Subband Analysis of Time Delay Estimation in STFT Domain

S. Wang, D. Sen and W. Lu

School of Electrical Engineering & Telecommunications
University of New South Wales, Sydney, Australia
sh.wang@student.unsw.edu.au, dsen@ee.unsw.edu.au, wenliang.lu@student.unsw.edu.au

Abstract

For decades, time delay estimation (TDE) has been a significant issue in areas of radar, speech and audio processing. In recent years, the accurate estimation of TDE has been an important topic in multichannel audio compression where the inter-aural time delay (ITD) between each channel and a reference channel is a natural parameter used to represent multiple audio channels. This paper analyzes and evaluates various TDE algorithms in terms of their accuracy and computational complexity.

1. Introduction

The estimation of TDE also known as the Time Difference of Arrivals (TDOA) in various fields (Carter 1987) is used extensively in radar, sonar, auditory localization (Blauert 1983) and varieties of other signal processing and telecommunications applications. Recent research on perceptual compression of multichannel audio (Baumgarte and Faller 2003), requires TDE algorithms to compute the inter-aural time delay(ITD) within each critical band. The most popular TDE algorithms are variations of the concept of Cross-Correlation (CC) (Hertz 1986). Among these, Generalized Cross-Correlation (GCC) (Knapp and Carter 1976) has been well accepted as an efficient method, and possesses integer sample resolution. However, higher precision of subsample resolution is typically required in various applications including high-fidelity audio applications where the effect of reverberation and sound spatialization is highly dependent on correct representation of delay as a function of frequency. The accurate estimation of time delay within frequency subbands is challenging both in terms of accuracy and computational complexity.

This paper focuses on the sub-band analysis of TDE in the frequency domain and is aimed at improving the accuracy and complexity of the method used in the Binaural Cue Coding (Baumgarte and Faller 2003) algorithm. Several algorithms with subsample resolution are investigated. In the first section of the paper, we discuss the theoretical basis of methods, such as Circular Cross-Correlation (CXC) and Linear Regression Modeling, followed by detailed analysis and results in subsequent sections. Various test signals including single and multiple sinusoidal signals are used to test the methods.

2. Methodology

The following four techniques for TDE, all implemented in the frequency domain are investigated:

2.1. Cross-correlation (CC) in frequency domain

Consider two discrete signals $x[k]$ and $y[k]$. The signals may be the two channels from a stereo recording or two signals from opposite sides of a reverberant chamber. The signals can be expressed as:

$$x[k] = s[k] + n_1[k]; \tag{1}$$

$$y[k] = s[k-d] + n_2[k]. \tag{2}$$

where $s[k-d]$ is a signal achieved by delaying original signal $s[k]$ by d samples. The ambient noise can be modeled by using Additive White Gaussian Noise as $n_1[k]$ and $n_2[k]$. Thus the cross-correlation of the two signals, $x[k]$ and $y[k]$, is given by:

$$c_{xy}[\tau] = \sum_{-\infty}^{\infty} x[k]y[k+\tau]. \tag{3}$$

Since both $x[k]$ and $y[k]$ are real, the above equation can also be written as:

$$c_{xy}[\tau] = x[-k] \otimes y[k]. \tag{4}$$

meaning of course that the cross-correlation of above two signals can be expressed as a linear convolution. Hence, its Discrete Time Fourier Transform (DTFT) can be expressed in terms of $X(\theta)$ and $Y(\theta)$, which are the DTFTs of $x[k]$ and $y[k]$ respectively:

$$C_{xy}(\theta) = \mathbf{F}\{c_{xy}[\tau]\} = \overline{X(\theta)}Y(\theta) \tag{5}$$

, where $\overline{X(\theta)}$ stands for the complex conjugate of $X(\theta)$.

Thus the inverse DTFT, $c_{xy}[\tau]$ in time domain is given by:

$$c_{xy}[\tau] = \frac{1}{2\pi}\Re\{\int_{-\pi}^{\pi} \overline{X[\theta]}Y[\theta]e^{j\tau\theta}d\theta\} \tag{6}$$

The delay *d* corresponds to the maximum of $c_{xy}[\tau]$, and thus can be computed from Equation (6). This means that the above cross-correlation of the two signals can be evaluated in the frequency domain. The delay, '*d*' can be determined by:

$$d = \mathrm{argmax}\{\mathrm{c_{xy}}[\tau]\} \tag{7}$$

2.2. Subband analysis of TDE in frequency domain

In the case of Binaural Cue Coding and other parametric audio compression techniques, the time delay needs to be determined within a relatively small frequency range

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

rather than the entire frequency domain. (Baumgarte and Faller 2002). When replacing the DTFT with a Discrete Fourier Transform (DFT), this subband analysis implies that only a few frequency components are available per band to compute the delay. The above technique lends itself very easily to this subband analysis. Time domain cross-correlation would require the conversion of each subband to the time domain and thus would prove computationally cumbersome.

Consider a subband with boundaries A_l and A_h which represent DFT indices and where $(A_h > A_l)$. The subband is thus a frequency bin containing frequency components ranging from $(A_l/N)F_s$ to $(A_h/N)F_s$, where F_s is the sampling frequency and N is fame size or the length of DFT. From Equation (6), the cross-correlation for the subband can be written as:

$$c_{lh}[\tau] = \frac{1}{N}\Re\{\sum_{A_l}^{A_h-1} \overline{X[\theta_k]}Y[\theta_k]e^{j\tau\theta_k}\} \quad (8)$$

It should be noted that A_h, the upper boundary of the subband is omitted. From Equation (7), the time delay in the subband can be determined from the following equation:

$$d_{lh} = \text{argmax}\{\text{c}_{\text{lh}}[\tau]\} \quad (9)$$

2.3. Circular cross-correlation (CXC)

This second method which we call Circular Cross-Correlation (CXC) is a variation of the above Cross-Correlation in frequency domain technique. The method is derived in order to better the integer resolution of the delay estimation in the previous technique while keeping computational complexity penalties as low as possible.

In order to determine a time delay with non-integer (or sub-sample) resolution, it would generally require interpolation of the original signals. In other words, upsampling the original sequences by a factor of '*m*', before computing the cross-correlation.

Thus, the upsampled delay '$m{*}d$' would be estimated by:

$$m{*}d = \text{argmax}\{\text{c}_{\hat{\text{x}}\hat{\text{y}}}[\tau]\} \quad (10)$$

where $c_{\hat{x}\hat{y}}[\tau]$ stands for the cross-correlation of the upsampled signals, $\hat{x}[k]$ and $\hat{y}[k]$. Consequently, the precision is increased to $\frac{1}{m}$ samples when using this method.

If instead of computing the cross-correlation in the time domain, a frequency domain approach is taken, where we can arbitrarily shift one signal by introducing a phase delay, the method would still possess the higher precision than a simple cross-correlation method, while maintaining a much lower complexity compared to a interpolated time-domain approach. The technique may however be prone to circular shifts and our paper discusses whether this produces any appreciable error in the delay computation.

2.4. Non-circular cross-Correlation (NCXC)

This method is designed to avoid the effects of circular shifting in the previous method. while maintaining the same level of accuracy as the circular cross correlation (CXC) approach.

To avoid any circular shifts, we pad the original signals with zeros at the end. While eliminating circular shifts, this approach could result in a dramatic increment in computational complexity, especially when exchanging between time domain and frequency domain, as the signal's length is consequently increased due to the padded samples. Later sections of the paper show results which make it possible to discuss whether the extra computation is warranted.

2.5. Linear regression modeling (LRM)

The DFT of the above original signals, $x[k]$ and $y[k]$,can be represented with the following relationship in the frequency domain (assuming Gaussian noise $n_1[k]$ and $n_2[k]$ have relatively low energy spectrum in frequency domain and thus can be ignored):

$$x[k] \Longrightarrow X[\theta]; \quad (11)$$

$$y[k] \approx x[k-d] \Longrightarrow X[\theta]e^{-jd\theta}. \quad (12)$$

In other words, a time delay contributes changes to the phases spectrum rather than their magnitude spectrum. Thus, the time delay *d* can be derived as:

$$d = \frac{\partial\Psi[\theta]}{\partial\theta} \quad (13)$$

where $\Psi[\theta]$ is the phase difference between two signals.

Multiple Linear Regression methods can thus be employed to estimate the slope of curve $\Psi[\theta]$ to θ, which can also be regarded as the group delay between two signals. However, because of the limited number of samples (when the subbands represent critical bands, and coding-delay considerations preclude the usage of long time frames, only a couple of frequency samples are available at the lower end of the spectrum), errors can be significant when using this technique.

2.6. Zero-padded linear regression modeling (ZPLRM)

To alleviate the above problem of limited number of frequency samples while not imposing a coding-delay penalty, we consider zero padding the original sequences in time domain. This has the effect of interpolating samples in frequency domain. Hence, the precision is expected to improved, especially at low frequency region, at the expense of computational complexity.

3. Experiments, results and discussion

Sinusoidal signals, which could be processed by human ears, are fundamental elements of natural or artificial sound signals (SmithIII 2003) . As a result, signals composed of single or multiple sinusoids are employed to evaluate each of the above algorithms.

Additionally, sub-band boundaries A_b as discussed previously could be decided by either uniform bandwidth segmentation, as designed in MPEG-I (Pan 1995), or critical bands partition, as shown in Table 1. The latter, whose bandwidth approximates 2ERB, is designed by Faller (Faller and Baumgarte 2003). Other factors involved in this paper are assigned as following: original sampling

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 1: Critical Band Boundaries

A_0	A_1	A_2	A_3	A_4	A_5	A_6
0	2	4	7	11	15	20
A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
26	34	44	56	71	90	113
A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}
142	178	222	277	345	430	513

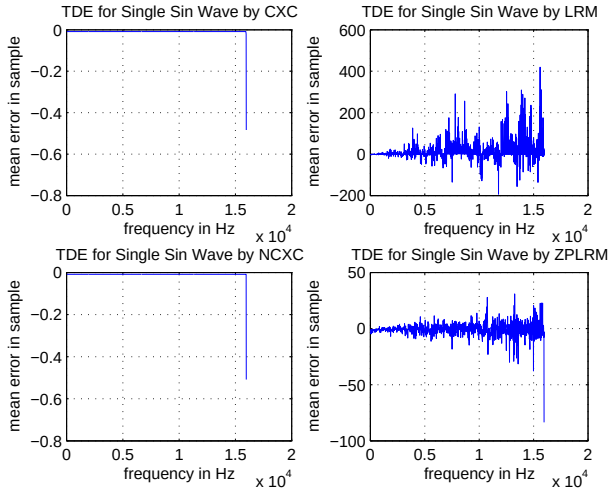

Figure 1: TDE for Single Sinusoid,
Mean Error vs Frequency, Uniform Subband BW

frequency $F_s = 32kHz$, signal length or Frames size $N = 1024$ and upsampling factor $m = 8$.

Delay is artificially introduced into our signals and is chosen from a random vector whose elements are real numbers varying between 0 and 2. The limits of this artificial delay are chosen with the highest subbands in mind to avoid phase ambiguity of the sinusoidal signals.

3.1. Tests on single sinusoids

For this experiment, the 1024 point DFT is divided into 32 subbands with uniform bandwidth. The experiment is carried out with the purpose of testing the performances of the different algorithms as a function of the frequency of the sinusoid. All four methods described above: CXC, NCXC, LRM and ZPLRM are tested. The frequency of the single sinusoidal is varied between $\frac{F_s}{N}$ to $512\frac{F_s}{N}$ by steps of $0.5\frac{F_s}{N}$. Thus the discrete spectrum of the sinusoid can be located exactly on an integer sample or between samples. Results are shown in Fig. 1, which contains curves of mean error versus frequency for each of the above four methods.

From Fig. 1 above, it is obvious that the algorithms, CXC and NCXC, are more accurate than the other two methods of evaluating different time delays. The mean error achieved by ZPLRM is much lower than that of LRM. Thus, as expected, zero padding does produce better performance. However, even with zero-padding, the linear regression methods do not compare with cross-correlation methods.

The sharp increment on mean error when using CXC and NCXC for extremely high frequencies can be explained

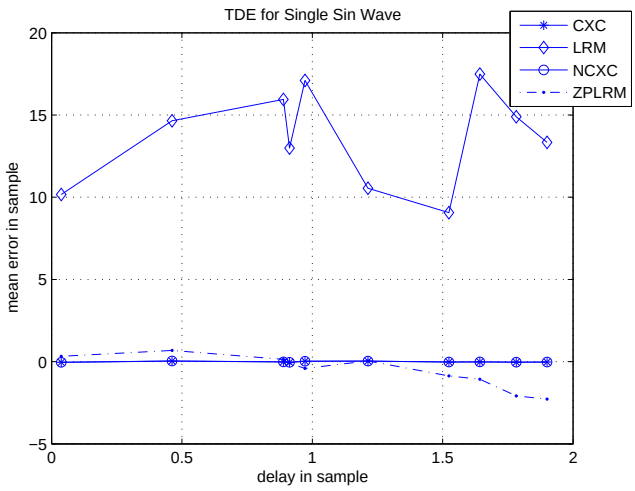

Figure 2: TDE for Single Sinusoid,
Mean Error vs Delay Values, Uniform Subband BW

by the periodicity of the cross-correlation. Since periods of high frequency sinusoids are close to the maximum delay value, the peak value is likely to be found in another period, which introduces such error at extremely high frequencies.

The error for LRM and ZPLRM tends to increase with increasing frequency and is a direct result of the limits of linear-regression for high frequency sinusoids.

The relationship between mean error and given delay value is shown in Figure 2. The performances of CXC and NCXC over various delay values are almost invariant. The results for LRM and ZPLRM also bear testament to the fact that the use of zero padding is capable of reducing estimation error, especially in methods using Linear Regression model.

Similar results are achieved when the subbands have non-uniform bandwidth given by Table1.

3.2. Tests on multiple sinusoids

Signals consisting of multiple sinusoidal signals are employed in this set of experiments. In addition to the same delay set as used in the previous section, several other conditions are addressed.

First, uniform bandwidth subbands are considered with 32 subbands. The sinusoid complex can either be chosen such that their frequency components correspond exactly to the DFT indices or alternatively, they can be chosen such they lie between the integer bins. Figure 3 represents the former case and shows similar results to the single sinusoid case presented in the previous section. Figure 4 presents the latter case with non-integer frequency components. For this case, the mean error is larger especially at low frequencies when using CXC, than any one of the previous cases. This is because the energy of each component is spread to adjacent frequency bins. This perhaps is a more realistic representation of what happens with natural audio signals whose frequency components will invariably lie on the frequency continuum rather than at distinct frequency bins.

A second case is considered, where the number of frequency components per subband are gradually increased to

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

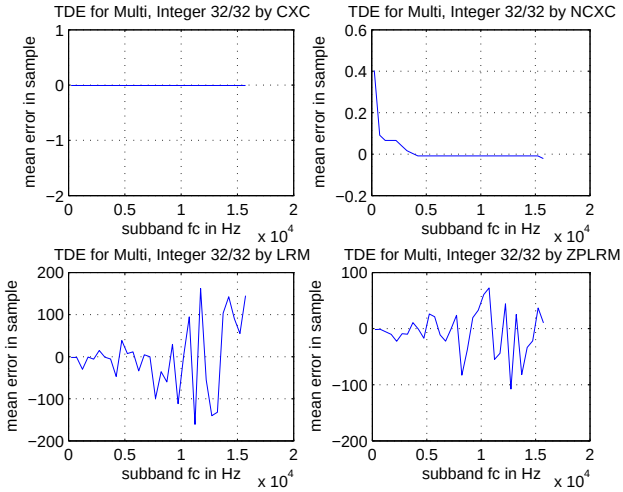

Figure 3: TDE for Multi-Sinusoids, Integer Samples Mean Error vs Subband Center Frequency fc

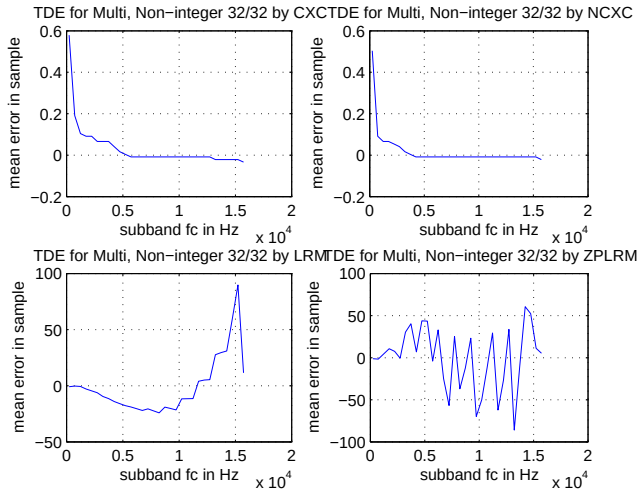

Figure 4: TDE for Multi-Sinusoids, Non-integer Samples Mean Error vs Subband Center Frequency fc

investigate the performance of different techniques depending on the number of frequency components per subband. As shown in Fig. 5, both CXC and NCXC maintain good TDE at relatively high frequencies above 4 kHz, while at low frequencies the mean error tends to increase with more sinusoidal components.

4. Computational complexity

The increase in length of the signals due to the use of upsampling results in a dramatic increase in computational complexity when compared to the simple cross-correlation (CC) method. The CXC method however does not require upsampling but produces sub-sample resolution time-delays meaning that the complexity of this algorithm is comparable with the CC method while producing higher accuracy.

Padding zeros in time domain to interpolate samples in frequency domain requires more computations and thus the NCXC is computationally more expensive compared with both CC and CXC.

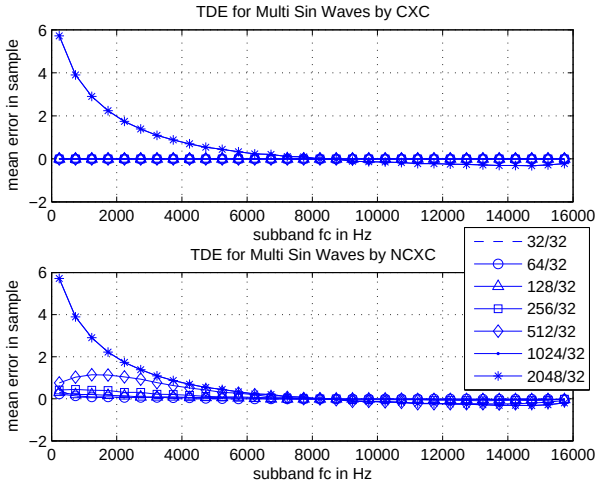

Figure 5: TDE for Multi-Sinusoids, Mean Error vs Subband Center Frequency fc

The same result could be observed between the LRM and ZPLRM methods, which means ZPLRM slightly improves the performance of LRM at the expense of computations.

5. Conclusions

In this paper, four different techniques of CXC, NCXC, LRM and ZPLRM, have been evaluated in terms of their ability to estimate time difference between two signals. The two signals are composed of additive sinusoids. The performance of the algorithms are compared in terms of their accuracy and computational complexity.

Generally speaking, CXC and NCXC perform better than methods of LRM and ZPLRM, which means Multiple Linear Regression might not be a suitable model for this application of estimating ICTDs for multichannel audio compression. By using CXC and NCXC, the precision can be improved to $1/m$ samples, where m is the upsampling factor. Additionally, zero padding in time domain, applied in NCXC and ZPLRM, is shown to reduce the mean error especially at high frequencies. Upsampling is also able to improve precision as expected. Zero-padding imposes a complexity penalty and thus the NCXC and ZPLRM algorithms are computationally more expensive, compared to CXC and LRM methods.

References

Baumgarte, F. and C. Faller (2003). Binaural Cue Coding–Part I: Psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing, 11*, 509–519.

Baumgarte, F. and C. Faller (May 2002). Estimation of Auditory Spatial Cues for Binaural Cue Coding. *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), 2*, 1801–1804.

Blauert, J. (1983). *Spatial Hearing. The Psychophysics of Human Sound Localization*. Cambridge, Mass.: MIT Press.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Carter, G. C. (1987). Coherence and Time Delay Estimation. *Proceedings of the IEEE, 75*, 236–255.

Faller, C. and F. Baumgarte (2003). Binaural Cue Coding–Part II: Schemes and applications. *IEEE Transactions on Speech and Audio Processing, 11*, 520–531.

Hertz, D. (1986). Time Delay Estimation by Combining Efficient Algorithms and Generalized Cross-Correlation Methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 34*, 1–7.

Knapp, C. H. and G. C. Carter (1976). The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 24*, 320–327.

Pan, D. (1995). A Tutorial on MPEG/Audio Compression. *IEEE Multimedia Journal, Summer*, 60–74.

SmithIII, J. O. (2003). *Mathematics of the Discrete Fourier Transform (DFT), with Music and Audio Applications.* Menlo Park, California: W3K Publishing.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review