STRATEGIES FOR COMBINING ACOUSTIC ECHO CANCELLATION AND ADAPTIVE BEAMFORMING MICROPHONE ARRAYS

Walter Kellermann
Fachhochschule Regensburg, Germany

ABSTRACT

New concepts for efficient combination of acoustic echo cancellation(AEC) and adaptive beamforming microphone arrays(ABMA) are presented. By decomposing common beamforming methods into a time-invariant part, which the AEC can integrate, and a separate time-variant part, the number of echo cancellers is minimized without rendering the system identification problem more difficult. Methods for controlling the interaction of ABMA and AEC are outlined and implementations for typical microphone array applications are discussed briefly.

1. INTRODUCTION

For acoustic echo control in conventional hands-free communication it is generally acknowledged that an echo canceller(EC) is desirable, which models the impulse response of the loudspeaker - enclosure - microphone system by an adaptive filter in order to remove echo components from the microphone signal. Other echo control methods, like loss insertion or nonlinear devices, are impairing full-duplex communication and, thus, are mostly considered as supplementary measures only. For applications such as teleconferencing between offices, studios, auditoria [2, 3, 4, 5, 6, 7] or car telephony [8, 9], convenience or safety aspects suggest that the personal microphone be replaced by a microphone array (MA) directing a beam of increased sensitivity at the active talker.

1.1. Acoustic Echo Path with Microphone Arrays

In contrast to single-microphone(SM) hands-free communication or multichannel teleconferencing [15], one might hope that for a MA no echo canceller(EC) is required, because the acoustic echo path from the loudspeaker could be sufficiently attenuated by the array directivity. Considering [10] as a guideline, echo attenuation should be at least 40dB during single-talk and 20dB during double-talk. Examining the echo attenuation provided by known MA implementations, we find:

1, The absolute gain of the MA has to increase along with the distance from the local talkers in order to compensate for the decay of the sound level ($\approx 6dB$ per doubling of distance in the far-field). This extra gain requires correspondingly more echo attenuation.

2, The directivity index – quantifying the gain of the desired direction over the average of all other directions – of fixed beamforming arrays does not exceed 20dB over a wide frequency range, and is much smaller at low frequencies [5, 7]. SNR improvement of adaptive beamforming arrays is limited to about 15dB for realistic conditions [3, 8, 11]. For reverberant environments, both quantities approximately express the echo attenuation provided by the MA.

3, Null-steering to the loudspeaker for maximum echo attenuation is only effective in nonreverberant environments [3]. Even in in carefully designed studios with optimized placement of sources, MA, and loudspeaker, unexpected reflections may reduce echo attenuation below 10dB [7].

For the echo path impulse response of an N-sensor MA in a reverberant environment, a simple model is supported by measurements: The array impulse response behaves like the sum of the N impulse responses for the individual microphones with the accumulated samples being mutually uncorrelated [12]. This implies an increased average echo attenuation for the MA on the order of about $10log_{10}N$ dB over a SM. This advantage of the MA must compensate for the additional gain due to the usually increased average talker-sensor distance and a possibly higher directivity of the SM compared to a single array sensor, if an EC of equal length should provide the same echo attenuation as for a SM. Thus, although the MA echo path could be further attenuated by loudspeaker arrays in combination with absorbing walls, AEC will in most cases remain desirable for full-duplex communications with MAs.

2. GENERIC CONCEPTS

In Fig.1 the structure of hands-free telecommunication using an ABMA is outlined. For the adaptive beamforming(BF), we allow here all spatially selective algorithms that extract the desired signal from the N microphone signals. This notion covers classical adaptive beamforming arrays [13] as well as beamsteering algorithms [2, 4, 6]. Only a single far-end-signal is allowed to avoid interference with the stereophonic AEC problem, which can be treated separately [14]. Two generic AEC approaches are discussed to illustrate the AEC problem[1]:

AEC-I operates directly on the microphone signals, i.e., for each of the N echo paths an acoustic echo canceller must be implemented. The AEC feels no repercussions by

[1] Note that this distinction is independent of the structure (fullband/subband/transform domain structures may be used) and of the adaptation algorithm for the AEC.

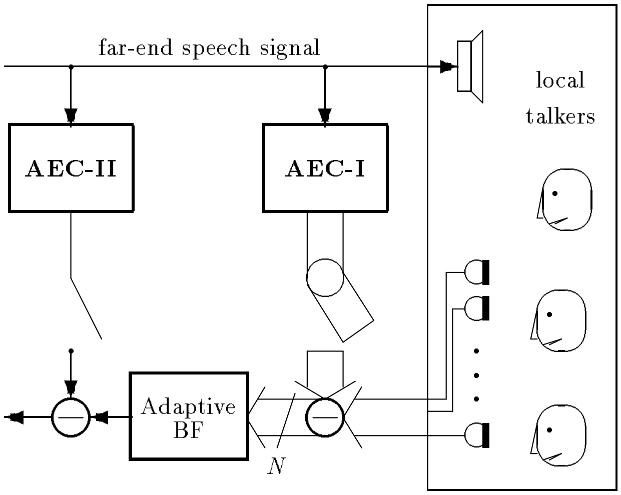

Figure 1. ABMA in a hands-free telecommunication system with two alternatives for AEC

the adaptive BF and, thus, the AEC problem is structurally the same as for a SM, duplicated by the number of sensors.

AEC-II operates on the output signal of the BF, requiring only a single EC. However, the AEC model has to incorporate the BF in addition to the acoustic echo path.

A major advantage of **AEC-I** is given by its structural simplicity as it only requires duplicating the established SM-AEC algorithms. However, for large N the computational load is considerable [1] and may be prohibitive for common teleconferencing and car telephony with $N = 7 \dots 23$ microphones [4, 7, 8, 9].

For **AEC-II**, only a single AEC is required, but this has to include the adaptive BF into its model of the echo path. As the unknown acoustic components cannot be identified separately from the known BF filtering system ('knapsack problem'), the time-variance of the BF poses a major problem: With the identification of the acoustic echo path being already difficult due to its large number of degrees of freedom and its unpredictable, potentially fast and severe changes of the impulse response [16], it becomes even more difficult if adaptive BF must be incorporated. Observing that the BF system must change its parameters whenever it 'switches' to a newly active local talker, severe fluctuations in the echo path impulse response occur at a time, when the adaptive EC is unable to track it, because the local source acts as interfering noise on the system identification. Hence, AEC-II will in general provide no echo attenuation until a far-end talker is in a single-talk period again and allows convergence of the EC. Thus, the benefits of AEC are often missing when they are desired most, i.e., during double-talk and at transitions from far-end activity to local activity and vice-versa (at other times loss insertion is less objectionable). As a result, the time-variance of the BF discourages the use of AEC-II.

3. NEW EFFICIENT CONCEPTS

From the previous section, we conclude that, for large N, new efficient concepts ideally should avoid the computational complexity of AEC-I and circumvent the time-variant BF in AEC-II. The key to this is to decompose the ABMA into a time-invariant stage followed by a time-variant stage. The time-invariant BF is to produce a minimum number of output signals, which the AEC can incorporate into its echo path model, and the time-variant part of the ABMA may not interfere with the AEC.

3.1. Beamforming methods

We distinguish two classes of BF methods which are common for MAs in telecommunications:

BF-I: For **beamsteering**, a set of M fixed beam signals is computed independently of the array input data, and the output of the beamformer is a weighted sum of these beams with time-variant weights accounting for the active talkers (**voting**) [2, 4, 6]^2.

BF-II: Classical **adaptive beamforming** methods aim at minimizing a statistical error criterion and filter the microphone signals accordingly [13]. Characteristically, the parameters of these systems are continually changing over time in order to converge to optimum filter coefficients [3, 8, 11]. (Note that tracking of moving or changing sources is usually not supported.)

3.2. AEC with BF-I

BF-I inherently provides the desired separation into a time-invariant and a time-variant stage. To minimize the number of signals for the AEC, we introduce a mapping of the M fixed-beam signals onto L 'talker beams' whenever $L < M < N$ (Fig.2). For maximum spatial selectivity, the mapping should select one fixed beam or a linear combination of two neighboring fixed beams per talker. The AEC has now to identify a time-invariant BF system as long as

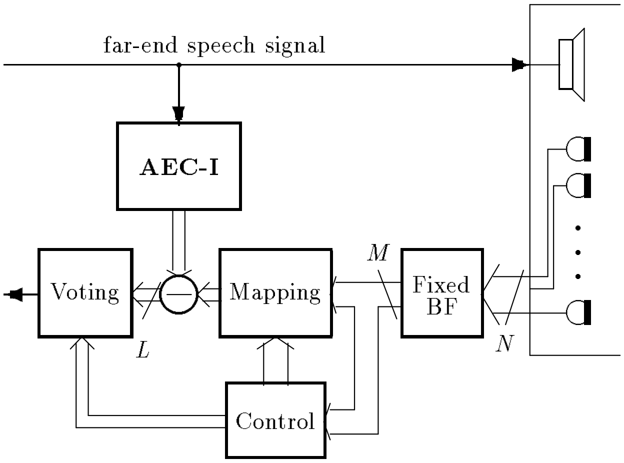

Figure 2. AEC combined with BF-I

[2] Note that all beam signals are meant to cover the entire frequency range of interest. Accounting for the wideband nature of speech and audio signals, nested arrays are usually employed whose outputs may be filtered as an ensemble [17] or as subarrays [2, 4, 7] before yielding a wideband beam signal. Fractional delay beamforming for increased spatial resolution is also covered by our model.

the mapping does not change and, thus, deals with an L-channel AEC-I problem.

3.3. AEC with BF-II

Similarly to the BF-I concept, we simultaneously apply L fixed sets of BF filters to the N microphone signals to account for each talker (Fig.3). Thus again, we obtain an L-channel AEC-I echo cancellation problem. The signal path of this structure is essentially the same as for BF-I, employing fixed beamforming and voting. The actual adaptive beamforming has been moved to the control path.

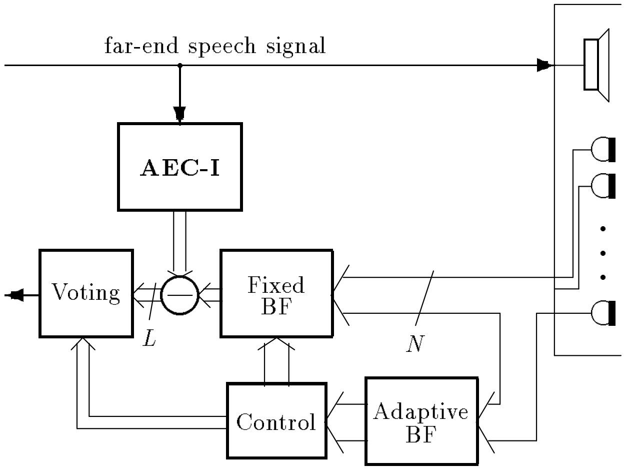

Figure 3. AEC combined with BF-II

For both BF-I and BF-II, the incorporation of the fixed beamforming into the echo path model requires a longer EC impulse response. The extra length is determined by the maximum delay realized in the delay and sum networks plus – for BF-I – interpolation and beam shaping filter order [4, 17], and – for BF-II – the length of the adaptive beamforming filter [8, 11].

3.4. Control mechanisms

With ABMAs and AEC being intensively researched areas on their own, we concentrate here on efficiently controlling their interaction. Unless referenced otherwise, the methods described below were verified and subjectively evaluated using recorded dialogues and measured impulse responses of MAs in cars, offices, and a videoconference studio.

3.4.1. Talker activity detection

The detection of talker activity is crucial for both AEC and BF. AEC relies on it for controlling the speed of adaptation, and BF needs it for voting and to identify periods when mapping for BF-I or optimum BF for BF-II can be learned. As in SM concepts, talker activity is classified by primarily evaluating the energies of loudspeaker and microphone signals, respectively [4, 6]. The spatial resolution of beamforming MAs provides additional information: E.g., for the BF-I concept, the M beam signal energies will show a typical pattern for each spatially fixed source such as the loudspeaker, which can then be distinguished from the patterns of other sources.

3.4.2. AEC

As long as computing resources allow, all L ECs should adapt in parallel during 'far-end talk only' periods. Alternatively, only the currently needed EC(s) (according to the voting) could be operated, while all others are kept frozen. As in the SM case, estimating the current echo path attenuation provided by AEC during far-end talk remains indispensable for determining the amount of required supplementary loss (notably during initial convergence, at changes of the acoustic path, and when the mapping for BF-I or the fixed BF of BF-II is updated).

3.4.3. BF during 'far-end talk only'

Experiments confirmed that using a BF configuration which simply minimizes echo feedback to avoid loss insertion, may give a disturbing spatial impression to the far-end party. Instead, we propose to use the BF configurations covering the local talkers and to insert supplementary loss.

3.4.4. Voting

The voting algorithm derives the array output signal from a weighted linear combination of L beam signals. Equally for BF-I and BF-II, the time-variant weights are chosen to allow a fast reaction to newly active local sources ($\approx$ 20msec) while at the same time avoiding the perception of switching noise [4]. For maximum spatial selectivity, for each talker only one beam signal should have a nonzero weight in the stationary case (for details see, e.g., [4]). When entering a far-end talk period we propose to start out with the weights for the most recently active local talker and gradually change weights to arrive at a beamforming averaging over all L talker beams.

3.4.5. Mapping for BF-I

For initialization, the results of a training procedure can be incorporated, or the dominant fixed beams during the first periods of local speech are used as initial talker beams. While applying the current fixed mapping to form the output signal, the control unit continuously monitors the short-term energies of the fixed beams and incorporates the beam energy patterns into a learning procedure – e.g., a first-order recursive filtering over time – for the currently active talker. The mapping should only be changed if a fixed beam or a combination of two neighboring fixed beams exhibits significantly more energy than the current mapping. A combination of two fixed beams is considered for the mapping only if the neighboring beams have about the same energy and their weighted sum produces clearly more energy than each of them. The mapping should preferably be updated during 'far-end talk-only' periods, as only then the AEC can identify the new echo path.

3.4.6. Fixed beamforming for BF-II

As with BF-I, the fixed beamforming for each of the L talkers must be initialized and should be updated only when the adaptive beamforming performs significantly better than the established fixed BF for the active talker. The initialization usually must include the localization of the desired sources and the convergence to an efficient BF configuration for each talker (c.f. [8]). The control unit is supported by an adaptive BF unit which is continually aiming at optimizing BF filters for the currently active local talker (not during double-talk or when several local talkers are active). For all L local talkers, the BF filter outputs must be computed for activity detection, if nothing else.

3.5. Examples

For illustration, the integration of our AEC concepts into various known ABMA implementations is considered.

For *car telephony,* MAs using GSC with typically 7 or 8 sensors [8, 9], have mainly be investigated for speech recognition applications so far. When using the BF-II concept for hands-free full-duplex telephony, the requirements for an EC are essentially the same as for a SM, as long as only a single local talker (e.g., the driver) is considered. Although the directivity gain of the array is not completely balanced by the increased average microphone distance compared to an optimally located SM, the incorporation of the beamforming into the echo path model leads to an EC impulse response of comparable length as for a SM.

For *desktop teleconferencing,* MAs compete with multichannel systems, offering the advantage of requiring less sensors when large groups communicate. The BF-II concept could be applied, e.g., to the AMNOR beamforming [3] based on $N = 4$ sensors. Assuming seated participants, the BF filters must be updated very infrequently and, as the echo paths will remain relatively stable most of the time, it will suffice to adapt one EC at a time. A realization of AEC with BF-I for desktop teleconferencing has been reported in [6]: Combining $N = 2$ dipole microphones, $L = 4$ beam signals are formed and only $min\{L, N\} = 2$ ECs need to be realized, acting directly on the microphone outputs.

For *videoconferencing,* MAs mounted to a wall or to the ceiling again compete with multi-channel systems (see, e.g. [15]). With nested beamsteering subarrays (BF-I) using a total of $N > 20$ microphones [4, 5, 7] up to $M = 7$ beams are formed, which cover typically $L = 2 \dots 5$ talkers. With a distance of $2 \dots 3$m between array and talker, an additional echo gain of at least 12 dB must be compensated by array directivity and AEC compared to SMs located at 0.5m from the talkers. Thus, the L ECs will in general be at least as complex as for SMs, unless the directivity of a loudspeaker array combined with absorbing surfaces provides additional echo attenuation.

For an *auditorium* as described in [2] using a planar array (BF-I, $N = 380$, $L = M = 27$), the echo cancellation problem is scaled up along three parameters compared to a teleconferencing studio: increased reverberation time demands longer EC impulse responses, increased talker-array distance provides extra echo gain demanding even longer EC impulse responses, and the large L requires more ECs. Thus, loudspeaker directivity and room design will remain of great importance for this application, if loss insertion is to be minimized.

4. CONCLUSIONS

Comparing the proposed concepts to AEC for a SM per local talker, the complexity of AEC for a MA is on the same order for car telephony and desktop teleconferencing, but increases along with array-talker distance for videoconferencing and auditoria. Many details of the outlined control methods call for further investigation, and more sophisticated approaches could be applied to key problems like beamforming training and voting. For spotting the most critical issues, however, real-life experiments using simple but complete implementations must be evaluated first.

5. ACKNOWLEDGMENT

The author wishes to thank Yannick Mahieux of CNET, Lannion, France, for inspiring discussions and for contributing helpful measurement data, and Gary W. Elko of Bell Labs, Murray Hill, NJ, who incited this work years ago.

REFERENCES

[1] M.M. Sondhi and W. Kellermann. Echo cancellation for speech signals. In S. Furui and M.M. Sondhi, eds., *Advances in Speech Signal Processing.* Marcel Dekker, 1991.

[2] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *JASA*, 78(5):1508–1518, 1985.

[3] Y. Kaneda and J. Ohga. Adaptive microphone-array system for noise reduction. *IEEE TR-ASSP*, 34(6):1391–1400, 1986.

[4] W. Kellermann. A self-steering digital microphone array. *Proc. ICASSP*, pp.3581–3584, Toronto, 1991.

[5] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58–71, 1991.

[6] P. Chu. Desktop mic array for teleconferencing. *Proc. ICASSP*, pp.2999–3002, Detroit, 1995.

[7] C. Marro, Y. Mahieux. Analysis of dereverberation and noise reduction techniques based on microphone arrays microphone with optimal filtering. *IEEE TR-SAP* (submitted).

[8] S. Oh, V. Viswanathan, and P. Papamichalis. Hands-free voice communication in an automobile with a microphone array. *Proc. ICASSP*, pp.I-281 – I-284, San Francisco, 1992.

[9] S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson. Noise reduction using an adaptive microphone array in a car. In *Conf. Rec. of IEEE ASSP Workshop on Appl. of DSP to Audio and Acoustics*, New Paltz, USA, 1993.

[10] ITU-T Recommendation G.167 - Acoustic Echo Controllers, March 1993.

[11] K. Farrell, R.J. Mammone, and J.L. Flanagan. Beamforming microphone arrays for speech enhancement. *Proc. ICASSP*, pp.I-285 – I-288, San Francisco, 1992.

[12] W. Kellermann. Some properties of echo path impulse responses of microphone arrays and consequences for acoustic echo cancellation. In *Conf. Rec. of the 4th Intern. Workshop on Acoustic Echo Control*, Røros, Norway, 1995.

[13] B.D. Van Veen, K.M. Buckley. Beamforming: A versatile approach to spatial filtering. *ASSP Mag.*, 5(2):4–24, 1988.

[14] M.M. Sondhi, D.R. Morgan, and J.L. Hall. Stereophonic echo cancellation: An overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8):148–151, 1995.

[15] N. Koizumi, S. Makino, and H. Oikawa. Acoustic echo canceller with multiple echo path. *JASJ - E*, 10(1):39–45, 1989.

[16] L.M. v.d. Kerkhoff and W.J.W. Kitzen. Tracking of a time-varying acoustic impulse response by an adaptive filter. *IEEE TR-SP*, 40(6):1285–1294, 1992.

[17] T. Chou. Frequency-independent beamformer with low response error. *Proc. ICASSP*, pp.2995–2998, Detroit, 1995.