Proceedings of ISSSE-92, Paris, 1-4 September 1992

TIME DELAY COMPENSATION FOR ADAPTIVE MULTICHANNEL SPEECH ENHANCEMENT SYSTEMS

K. U. Simmer, P. Kuczynski, A. Wasiljeff

University of Bremen,

Department of Physics and Electrical Engineering,

P.O. Box 330 440, D-2800 Bremen 33, FRG

ABSTRACT

Several algorithms for adaptive multichannel speech enhancement have been tested in an office room and in an anechoic chamber using different noise sources. Temporal synchronisation requires time delay estimation and compensation of the desired speech signal received at different sensors.

1. INTRODUCTION

As most digital speech processing systems have been designed for noise-free environments, the presence of background noise can seriously degrade their performance. In this paper we discuss algorithms for adaptive noise reduction of speech signals using multichannel systems with several acoustic sensors. Non - stationarity of the speech signal and the low spatial coherence of the received noise signals create problems as well as reverberation and echoes that predominate over direct path noise signals in a typical office environment.

2. SYSTEM DESCRIPTION

The receiving system consists of a planar array of four omnidirectional microphones placed at the corners of a square of 60×60 cm. The desired signals are produced by different persons. Either white random noise emitted from a loudspeaker or a hair dryer are used as noise sources. The signals are digitized by a DSP32C signal processor board using a sampling rate of 16 kHz. They pass a beamstearing unit with four delays. Automatic adjustment of these delays is part of the following study.

3. ALGORITHMS FOR ADAPTIVE MULTICHANNEL PROCESSING OF NOISY SPEECH SIGNALS

Algorithms due to Frost [1], Griffiths-Jim [2], Duvall [3] and Compernolle [4] have been investigated in this paper. The noise suppression filter of Kaneda and Tohyama [5] has been generalized to N sensors by the authors.

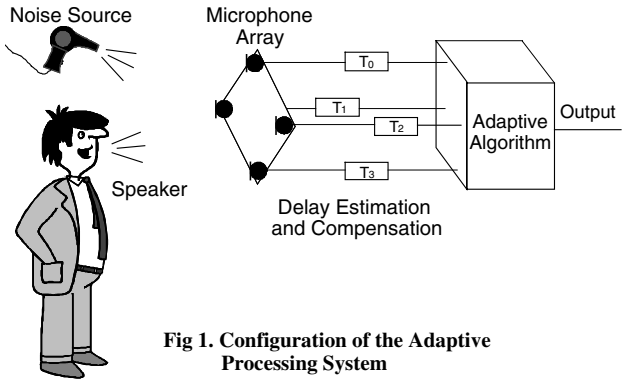

Fig 1. Configuration of the Adaptive Processing System

The algorithm is based on the following assumptions. The speaker has to be close to the sensor array, the distance d between two sensor elements has to obey the spatial sampling theorem $d \le \lambda_{pitch} / 2$ with regard to the pitch-wavelength $\lambda_{pitch} = c_{sound} / f_{pitch}$, desired speech signal and noise signals are assumed to be uncorrelated, the noise signals are spatially incoherent, i.e. the correlation of the noise signals at two different sensors is close to zero. Cross-spectra $X_i \cdot X_j^*$ of the input signals on sensor pairs of the array are used to estimate the transfer function:

$$W(f) = \frac{\frac{2}{N \cdot (N-1)} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \mathrm{Re}\{X_i(f) \cdot X_j^*(f)\}}{\frac{1}{N} \sum_{i=0}^{N-1} |X_i(f)|^2} \quad (1)$$

75% overlapping Hanning windows (256 samples) are used for short-time FFT of the signals. The short-time spectra of the N channels are averaged and multiplied by the transfer function (1). Inverse transformations and overlap and add technique yield the filtered output.

4. TIME DELAY ESTIMATION

An important problem in practical systems is the temporal synchronisation of the desired speech signals. The signals received by N sensors pass a beam steering unit (see fig. 1). The delays T_i of the unit should be automatically adjusted to steer the beam of the array into the direction of the speaker. Time delay estimation of the received signals has to be performed under difficult conditions. The signal source is subject to irregular motions, as the speaker moves his head. In time intervals with low signal to noise ratio as unvoiced speech or speech-pauses the array may look into a wrong direction if no counter-measures are taken. The basic idea proposed in this paper is a switched beam steering unit. The position of the speaker is determined by delay estimation between signal arrivals at different sensors during intervals with maximum signal to noise ratio. The steering direction of the array is hold during intervals with low signal to noise ratio. The input signals can be modelled as

$$\begin{aligned} x_i(t) &= s(t) + n_i(t) \\ x_j(t) &= s(t + D_{ij}) + n_j(t) \end{aligned} \qquad (2)$$

where the speech signal s(t) is assumed to be uncorrelated with the spatially incoherent noise sources $n_i(t)$ and $n_j(t)$. To determine the time delay D_{ij} we compute the cross-correlation function:

$$\begin{aligned} R_{x_i x_j}(\tau) &= E\,[x_i(t)\, x_j(t - \tau)] \\ &= E\,[(s(t) + n_i(t)) \cdot (s(t + D_{ij} - \tau) + n_j(t - \tau))] \\ &= E\,[s(t) \cdot s(t - (\tau - D_{ij}))] \\ &= R_{ss}(\tau - D_{ij}) \end{aligned} \qquad (3)$$

since

$$\begin{aligned} E\,[n_i(t) \cdot n_j(t - \tau)] &= E\,[n_i(t) \cdot s(t + D_{ij} - \tau)] = E\,[s(t) \cdot n_j(t - \tau)] \\ &= 0 \end{aligned} \qquad (4)$$

To estimate the time-delay D_{ij} it is necessary to find the argument τ, that maximizes the value of (3). The maximum value of (3) can also be interpreted as the mean power of the desired speech signal. To distinguish between speech intervals with maximum SNR and speech-pauses with a low SNR the temporal coherence function is used:

$$\gamma_{x_i x_j}(\tau) = \frac{R_{x_i x_j}(\tau)}{\sqrt{R_{x_i x_i}(0) \cdot R_{x_j x_j}(0)}} \qquad (5)$$

Same noise power at the different channels gives:

$$R_{x_i x_i}(0) = R_{x_j x_j}(0) = R_{ss}(0) + R_{nn}(0) \qquad (6)$$

Therefore it is possible to estimate the signal to noise ratio by computing the coherence function at $\tau = D_{ij}$:

$$\gamma_{x_i x_j}(D_{ij}) = \frac{R_{ss}(0)}{R_{ss}(0) + R_{nn}(0)} = \frac{1}{1 + \dfrac{1}{R_{ss}(0) \,/\, R_{nn}(0)}} \qquad (7)$$

The temporal coherence function (5) is used to estimate the time delay D_{ij} during the speech-sequence. The maximum magnitude of the coherence function as given by (7) determines whether a speech signal is present or not. Formula (7) shows that a coherence function close to one corresponds to a high signal to noise ratio and yields reliable estimates of the time delays D_{ij}. The proposed coherence detector works as follows:
The time delays D_{ij} are estimated only if the maximum of the coherence function (7) exceeds a preset threshold level (e.g. 0.6). The steering direction of the array is hold during intervals with low SNR.

5. RESULTS

Figure 2a shows a three dimensional plot of a short time coherence function of a segmented speech signal, with 1.4 sec duration. Each segment has a length of 1024 samples, which is equivalent to 1/16 second at a sampling rate of 16kHz. The segments are windowed using 75% overlapping Hanning windows. The speaker is moving from right to left as it is clearly seen in the contour plot of figure 2b. In the case where the speaker does not move and stays at a fixed position, the maximum of the coherence function remains at the same delay (12 samples), as is shown in figure 3a and b. Regions with low coherence maximum occur at time intervalls with unvoiced speech. The delay and hence the direction of the speaker cannot be determined in these coherence valleys. Figure 4 shows estimated delays for a single word of 1 sec duration as a function of time. The upper figure 4a shows delay estimation without coherence detector. Figure 4b shows the excellent performance of the proposed coherence detector for a threshold level of 0.6. Figure 4c gives the corresponding maximum of the coherence function. Figure 5 gives the standard deviation of the delay estimation error of the described coherence detector level for different algorithms. Above a coherence threshold of 0.6 the simple short time coherence function as defined in (5) gives the lowest delay estimation error of all algorithms investigated in this paper. Figure 6 shows the noise reduction as defined in [10] as function of frequency for different microphone distances d for Frost´s algorithm and figure 7 for the algorithm proposed by the authors. A large microphone distance d gives better noise reduction for lower frequencies. However one should keep in mind that violating the spatial sampling condition with regard to the pitch wavelength as mentioned in chapter 3 leads to ambiguities in the measured delays. Although all investigated algorithms work rather well under the artificial conditions of an anechoic chamber (fig. 8) the proposed algorithm (1) seems to give better performance in an office environment (fig. 9).

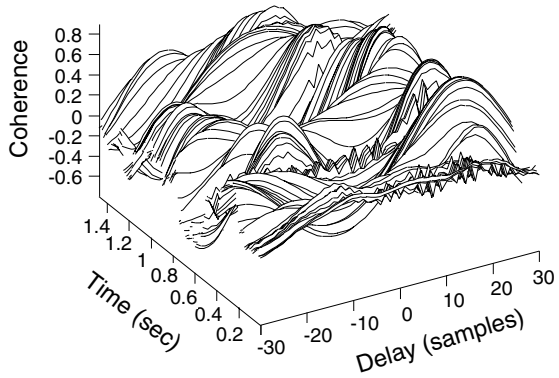

Fig. 2a Short time coherence of segmented speech signal with a moving speaker.

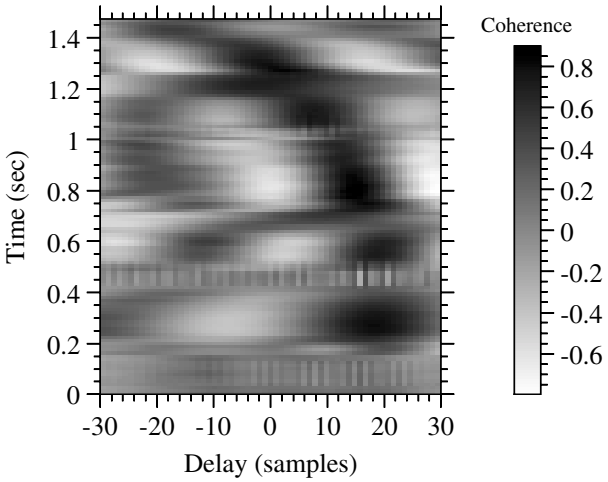

Fig. 2b Grey scale contour plot of **2a**.

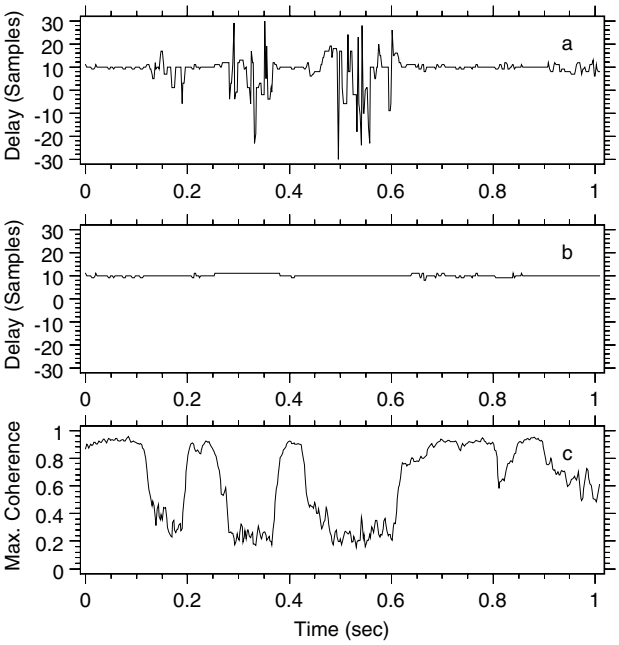

Fig. 4 Delay estimation with corresponding coherence. (True delay D = 10).
a) Delay estimation without coherence detector.
b) Delay estimation with coherence detector.
c) Maximum of the coherence function.

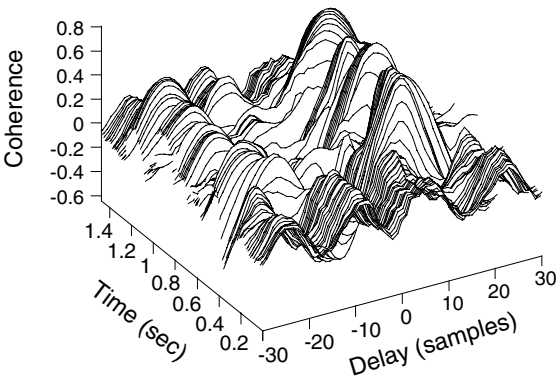

Fig. 3a Short time coherence of a segmented speech signal. Speaker at a fixed position.

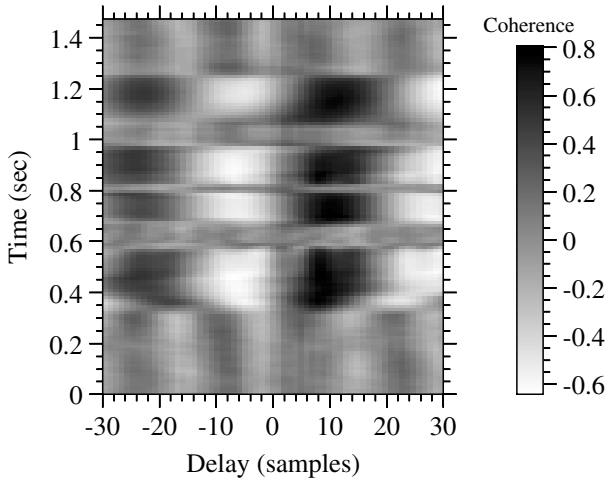

Fig. 3b Grey scale contour plot of **3a**.

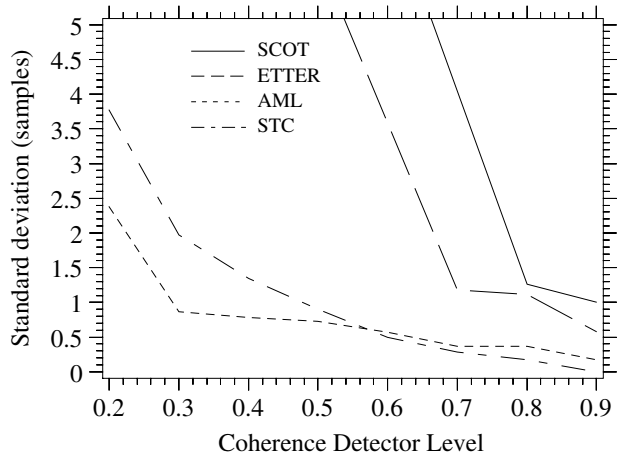

Fig. 5 Standard deviation of time delay estimates as function of the described coherence detector level. (Single spoken word, average SNR 6dB).

SCOT: Smoothed Coherence Transform. Estimated using 4 disjoint data segments of 512 samples [6],[7].
ETTER: Adaptive Estimation of Time Delay [8].
AML: Approximate Maximum Likelihood Estimator [6]. Estimated using 8 data segments of 256 samples
STC: Short Time Coherence as defined in (5) using one Hanning window of 2048 samples.

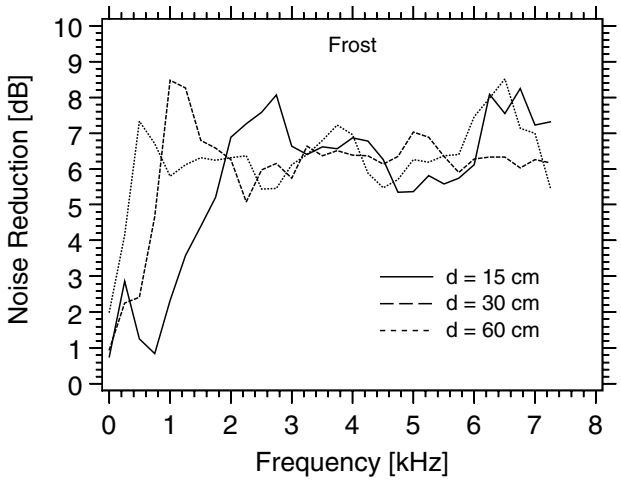

Fig. 6 Noise reduction as function of frequency, different microphone distances d as parameter. (Frost´s algorithm).

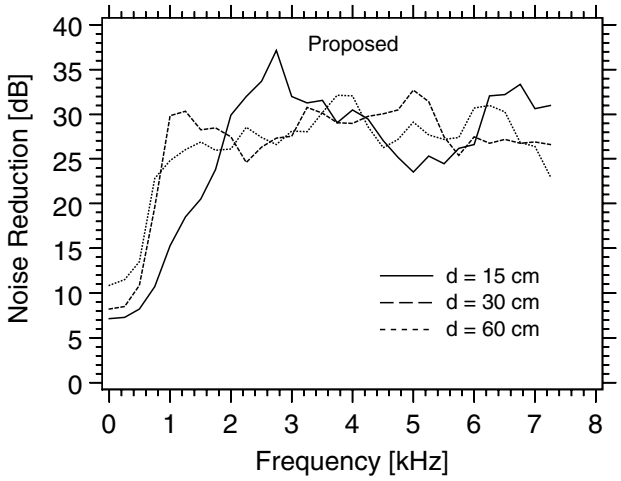

Fig. 7 Noise reduction as function of frequency, different microphone distances d as parameter. (Algorithm proposed by the authors).

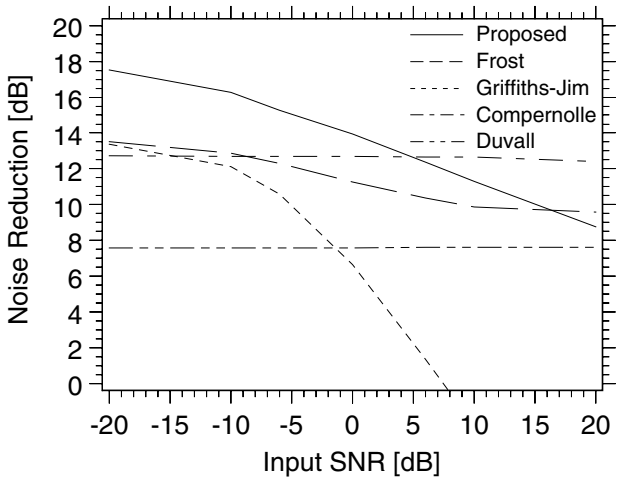

Fig. 8 Noise reduction as function of SNR in an anechoic chamber.

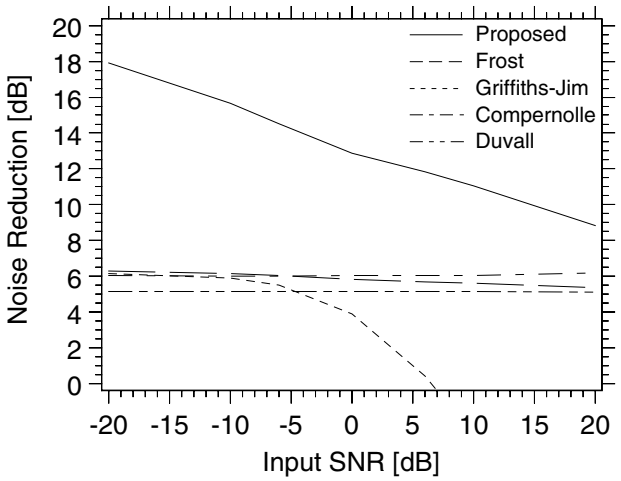

Fig. 9 Noise reduction as function of SNR in an office room.

References

[1] O.L. Frost, III, "An Algorithm for Linearly- Constrained Adaptive Array Processing," Proc. IEEE, vol. 60, no.8, pp. 926-935, Aug. 1972.

[2] L. J. Griffiths, C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming" IEEE Trans. Antennas Propagat. vol. 30, pp. 27-34, Jan. 1982.

[3] B. Widrow et al., "Signal Cancellation Phenomena in Adaptive Antennas: Causes and Cures," IEEE Trans. Antennas Propag., vol AP-30, p. 469, May 1982.

[4] D. Van Compernolle: "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP-90, Albuquerque, pp. 833-836.

[5] Y. Kaneda, M. Tohyama, "Noise Suppression Signal Processing Using 2-Point Received Signals", Electronics and Communications in Japan, Vol.67-A, pp.19-28, Apr. 1984.

[6] C. H. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 320-327, Aug. 1976.

[7] K. Scarbrough, N. Ahmed, G. C. Carter, "On the Simulation of a Class of Time Delay Estimation Algorithms", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, pp. 534-540, June 1981.

[8] D. M. Etter, S. D. Stearns: "Adaptive Estimation of Time Delay in Sampled Data Systems", IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 582-587, June 1981.

[9] R. Zelinski: "Adaptive Einstellung der Signallaufzeit für ein Geräuschunterdrückungssystem mit Mikrofon-gruppe", ITG - Fachbericht 107, pp.307 - 312, 1989.

[10] K. U. Simmer, A. Wasiljeff: "Analysis and Comparison of Systems for Adaptive Array Processing of Speech Signals in a Noisy Environment", Treizième Colloque GRETSI, Juan-Les-Pins, pp. 529 - 532, 1991.