JOINT DESIGN OF ACOUSTIC ECHO CANCELLATION AND ADAPTIVE BEAMFORMING FOR MICROPHONE ARRAYS

Walter Kellermann

Institute of Applied Research at the Fachhochschule Regensburg,
Hermann-Geib-Str.18, 93053 Regensburg, Germany
email: walter.kellermann@e-technik.fh-regensburg.de

ABSTRACT

For a recently proposed concept combining acoustic echo cancellation (AEC) and adaptive beamforming microphone arrays (ABMAs), crucial design and control issues are discussed. For ABMAs, data-independent and data-dependent beamforming algorithms are considered. While the actual signal processing of ABMA and AEC can be largely decoupled, efficient implementations benefit from control mechanisms overviewing the entire system. Key design parameters for typical microphone array applications are discussed.

1 INTRODUCTION

For hands-free communication it is sometimes desirable to employ a microphone array (MA) instead of a single microphone (SM) because a MA can direct a 'beam' of increased sensitivity to the desired source and suppress unwanted sources from other directions [1, 2, 3]. However, for full-duplex communication, MAs usually do not eliminate the need for AEC for three reasons [4]: First, due to the larger distance between local talkers and MA, the array input gain must be larger than for a SM positioned next to the talker, and the acoustic echo is amplified correspondingly. Second, the directivity gain of a MA is limited, especially for low frequencies and near-field conditions [2, 3, 5, 6, 7]. Third, null-steering to the loudspeaker for maximum echo attenuation is effective in nonreverberant environments [2] or for noise sources in the near-field [8], but generally not in reverberant environments [7].

A generic scheme for combining AEC and ABMA is outlined in Fig.1 [4]: The basic idea of the concept is the decomposition of the beamforming(BF) into a time-invariant beamforming and a time-variant voting stage. As detailed in [4], this decomposition is necessary to prevent the time-variance of the BF from obstructing the identification of the echo path: With a time-variant BF as part of the echo path, the AEC would have to find a new echo path model whenever

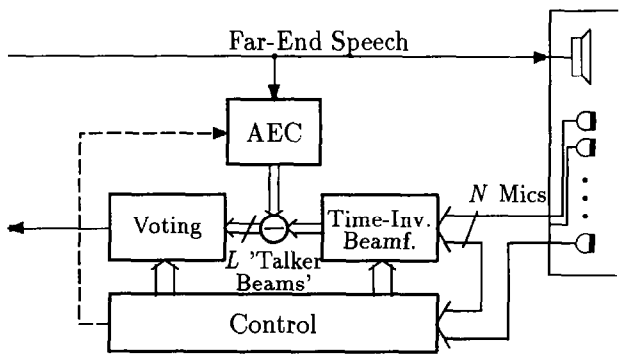

Fig.1: General structure combining ABMAs and AEC

the adaptive BF changes its parameters. Thus, no echo attenuation could be assured when it is needed most: during the transition from remote to local talker activity, or vice-versa, and during double-talk situations.

The time-invariant BF produces L so-called 'talker beams' from $N \geq L$ microphone signals, so that the AEC unit sees L acoustic echo paths which incorporate a time-invariant BF filtering each. Thus, the AEC corresponds simply to an L-fold SM echo cancellation (EC) problem. As a consequence, we can apply all AEC structures which are known for the SM case. This includes subband/frequency/transform-domain structures for modelling the echo path and the respective varieties of adaptation algorithms (see e.g. [9, 10]).

2 GENERIC STRUCTURES

For both data-dependent and data-independent BF algorithms [1] we present general structures for the combination of AEC and ABMAs and discuss their basic properties.

2.1 Beamsteering (BF-I) with AEC

For beamsteering, a set of M fixed beam signals is computed independently of the array input data ('data-independent BF' [1]), and the output of the beamformer is a weighted sum of these beams with time-

variant weights accounting for the active talkers (**voting**) [11, 12, 13][1]. BF-I inherently provides the desired separation into a time-invariant and a time-variant stage. To minimize the number of beams for the AEC, we introduce a mapping of the M fixed-beam signals onto the L talker beams whenever $L < M < N$ (Fig.2). While the fixed BF will be designed to cover all possible talker positions in the given environment, the mapping should only pass on beams which are used in the current session. For maximum spatial selectivity, the

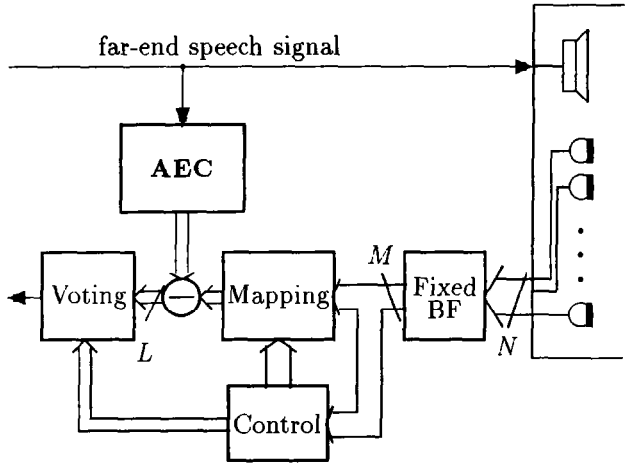

Figure 2: BF-I with AEC

mapping should select one fixed beam or a linear combination of two neighboring fixed beams per talker[2]. For details on initialization and constraints see [4].

2.2 Statistically optimum beamforming (BF-II) with AEC

Data-dependent adaptive BF methods aim at minimizing a statistical error criterion and filter the microphone signals accordingly (e.g. Generalized Sidelobe Canceler (GSC), Frost beamformer, c.f. [1]). Characteristically, the parameters of these systems are continually changing over time in order to converge to optimum filter coefficients. A constraint function assures that the desired signal is not cancelled [2, 5, 6]. If the current talker position is not known to the BF, the parameters of the constraint function (e.g. steering delays) have to be identified before an optimum beam can be formed.

For combining BF-II with AEC (Fig.2.2), we move the adaptive part of the data-dependent BF into a control path, where L optimum BF filters sets are continuously learnt, one for each talker. After major changes, the coefficients of the fixed filters in the signal path are updated accordingly. In parallel, L fixed sets of BF filters are simultaneously operating on the N microphone signals to produce the desired 'talker beam' signals. The L-channel AEC problem and the voting remains the same as with the BF-I structure. As with BF-I, the fixed BF for each of the L talkers

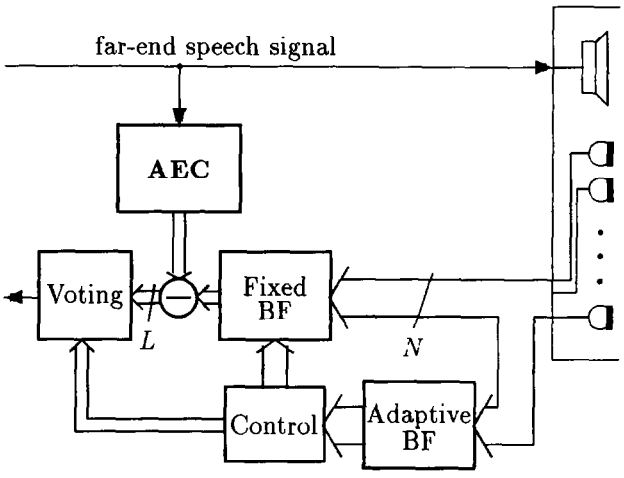

Figure 3: BF-II with AEC

must be initialized and should be updated only when the adaptive BF performs significantly better than the established fixed BF for the active talker. Initialization usually includes the localization of the desired sources before the adaptive BF algorithm converges to an efficient BF configuration for each talker (c.f. [5, 17]). Other aspects were already discussed in [4].

2.3 AEC

As the AEC is reduced to an L-channel system identification problem, the choice of the model structure and the corresponding adaptation algorithm is determined by the same factors as in the SM case: desired convergence speed and echo attenuation requirements must be balanced with constraints on group delay and hardware. The length of the impulse reponse, which has to be modeled for each talker beam, is determined by the acoustic echo path and the fixed BF filter. While the impulse response length for the acoustic path should be slightly shorter than in the SM case [4, 15], the impulse response length of the BF filter will lead to an increased total length of the adaptive filter: The BF filters will include steering delays, interpolation filtering for fractional delay realization, and possibly transform delays if the BF is performed in a transform domain (e.g. [16]). Considering that the steering delays simulate a physical movement of the array, the largest distances (of the outermost sensors) will still be short relative to the acoustic echo path lengths that have to be modeled. For interpolation usually not more than 8 samples are used [12]. In addition, the BF filter has to include frequency-selective filtering for BF-I systems with

[1] 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 [14] or as subarrays [11, 12, 7] before yielding a wideband beam signal. Fractional delay BF for increased spatial resolution is also covered by our model.

[2] If in some applications more talkers than beam signals may be active ($L \geq M$), e.g. in an auditorium, the mapping is omitted.

nested arrays, or the optimum BF filters for BF-II systems. Both kinds of filters are generally FIR filters of length $L_h \leq 128$ (e.g. [7, 16]) or low order IIR filters [12]. As a result, the impulse response for the fixed BF will still remain short relative to the acoustic path, as long as no high-resolution transform is used for BF in conjunction with a short acoustic path.

3 CONTROL MECHANISMS

While ABMAs and AEC are extensively researched areas on their own, we concentrate here on control mechanisms for their combination, elaborating some of the methods described in [4].

3.1 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 short-term and mid-term average power of loudspeaker and microphone signals, respectively [12, 13]. The spatial resolution of beamforming MAs provides additional information: The averaged power of the beam signals will show a typical pattern for each spatially fixed source such as a loudspeaker, which can then be distinguished from the patterns of other sources. For detecting double-talk situations, the spatial information proved more reliable than correlation techniques as commonly used for SM systems. Moreover, evaluating the envelope of the beam signals over time, high-level noise sources can be spatially separated from speech-like signals as long as they produce stationary envelopes.

3.2 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, while at the same time avoiding the perception of switching noise [12]. Typically, a sigmoid-like gain function provides a fast and smooth attack, whereas the time constant for releasing a beam must be considerably longer. For maximum spatial selectivity, only one beam signal per talker should have a nonzero weight in the stationary case (for details see, e.g. [12]). The weights will in practice also incorporate other gain factors as determined by loss insertion algorithms and automatic gain control (AGC) devices.

When entering a 'far-end talk only' period, voting strategies between two extremes will be used: Either one stays with the most recently active local talker beam or the weights are gradually changed to arrive at an average over all L talker beams. The latter provides the often desired uniform coverage of the entire local environment to the remote party. Weighting strategies which exploit varying echo attenuation over different talker beams to maximize overall echo attenuation subject to a constant gain constraint, may give a disturbing spatial impression to the remote end party. Instead, we propose to simply ensure the required echo attenuation by inserting uniform loss for all beams with nonzero weights.

3.3 AEC

Let us recall that AEC algorithms basically consist of a filtering part for producing echo estimates and an adaptation part to identify the current echo path, both of which imply considerable computational load. If computational cost is not an issue, all L ECs should produce echo estimates and be adapted in parallel during all 'far-end talk only' periods. For demanding adaptation algorithms like RLS [10] and/or with limited resources, alternating or reduced-rate adaptation of the L adaptive filters (or some of them) may be required to save computational load. Larger savings are obtained if the voting information is taken into account: Echo estimates are only needed for talker beams with nonzero weight. (However, if the echo path of currently unused beams should always be identified for later use, filtering is still a prerequisite.) For minimum computational cost for AEC, one will select only the dominant talker beam in the voting stage and correspondingly operate only one adaptive filter of the AEC unit at a time. Smooth transitions of the voting from one talker beam to another can still be obtained, but require extra loss insertion: Instead of producing echo estimates for several talker beams, all talker beams except for the dominant are simply attenuated during the weight transition in order to ensure a prescribed amount of echo attenuation. As for 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).

4 DESIGN EXAMPLES

For efficient designs minimum computational complexity has to be combined with optimum performance. This means here that a minimum number of talker beams should provide perfect coverage of the local sources, and that the attenuation of the acoustic echo should be achieved at minimum computational cost by the AEC with a minimum of loss insertion.

For **car telephony**, MAs using GSC with typically 7 to 10 sensors [5, 17], 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 AEC are essentially the same as for a SM, as long as only a single adaptive filter is operated

at any time. Although the directivity gain of the array is not completely outweighed by the increased average microphone distance relative to an optimally placed SM, the incorporation of the BF into the echo path model leads to an AEC impulse response of comparable length as for a SM.

For **desktop teleconferencing**, MAs compete with multi-channel systems, offering the advantage of requiring less sensors when large groups communicate. Both BF concepts have already been applied (c.f. [13] for BF-I or [2] for BF-II) with a small number of sensors ($N = 2 \ldots 4$) so that no mapping for BF-I is needed. If $N \leq L$ as suggested in [13], the AEC unit should act directly on the microphone outputs. 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 the AEC at a reduced rate. As high quality is required for the AEC, the filtering should be performed for all beams with nonzero weights.

For **videoconferencing**, MAs mounted to a wall or to the ceiling again compete with multi-channel systems (see, e.g. [18]). With nested beamsteering subarrays (BF-I) using $N = 15, \ldots, 25$ microphones [12, 3, 7] up to $M = 7$ beams are formed, which cover typically $L = 2 \ldots 5$ talkers. In [4] we showed that the requirements for each AEC channel will be at least as complex as for an individual microphone given to each participant, and the quality demands will necessitate continuous adaptation of the AEC for each talker beam.

For an **auditorium** as described in [11] using a planar array (BF-I, $N = 380$, $L = M = 27$), the AEC problem is scaled up along three parameters compared to a teleconferencing studio: increased reverberation time demands longer impulse responses, increased talker-array distance provides extra echo gain demanding even longer impulse responses, and the large L requires more adaptive filters. Voting-selected filtering and reduced-rate adaptation are needed to keep the computational load within realizable dimensions. Nevertheless, loudspeaker directivity and room design will remain of great importance for this application, if loss insertion is to be minimized.

REFERENCES

[1] B.D. Van Veen and K.M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2):4-24, April 1988.

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

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

[4] W. Kellermann. Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. ICASSP 97, pp.219–222, Munich, Germany, April 1997.

[5] S. Oh, V. Viswanathan, and P. Papamichalis. Hands-free voice communication in an automobile with a microphone array. ICASSP 92, pp.I-281 - I-284, San Francisco, CA, USA, March 1992.

[6] K. Farrell, R.J. Mammone, and J.L. Flanagan. Beamforming microphone arrays for speech enhancement. ICASSP 92, pp.I-285 - I-288, San Francisco, CA, USA, March 1992.

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

[8] P. Chu. Superdirective microphone array for a set-top videoconferencing system. ICASSP 97, pp.235–238, Munich, Germany, April 1997.

[9] 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, Inc., 1991.

[10] E. Hänsler. The hands-free telephone problem - an annotated bibliography, September 1993. 3rd Int. Workshop on Acoustic Echo Control, Lannion, France.

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

[12] W. Kellermann. A self-steering digital microphone array. ICASSP 91, pp. 3581–3584, Toronto, Canada, May 1991.

[13] P. Chu. Desktop mic array for teleconferencing. ICASSP 95, pp. 2999-3002, Detroit, MI, May 1995.

[14] T. Chou. Frequency-independent beamformer with low response error. ICASSP 95, pp. 2995-2998, Detroit, MI, May 1995.

[15] W. Kellermann. Some properties of echo path impulse responses of microphone arrays and consequences for acoustic echo cancellation. In *Conf. Rec. 4th IWAENC, Røros, Norway, June 1995.*

[16] I. Claesson, S.E. Nordholm, B.A. Bengtsson, and P. Eriksson. A multi-DSP implementation of a broadband adptive beamformer for use in a hands-free mobile radio telephone. *IEEE TR-VT*, 40(1):194–202, February 1991.

[17] M. Dahl, I. Claesson, and S. Nordebo. Simultaneous echo cancellation and car noise suppression employing a microphone array. ICASSP 97, pp.239–242, Munich, Germany, April 1997.

[18] N. Koizumi, S. Makino, and H. Oikawa. Acoustic echo canceller with multiple echo path. *Journal Acoustical Society of Japan - English*, 10(1):39–45, 1989.