CEPSTRAL PREFILTERING FOR TIME DELAY ESTIMATION IN REVERBERANT ENVIRONMENTS

Alex Stéphenne and Benoît Champagne

INRS-Télécommunications, Université du Québec, 16 place du Commerce
Verdun, Québec, Canada H3E 1H6
email: {stephenn,bchampgn}@inrs-telecom.uquebec.ca

ABSTRACT

Time delay estimation (TDE) between the signals received by two or more spatially separated microphones can be used as a means for the passive localization of the dominant talker in applications such as audio-conference. However, in a recent study, it has been shown that reverberation can have disastrous effects on TDE performance. In this paper, we develop and evaluate a new cepstral prefiltering technique which can be applied on the microphone signals before the actual TDE in order to obtain a more accurate estimate of the position of a source in a typical reverberant environment. The performance of a TDE system with and without cepstral prefiltering is investigated under controlled conditions via Monte-Carlo simulations. The results clearly demonstrate the beneficial effects of the new cepstral prefiltering technique on TDE performance (i.e., reduction of bias, variance and number of anomalous estimates).

1. INTRODUCTION

Speech signal acquisition, in applications such as audio-conference or hands-free telephony in small office room, is usually corrupted by reverberation and directional noise sources. These interferences can be suppressed by the use of a microphone array properly oriented (or steered) in the direction of the dominant talker [1]. In order for the microphone array to effectively filter out the interfering signals, it is necessary to know the precise location of the dominant talker. Time delay estimation (TDE) between the signals received by two or more spatially separated microphones has been proposed as a means for the passive localization of the dominant talker [2].

The generalized cross-correlation (GCC) method is one of the most popular techniques for TDE. In this method, the delay estimate is obtained by maximizing the output of the cross-correlation between filtered versions of the input signals over an *a priori* search interval. In the absence of reverberation, with the proper choice of filters, the GCC method reduces to the maximum likelihood time delay estimator and is nearly optimal. However, it has been shown in [3] that the presence of reverberation in the received signals has disastrous effects on the performance of the GCC method.

In this paper, we develop and evaluate a new cepstral prefiltering technique to be used in connection with a standard TDE method like GCC in the presence of reverberation. The paper begins with a brief review of the GCC method and a discussion of its performance in a reverberant environment. Next, we introduce relevant concepts of cepstral analysis and homomorphic deconvolution. These concepts are then used to devise a new cepstral prefilter 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 standard GCC, is named GCC-CEP.

Simulations are carried out and the results clearly demonstrate the beneficial effects of the cepstral prefilters in TDE under reverberant conditions. In particular, we note a strong reduction of the number of anomalous estimates and an improvement of the variance and the bias of the non-anomalous estimates.

2. THE GCC METHOD

In the classical TDE problem between two-channels, the received signals are modeled as follows:

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

where the $x_i(t)$ are the output signals of the ith receiver, the $n_i(t)$ are additive noises, $s(t)$ is the unknown source signal, τ is the unknown delay and T denotes the duration of the observation interval. It is further assumed that $s(t)$, $n_1(t)$ and $n_2(t)$ are real zero-mean, uncorrelated, stationary Gaussian random processes. This model corresponds to a lossless non-dispersive medium with a single propagation path from the source to the receivers.

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

$$R_{12}(\tau) = \int_{-\infty}^{\infty} |G(f)|^2 X_1(f) X_2^*(f) e^{j2\pi f \tau} df. \tag{2}$$

where $X_i(f)$ denotes the Fourier transform of $x_i(t)$ over the interval $0 \leq t \leq T$, "$*$" 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,

Support for this work has been provided by NSERC operating grant OGP0105533 and by FCAR grant 93-ER-1577.

0-7803-2431-5/95 $4.00 © 1995 IEEE

the time delay estimate obtained by maximization of (2) is the maximum likelihood (ML) estimate if

$$|G(f)|^2 = \frac{S(f)}{N_1(f)N_2(f)}\{1 + \frac{S(f)}{N_1(f)} + \frac{S(f)}{N_2(f)}\}^{-1}, \qquad (3)$$

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.

For the classical TDE model (1), the ML estimator of time delay is asymptotically unbiased and efficient in the limit $T \rightarrow \infty$. In a reverberant environment, however, the noises are highly correlated with the source signal and the ML estimator, which was developed for the case of uncorrelated signal and noises, is no longer optimal. Effects of room reverberation on time-delay estimation performance have been investigated in [3]. The results of this study clearly demonstrate the adverse effects of revererberation on TDE. In particular, it has been shown that the percentage of anomalous estimates 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). These peaks result from the correlation existing between pairs of echos on the different channels. As the level of reverberation increases, the amplitudes of the erroneous peaks increase, eventually making the ML estimator unreliable. Furthermore, the bias and the standard deviation of the estimates increase with the level of reverberation.

3. CEPSTRAL ANALYSIS AND HOMOMORPHIC DECONVOLUTION

We can no longer use the simple model (1) to develop an efficient time delay estimator in the presence of reverberation. Instead, let us assume for the time being that the acoustic transmission channel between the source and each of the receivers is linear and time-invariant. Then a more complete model for the receiver signals can be expressed as follows:

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

where $*$ denotes the operation of convolution and the $h_i(t)$ are the acoustic impulse responses between the source and the ith receiver. 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 interferences, such as external noise, are modeled as an additive uncorrelated noise denoted by $n_i(t)$. In applications, sampled versions of the signals are used so we shall consider a discrete version of model (4).

The complex cepstrum (simply called cepstrum hereafter) of a discrete signal $x[n]$ is defined as

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

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 as defined in [4], and the variable k is called quefrency. If we compute the cepstrum of (4), we find that

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

where

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

and $N_i(\omega)$, $H_i(\omega)$ and $S(\omega)$ are the Fourier transforms of $n_i[n]$, $h_i[n]$ and $s[n]$ respectively. Note that the error term $\widehat{\eta}_i[k] = 0$ if $n_i[n] = 0$.

It is well known that the cepstrum of a signal can be decomposed into a minimum phase component (MPC) and an all-pass component (APC) [4]. More precisely, we have

$$\widehat{h}_i[k] = \widehat{h}_{i,ap}[k] + \widehat{h}_{i,min}[k] \qquad (8)$$

where

$$\widehat{h}_{i,min}[k] = \begin{cases} 0\,, & n < 0 \\ \widehat{h}_i[0]\,, & n = 0 \\ \widehat{h}_i[k] + \widehat{h}_i[-k]\,, & n > 0 \end{cases} \qquad (9)$$

$$\widehat{h}_{i,ap}[k] = \begin{cases} \widehat{h}_i[k]\,, & n < 0 \\ 0\,, & n = 0 \\ -\widehat{h}_i[-k]\,, & n > 0 \end{cases} \qquad (10)$$

Using this decomposition in (6), we obtain

$$\widehat{x}_i[k] = \widehat{h}_{i,min}[k] + \widehat{h}_{i,ap}[k] + \widehat{s}[k] + \widehat{\eta}_i[k]. \qquad (11)$$

Our objective is to attenuate the reverberation which is entirely characterized by $\widehat{h}_i[k]$, an additive component of the microphone signal cepstrum (neglecting the noise term). Intuitively we would like to estimate and subtract from $\widehat{x}_i[k]$ in (6) the part of $\widehat{h}_i[k]$ due to reverberation. As we will see in the next section, the MPC-APC decomposition is very useful since it is highly advantageous to treat the APC and MPC differently in the cepstral prefiltering procedure.

4. THE NEW CEPSTRAL PREFILTERING APPROACH

Special care must be taken in the cepstral prefiltering not to introduce phase distortions which would make our time delay estimate useless. Important information about the delay between the two channels is contained in $\widehat{h}_{i,ap}[k]$. Modification to $\widehat{h}_{i,ap}[k]$ is therefore susceptible to introduce serious error in the final delay estimate. This observation is supported by experimental evidence, which indicate that it is preferable not to affect this component. On the other hand, the same experiments demonstrated that the time delay estimate is relatively insensitive to small modifications on $\widehat{h}_{i,min}[k]$. Furthermore, we noted that it was possible to improve time delay estimates by subtracting only the MPC of the cepstrum of the channels from the microphone signal cepstra. These observations led us to develop a new dereverberation approach based on the estimation and the subtraction of $\widehat{h}_{i,min}[k]$ from the total cepstrum.

In the proposed approach, the cepstral prefiltering is done on a frame by frame basis. 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. Based on this assumption, a recursive cepstral averaging technique is developed to estimate the MPC of the channel cepstrum, which is then subtracted from the microphone signal cepstrum. The resulting cepstrum domain information is then reconverted in the time domain where it can be fed into the GCC.

An exponential window is applied to each frame before the cepstrum computation. The exponential window has the following form:

$$w[n] = \alpha^n \ , \quad 0 \leq n \leq K-1, \tag{12}$$

where K is the frame size and $0 < \alpha \leq 1$. The effect of such a windowing operation on $x[n]$ is to move the poles and zeros of $X(\omega)$ towards the interior of the unit circle. The purpose of the exponential window is to increase the relative importance of the MPC over the APC. Since we are only allowed to modify the MPC in order to attenuate the effect of reverberation on TDE, it is beneficial to use an exponential window.

The mth frame of 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 (12) with coefficient α to $x_i[n;m]$ and compute the corresponding cepstra $\widehat{x}_i[k;m]$.

2) Compute the MPC of $\widehat{x}_i[k;m]$ as defined in (9). Denote it by $\widehat{x}_{i,min}[k;m]$.

3) Compute the average of $\widehat{x}_{i,min}[k;m]$ over successive frames in order to obtain an estimate of $\widehat{h}_{i,min}[k]$ which is denoted by $\overline{h}_{i,min}[k;m]$. The averaging is done according to the following recursive equation:

$$\overline{h}_{i,min}[k;m] = \begin{cases} \widehat{x}_{i,min}[k;m], & m = 1, \\ (1-\mu)\overline{h}_{i,min}[k;m-1] + \mu\widehat{x}_{i,min}[k;m], & m > 1, \end{cases} \tag{13}$$

where the parameter μ $(0 \leq \mu \leq 1)$ controls the memory of the recursive averaging procedure.

4) Subtract $\overline{h}_{i,min}[k;m]$ from $\widehat{x}_i[k;m]$ in order to obtain a new microphone signal cepstrum with less contribution from the reverberation. Denote the results by $\tilde{x}_i[k;m]$:

$$\tilde{x}_i[k;m] = \widehat{x}_i[k;m] - \overline{h}_{i,min}[k;m]. \tag{14}$$

5) Transform the cepstra obtained in the previous step, $\tilde{x}_i[k;m]$, back to the time domain and 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.

Remember that in order for the cepstral prefiltering procedure to be effective, the MPC of the source signal cepstrum must be zero-mean. In practice, such an assumption is often too restrictive. It is possible to modify the prefiltering procedure to consider source signal 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 procedure. To do so, we simply set $\overline{h}_{i,min}[k;m] = 0$ for these intervals in the third 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 con^centrated in the lower portion of the quefrency domain [4]) and if these intervals do not coincide with quefrency intervals for which the MPC of the channel cepstrum is important.

5. SIMULATION AND IMPLEMENTATION DETAILS

The performance of the new GCC-CEP time delay estimator has been investigated via Monte-Carlo simulations. The simulation methodology is identical to that described in [3]. For this reason, we only briefly described in this section the simulation scenario and focus our attention on the specific implementation details associated with the new cepstral prefiltering method.

For the simulation we consider a rectangular room with uniform wall reflexion 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 $3.0m$ respectively. The source position is $(2.4835, 2.0, 1.8)m$ and the microphone positions are $(6.5, 2.8, 1.0)m$ and $(6.5, 3.8, 1.0)m$. The source signal $s[n]$ is obtained by passing a Gaussian white noise sequence through a band-pass filter with cut-off frequencies $f_l = 450$ and $f_u = 3475$Hz. The sampling frequency is $f_s = 10$kHz. Digital versions of the acoustic impulse responses $h_i[n]$ are obtained with the image method (properly modified for cardioid microphones). These responses are truncated to about 6000 samples $(0.6s)$. Even for the worst case considered, the truncated tail of the response is more than 40dB below the main peak corresponding to the direct path signal.

The frame size for the cepstral analysis is set to $K = 2048$ samples $(204.8ms)$. In theory, the frame size should be sufficiently large to avoid signal segmentation effects on the cepstrum computation. On the other hand, the frame size should not be too large if we want to be able to consider the channel as time-invariant over a few frames.

The exponential window coefficient α in (12) is set to 0.9985. A smaller value would increase the relative importance of the MPC but would also reduce the effective size of the frame. The optimal value of α is strongly dependent on the frame size K and was found empirically.

The memory parameter μ of the cepstral averaging (13) is set to 0.06. A larger value would result in a faster convergence, but the larger estimation error after convergence would make the cepstral prefiltering less advantageous. For the selected value of $\mu = 0.06$, the convergence time was of the order of $2s$. Note 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 procedure. 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 (13). The results presented in the next section were obtained with a frame overlap of 0%.

For the type of source signals considered here, the zero-mean assumption of the MPC of the source signal cepstrum was only noticeably violated for values of quefrency inferior to about 12. This value was found experimentally. Accordingly, we set the estimated MPC of the channel cepstrum in

(13) equal to zero for values of quefrency inferior to $k_1 = 12$.

For the given room configuration, 300 independent time delay estimates are calculated with the new GCC-CEP system and with the standard GCC method. The integration time T of the TDE procedure is set to $204.8ms$. Estimates for which the absolute error exceeds $3T_s$ $(0.3ms)$ are classified as anomalies. 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 are calculated.

6. RESULTS AND DISCUSSION

Typical results, as a function of the reverberation time, T_r, are shown in Fig. 1 to 3. The vertical bar superimposed on each data point represents the 95 percent confidence interval for that measurement. It can be seen from Fig. 1 that the cepstral prefiltering greatly reduces the percentage of anomalous time delay estimates. For instance, if we fix the maximum acceptable level of anomalies to 10 percent, we conclude that the cepstral prefiltering rises the acceptable T_r from 0.18s to 0.33s, which is not uncommon in audio-conference applications.

The bias and standard deviation of the non-anomalous estimates are illustrated in Fig. 2 and 3, respectively. We note the great reduction of bias when cepstral prefiltering is used for all values of T_r. We also notice 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 $T_r < 0.03$s is due to non-zero estimation errors of the channel MPC cepstrum in (13). However, such small values of T_r do not correspond to practical situations.

Note that the cepstral prefiltering procedure takes a finite amount of time before being effective. 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 too rapidly, the averaging procedure in (13) will be ineffective. Further research is still necessary to understand the effects of source motion on the performance of GCC-CEP. It might also be possible to improve the cepstral prefiltering in order to have faster convergence and to be able to track a rapidly moving source.

REFERENCES

[1] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West and M. M. Sondhi, "Autodirective microphone systems", *Acustica*, vol. 73, pp. 58-71, 1991.

[2] H. F. Silverman *et al.* , "A two-stage algorithm for determining talker location from linear microphone array data", *Computer Speech and Language*, vol. 6, pp. 129-152, 1992.

[3] S. Bédard, B. Champagne and A. Stéphenne, "Effects of room reverberation on time delay estimation performance", *Proc. IEEE Int. Conf. ASSP*, Adelaïde, Australia, April 1994, pp. 2.261-2.264.

[4] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*, Englewood Cliffs, NJ: Prentice-Hall, 1975.

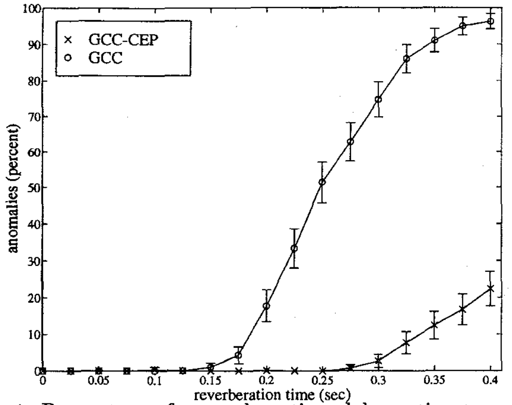

Figure 1: Percentage of anomalous time delay estimates versus reverberation time obtained with and without cepstral prefiltering.

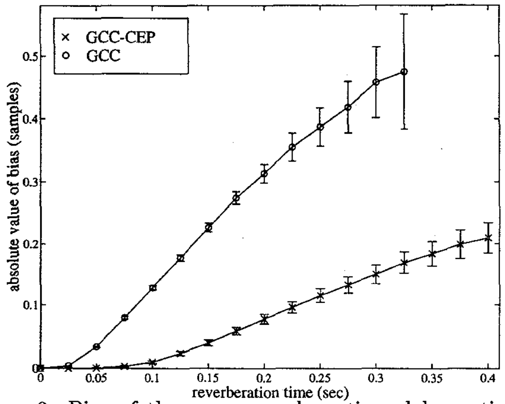

Figure 2: Bias of the non-anomalous time delay estimates versus reverberation time obtained with and without cepstral prefiltering.

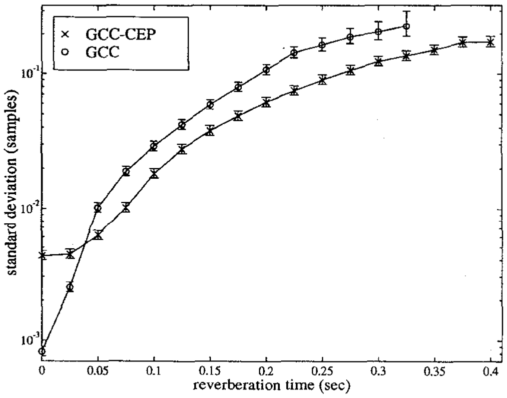

Figure 3: Standard deviation of the non-anomalous time delay estimates versus reverberation time obtained with and without cepstral prefiltering.