Free Terminal Testing,* Geneva, Switzerland, 2000.

[25] C. H. Knapp, G. C. Carter: *The Generalized Correlation Method for Estimation of Time Delay,* IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-24, no. 4, pp. 320-327, 1976.

[26] E. Lleida, E. Masgrau, A. Ortega: *Acoustic Echo and Noise Reduction for Car Cabin Communication,* Proc. EUROSPEECH '01, vol. 3, pp. 1585-1588, 2001.

[27] A. Mader, H. Puder, G. Schmidt: *Step-Size Control for Acoustic Echo Cancellation Filters – An Overview*, Signal Process., vol. 80, no. 9, pp. 1697-1719, 2000.

[28] A. Ortega, E. Lleida, E. Masgrau, F. Gallego: *Cabin Car Communication System to Improve Communication Inside a Car,* Proc. ICASSP '02, vol. 4, pp. 3836-3839, 2002.

[29] M. R. Schroeder: *Improvement of Acustic-Feedback Stability by Frequency Shifting,* J. Acoust. Soc. Am., vol. 36, no. 9, pp. 1718-1724, 1964.

[30] A. Sugiyama, Y. Joncour, A. Hirano: *A Stereo Echo Canceller with Correct Echo-Path Identification Based on an Input-Sliding Technique,* IEEE Signal Process. Mag., vol. 49, no. 1, pp. 2577-2587, 2001.

[31] B. D. Van Veen, K. M. Buckley: *Beamforming: A Versatile Approach to Spatial Filtering,* IEEE Signal Process. Mag., vol. 5, no. 2, pp. 4-24, 1988.

[32] U. Zölzer (editor): *DAFX Digital Audio Effects,* Wiley, New York, 2002.