SPEAKER LOCALIZATION AND ITS APPLICATION TO TIME DELAY ESTIMATORS FOR MULTI-MICROPHONE SPEECH ENHANCEMENT SYSTEMS

Martin Drews

Institut für Fernmeldetechnik, Technische Universität Berlin
Einsteinufer 25, D-10587 Berlin, Germany
phone: +49 30 31424573 fax: +49 30 31425799
e-mail: drews@ftsu00.ee.tu-berlin.de

ABSTRACT

A time delay estimator for a multi-microphone speech enhancement system with 16 microphones is presented. It is based on a generalized cross-correlator and an improved peak detector. The problems associated with delay estimation in noisy speech signals are solved by performing a speaker localization and a plausibility check of the time delays derived from the speaker position. By applying these techniques to the time delay estimator, a significant reduction of the computational load is achieved, and the TDOA estimation errors are reduced.

1 INTRODUCTION

When using hands-free speech communication systems indoors, the speech signal acquisition is usually corrupted by reverberation and additional background noise. Many techniques which are efficient at enhancing noisy speech make use of more than one microphone for speech input. A well known class of speech enhancement techniques is based on a conventional delay-and-sum beamformer [6] extended by an adaptive post-filter.

The delay-and-sum beamformer estimates the time differences of arrival (TDOAs) between the speech signals received by the microphones, compensates for these TDOAs, and sums the resulting signals. The summation of the delay-compensated input signals causes an attenuation of the uncorrelated components while the correlated components are retained. In detail, with M microphones the noise power is reduced by the factor of $10 \cdot \lg M$ dB, provided that the received background signals are uncorrelated. After beamforming, the adaptive post-filter achieves further noise power reduction, but it also affects the speech quality since "musical" distortions remain in the speech signal after filtering.

For high quality speech enhancement, a technique is required which provides a sufficiently high noise power reduction together with a very low impairment of the speech quality. To meet these requirements in a straightforward implementation, the speech enhancement is done without any filtering and the delay-and-sum beamformer is driven with a large number of microphones (here: $M = 16$). This beamformer is illustrated in fig. 1.

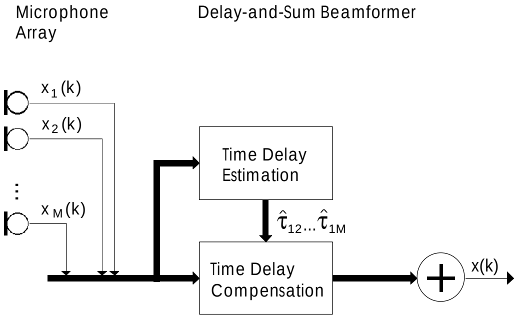

Fig. 1: Structure of the speech enhancement system

The main problems in using delay-and-sum beamformers for speech enhancement are associated with the time delay estimator. For time delay compensation, M-1 TDOA estimates are necessary. However, the estimation of all TDOAs will be computationally very expensive if the number of microphones is large. Due to TDOA estimation errors, the speech signals are out of phase after delay compensation, and their mutual correlation decreases, especially at high frequencies. Consequently, the summation of the delay-compensated signals results in a noise-reduced speech signal which is more or less attenuated towards high frequencies. For this reason, the output speech quality of the delay-and-sum beamformer depends on the exactness of the TDOA estimation.

The present work has focused on time delay estimators which show only small errors in order to utilize delay-and-sum beamformers for high quality speech enhancement. Since the microphone array used for speech acquisition consists of more than 4 microphones, two channel selectors and a speaker localization method are added to the time delay estimator. By performing speaker localization with less than M-1 TDOA estimates and subsequent calculation of all TDOAs, the computational load as well as the TDOA estimation errors are reduced.

The work presented in this paper was funded by Deutsche Telekom.

2 SPEAKER LOCALIZATION

The time delay estimator required for delay-and-sum beamforming has to provide *M*-1 TDOA estimates. However, by evaluating more than three TDOA estimates, speaker localization results in a speaker position estimate from which all *M*-1 TDOA estimates are derived. Consequently, the time delay estimator achieves a computationally less expensive implementation using only *D* microphones for TDOA analysis ($4 \leq D \leq M$, here: $D = 8$) and subsequently performing speaker localization.

2.1 Localization Problem

Arranging the microphone array used here, *M* microphones are placed in a plane with positions $\{x_i, y_i, 0\}$. At the location of the reference microphone, the origin of a Cartesian coordinate system is established. The speaker is assumed to be a point source whose location is denoted by the spatial coordinates $\{x_S, y_S, z_S\}$. Accordingly, the distances r_i between the speaker and the *i*-th microphone are given by:

$$r_i^2 = \left(x_S - x_i\right)^2 + \left(y_S - y_i\right)^2 + z_S^2 \quad \text{for } i = 1 \ldots M \qquad (1)$$

Due to speaker movement, the speaker position and consequently the distances to the microphones are time-varying. This geometry is sketched in fig. 2.

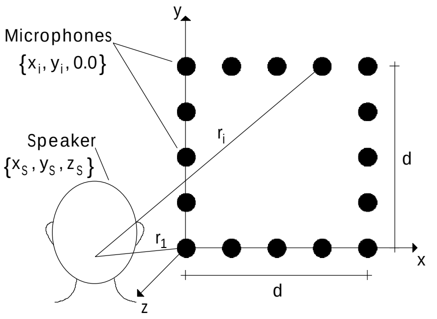

Fig. 2: Speaker-microphone geometry with $M = 16$ microphones (here: $d = 0.4$ m)

For speaker localization, those TDOAs are available which correspond to the reference microphone and the *i*-th microphone. The true TDOAs τ_{1i} are related to the difference r_{1i} between the two distances r_1 and r_i according to the following equation:

$$r_{1i} = \frac{c}{f_T} \cdot \tau_{1i} = r_1 - r_i \quad \text{for } i = 2 \ldots D \qquad (2)$$

where f_T is the sampling rate, *c* is the speed of sound propagation, and τ_{1i} is expressed in sample periods. As the TDOAs are assumed to be corrupted by additive noise, only estimates of the TDOAs are available. Furthermore, the TDOA estimation errors differ significantly depending on the relevant microphone. This characteristic is clearly visible from the probability density function of the TDOA estimation errors depicted in fig. 3.

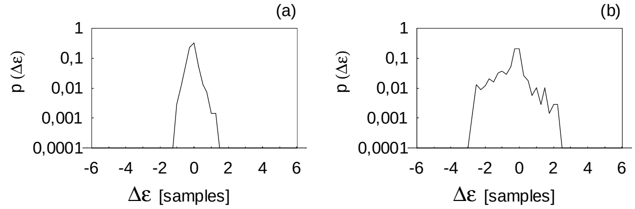

Fig. 3: Probability density function $p(\Delta\varepsilon)$ of the TDOA estimation errors measured only during speech activity at a signal-to-noise ratio of 0 dB: (a) microphone no. 5 and (b) microphone no. 7

When using more than three TDOA estimates, (1) and (2) formulate an over-determined set of non-linear equations whose solution results in a speaker position estimate. The Maximum Likelihood (ML) estimate of the speaker position is then found by the least-squares estimate [2] of the TDOAs $\hat{\tau}_{S,1i}$ calculated from the speaker position estimate and the true TDOAs:

$$\mathbf{p} = \min_{\mathbf{p}} \left[\sum_{i=2}^{L} \frac{1}{\sigma_i^2} \cdot \left(\tau_{1i} - \hat{\tau}_{S,1i}\right)^2 \right] \qquad (3)$$

where **p** denotes the vector of the speaker position coordinates $\{x_S, y_S, z_S\}$. If the speaker position estimate is derived using the non-linear equation (1), however, the solution of the ML error criterion (3) will require numerical search methods which are rather problematic in this application. For this reason, closed-form solutions are developed which show performance features comparable to those of the ML estimator [2, 4, 5].

2.2 Closed-Form Speaker Position Estimation

For solving the localization problem addressed here, the closed-form solution according to [4] is favoured which approximates ML estimation by hyperbolic location estimation. This closed-form solution is found by linearizing the set of non-linear equations assembled from (1) and (2). In detail, linearization is accomplished by: (i) squaring (2), (ii) inserting the result of the squared distance r_i^2 in (1), and (iii) subtracting the resulting equation from (1) at $i = 1$. After these steps, a set of linear equations is obtained which is expressed by the matrix equation:

$$\mathbf{G} \cdot \mathbf{p_a} - \mathbf{h} = \mathbf{0} \qquad (4)$$

where

$$\mathbf{p_a} = \begin{bmatrix} x_S \\ y_S \\ r_1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} x_1 - x_2 & y_1 - y_2 & r_{12} \\ \vdots & \vdots & \vdots \\ x_1 - x_D & y_1 - y_D & r_{1D} \end{bmatrix}$$

$$\mathbf{h} = \frac{1}{2} \cdot \begin{bmatrix} r_{12}^2 + \left(x_1^2 - x_2^2\right) + \left(y_1^2 - y_2^2\right) \\ \vdots \\ r_{1D}^2 + \left(x_1^2 - x_D^2\right) + \left(y_1^2 - y_D^2\right) \end{bmatrix} \qquad (5)$$

This set of linear equations will be over-determined if more than three TDOA estimates are involved, and linear independent due to TDOA estimation errors. As a disadvantage of linearization, the z-coordinate of the speaker position is cancelled. Solving the matrix equation (4) results in the desired speaker position. Assuming similar distances r_i, the closed-form solution is approximated by the ML estimate of $\mathbf{p_a}$ [4]:

$$\mathbf{p_a} = \left(\mathbf{G}^{\mathrm{T}} \cdot \mathbf{Q}^{-1} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathrm{T}} \cdot \mathbf{Q}^{-1} \cdot \mathbf{h} \qquad (6)$$

For simplification, the covariance matrix $\mathbf{Q}$ of the TDOA estimates is set to:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & \cdots & 1 \end{bmatrix} \qquad (7)$$

Finally, the remaining coordinate of the speaker position z_S is calculated from the equation:

$$z_S = \pm\sqrt{r_1^2 - x_S^2 - y_S^2} \qquad (8)$$

Considering the different error levels of each TDOA estimate, selecting those TDOA estimates which show small errors may improve speaker localization. Since the TDOA estimation errors are unknown in practice, the set of D-1 linear equations is reduced to several over-determined subsets, each containing less than D-1 but more than 3 equations. Due to TDOA estimation errors, all these subsets result in several different positions. After selecting and averaging those positions which are closest to each other, a more reliable speaker position is obtained. In detail, from 7 positions, derived from 7 sets of 6 equations, four positions are selected and averaged to form the final speaker position.

2 TIME DELAY ESTIMATION

The basic approach to estimating TDOAs is to cross-correlate the relevant two signals. For this process, usually a generalized cross-correlator [3] is applied. A subsequent peak detector provides an estimate for the TDOA by evaluating the obtained cross-correlation function.

The TDOA estimates are subject to anomalous and normal errors. Moreover, TDOA estimates which are measured during speech pauses or periods with dominating noise signals are non-plausible. Anomalous errors result from ambiguous peaks which occur when early echoes of the speech signal [1], periodic speech components [5], or background speech components cause an periodical nature of the cross-correlation function. In all other cases of speech activity with sufficient high signal-to-noise ratios, the TDOAs show normal errors, as the true cross-correlation peak is now only slightly displaced by noise components.

Taking into account the TDOA estimation errors mentioned above, a time delay estimator has been developed. The structure is shown in fig. 4.

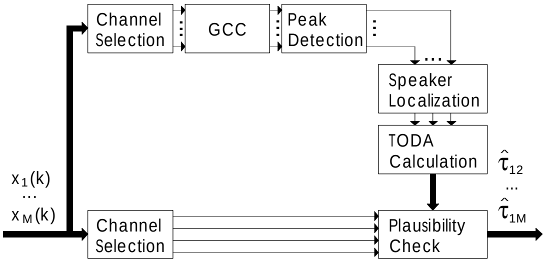

Fig. 4: Structure of the time delay estimator

The channel selector placed in front of the generalized cross-correlator (GCC) selects those D input signals which are received by the most distant microphones of the array. With the generalized cross-correlator and a subsequent peak detector, a separate analysis for each TDOA is performed. The resulting D-1 TDOAs are used for speaker localization. The final M-1 TDOA estimates are obtained after applying the plausibility check to those TDOAs which are directly calculated from the speaker position estimate. Since the final TDOA estimates are derived from a speaker position that best fits the analysed TDOAs in the sense of Maximum Likelihood, normal errors are reduced.

3.1 Generalized Cross-Correlator (GCC)

The generalized cross-correlator calculates the cross-correlation function in the frequency domain by performing adaptive pre-correlation filtering, cross spectral density measurement, and inverse Fourier transform. The adaptive pre-correlation filter modifies the Fourier-transformed input signals in order to reduce normal errors which are caused by the noise influence on the cross-correlation function. It is adapted according to the following equations:

$$H(n) = \begin{cases} 1 & \text{if} \quad E(n) < \dfrac{1}{N}\sum\limits_{n=0}^{N} E(n) \\ 0 & \text{else} \end{cases}$$

$$E(n) = \frac{\sum\limits_{i=0,\ j=i+1}^{D} \left[A_{ii}(n) - A_{jj}(n)\right]^2}{\left[\sum\limits_{i=0}^{D} A_{ii}(n)\right]^2} \qquad (9)$$

where A_{ii},(n) is the power spectral density of the i-th input signal and N is the number of spectral lines. With this frequency-selective filter, only those frequencies are retained for cross-spectral density measurement which show small mean square differences of the power spectral densities, thus indicating a high level of correlated speech signal components and a relatively low noise influence.

3.2 Peak Detector

The peak detector searches for the dominant peak of the cross-correlation function and interprets its location as an estimate for the actual TDOA. The peak detector used here is the same as proposed in [5]. It avoids the occurrence of anomalous errors by limiting the peak search region during speech activity and short speech pauses. In order to reduce normal errors, the resolution of the cross-correlation function is improved by means of interpolation.

3.3 Plausibility Check

The plausibility check applied here is similar to the one presented in [5]. It is based on speaker position evaluation and coherence measurement. With this plausibility check, non-plausible TDOAs are detected and are subsequently replaced by the TDOA estimates accepted previously.

For an improved plausibility check, the selecting-and-averaging procedure performed during speaker localization is carried out four times in order to obtain four different speaker positions. By using these position estimates, the subsequent coherence measurement results in four different coherence values. In this process, a four-dimensional cross-correlation function is determined from selected input signals and evaluated at the location of the relevant TDOA estimates derived from each speaker position estimate using (1) and (2). The channel selector placed in front of the plausibility check (fig. 4) selects those four input signals which are acquired by the microphones situated at the corners of the array.

The plausibility check is performed through the following decision process: if all of the speaker positions are real-valued *and* if the maximum of all coherence values exceeds the preset threshold level of 0.25, the TDOA estimates are regarded as plausible and the relevant M-1 TDOAs are taken as the final estimates $\hat{\tau}_{S,1i}$. In case of complex-valued speaker positions or too low coherence values the TDOA estimates are found to be non-plausible and the TDOAs previously accepted are retained. The decision process is biased towards non-plausibility for higher reliability.

4 RESULTS AND CONCLUSION

For comparative purposes, the time delay estimator presented in [5] was extended to provide M-1 TDOAs. This time delay estimator, treated here as reference, analyses for only D-1 = 3 TDOAs, evaluates them by performing a less accurate speaker localization and virtually the same plausibility check, and finally derives the required M-1 TDOAs from the speaker position obtained.

The reference time delay estimator was compared to the one presented here by evaluating the overall mean square TDOA estimation error $\Delta\tau$ introduced in [5]. For this purpose, a series of simulations were carried out using different noisy speech signals at various input SNRs. The resulting TDOA estimation errors are summarized in fig. 5. It is clearly shown that the proposed time delay estimator yields some improvement with respect to the estimation error. Compared to standard time delay estimators which do not include speaker localization nor plausibility check, the TDOA estimation errors of both time delay estimators are reduced by a factor of up to six [5].

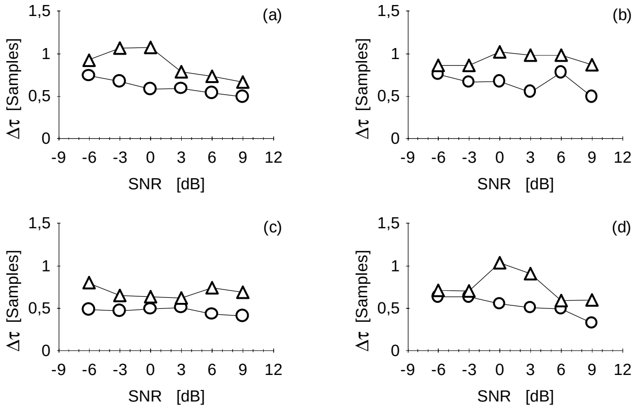

Fig. 5: Mean square TDOA estimation error $\Delta\tau$ of the proposed time delay estimator (—○—) and the reference time delay estimator (—△—): (a) speech with a longer silence period, decorrelated background noise, (b) speech with periodical parts, decorrelated background noise, (c) speech with a longer silence period, background speech, (d) speech with periodical parts, background speech

A time delay estimator including a speaker localization method and a plausibility check has been presented. It is an efficient, robust and accurate estimator when applied to a delay-and-sum beamformer driven with 16 microphones. Moreover, it requires significantly less computation power compared with estimation of all TDOAs since only 8 of the 16 input signals are involved in TDOA analysis, and the speaker localization method is computationally not expensive. However, the reference time delay estimator achieves further reduction of the computational load as only 4 input signals are analysed for TDOAs, but it shows less accurate TDOA estimates. As a final conclusion, the proposed time delay estimator enables delay-and-sum beamformers to attain speech enhancement with an improved output speech quality.

REFERENCES

[1] S. Bédard, B. Champagne, A. Stéphenne: Effects of room reverberation on time delay estimation performance; Proc. ICASSP 1994, pp. 2261-2264

[2] M. S. Brandstein, J. E. Adcock, H. F. Silverman: A closed-form method for finding source locations from microphone-array time-delay estimates; Proc. ICASSP 1995, pp. 3019-3022

[3] G. C. Carter: Coherence and time delay estimation; Proc. IEEE, vol. 73, Feb. 1987, pp. 236-255

[4] Y. T. Chan, K. C. Ho: A simple and efficient estimator for hyperbolic location; IEEE Trans. on Signal Processing, vol. 42, no. 8, August 1994, pp. 1905-1915

[5] M. Drews: Time delay estimation for microphone array speech enhancement systems; Proc. Eurospeech 1995, pp. 2013-2016

[6] S. Haykin: Array signal processing; Prentice Hall, 1985