ELSEVIER

Signal Processing 59 (1997) 253–266

SIGNAL
PROCESSING

A new cepstral prefiltering technique for estimating time delay under reverberant conditions

Alex Stéphenne, Benoît Champagne*

Institut National de la Recherche Scientifique, INRS-Télécommunications, 16 place du Commerce, Verdun, Québec, Canada H3E 1H6

Received 9 February 1996

Abstract

A microphone array can be used for hands-free acquisition of speech under reverberant conditions. This requires knowledge about the desired talker location, which can be obtained by estimating the time delays between the signals received by one or more pairs of spatially separated microphones. However, in a typical audio-conference room, strong reverberation is usually present and can have disastrous effects on the performance of conventional time delay estimation (TDE) methods. In this article, we present and evaluate a new cepstral prefiltering technique which can be applied on the received signals before the actual TDE in order to obtain a more accurate estimate of the delay in a typical reverberant environment. The technique is based on the estimation and the subtraction of the minimum-phase component (MPC) of the channel cepstrum from the total cepstrum of each microphone signal. So, in the same way that it is necessary in certain TDE methods to estimate the power spectral densities of the signals of interest from the received data, the new method requires the estimation of the channel MPC in the cepstral domain. The performances of a TDE system with and without cepstral prefiltering are compared via Monte-Carlo simulations for fixed random and speech sources as well as for a moving random source. The results clearly demonstrate the beneficial effects of the new cepstral prefiltering technique on TDE performance when the source is fixed or slowly moving. © 1997 Elsevier Science B.V.

Zusammenfassung

Ein Mikrofonarray kann für die Aufnahme von Sprache mit Freisprecheinrichtungen unter verhallten Bedingungen eingesetzt werden. Dies setzt Kenntnisse über die gewünschte Sprecherposition voraus, die durch eine Schätzung der Zeitverzögerungen zwischen den empfangenen Signalen von einem oder mehreren räumlich getrennten Mikrofonpaaren erhalten werden kann. In einem typischen Audiokonferenz-Raum ist jedoch gewöhnlich großer Nachhall zu erwarten, der katastrophale Auswirkungen auf die Arbeitsweise einer konventionellen Zeitverzögerungsschätzung (TDE) hat. In diesem Artikel zeigen und entwickeln wir ein neues Cepstrum-Vorfilterungsverfahren, das auf die empfangenen Signale vor der eigentlichen TDE angewandt werden kann, um eine genauere Schätzung der Zeitverzögerung in einer typischen verhallten Umgebung zu erhalten. Das Verfahren basiert auf der Schätzung und Subtraktion des minimalphasigen Anteils (MPC) des Kanalcepstrums vom Gesamtcepstrum jedes einzelnen Mikrofonsignals. Ebenso wie es bei verschiedenen TDE-Methoden nötig ist, das Leistungsdichtespektrum des interessanten Signals zu schätzen, benötigt diese neue Methode eine Schätzung der Minimalphasenkomponente des Kanals im Cepstral-Bereich. Die Güte eines TDE-Systems mit und ohne cepstraler Vorfilterung werden mit Monte-Carlo Simulationen sowohl für feste Zufalls- und Sprachquellen, als auch für eine bewegliche Zufallsquelle verglichen. Die Ergebnisse zeigen

* Corresponding author. Tel.: 514 765 7773; fax: 514 761 8501; e-mail: bchampgn@inrs-telecom.uquebec.ca.

PII S0165-1684(97)00051-0

deutlich die vorteilhaften Effekte der neuen cepstralen Vorfilterungsmethode auf die TDE Genauigkeit, wenn die Quell fest ist oder sich langsam bewegt. © 1997 Elsevier Science B.V.

Résumé

Un réseau de microphones peut être utilisé lors de la réception mains-libres de signaux de parole en milieu réverbérant. Ceci nécessite la connaissance de la position du locuteur, qui peut être obtenue en estimant les délais de propagation entre les signaux reçus par plusieurs paires de microphones. Cependant, dans une salle de télé-conférence typique, un fort niveau de réverbération est habituellement présent et peut avoir des effets désastreux sur la performance des méthodes d'estimation de délai (ED) conventionnelles. Dans cet article, nous présentons et évaluons une nouvelle technique de préfiltrage cepstral pouvant être appliquée aux signaux reçus avant l'ED de façon à obtenir des estimés de délai plus précis en milieu réverbérant. Cette technique est basée sur l'estimation de la composante en phase minimale (CPM) du cepstre du canal de transmission, que l'on soustrait ensuite du cepstre du signal reçu à chaque microphone. Donc, de la même façon qu'il est nécessaire, pour certaines méthodes d'ED, d'estimer la densité spectrale de puissance des signaux d'intérêt à partir des signaux reçus, la nouvelle technique nécessite l'estimation de la CPM du canal de transmission dans le domaine cepstral. Les performances d'un système d'ED avec et sans préfiltrage cepstral sont comparées à l'aide de simulations Monte-Carlo pour une source aléatoire fixe ou en mouvement, ainsi que pour une source fixe de parole. Les résultats démontrent clairement les effets bénéfiques de la technique de préfiltrage cepstral sur la performance du système d'ED lorsque la source est fixe ou bouge lentement. © 1997 Elsevier Science B.V.

Keywords: Room reverberation; Cepstrum; Time delay estimation

1. Introduction

When using an audio-conference system or a hands-free telephone in office rooms, speech signal acquisition is usually corrupted by reverberation and other directional noise sources. Significant suppression of these interferences is possible if one uses a microphone array properly steered in the direction of the desired talker [6]. This implies that it is necessary, at first, to estimate the precise location of this talker if the microphone array is to effectively filter out the interfering signals. One way to achieve this is to first estimate the time delays between the signals received by several pairs of spatially separated microphones, and then to triangulate the source using these estimates [15].

The generalized cross-correlation (GCC) method is one of the most popular techniques for time delay estimation (TDE) [8]. In this method, the delay estimate is obtained as the time-lag which maximizes the cross-correlation between filtered versions of the input signals. If the received signals are free of reverberation and are properly filtered, the GCC method reduces to the maximum likelihood time delay estimator and is asymptotically efficient in the limit of long observation time. However, a recent study [3] has shown that the presence of reverberation in the received signals can have disastrous effects on the performance of the GCC method. Thus there is a need for a new TDE method that is efficient under reverberant conditions.

Until now, few papers have dealt with the problem of TDE in the presence of reverberation. Some studies have been done on TDE when only one of the received signals is corrupted with a single [4, 5] or at most a few [19] echos while the other signal is completely free of any echo. A TDE algorithm has also been developed in the special simple case that arise when only one echo is present at each receiver [17]. Other studies have concentrated on identifying reverberant channels ([7, 10, 16], for example), the result of which can then be used to devise a dereverberation procedure via an inverse filter prior to the actual TDE. In fact, in [10, 16], dereverberation techniques are developed for simple channels with only a few paths. It is tempting to try to use such an approach in developing a new TDE technique under reverberant conditions but, in a typical audio-conference environment, the channel is often not minimal phase so that a causal and stable inverse filter might not exist [12, 11]. Furthermore, the process can be highly sensitive to variations in the estimate of the channel [11]. As we can see, TDE in a reverberant environment remains a challenging problem.

In this article, we present and evaluate a new cepstral prefiltering technique which is meant to be used in connection with standard GCC methods for TDE when reverberation is present. The devised cepstral prefilter is able to attenuate the effect of reverberation on signals received by individual microphones before feeding them into a GCC. The resulting TDE system, consisting of a set of cepstral prefilters followed by the GCC, is named GCC-CEP. Simulations are carried out for fixed random and speech sources as well as for moving random sources. The results suggest that the new cepstral prefilter can significantly improve TDE performance under reverberant conditions when the source is standing still or is slowly moving. In particular, a reduction of the percentage of anomalous estimates and of the bias and the standard deviation of the non-anomalous estimates is observed.

The paper is organized has follows. Section 2 contains a brief review of the GCC method and a discussion of its performance in a reverberant environment. Section 3 introduces the motivation for using cepstral processing and describes the approach used in the new cepstral prefiltering technique. The actual cepstral prefiltering algorithm and some implementation details are also presented. Simulation results are reported in Section 4. A summary and some final remarks are contained in Section 5.

2. Overview of the GCC method

In the classical formulation of the TDE problem between two channels, a lossless non-dispersive medium with a single propagation path from the source to the receivers is assumed. Thus, the received signals are modeled as follows:

$$\begin{aligned} x_1(t) &= s(t) + n_1(t), \quad 0 \leqslant t \leqslant T, \\ x_2(t) &= s(t+\tau) + n_2(t), \end{aligned} \tag{1}$$

where $x_i(t)$ $(i = 1,2)$ is the output signal of the ith receiver, $n_i(t)$ is an additive noise term, $s(t)$ is the unknown source signal, τ is the unknown delay and T is the length of the observation interval. The signals $s(t)$, $n_1(t)$ and $n_2(t)$ are themselves modeled as real, zero-mean, uncorrelated, stationary Gaussian random processes.

In the GCC method, the time-delay estimate is obtained as the value of the time-lag, τ, which maximizes the generalized cross-correlation function given by [8]

$$R_{12}(\tau) = \int_{-\infty}^{\infty} |G(f)|^2 X_1(f) X_2^*(f) \mathrm{e}^{\mathrm{j}2\pi f\tau}\,\mathrm{d}f, \tag{2}$$

where $X_i(f)$ denotes the Fourier transform of $x_i(t)$ over the interval $0 \leqslant t \leqslant T$, the superscript asterisk denotes the complex conjugate and $G(f)$ is a filter transfer function. The filter $G(f)$ is typically chosen so as to attenuate the signals in spectral regions where the signal-to-noise ratio is the lowest. In particular, the time delay estimate obtained by maximization of (2) is the maximum likelihood (ML) estimate if

$$|G(f)|^2 = \frac{S(f)}{N_1(f)N_2(f)} \left[1 + \frac{S(f)}{N_1(f)} + \frac{S(f)}{N_2(f)}\right]^{-1}, \tag{3}$$

where $S(f)$, $N_1(f)$ and $N_2(f)$ denote the power spectral densities of $s(t)$, $n_1(t)$ and $n_2(t)$, respectively. In a practical application, the above spectral densities are generally unknown and must be estimated from the data.

For the classical TDE model (1), the ML estimator of time delay is asymptotically unbiased and efficient in the limit $T \to \infty$. In a reverberant environment, however, the noises are highly correlated with the source signal and the GCC estimator based on (1)–(3), which was developed for the case of uncorrelated signal and noises, is no longer optimal. Effects of room reverberation on GCC performance have been investigated in [3]. The results of this study demonstrate the adverse effects of reverberation on TDE performance. In particular, it is shown that the percentage of anomalous estimates (or large errors) is characterized by an abrupt increase (from 0%) as the level of reverberation reaches a critical value. This behavior is due to the presence of erroneous peaks in the output of the ML cross-correlator (2)–(3), which result from the correlation existing between various pairs of echos signals on the two channels. As the level of reverberation increases, the amplitudes of the erroneous peaks increase, eventually making the ML estimator totally unreliable. Furthermore, below the critical value of reverberation mentioned above, the bias and the standard deviation increases with the level of reverberation. The increase in the bias is mainly due to the existence of strong initial echos and can greatly vary from one room configuration to another (room geometry, microphones and talker locations). The increase in the standard deviation is associated with the local effects of reverberation on the GCC around the true time delay value. This increase can be closely predicted theoretically by using an equivalent white noise model for the reverberation.

3. The new cepstral prefiltering technique

In this section we describe the new cepstral prefiltering technique that we propose for TDE between the direct path signals received by a pair of microphones in the presence of reverberation. The first subsection introduces the motivation for using cepstral processing and describes the approach used in the new cepstral prefiltering technique. The second and third subsections present the algorithm and some implementation considerations, respectively.

3.1. Approach and motivation

According to the discussion in Section 2, the classical signal model (1) is not appropriate to develop an efficient TDE method for reverberant environments. To overcome this limitation, let us assume for the time being that the acoustic transmission channel between the source and each of the microphones is linear and time-invariant. A more general mathematical model for the microphone signals can then be expressed as follows:

$$\begin{aligned} x_1(t) &= [h_1 * s](t) + n_1(t), \quad 0 \leqslant t \leqslant T, \\ x_2(t) &= [h_2 * s](t) + n_2(t), \end{aligned} \tag{4}$$

where $*$ denotes the operation of convolution, $h_i(t)$ is the acoustic impulse response between the source and the ith microphone, and the signals $s(t)$, $n_1(t)$ and $n_2(t)$ are defined as in the previous section. The presence of reverberation in each channel is entirely accounted for by $h_i(t)$. Other external interferences are modeled by the additive, uncorrelated noise term $n_i(t)$.

In this work we are mainly interested in the effect of reverberation on TDE. Accordingly, we assume that the reverberation level is high enough so that $n_i(t)$ is negligible. In practice, this assumption is valid when one talker speaks at a time and when the ambient noise level is sufficiently low as to permit effective communication. We therefore assume that the interference is essentially due to $h_i(t)$. This type of interference is commonly called convolutional smearing but can be transformed into an additive component by using the complex cepstrum. The motivation for transforming the interference into an additive component is that we can effectively deal with such a component via linear filtering techniques. This is the basic idea behind homomorphic deconvolution and we chose to follow this type of approach in devising our new cepstral prefiltering technique. It is now necessary to formally define the cepstrum and to give some additional details relevant to the problem of reverberation attenuation for TDE.

In applications, sampled versions of the signals are used for processing so we shall consider a discrete version of model (4). The complex cepstrum (simply called cepstrum hereafter) of a discrete-time signal $x[n]$ is defined as [13]

$$\widehat{x}[k] = F^{-1}\{\log X(\omega)\}, \tag{5}$$

where $X(\omega)$ is the Fourier transform of $x[n]$, $F^{-1}\{\cdot\}$ represents the inverse Fourier transform, the log operator is the complex logarithm, and the integer variable k is called quefrency. Note that a phase unwrapping procedure must be used when computing the complex logarithm in order to ensure its uniqueness and its analyticity.

One important property of the complex cepstrum is that it transforms convolution in the time domain into addition in the quefrency-domain. Thus, if we compute the cepstrum of (4), we find that

$$\widehat{x}_i[k] = \widehat{h}_i[k] + \widehat{s}[k] + \widehat{\eta}_i[k], \tag{6}$$

where $\widehat{x}_i[k]$, $\widehat{h}_i[k]$ and $\widehat{s}[k]$ are the cepstrum of the discrete-time signals $x_i[n]$, $h_i[n]$ and $s[n]$, respectively, and

$$\widehat{\eta}_i[k] = F^{-1}\left\{\log\left(1 + \frac{N_i(\omega)}{H_i(\omega)S(\omega)}\right)\right\}. \tag{7}$$

Here (and subsequently), $N_i(\omega)$, $H_i(\omega)$ and $S(\omega)$ denote the Fourier transforms of $n_i[n]$, $h_i[n]$ and $s[n]$, respectively. Note that, since $n_i[n]$ is negligible, $\widehat{\eta}_i[k]$ is also negligible. Our objective is to attenuate the reverberation which is entirely characterized by $\widehat{h}_i[k]$, an additive component of the microphone signal cepstrum. Clearly, this can be achieved by estimating and subtracting from $\hat{x}_i[k]$ in Eq. (6) the part of $\hat{h}_i[k]$ due to reverberation.

In this work, we shall find it useful to decompose the room impulse response $h_i[n]$ (or equivalently its Fourier transform $H_i(\omega)$) into a minimum-phase component (MPC) and an all-pass component (APC). More precisely, it is always possible to write [13]

$$H_i(\omega) = H_{i,\mathrm{ap}}(\omega) H_{i,\mathrm{min}}(\omega), \tag{8}$$

where $H_{i,\mathrm{ap}}(\omega)$ and $H_{i,\mathrm{min}}(\omega)$ are the frequency-domain APC and MPC, respectively. By construction, $|H_{i,\mathrm{ap}}(\omega)| = 1$ while all the poles and zeros of the Z-transform $H_{i,\mathrm{min}}(z)$ are inside the unit circle. Due to the convolution property of the cepstrum, it follows that

$$\hat{h}_i[k] = \hat{h}_{i,\mathrm{ap}}[k] + \hat{h}_{i,\mathrm{min}}[k], \tag{9}$$

where $\hat{h}_{i,\mathrm{ap}}[k]$ and $\hat{h}_{i,\mathrm{min}}[k]$ are the quefrency-domain APC and MPC, respectively. In practice, these components can be computed via the following formulae [13]:

$$\hat{h}_{i,\mathrm{min}}[k] = \begin{cases} 0, & k<0, \\ \hat{h}_i[0], & k=0, \\ \hat{h}_i[k] + \hat{h}_i[-k], & k>0, \end{cases} \tag{10}$$

$$\hat{h}_{i,\mathrm{ap}}[k] = \begin{cases} \hat{h}_i[k], & k<0, \\ 0, & k=0, \\ -\hat{h}_i[-k], & k>0. \end{cases} \tag{11}$$

Using the decomposition (9) in (6), we obtain

$$\hat{x}_i[k] = \hat{h}_{i,\mathrm{min}}[k] + \hat{h}_{i,\mathrm{ap}}[k] + \hat{s}[k] + \hat{\eta}_i[k]. \tag{12}$$

The introduction of a time delay d_0 on a microphone signal can be seen as a complex multiplication by a factor $\mathrm{e}^{-\mathrm{j}\omega d_0}$ in the frequency domain and thus it only affects the APC, not the MPC. Similarly, the relative time delay between the signals at two spatially separated microphones has no effect on the MPC of the two microphone signals. One might therefore be tempted to use only the all-pass component to make the TDE via GCC. It is simple to see why such an approach would not be good. Assume for the moment that we only use the APC for the TDE. This would be equivalent to subtracting the MPCs of the channel and the source signal cepstra from the total cepstrum of each microphone signal. Subtraction in the quefrency domain is equivalent to filtering in the time domain. Remember that our goal here is to reduce the effect of reverberation which is entirely characterized by the channel. The quefrency domain subtraction of the MPC of the channel is equivalent to filtering each microphone signal with a filter that is channel dependent. It is reasonable to expect such an operation to affect the part of the signal due to reverberation. Subtraction of the MPCs of the microphone signal cepstra reduces the signal to reverberation ratio in both channels. This translates into a loss of correlation between the two microphone signals so that the subsequent TDE via GCC is less effective. We therefore note that the cepstral prefiltering technique cannot be based on the subtraction of the MPCs from the total microphone cepstra. This conclusion was verified experimentally.

Special care must also be taken in the cepstral prefiltering not to introduce phase distortions which would make our time delay estimate useless. Experimental evidence has shown that modifications to $\hat{h}_{i,\mathrm{ap}}[k]$ ($i = 1,2$) are susceptible to introduce serious errors in the final delay estimates. This observation indicates that it is preferable not to affect this component. On the other hand, the same experiments demonstrated that it was possible to significantly improve time delay estimates by subtracting only the MPC of the channel cepstra from the received signal cepstra. Furthermore, we noted that the time delay estimate is relatively insensitive to small random perturbations of the subtracted $\hat{h}_{i,\mathrm{min}}[k]$. These observations led us to develop a new dereverberation approach based on the estimation and the subtraction of $\hat{h}_{i,\mathrm{min}}[k]$ from the total cepstrum.

In the proposed approach, the cepstral prefiltering is done on a frame by frame basis so that sequences of K samples are considered for cepstral prefiltering in each channel. The underlying assumption is that the MPC of the source signal cepstrum varies from frame to frame and is zero mean, while the MPC of the channel cepstrum is only slowly varying. We can expect such an assumption to be valid in practice since a typical source signal would vary from frame to frame while the channel impulse response would be almost fixed if the source was only slowly moving. Based on this assumption, a recursive cepstral averaging technique is proposed to estimate the MPC of the channel cepstrum, which is then subtracted from the microphone signal cepstrum. The resulting quefrency-domain information is then reconverted in the time domain where it can be fed into a conventional GCC. In the same way that it is necessary with certain GCC methods to estimate the power spectral densities of the signals of interest from the received data, the GCC-CEP method requires the estimation of the channel MPC in the quefrency domain.

Since we chose to modify only the MPCs in order to attenuate the effects of reverberation on TDE, it is beneficial to try to increase the relative importance of the MPCs over the APCs prior to the actual cepstral prefiltering operation. To do so, we can simply apply an exponential window to each frame before the cepstrum computation. The exponential window has the following form:

$$w[n] = \alpha^n, \quad 0 \leqslant n \leqslant K - 1, \tag{13}$$

where K is the frame size and $0 < \alpha \leqslant 1$. The effect of such a windowing operation on an arbitrary signal $x[n]$ is to move the poles and zeros of its Z-transform, $X(z)$, towards the interior of the unit circle, thus increasing the relative importance of the MPC over the APC [13].

Note that, in the case of a source in motion, the MPC of the channel is known to vary more slowly than the APC [18]. Our new cepstral prefiltering technique takes advantage of this fact since it only tries to estimate and subtract the MPC of the channel.

3.2. The algorithm

The nth sample of the mth frame for the ith microphone channel is denoted $x_i[n;m]$, where $n = 0, 1, \ldots, K-1$. For each frame and for each channel (i.e., for $m = 1, 2, \ldots$ and $i = 1, 2$), the cepstral prefiltering consists of the following steps:

1. Apply the exponential window (13) with coefficient α to $x_i[n;m]$.
2. Compute the corresponding cepstra, denoted $\hat{x}_i[k;m]$ $(k = 0, 1, \ldots, K-1)$.
3. Compute the MPC of $\hat{x}_i[k;m]$ as defined in (10). Denote it by $\hat{x}_{i,\min}[k;m]$.
4. Compute the average of $\hat{x}_{i,\min}[k;m]$ over successive frames in order to obtain an estimate of $\hat{h}_{i,\min}[k]$, which is denoted by $\bar{h}_{i,\min}[k;m]$. The averaging is done according to the following recursive equation:

$$\bar{h}_{i,\min}[k;m] = \begin{cases} \hat{x}_{i,\min}[k;m], \quad m = 1, \\ (1-\mu)\bar{h}_{i,\min}[k;m-1] + \mu\hat{x}_{i,\min}[k;m], \\ \quad m > 1, \end{cases} \tag{14}$$

where the parameter μ $(0 \leqslant \mu \leqslant 1)$ controls the memory of the recursive averaging procedure.
5. Subtract $\bar{h}_{i,min}[k;m]$ from $\hat{x}_i[k;m]$ in order to obtain a new microphone signal cepstrum with less contribution from the reverberation. Denote the results by $\tilde{x}_i[k;m]$:

$$\tilde{x}_i[k;m] = \hat{x}_i[k;m] - \bar{h}_{i,\min}[k;m]. \tag{15}$$

6. Transform the cepstra obtained in the previous step, $\tilde{x}_i[k;m]$, back to the time domain.
7. Apply the inverse exponential window.

The resulting frames are then ready to be fed into the GCC. The combined system consisting of the above cepstral prefilters followed by a GCC is called GCC-CEP. A diagram of the cepstral prefiltering algorithm is shown in Fig. 1.

3.3. Implementation considerations

Below, we briefly comment on the selection of the various parameters used in the cepstral prefiltering algorithm.

The frame size K should not be too large if we want to be able to consider the channel as time-invariant (or at least slowly varying) over a few frames. On the other hand, the frame size should be sufficiently large to avoid signal segmentation effects on the cepstrum computation [13]. Such effects include time aliasing and edge errors caused by the fact that reverberation due to the signal in the present frame appears in the next frames while reverberation due to the signal in the previous frames appears in the present frame.

The exponential window coefficient α in (13) also has to be chosen with care. A small value would increase the relative importance of the MPC but would also reduce the effective size of the frame. The

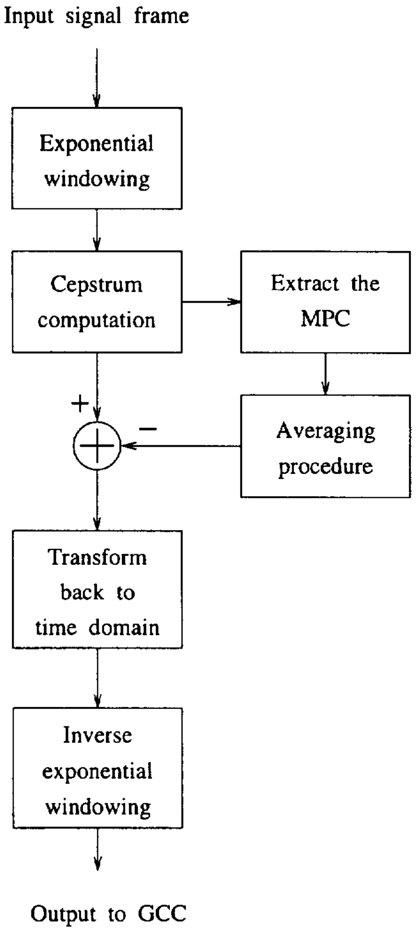


Fig. 1. Diagram of the cepstral prefiltering algorithm.

optimal value of α is strongly dependent on the frame size K and must be found empirically.

A very small value of the memory parameter μ would make the cepstral prefiltering more beneficial for a fixed source but the convergence time of the cepstral prefilters would be larger. The value of μ has to be large enough so that fast convergence is possible. On the other hand, if it is too large then the estimation error after convergence would also be large and the cepstral prefiltering would be less advantageous.

Remember that for the cepstral prefiltering technique to be effective, the MPC of the source signal cepstrum must be zero-mean. In practice, such an assumption may be too restrictive. It is possible to modify the prefiltering technique to consider source signals with non-zero-mean MPC cepstrum in certain quefrency intervals. In fact, if we know these intervals a priori, it is possible to leave them unmodified by the cepstral prefiltering technique. To do so, we simply set $\bar{h}_{i,\min}[k;m] = 0$ for these intervals in the fourth step of the cepstral prefiltering. This approach is especially effective if the average MPC of the source signal cepstrum is confined to a few small quefrency intervals (e.g., the MPC of voiced speech cepstra is concentrated in the lower portion of the quefrency domain [13]) and if these intervals do not coincide with quefrency intervals for which the MPC of the channel cepstrum is important.

For the random sources as well as for the speech signals considered here (see Section 4), the zero-mean assumption of the MPC of the source signal cepstrum was only noticeably violated for values of quefrency inferior to about 12 (with a sampling rate of 10 kHz). This value was found experimentally. Accordingly, we set the estimated MPC of the channel cepstrum in (14) equal to zero for values of quefrency $k_1 < 12$.

Note finally that if the convergence rate is of prime importance (e.g., when tracking a moving source), then it is possible to use overlapping frames in the cepstral prefiltering technique. Frame overlapping increases the convergence rate but the channel MPC estimation error after convergence is slightly increased because of the reduced amount of time considered in the cepstral averaging procedure (14).

4. Evaluation

The performance of the new GCC-CEP time delay estimator was investigated via Monte-Carlo simulations and compared with the performance of a conventional GCC (i.e., without the cepstral prefilters). Three different source signals were used: a fixed random source, a fixed speech source and finally a moving random source.

We consider a rectangular room with uniform wall reflection coefficients. A rectangular coordinate system with the origin in one corner and axes parallel to the walls is used to reference points in the room. The dimensions of the room along these axes are 10.0, 6.6 and 3.0 m, respectively. The source is omnidirectional. The microphone have cardioïd directivity patterns pointing in the negative x direction and their positions are $(6.5, 2.8, 1.8)$ and $(6.5, 3.8, 1.8)$ m.

The sampling frequency of the synthetic microphone signals is $f_s = 10$ kHz. The background noise is 30 dB below the microphone signal direct path power in each channel. The results presented below for the cepstral prefiltering are obtained with a frame overlap of 0%, a frame size $K = 2048$ samples (204.8 ms) and an exponential window coefficient α in (13) set to 0.9985. The integration time T of the TDE procedure is set to 204.8 ms for both the GCC and the GCC-CEP method. Based on the shape of the received random signals autocorrelation function it is reasonable to classify time delay estimates for which the absolute error exceeds $3T_s$ (0.3 ms) as anomalies. Based on a priori knowledge of the signal and noise spectra, the weighting function $G(f)$ (3), used in the correlator of GCC and GCC-CEP, is set to one inside the source signal passband and to zero outside.

4.1. Fixed random sources

The position of the source is $(2.4835, 2.0, 1.8)$ m. The source signal $s[n]$ is obtained by passing a Gaussian white noise sequence through a bandpass filter with cut-off frequencies $f_l = 450$ and $f_u = 3475$ Hz. Digital versions of the room impulse responses $h_i[n]$ are generated with Allen and Berkley's implementation of the image model technique [1] (properly modified for cardioïd microphones) with Peterson's modification [14]. Impulse responses are generated for each value of reverberation time, which is defined as the time required for the energy of a signal to decrease by 60 dB and is obtained with Eyring's reverberation formula [9]. These responses are truncated to about 6000 samples (0.6 s). Even for the worst case considered here, the truncated tail of the response is more than 40 dB below the main peak corresponding to the direct path signal.

The performance of the GCC-CEP algorithm after convergence is relatively insensitive to the value of the memory parameter μ of the cepstral averaging (14) which could vary from 0.03 to 0.2 without great impact on the simulation results. In the results described below, we have set μ to 0.06. For this selected value of μ, the convergence time for the estimate of $\widehat{h}_{i,\min}[k; m]$ (14) was of the order of 2 s.

For the given room configuration and for each value of reverberation time, 300 independent time delay estimates were calculated with the new GCC-CEP system and with the standard GCC method. Using the 300 time delay estimates (obtained after convergence of the cepstral prefilters), the percentage of anomalies and the sample bias and variance of the non-anomalous estimates were calculated.

Results, as a function of the reverberation time, T_r, are shown in Fig. 2. The vertical bar superimposed on each data point represents the 95% confidence interval for that measurement. It can be seen from Fig. 2(a) that the cepstral prefiltering greatly reduces the percentage of anomalous time delay estimates.

The bias and standard deviation of the non-anomalous estimates are illustrated in Figs. 2(b) and (c), respectively.[1] We note the great reduction of bias when cepstral prefiltering is used for all values of T_r. We also note that, for almost all values of T_r, the standard deviation is reduced by approximately 5 dB. The larger variance of the GCC-CEP method for $T_r < 0.06$ s is due to non-zero estimation errors of the channel MPC cepstrum in (14). However, such small values of T_r do not correspond to practical situations.

The effect of the cepstral prefiltering on the cross-correlation function in GCC has also been investigated. The presence of reverberation raises the sidelobe levels of the cross-correlation function, eventually making the TDE unreliable. Figs. 3(a) and (b) show waterfall graphs illustrating three realizations of the cross-correlation function with GCC and GCC-CEP, respectively, and for $T_r = 0$, 0.27 and 0.39 s. The maxima are indicated by "o" and correspond to the time delay estimates. The true time delay value is -9 samples. We note that the cepstral prefiltering significantly reduces the sidelobe levels in the presence of strong reverberation, making the TDE more reliable.

4.2. Fixed speech sources

The source signals we are considering here are coming from two audio files. The first is from a male speaker, the other from a female. The rest of the simulation scenario and in particular the room impulse

[1] Note that, in Figs. 2(b) and (c), the results for the GCC are given only for $T_r < 0.65$. Beyond this value, delay estimates are dominated by anomalies.

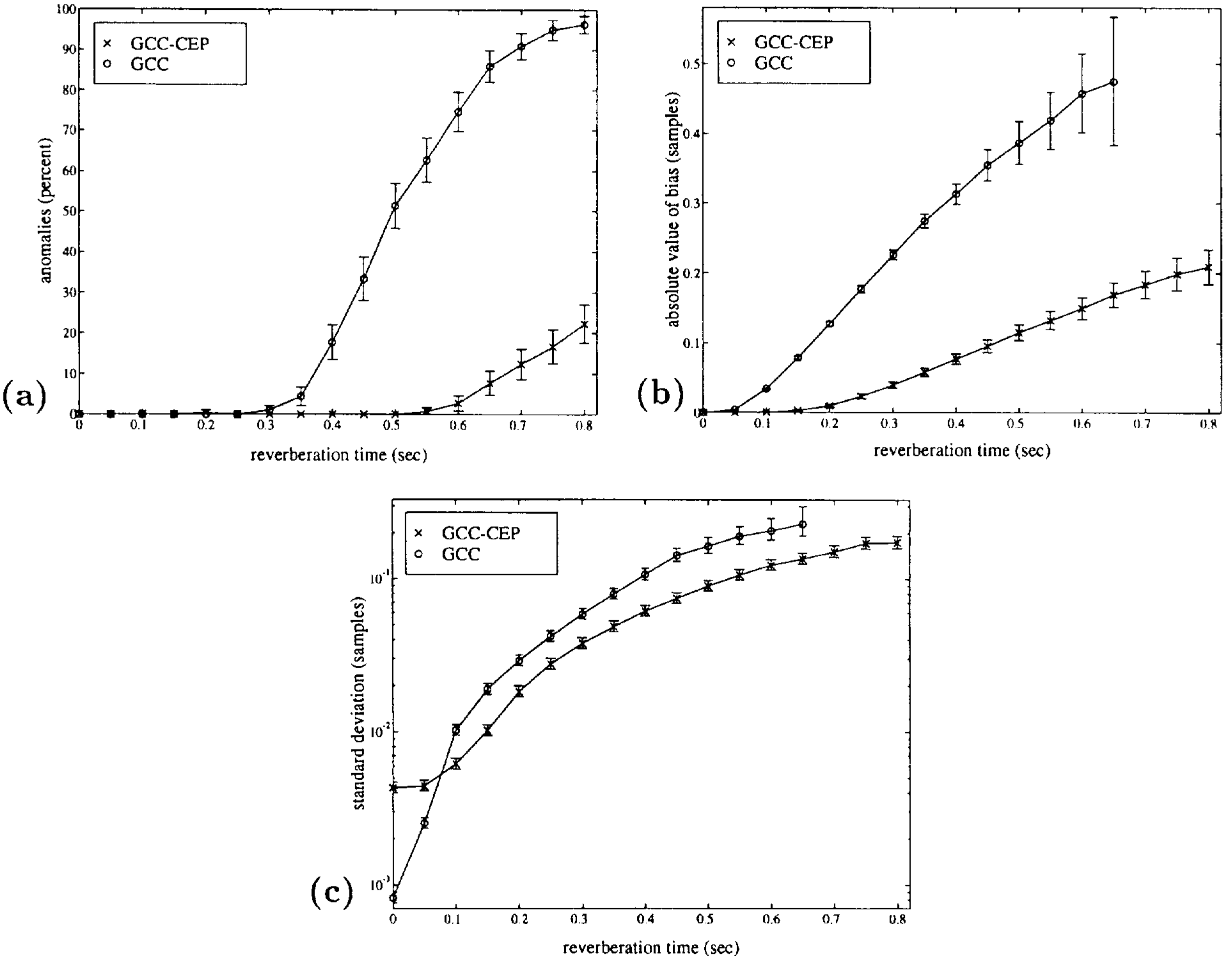


Fig. 2. (a) Percentage of anomalous time delay estimates, (b) bias of the non-anomalous time delay estimates and (c) standard deviation of the non-anomalous time delay estimates versus reverberation time, obtained with and without cepstral prefiltering.

responses are identical to the ones used in the preceding section.

In order for the cepstral prefiltering to be effective, the cepstral averaging procedure (14) must not be carried out during silent segments of the microphone signals. Note that even though the cepstral average is not updated for silent frames, the cepstral prefiltering is still done. In a practical system, a speech detector would therefore be indispensable. For the sake of simulation, continuous speech was used so that no silent frame was present.

Figs. 4(a) and (b) show the time delay estimates after convergence for a male and a female speech source, respectively ($T_r = 0.27$ s). The circles and asterisks (joined with a continuous line) are obtained with GCC-CEP for high- and low-energy frames, respectively (the classification into low- and high-energy frames was done manually). The crosses (joined with a dashed line) are obtained with GCC. The horizontal dotted line at −9 indicates the true delay. The two horizontal bands below the curves show the occurrence of anomalies for each method. Below these bands, some performance measures are indicated: the percentage of anomalies and the bias and standard deviation of the non-anomalous estimates. Note that these graphs indicate that TDE is

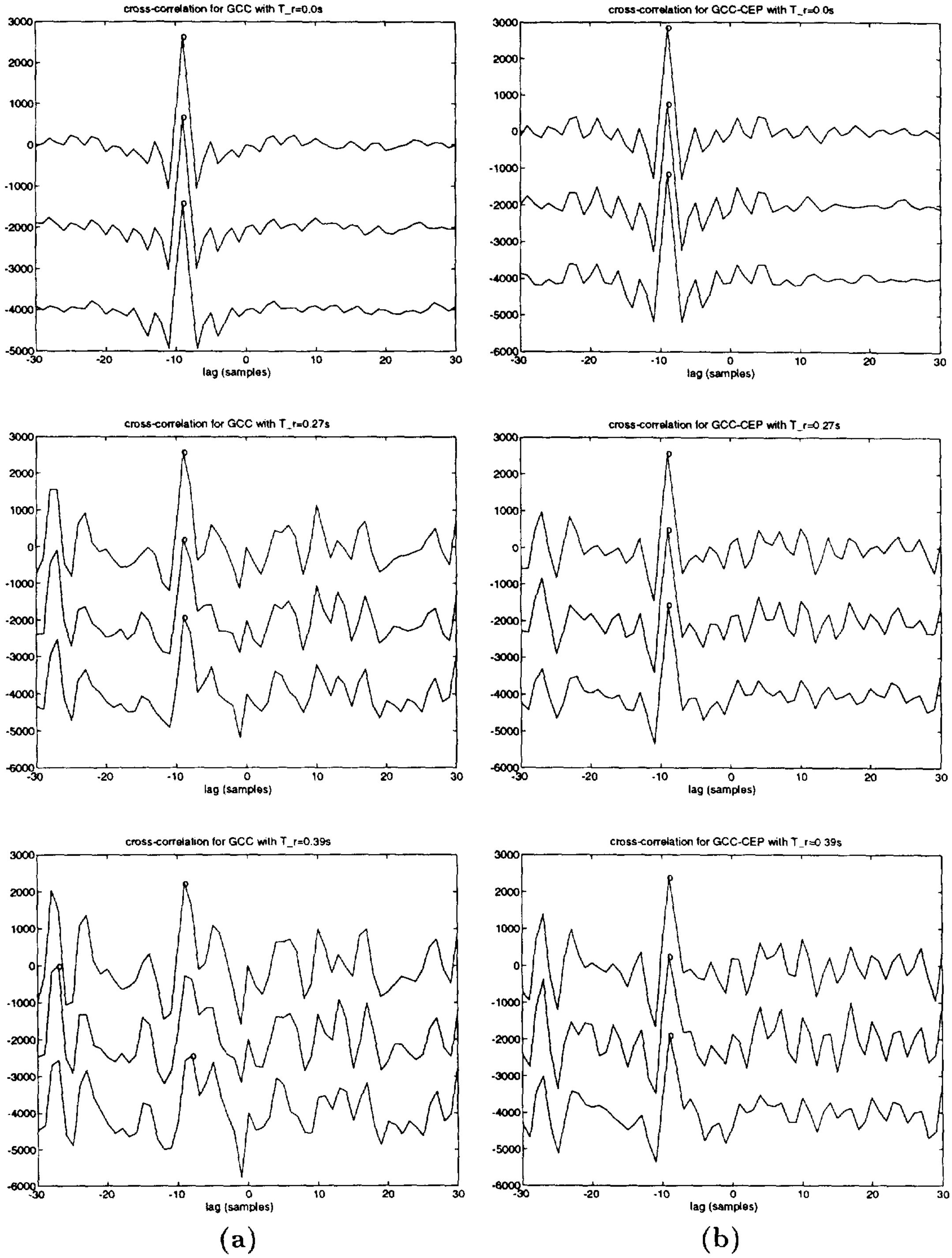
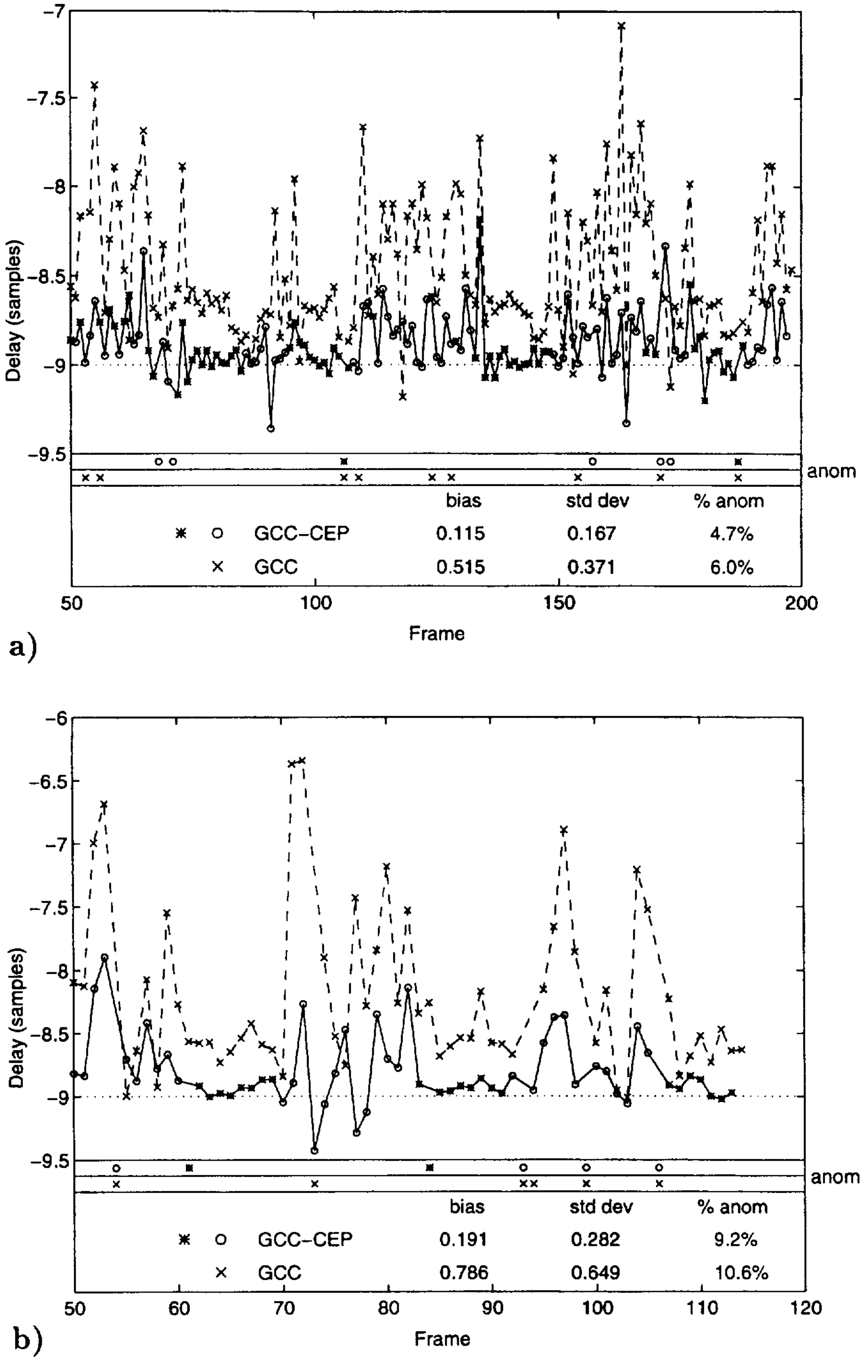


Fig. 3. Waterfall graphs illustrating three realizations of the cross-correlation function with (a) GCC and (b) GCC-CEP for $T_r = 0$, 0.27 and 0.39 s. The maxima are indicated by "o" and correspond to the time delay estimates. The true value is −9 samples.



a)

		bias	std dev	% anom
✱ ○	GCC-CEP	0.115	0.167	4.7%
×	GCC	0.515	0.371	6.0%

b)

		bias	std dev	% anom
✱ ○	GCC-CEP	0.191	0.282	9.2%
×	GCC	0.786	0.649	10.6%

Fig. 4. TDE (for frame indices greater than 50) for a fixed (a) male and (b) female speech source ($T_r = 0.27$ s). The circles and asterisks are obtained with GCC-CEP for high- and low-energy frames, respectively. The crosses are obtained with GCC.

generally more precise for low-energy frames than for high-energy frames. To explain this, we first note that frames with low-energy often correspond to unvoiced speech while high-energy frames can generally be associated with voiced speech. The inferior performance of GCC for high-energy frames could be explained by the quasi-periodic nature of voiced speech which can adversely affect the accuracy of TDE via GCC. Indeed, recall that the selected weighting function, $G(f)$ (3), is more appropriate for unvoiced speech characterized by an almost flat spectrum than for voiced speech which clearly has a non-flat spectrum.

Overall we can observe that most of the estimates are more accurate for GCC-CEP than for GCC, even for the frames with low source signal energy. The number of anomalies is only slightly reduced when the cepstral prefiltering is used, but there is a strong reduction of bias (by a factor of about 4) and standard deviation (about 6 dB). Note that the convergence time of the GCC-CEP method is increased compared to the fixed random source case since the cepstral average is not updated during silent frames.

4.3. *Moving random sources*

The source signal is the same as for the fixed source case but the source is now in motion. The microphone signals are generated by the method of [2], which is based on an extension of the image model technique to moving point sources. Since the rectangular room is of finite dimension, the source can only be moving in one specific direction for a limited time. A sinusoïdal trajectory has therefore been used. The position of the source at a given time, t, is given by $(2.4835, 2\sin(v_{\max}t/2) + 3.3, 1.8)$ m, where $v_{\max}$ is the maximal speed of the source.

The evaluation of the performance of a given TDE method for moving sources necessitate the introduction of a reference time delay (RTD), which is defined here as the delay estimate obtained with GCC in the absence of reverberation. The difference between this RTD and the delay obtained under reverberant conditions with different TDE methods can be used for performance assessment. This difference will simply be called an estimation error from now on, even though this is not a rigorous definition.

In the case of a moving source, the parameter μ (14) must now be sufficiently large as to allow the tracking of variations in the reverberation structure due to motion. But then again the value must be small enough to actually average out the cepstrum. Intuitively we should therefore expect the optimal μ value to increase as $v_{\max}$ increases. This was verified experimentally. The optimal μ for $v_{\max} = 0.25$ m/s is 0.14 and for $v_{\max} = 0.5$ m/s it is 0.2. We observed experimentally that the choice of μ had relatively small effects on the GCC-TDE performance after convergence as long as it was maintained between 0.06 and 0.2 for values of $v_{\max} < 0.5$ m/s. After convergence of the cepstral prefilters and for values of μ and $v_{\max}$ inside the mentioned ranges, anomalies are almost inexistent and the bias and standard deviation are, in the worst case, the same as with GCC. For the same speed but for values of μ larger than 0.2, anomalies begin to appear. Even if the choice of the optimal μ is dependent on the speed of the source, for a typical system we would not be expected to know the speed of the source in advance. The results presented in this section were therefore obtained with μ fixed to 0.06.

Simulation results for $v_{\max} = 0.25$ m/s are shown in Fig. 5. Part (a) and (b) illustrate the delay estimates and the estimation errors versus frame number for $T_r = 0.27$ s, while (c) and (d) are for $T_r = 0.30$ s. The anomalous estimates are not shown on (b) and (d). We can note that there is an adaptation period for GCC-CEP characterized by a large number of anomalies. Note that this adaptation time would be smaller for larger values of μ. After this adaptation period, we note that there is a reduction in the number of anomalies when the cepstral prefiltering is used. The bias of the non-anomalous estimates is also greatly reduced while the standard deviation remains about the same. Other simulations were conducted and we noted that the cepstral prefiltering became less advantageous for $v_{\max} > 0.5$ m/s.

5. Summary and discussion

We have presented a new cepstral prefiltering technique to be used in connection with a standard time delay estimator in the presence of reverberation. The technique is based on the estimation and the subtraction of the MPC of the channel cepstrum from the total cepstrum of each microphone. Simulation results were presented for both fixed random and speech sources

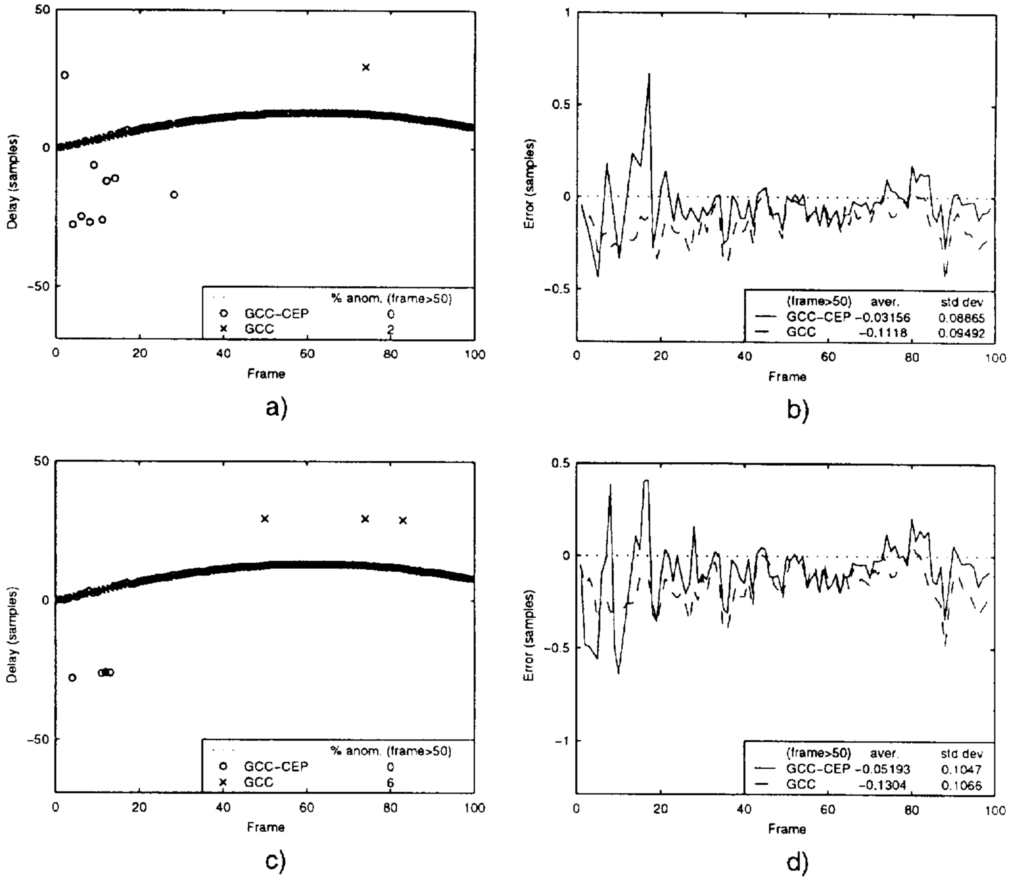


Fig. 5. TDE for a moving random source with $v_{\max} = 0.25\,\mathrm{m/s}$. (a) and (b) represent, respectively, the delay estimates and the errors versus frame number for $T_r = 0.27\,\mathrm{s}$, while (c) and (d) are for $T_r = 0.30\,\mathrm{s}$.

as well as for a moving random source. In all of the studied cases, the use of the cepstral prefiltering technique increased the accuracy of the time delay estimates.

The cepstral prefiltering technique takes a finite amount of time before being effective. Nevertheless, if the source is motionless or only slowly moving, then it is possible to adjust the various parameters of the cepstral prefiltering in order for the GCC-CEP system to be effective. On the other hand, if the source is moving faster than about 0.5 m/s, the averaging procedure in (14) will become less effective. One way to reduce the convergence time of the cepstral prefiltering so that the GCC-TDE system could track a rapidly moving source would be to adaptively adjust the parameter μ as the speed of the source varies. Another way would be to use some kind of cepstral prediction instead of cepstral averaging in (14).

References

[1] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Amer. 65 (1979) 943–950.

[2] B. Champagne, Simulation of the response of multiple microphones to a moving point source, Applied Acoust. 42 (1994) 313–332.

[3] B. Champagne, S. Bédard, A. Stéphenne, Performance of time delay estimation in the presence of room reverberation, IEEE Trans. Speech Audio Process. 4 (2) (1996) 148–152.

[4] Y.T. Chan, P.C. Ching, Non-stationary time delay estimation with a multipath, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1989, pp. 2736-2739.

[5] P.C. Ching, K.C. Ho, Y.T. Chan, Constrained adaptation for time delay estimation with multipath propagation, IEE Proc. F (Radar Signal Process.) 138 (1991) 453-458.

[6] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, M.M. Sondhi, Autodirective microphone systems, Acustica 73 (1991) 58-71.

[7] J.P. Ianniello, High resolution multipath time delay estimation for broadband random signals, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1987, pp. 12.4.1-12.4.4.

[8] C.H. Knapp, G.C. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (1976) 320-327.

[9] H. Kuttruff, Room Acoustics, 3rd ed., Elsevier Applied Science, 1991, Chapter 5, pp. 114-118.

[10] V.A. Margo, D.M. Etter, N.C. Carlson, Multiple short-length adaptive filters for time-varying echo cancellation, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1993, pp. 161-163.

[11] J. Mourjopoulos, On the variation and invertibility of room impulse response functions, J. Sound Vibration 102 (1985) 217-228.

[12] S.T. Neely, J.B. Allen, Invertibility of a room impulse response, J. Acoust. Soc. Amer. 66 (1979) 165-169.

[13] A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1975, Chapter 10.

[14] P.M. Peterson, Simulating the response of multiple microphones to a single acoustic source in a reverberant room, J. Acoust. Soc. Amer. 80 (5) (1986) 1527-1529.

[15] H.F. Silverman et al., A two-stage algorithm for determining talker location from linear microphone array data, Comput. Speech and Language 6 (1992) 129-152.

[16] J.O. Smith, B. Friedlander, Adaptive multipath delay estimation, IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (1985) 812-822.

[17] H.C. So, P.C. Ching, Y.T. Chan, A novel constrained algorithm for time delay estimation in the presence of multipath transmissions, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1993, pp. 313-315.

[18] M. Tohyama, R.H. Lyon, T. Koike, Source waveform recovery in a reverberant space by cepstrum dereverberation, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 1993, pp. 1.157-1.160.

[19] R.J. Tremblay, G.C. Carter, D.W. Lytle, A practical approach to the estimation of amplitude and time-delay parameters of a composite signal, IEEE J. Oceanic Eng. OE-12 (1979) 273-278; J. Acoust. Soc. Amer. 80 (1986) 1527-1529.