Using Delay Estimation to Reduce Comb Filtering of Arbitrary Musical Sources

ALICE CLIFFORD,[1] *AES Member,* **AND JOSHUA D. REISS,**[1] *AES Member*
(alice.clifford@eecs.qmul.ac.uk) (josh.reiss@eecs.qmul.ac.uk)

Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

Comb filtering occurs when a signal is summed with a delayed version of itself. This can occur in live or studio sound production when multiple microphones reproduce a single source. The delay between microphone signals can be estimated using signal processing techniques and the signals aligned by applying a compensating delay. Accurate delay estimation is important for comb filter reduction as errors will lead to flanging effects on the input sources. This paper offers a novel analysis of the accuracy of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) delay estimation technique when applied to arbitrary music signals whereas previous research is mostly concerned with speech signals.

We show that the performance of GCC-PHAT is dependent on the choice of window used for the Discrete Fourier Transform calculation and, for poor choice of window, may also be highly dependent on the bandwidth of the incoming signal. This has not been explored previously in the literature. Analysis is provided that shows that the side lobe characteristics affect the accuracy of delay estimation, and windows that taper to 0 will provide the highest accuracy. The derived results are further confirmed through analysis and experimentation with simulated and real signals. In particular, the Hann or Blackman windows offer the highest performance for a variety of musical signals, with over 90% accuracy for frame sizes over 256 samples, and are unaffected by input signal bandwidth.

0 INTRODUCTION

A common technique in live and studio production is to use multiple microphones to reproduce a single source, for example as in Fig. 1. This commonly occurs with instruments such as guitar and pianos to get an accurate reproduction of different aspects of a single instrument that then gives the sound engineer the flexibility to mix different microphone signals together to produce the sound of the instrument they specifically want.

It is difficult, and often undesired, to place the microphones equidistant from the sound source, therefore the sound from the instrument will arrive at each microphone at a different time. When the microphones are mixed, this is equivalent to summing a signal with a delayed version of itself, which is known to cause comb filtering.

It is possible to reduce the effect of comb filtering by applying a compensating delay to one of the microphone signals to give the impression the source is arriving at each microphone at the same time. This is traditionally done by ear until the "phasiness" is reduced or by measuring the distances between sources and microphones and calculating the difference in delays. With modern audio editing software it is also possible to manually nudge audio regions in line by eye or by ear. The problem with these methods is they are unlikely to be accurate. Assuming a sampling frequency of 44.1 kHz and a speed of sound of 344 m/s, one sample delay is enough to cause a comb filter that is a 1st order low pass filter. This is equivalent to a difference in source to microphone distance of just 0.0078 m. Therefore, sample-accurate manual delay correction is almost impossible. Adjusting delays by ear means that the comb filtering may appear to be reduced for the sample of audio you are listening to but if the audio changes, for example if an instrument plays a different range of notes, the comb filtering could reappear in the frequency range of the new set of notes. Estimating delays by measuring distances has its own problems as the speed of sound is not constant and can easily be changed by temperature and humidity [1]. In both cases if the source moves, the delays will change and comb filtering will once again occur.

For accurate delay estimation signal processing has to be employed using time delay estimation, or time difference of arrival, methods. A number of methods have been proposed and an overview can be found in [2]. This paper is concerned with the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [3], which is a common method in microphone array signal processing for beam forming

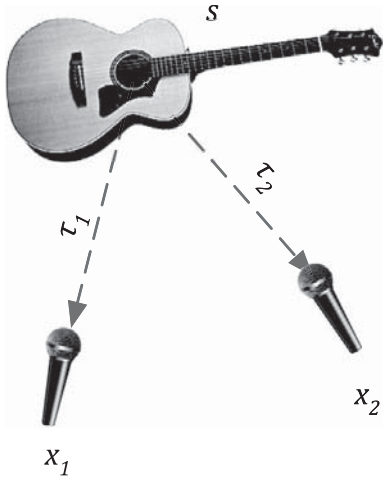


Fig. 1. A common layout for reproducing a single source s with multiple microphones x_1 and x_2.

and source localization [4], mostly for speech communication [5]. It has been adopted due to its low computational complexity and ability to have different weightings applied for different uses, such as the Phase Transform. The GCC-PHAT allows sample accurate delay estimation while also being able to track moving sources on a block-by-block basis.

When using the GCC-PHAT for comb filter reduction high accuracy is needed. If errors occur there is a risk of introducing a flanging effect onto the source signal as the delay compensation changes rapidly. There is a wide body of research on this method, and testing its abilities in noisy and reverberant environments, or with additional uncorrelated or correlated noise and the GCC-PHAT is generally considered adequate in both cases [6]. There is little research in the literature investigating how other properties such as the window shape used and the input signal affects the accuracy.

More recently delay estimation has been extended to musical settings , for example in loudspeaker system alignment [7], system measurement tools [8], as well as proposals for use in comb filter reduction [9,10, 11]. Work in [7] details considerations that need to be taken when using arbitrary signals, instead of traditional noise sources, for transfer function calculation, such as averaging, accumulation, coherence measurement, and noise reduction. Many of these techniques are applicable to delay estimation of musical signals but have not been applied to this problem.

Other delay estimation techniques include Adaptive Eigenvalue Decomposition [12] and Least Mean Square estimation [13] that are based on using adaptive filters to converge to a solution. These methods are more computationally complex and do not provide a significant increase in accuracy [14].

This paper examines how the accuracy of GCC-PHAT changes depending on the bandwidth of the incoming signal, which is unknown prior to calculation and how the window function used in the GCC-PHAT calculation plays an important part in achieving high accuracy of delay estimation for comb filter reduction. It expands on previous work by the authors in [11] by providing a theoretical explanation of the effect of window size and bandwidth on delay estimation accuracy and expands the experimental analysis to a wider variety of instruments.

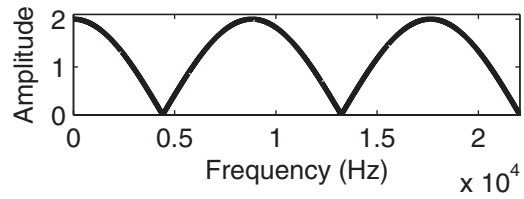


Fig. 2. Transfer function of a comb filter with a relative delay of 5 samples at 44.1 kHz sampling rate.

1 BACKGROUND

1.1 Comb Filtering

Comb filtering occurs when a signal is summed with a duplicated, delayed version of itself. The sound of a comb filter is usually described as being "phasey," most likely due to the fact that comb filtering forms the basis of flanging and phasing effects. In music production comb filtering can also occur when audio is duplicated, processed, and mixed with the original signal, such as recording a guitar both direct and through an amplifier and microphone. Additionally, it can occur when stereo recordings are mixed to mono.

A single source, s being reproduced by two microphones x_1 and x_2, as in Fig. 1 can be described as

$$x_1[n] = s[n - \tau_1] \tag{1}$$

$$x_2[n] = s[n - \tau_2] \tag{2}$$

where n is the current time step and τ_1 and τ_2 are the delays associated with the sound source traveling from the source position to the position of x_1 and x_2. Uncorrelated noise, reverberation, and attenuation due to distance are not considered. When the microphones are summed to become x, in terms of s this is

$$x[n] = s[n - \tau_1] + s[n - \tau_2]. \tag{3}$$

It can also be stated that

$$x_2[n] = x_1[n - \tau] \tag{4}$$

assuming $\tau_2 > \tau_1$ where $\tau = \tau_2 - \tau_1$.

Comb filtering is so called due to the "comb" shaped frequency response it produces, as seen in Fig. 2. It is characterized by the peaks and troughs associated with the filter that occur due to the cancellation and reinforcement of frequencies along the audible spectrum. As a signal is delayed in time, all frequencies are delayed by the same time, which results in a linear phase shift across the spectrum, causing some frequencies to cancel and others to reinforce. The period of this reinforcement and cancellation is directly related to the amount of delay that is occurring.

Amplitude difference between the microphones signal also changes the frequency response of the filter. Equal amplitude will result in complete rejection at the troughs

whereas if the delayed signal is of a lower amplitude than the direct signal, the filter will be less severe. Previous research suggests comb filtering can be heard when the delayed signal is as much as 18 dB lower in amplitude than the direct signal [15].

1.2 Delay Estimation Using GCC-PHAT

It is not possible to estimate τ_1 and τ_2 directly from Eq. (2) without any prior knowledge of s. Delay estimation methods are commonly referred to as time difference of arrival as it is only possibly to estimate τ, the relative delay of a source between microphones.

The Generalized Cross Correlation, or GCC, is defined by

$$\Psi_G[k] = X_1^*[k] \cdot X_2[k] \tag{5}$$

in the frequency domain and

$$\psi_G[n] = \mathcal{F}^{-1}\{\Psi_G[k]\} \tag{6}$$

in the time domain where $\mathcal{F}^{-1}$ is the Inverse Fourier Transform, X_1 and X_2 are x_1 and x_2 in the frequency domain, $k = 0, \ldots, N-1$ is the frequency bin number, and |*| denotes the complex conjugate. The delay, τ, is estimated by finding the position of the maximum of the output function, where

$$\tau = \arg\max_n \quad \psi_G[n]. \tag{7}$$

It is well known that the GCC is susceptible to uncorrelated noise and reverberation that can reduce the accuracy of the estimation and it is an open problem to improve the robustness of the method [2,6, 16,17]. An accurate and stable estimation of delay is imperative to reduce errors in the subsequent usage of the estimation. For example it is important in comb filter estimation as sudden changes in the estimated delay produce audible artifacts. There are a variety of weighting functions suggested in the literature. The most commonly used is the Phase Transform, which has been shown to improve performance in noisy and reverberant conditions [18,19]. The Phase Transform uses only the phase of the GCC in the frequency domain to become the GCC-PHAT. Therefore Eq. (6) becomes

$$\Psi_P[k] = \frac{X_1^*[k] \cdot X_2[k]}{\left|X_1^*[k] \cdot X_2[k]\right|} \tag{8}$$

in the frequency domain and

$$\psi_P[n] = \mathcal{F}^{-1}\{\Psi_P[k]\} \tag{9}$$

in the time domain. The delay is estimated by

$$\tau = \arg\max_n \psi_P[n]. \tag{10}$$

The GCC-PHAT calculates the difference in phase between each microphone signal in the frequency domain before being transformed back to the time domain to estimate the delay. This method is used because the delay between two signals is contained within the phase difference. The shift theorem states that when a signal is delayed, a linear phase component is added. The slope of the linear phase is equal to the delay, otherwise known as group delay. The Discrete Fourier Transform X_2 of the microphone signal x_2 is

$$X_2[k] = \sum_{n=0}^{N-1} w[n]x_2[n]e^{-j\omega_k n} \tag{11}$$

where $\omega_k = 2\pi k/N$ and w is a window function. Assuming a rectangular window function where $w[n] = 1$, using Eq. (4) this becomes

$$X_2[k] = \sum_{n=0}^{N-1} x_1[n-\tau]e^{-j\omega_k n} \tag{12}$$

$$= e^{-j(n-\tau)\omega_k} X_1[k] \tag{13}$$

The term $e^{-j(n-\tau)\omega_k}$ is the linear phase $\Phi[k]$ introduced to the output spectrum which can be calculated by

$$\Phi[k] = arg(X_2[k]) - arg(X_1[k]) \tag{14}$$

that is also performed in Eq. (9). It should be noted that this is equivalent to estimating the impulse response and applying the PHAT, which is the technique recommended in [7].

Techniques exist to estimate the delay simply by calculating the gradient of the linear phase term [20]. This technique is highly susceptible to uncorrelated noise and requires smoothing of results. Other methods exist for using just the phase to estimate the delay [21,22] although these have been shown to exhibit poor performance. Work in [23] outlines a method for estimating delay using a combination of frequency content and phase offset but is specific to a certain type of signal.

Studies in [24] and [25] suggest that with a harmonic input signal the Phase Transform is detrimental to the delay estimation accuracy and outline a method for varying the degree in which the Phase Transform is applied, depending on how harmonic the signal is. We address this claim and it is discussed with analysis in Section 3.

1.3 Windowing

The GCC-PHAT is still commonly used in the same form as when first introduced in [3]. It has consistently been shown to perform adequately, and therefore no significant adaptations of the algorithm have been widely accepted.

The main variables that can be changed in the algorithm are the weighting function, window shape, window size, and hop size. This paper uses the Phase Transform weighting function. The window shape used with the DFTs in the GCC-PHAT has not been discussed in the literature and is an important, often overlooked stage of the calculation. This section proceeds to investigate the effect different window shapes have on delay estimation and how this relates to musical signals.

The GCC-PHAT requires that the Discrete Fourier Transform (DFT) of each microphone signal is calculated over a discrete window of data. It is common for the data to be weighted with a function such as the Kaiser or Hamming window. A survey of the literature on delay estimation suggests no justification for the window function chosen. Research into speech source localization [20] uses phase

differences to calculate delay and mentions the use of a Hann window in preceding work [26]. An overview of delay estimation methods [2] uses the Kaiser window for the cross correlation. Other works use the Hann window [9,27] or the Hamming window [28] without justification. Work into the differences on perception of synthesized speech using either magnitude or phase spectrum [29] compares two window functions, rectangular and Hamming. The GCC-PHAT relies on accurate phase measurement, but this work does not provide an explanation for how the Hamming window changes the phase and therefore alters the result compared to the rectangular window. Other examples using the GCC-PHAT in the literature do not describe the window function used.

As each window function has its own characteristics, such as the type of spectral leakage that occurs, this may affect the delay estimation, and the window function should not be an arbitrary decision. A theoretical study of the effect of window function on delay estimation [30] leads to the conclusion that the error is independent of the window, if the window is sufficiently wide. In reality, the window size is restrained by computation and sufficiently large windows are not necessarily available. It also does not discuss the effect that the input signal has on delay estimation. Other work investigates the effect window side lobes have on multifrequency signal measurement [31] but does not detail how this affects the phase, which is significant when discussing time delay. In this paper we provide a novel theoretical and experimental analysis of the effect of window shape on delay estimation accuracy with real, arbitrary musical signals.

Section 2 of this paper investigates the considerations that need to be taken to gain maximum efficiency from the GCC-PHAT through window shape selection and how the input signal affects the accuracy of the delay estimation. Section 3 provides an analysis of the theory on window shape selection using simulated and real recordings. Section 4 concludes the paper and offers recommendations on the best practices for performing delay estimation of musical signals.

2 WINDOWING AND SIGNAL BANDWIDTH

As mentioned previously, the GCC-PHAT estimates the linear phase shift between X_1 and X_2 with the individual phase shift θ_k of each frequency bin k linearly related with the sample delay τ. Taking Eq. (8) and assuming X_1 and X_2 are full bandwidth signals with significant data for all k, the phase difference using the GCC-PHAT then becomes

$$\Psi_P[k] = e^{j\theta_k} = e^{-j\omega\tau} \tag{15}$$

The inverse DFT yields the final result

$$\psi_P[n] = \frac{1}{N}\sum_{k=0}^{N-1} e^{-j\omega\tau} e^{jn\omega_k} \tag{16}$$

$$= \frac{1}{N}\sum_{k=0}^{N-1} e^{j(n-\tau)(\omega-\omega_k)} \tag{17}$$

$$= \begin{cases} 1 \text{ if } n = \tau \\ 0 \text{ if } n \neq \tau \end{cases} \tag{18}$$

that is equal to Eq. (9) and the delay can be accurately estimated as τ. For Eq. (15) to hold, θ_k has to be correct for all values of k.

A real signal, such as musical signals, will not be full bandwidth. Different instruments produce notes that occupy different areas of the frequency spectrum. Percussive instruments may produce a more noise-like sound that occupies a large part of the spectrum whereas a harmonic instrument, such as a flute, will only produce harmonics of a fundamental frequency. There will also be a limit to the range of notes it can produce and therefore the fundamental frequency.

In the extreme case of this, taking a single complex sinusoid $s = e^{j\omega n}$ where $\omega = 2\pi l/N$, l is an integer $0 \leq l < N$ and $s_\theta = e^{j(\omega n+\theta)}$ we know from the shift theorem that

$$S_\theta[k] = e^{j\theta} S[k] \tag{19}$$

where S is s in the frequency domain. S will have a single non-zero value when $k = l$. Hence when $k \neq l$

$$\frac{S_1^*[k] \cdot S_2[k]}{\left|S_1^*[k] \cdot S_2[k]\right|} \neq e^{j\theta} \tag{20}$$

as this leads to division by 0 therefore it is undefined.

The delay cannot be simply estimated from the value of θ as this is only correct for when $k = l$ so gives no context as to the slope of the phase and thus the corresponding delay in samples.

In Eq. (19), s is assumed to contain an integer number of periods within N. Spectral leakage occurs when the input signal contains a non-integer number of periods within the window. This is often the case with real signals. The result of this is that for a single sinusoid the frequency domain signal is no longer a delta function but resembles the frequency spectrum of the particular window.

The spectral leakage also implies that all values of k will be defined, which is not the case in Eq. (20). If $s = e^{j\omega n}$ where $\omega = 2\pi l/N$ and l is not an integer then all k will be defined and the GCC-PHAT can be calculated. Despite this, the correct delay will still not be estimated as the phase from the nearest value of k to l will spread into neighboring bins. If $\theta_k = \theta$ for all k due to the leakage, Eq. (15) does not hold. As θ_k is a single value, the slope is 0. Therefore the delay estimate is 0, which is incorrect.

The more values of θ_k that are the correct estimate of real phase difference, the more likely the estimation of delay will be correct. The errors are caused by spectral leakage and become more apparent when considering a real signal as a sum of sinusoids at different amplitudes and frequencies. This is due to the interference between side lobes of high amplitude sinusoids and low amplitude sinusoids that is also known to effect multifrequency signal measurement [31]. If a sinusoid is of lower amplitude than the side lobe of another sinusoid in the frequency domain it will be distorted or completely masked in both magnitude and phase.

It stands that if the bandwidth of the signal is increased, with more higher amplitude sinusoids, more values of θ_k will be correct. Equally, if the side lobes are lower amplitude either by the window shape producing lower maximum amplitude side lobes or having a steeper side lobe roll off rate, then less lower amplitude side lobes will be masked and accuracy will be improved.

From this we hypothesise that delay estimation accuracy is dependent on the incoming signal bandwidth and the characteristics of the window shape chosen.

3 ANALYSIS

This section outlines an experimental analysis with simulated and real musical signals of how the bandwidth of the input signal and the window used when performing the GCC-PHAT affects the accuracy of the subsequent delay estimation.

3.1 Bandwidth Limited White Noise

The variation between musical signals in the frequency domain can be simplified as stating that different instruments will produce sounds that occupy different areas of the frequency spectrum with different bandwidths. The effect this has on the GCC-PHAT can be observed under controlled conditions, not taking into account amplitude or temporal changes, by using filtered white noise as an input signal. This was used as an input to simulate microphone signals by duplicating the filtered input signal and delaying the duplicate by 10 samples at 44.1 kHz sampling rate. The audio excerpts were 10 seconds in length.

The white noise was filtered using low pass, high pass, and band pass 4th order Butterworth filters centered at 11.25 kHz to investigate whether the centroid of the spectrum altered the accuracy. For each execution of the simulation the bandwidth of the 3 filters was altered. In the case of the low and high pass filters the cut off frequency was altered to achieve the desired bandwidth. The bandwidth of each filter was then varied between 50 Hz and $\frac{F_s}{2}$ where F_s is the sampling frequency. The delay was estimated at each execution with the GCC-PHAT using 7 of the most common window shapes: Blackman, Blackman-Harris, Flat Top, Gaussian, Hamming, Hann, and rectangular, with a frame size of 2048 samples. The accuracy is determined as a percentage of frames over the 10 second sample in which the delay was estimated correctly with an error of ± 2 samples.

Fig. 3 shows the results using the rectangular window. It can be seen that for all filters at the same bandwidth the results are similar and the point at which 100% accuracy is achieved is the same for all filters. This leads to the conclusion that the centroid of the spectrum has only a minor effect on the accuracy of delay estimation. Therefore the low pass filter results are used for the analysis in the rest of the paper.

Fig. 4 shows the results for all windows tested for the low pass filter with increasing bandwidth. This shows that each window offers a different level of performance in delay

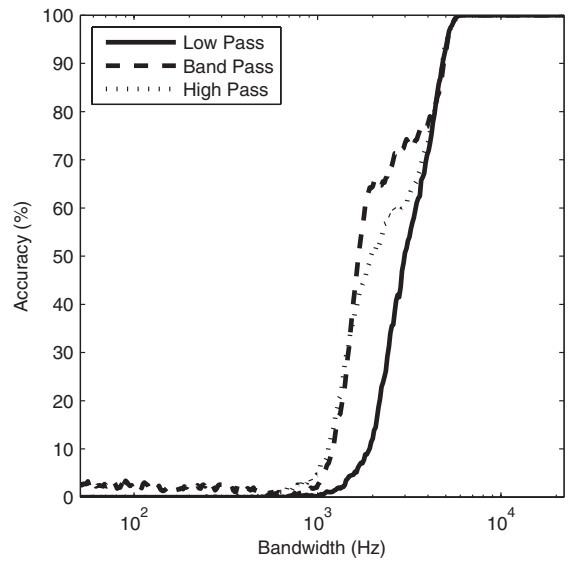


Fig. 3. Accuracy of delay estimation as a percentage of correct frames with an error of ± 2 samples using a rectangular window with increasing bandwidth using low pass, high pass and band pass filter centred at 11.25 kHz.

estimation and therefore the choice of window should not be trivial. The rectangular window reaches 100% accuracy at a bandwidth of 5937 Hz, whereas the Blackman window reaches 100% accuracy at a bandwidth of 128 Hz. The accuracy increases as bandwidth increase for all window shapes.

Table 1 shows the mean accuracy for each window shape over all bandwidths ranked in descending order from most accurate to least accurate. The side lobe height, side lobe roll-off, and start and end values are also shown. The window shapes with a 60 dB/decade side lobe slope outperform the windows with 20 dB/decade slope. The Blackman window also appears more accurate than the Hann window by 4% since it has a lower side lobe maximum height. The accuracy of the windows that do not taper to 0 then decreases according to the start value. This confirms the hypothesis that windows with a steeper side lobe roll off slope or lower side lobe maximum height result in higher accuracy.

To explain this further, Fig. 5 shows the GCC-PHAT output using a rectangular window and equivalent phase spectrum for white noise low pass filtered with a cut off frequency of 1000 Hz using a 4th order Butterworth filter and unfiltered white noise delayed by 10 samples. Fig. 5a shows the GCC-PHAT output of the low pass filtered and unfiltered white noise. The unfiltered GCC-PHAT shows a very clear peak at the delay value of 10 samples. The filtered GCC-PHAT has a peak at the correct delay value but also a peak at 0, which is the maximum and therefore the estimated delay. It is not possible to simply ignore the values at $\tau = 0$ when performing the GCC-PHAT as it is possible that no delay occurs and these need to be estimated. This is explained by examining the corresponding phase spectrum in Fig. 5b. The unfiltered example shows a distinct linear phase whereas the filtered example shows linear phase for

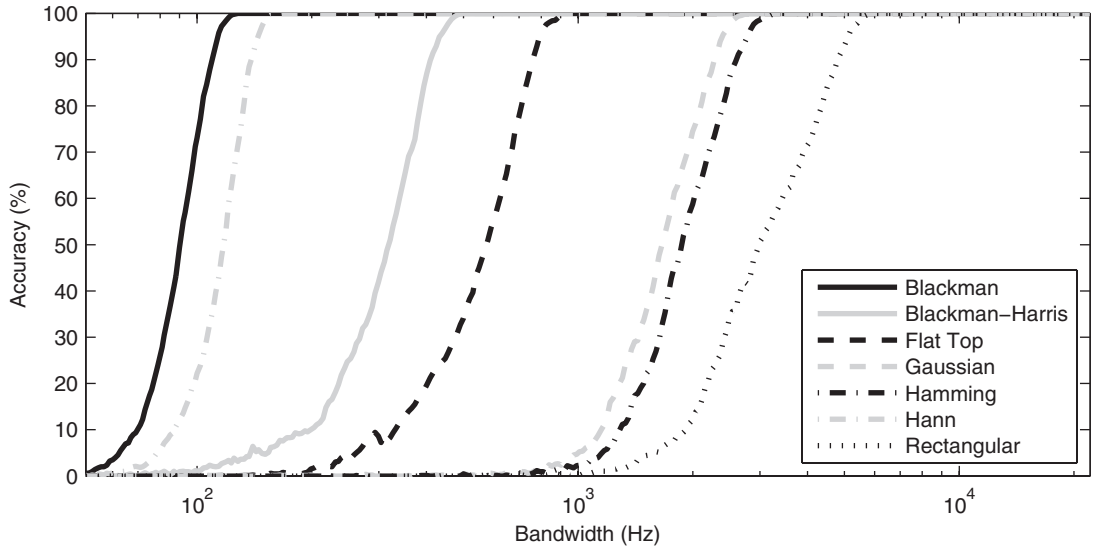
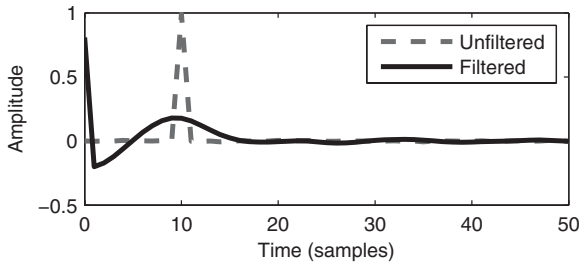
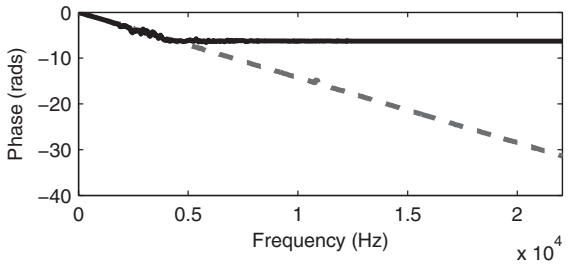


Fig. 4. Accuracy of delay estimation as a percentage of correct frames with an error of ±2 samples using a selection of windows with increasing bandwidth using a low pass filter.



(a) GCC-PHAT output of white noise



(b) Phase spectrum of white noise

Fig. 5. The GCC-PHAT output and corresponding unwrapped phase spectrum of unfiltered and low pass filtered white noise.

the pass band of the filter, up to 1000 Hz, but in the cut band of the filter the phase is horizontal, corresponding to the significant 0 peak in the GCC-PHAT output. This is a result of the higher amplitude spectral leakage of the rectangular window. With the Blackman or Hann windows this does not occur and hence the GCC-PHAT output is the same for both filtered and unfiltered signals.

3.2 Real Recordings

The window shapes being evaluated were tested on real recordings. The recordings were made using two microphones placed at arbitrary distances from a loudspeaker to incite a delay between the microphone signals and were recorded in an acoustically treated recording studio. In this paper we assume the sources are point sources to primarily investigate the effect of source bandwidth on delay estimation accuracy rather than the effect of different instrument sound transmission.

The microphone signals were analyzed using the GCC-PHAT with various window shapes. Twenty different musical audio samples were tested, each of 30 seconds in length. The audio samples were a selection of instrument recordings that occupy different frequency ranges.

Table 1. Mean accuracy over all filter bandwidths for low pass filtered noise for each window shape showing window features.

Window	Mean accuracy (%)	Maximum side lobe height (dB)	Side lobe roll-off (dB/decade)	Start/end value
Blackman	90.74	−58.1	60	0
Hann	86.67	−31.5	60	0
Blackman-Harris	71.00	−71.5	20	6.00×10^{-5}
Flat Top	61.34	−93.6	20	-4.2×10^{-4}
Gaussian	43.00	−43.3	20	4.3×10^{-2}
Hamming	40.82	−42.7	20	0.08
Rectangular	32.85	−13.3	20	1

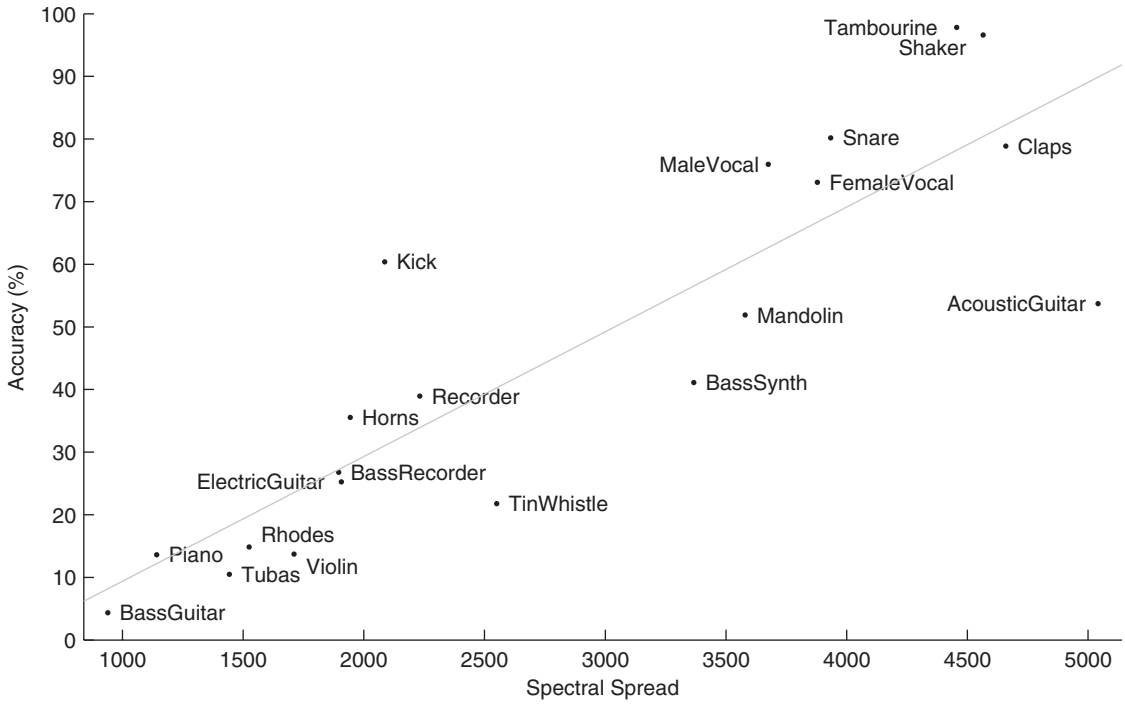


Fig. 6. Delay estimation accuracy for 20 audio excerpts using a rectangular window plotted against spectral spread.

The bandwidth of each audio sample was measured by calculating spectral spread, or standard deviation, defined by

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}(|X[k]| - \mu)^2} \tag{21}$$

where

$$\mu = \sum_{k=0}^{N-1}|X[k]|. \tag{22}$$

and X is the input signal x in the frequency domain. The spectral spread was estimated over the whole duration of the audio sample.

Figs. 6 and 7 show the accuracy of delay estimation for each audio sample plotted against the spectral spread. Fig. 6 shows the results of delay estimation using the rectangular window and Fig. 7 the results using the Hann window. In Fig. 6 it is apparent that as the spectral spread (and thus the bandwidth of the signal) increases the accuracy of the delay estimation increases. As expected, this is not the case for the Hann window, which gives the better performance for all test audio samples, although 100% accuracy is not achieved due to the recording environment.

This can be further explained by analyzing the estimation data over time of different inputs. Figs. 8a and 8b show the output of the GCC-PHAT using a rectangular window showing the delay estimation for each frame of data of two example audio samples, a bass guitar and an acoustic guitar. The estimation for the bass guitar is inaccurate with the correct delay rarely being estimated and an estimate of 0 is more likely. This is due to that shown in Fig. 5. In comparison, the acoustic guitar estimates a delay of either 0 or the correct delay per frame. All signals processed with the Hann window show an improvement in accuracy toward 100%.

Fig. 9 shows the mean estimated delay and standard deviation over the entirety of each audio excerpt. For the Hann window the mean delay for every instrument is within 1 sample of the correct value, indicated by a horizontal dashed line. In the rectangular window case, only the shaker and tambourine audio excerpts result in a mean delay within 1 sample of the correct value. This agrees with the result in Fig. 6, where these excerpts exhibit the highest accuracy in the rectangular window case.

From Fig. 9 it can be seen that the standard deviation for the rectangular window is higher for every instrument under test than the Hann window. There is a mean decrease in standard deviation using the Hann window over the rectangular window of 26.65. This means the spread of estimated delays for the Hann window is much smaller and around the correct mean, thus showing the accuracy of the Hann window is higher than the rectangular window but that also the error of the Hann window is lower.

Fig. 10 shows the mean accuracy of all 20 test recordings for frame sizes from 128 samples to 8192 samples for each window shape. There is a general trend of increasing

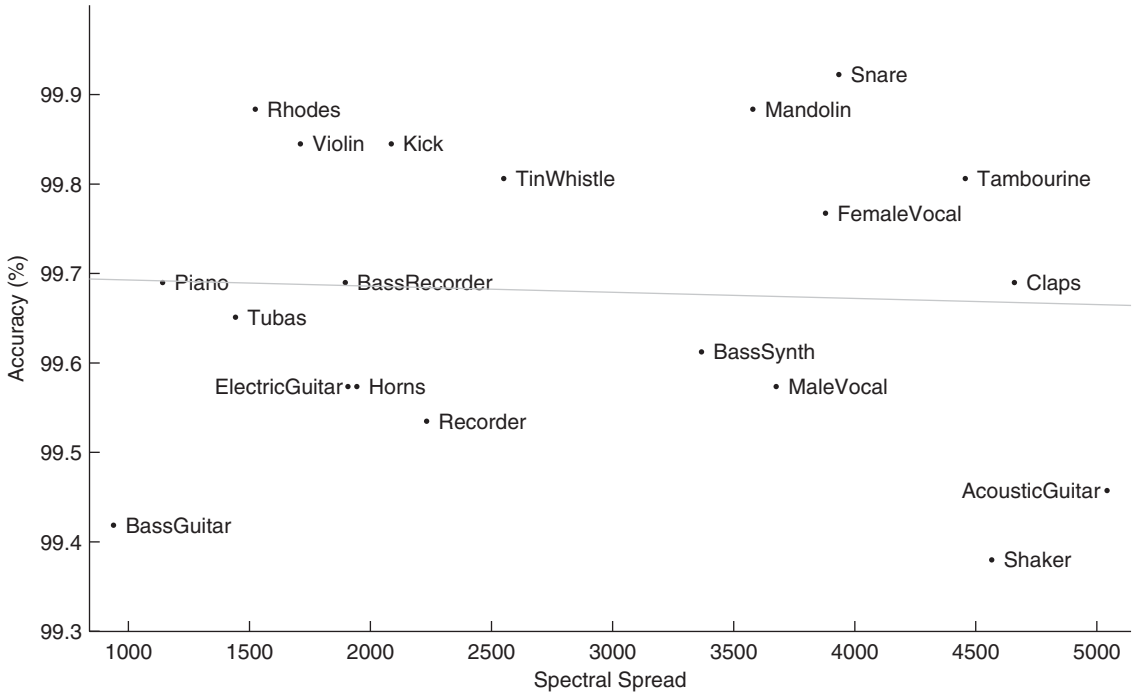
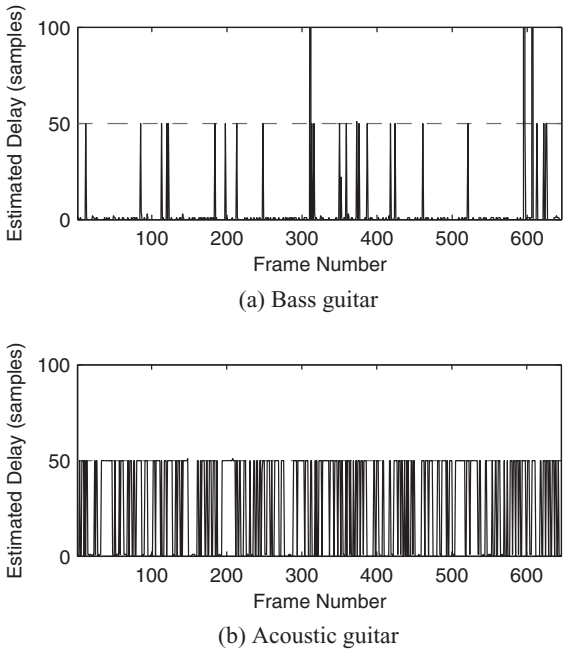


Fig. 7. Delay estimation accuracy for 20 audio excerpts using a Hann window plotted against spectral spread showing accuracy $> 99\%$.



(a) Bass guitar

(b) Acoustic guitar

Fig. 8. Output of the GCC-PHAT using the rectangular window shown as delay estimation for each frame of data. The dashed horizontal line indicates the correct delay.

accuracy as frame size increases. This is expected since with increasing frame size there is more data available to perform the GCC-PHAT. But the differences in performance from each window remains even at large frame sizes.

Table 2 shows the mean of all frame sizes for each window. The results follow a similar trend as that for the filtered white noise. The Hann and Blackman windows provide the highest accuracy with a side lobe roll of 60 dB/decade followed by windows with low amplitude side lobes. The rectangular window continues to perform the worst.

The MATLAB code and audio data for the analysis are freely available.[1] The audio data is available under a Creative Commons license.

4 CONCLUSION

This paper has investigated the effect of using the GCC-PHAT to estimate the delay between microphone recordings of the same musical source for use in alignment and comb filter reduction. It has been shown that considerations need to be taken into account when applying delay estimation to musical signals as opposed to speech signals and recommendations have been made for best practice to achieve the highest accuracy. Prior research is focused on speech signals and does not address inaccuracies in delay estimation using the GCC-PHAT with a variety of input signals and arbitrary DFT window shape. This is important for comb filter reduction as errors in delay estimation will cause further flanging effects on the input sources.

We have shown that the window function used during the GCC-PHAT calculation plays a large role in the ultimate performance of the method, which had not previously been examined in the literature. This is due to the interference between frequency components with different amplitudes caused by spectral leakage, leading to errors in the GCC-PHAT calculation. This interference is greatest when the

[1] https://code.soundsoftware.ac.uk/projects/gccphat-windowing

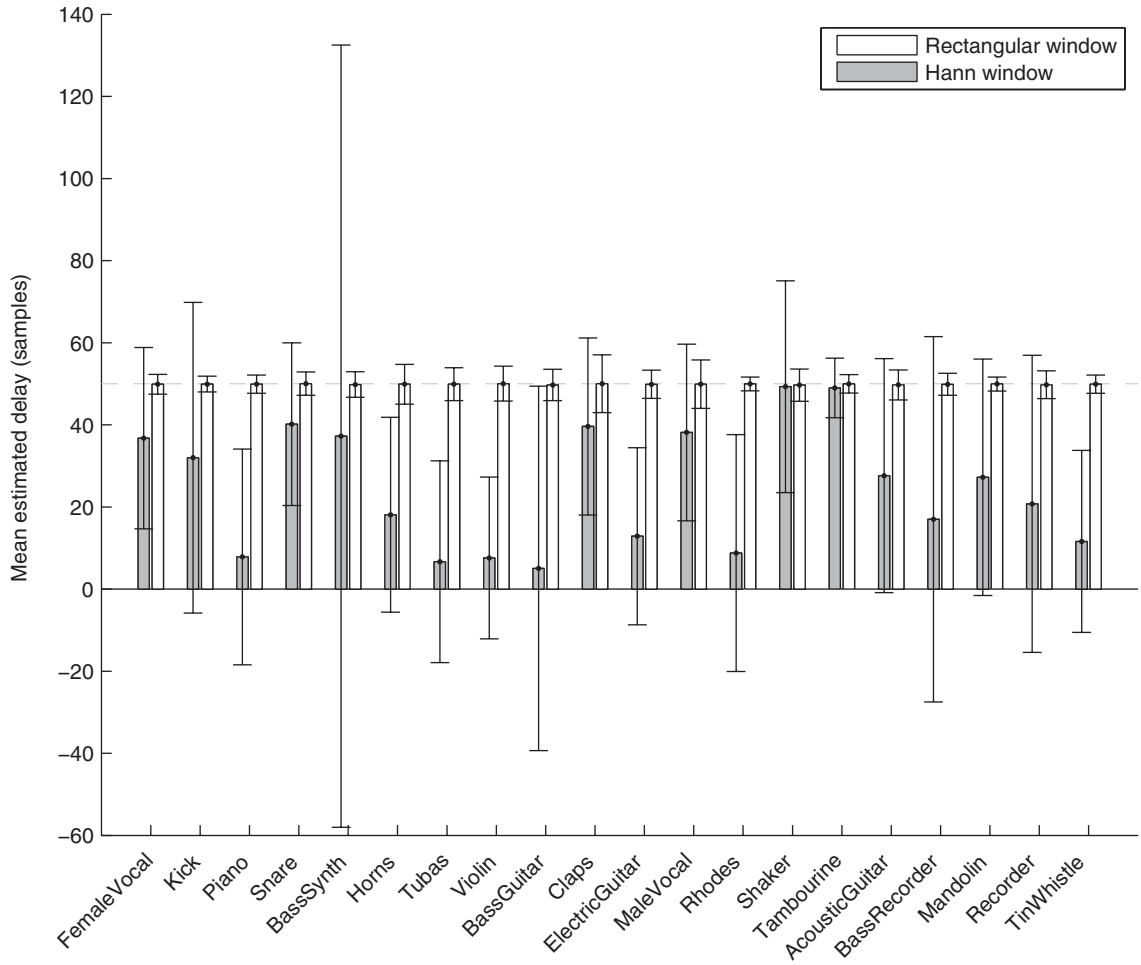


Fig. 9. Mean delay estimated for each instrument showing the standard deviation of all calculated delays. The correct delay of 50 samples is indicated with a horizontal dashed line.

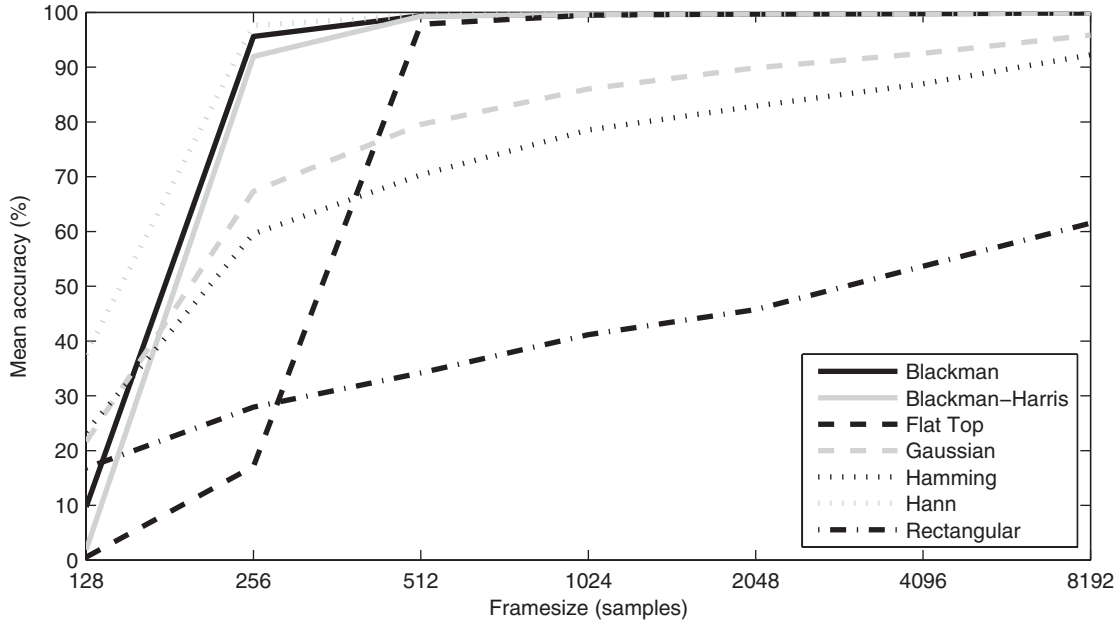


Fig. 10. Mean accuracy of delay estimation over all audio excerpts using a selection of common frame sizes and windows.

Table 2. Mean accuracy over all audio excerpts and frame sizes for each window shape showing window features.

Window	Mean accuracy (%)	Maximum side lobe height (dB)	Side lobe roll-off (dB/decade)	Start/end value
Hann	90.52	−31.5	60	0
Blackman	86.24	−58.1	60	0
Blackman-Harris	84.58	−71.5	20	6.00×10^{-5}
Gaussian	76.11	−43.3	20	-4.2×10^{-4}
Flat Top	73.49	−93.6	20	4.3×10^{-2}
Hamming	70.57	−42.7	20	0.08
Rectangular	40.14	−13.3	20	1

input signal is of a narrow bandwidth and when the window function has high amplitude side lobes with a shallow roll off. A theoretical analysis has been presented, leading to the conclusion that window functions which reach 0 at the extremities will offer the greatest performance.

An experimental analysis of simulated and real musical signals was outlined that shows that the higher the bandwidth, or spectral spread, of an input signal, the higher the accuracy of the delay estimation. A number of window functions were compared and it was found that the Hann or Blackman windows offer the greatest performance for all input signals, resulting in 100% accuracy of simulated signals and 50% increase in accuracy for real low bandwidth audio excerpts compared to the rectangular window, the worst performing window function.

5 ACKNOWLEDGMENT

This research was funded by an EPSRC DTA studentship.

6 REFERENCES

[1] D. Howard and J. Angus, *Acoustics and Psychoacoustics* (Oxford, UK: Focal Press , 2000).

[2] J. Chen, J. Benesty and Y. A. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview," *EURASIP J. on Applied Signal Processing*, vol. 2006, pp. 1–19 (2006).

[3] C. H. Knapp and G. C. Carter, "Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327 (1976).

[4] J. Benesty, J. Chen and Y. Huang, *Microphone Array Signal Processing* (Germany : Springer, 2008).

[5] J. Benesty, M. Sondhi and Y. Huang, *Springer Handbook of Speech Processing*(Springer, 2008).

[6] J. Chen, J. Benesty, and Y. Huang, "Performance of GCC- and AMDF-Based Time-Delay Estimation in Practical Reverberant Environments," *EURASIP J. Applied Signal Processing*, vol. 1, pp. 25–36 (2005).

[7] J. Meyer, "Precision Transfer Function Measurements Using Program Material as the Excitation Signal," in *Proceedings of the 11th International Conference of the Audio Engineering Society: Test and Measurement* (1992 May), paper 11-040.

[8] Meyer Sound, *SIM System II V.2.0 Operation Manual* (1993).

[9] E. Perez Gonzalez and J. Reiss, "Determination and Correction of Individual Channel Time Offsets for Signals Involved in an Audio Mixture," presented at *the 125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7631.

[10] A. Clifford and J. Reiss, "Calculating Time Delays of Multiple Active Sources in Live Sound," presented at the *129th Convention of the Audio Engineering Society (*2010 Nov), convention paper 8157.

[11] A. Clifford and J. Reiss, "Reducing Comb Filtering on Different Musical Instruments Using Time Delay Estimation," *J. Art of Record Production*, vol. 5, July 2011.

[12] J. Benesty, "Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization," *J. Acous. Soc. Am.*, vol. 107, no. 1, pp. 384–391 (2000).

[13] F. Reed, P. Feintuch and N. Bershard, "Time Delay Estimation Using the LMS Adaptive Filter—Static Behaviour," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 561–571 (1981).

[14] A. Brutti, M. Omologo and P. Svaizer, "Comparison between Different Sound Source Localization Techniques Based on a Real Data Collection," in *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays* (Trento, Italy), 2008.

[15] S. Brunner, H.-J. Maempel, and S. Weinzierl, "On the Audibility of Comb-Filter Distortions," presented at the*122nd Convention of the Audio Engineering Society* (2007 May), convention paper 7047.

[16] B. Champagne, S. Bédard and A. Stéphenne, "Performance of Time-Delay Estimation in the Presence of Room Reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 148–152 (1996 Mar.).

[17] M. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas and J. Escolano, "Evaluation of Generalized Cross-Correlation Methods for Direction of Arrival Estimation Using Two Microphones in Real Environments," *Applied Acoustics*, vol. 73, pp. 698–712(2012 Aug.).

[18] L. Chen, Y. Liu, F. Kong and N. He, "Acoustic Source Localization Based on Generalized Cross-Correlation Time-Delay Estimation," *Procedia Engineering*, vol. 15, no. 0, pp. 4912–4919 (2011), CEIS 2011.

[19] J. Hassab and R. Boucher, "Performance of the Generalized Cross Correlator in the Presence of a Strong Spectral Peak in the Signal," *IEEE Transactions on Acoustics,*

Speech and Signal Processing, vol. 29, pp. 549–555 (1981 Jun.).

[20] M. S. Brandstein and H. F. Silverman, "A Practical Methodology for Speech Source Localization with Microphone Arrays," *Computer, Speech and Language*, vol. 11, pp. 91–126 (1997 Apr.).

[21] S. Björklund and L. Ljung, "An Improved Phase Method for Time-Delay Estimation," *Automatica*, vol. 45, no. 10, pp. 2467–2470 (2009).

[22] S. Assous, C. Hopper, M. Lovell, D. Gunn, P. Jackson and J. Rees, "Short Pulse Multi-Frequency Phase-Based Time Delay Estimation," *J. Acous. Soc. Am.*, vol. 127, no. 1, pp. 309–315 (2009).

[23] S. Assous and L. Linnett, "High Resolution Time Delay Estimation Using Sliding Discrete Fourier Transform," *Digital Signal Processing*, vol. 22, pp. 820–827 (2012 Sept.).

[24] K. D. Donohue, J. Hannemann and H. G. Dietz, "Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments," *Signal Processing*, vol. 87, no. 7, pp. 1677–1691(2007).

[25] D. Salvati, S. Canazza, and A. Roda, "A Sound Localization Based Interface for Real-Time Control of Audio Processing," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)* (2011).

[26] M. S. Brandstein and H. F. Silverman, "A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (1997).

[27] C. Tournery and C. Faller, "Improved Time Delay Analysis/Synthesis for Parametric Stereo Audio Coding," presented at the *120th Convention of the Audio Engineering Society* (2006 May), convention paper 6753.

[28] D. Bechler and K. Kroschel, "Considering the Second Peak in the GCC Function for Multi-Source TDOA Estimation with a Microphone Array," in *International Workshop on Acoustic Echo and Noise Control* (2003).

[29] K. K. Paliwal and L. D. Alsteris, "On the Usefulness of STFT Phase Spectrum in Human Listening Tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170 (2005).

[30] R. Balan, J. Rosca, S. Rickard and J. O'Ruanaidh, "The Influence of Windowing on Time Delay Estimates," in *International Conference on Information Sciences and Systems* (2000).

[31] M. Novotny and M. Sedlacek, "The Influence of Window Sidelobes on DFT-Based Multifrequency Signal Measurement," *Computer Standards and Interfaces*, vol. 32, pp. 110–118 (2010 Mar.).

THE AUTHORS

Alice Clifford

Josh Reiss

Alice Clifford is a Ph.D. research student with the Centre for Digital Music in the School of Electronic Engineering and Computer Science at Queen Mary, University of London. Her research focuses on removing microphone artifacts in live sound. In 2008 she graduated from De-Montfort University, Leicester with a BSc in audio and recording technology and in 2009 graduated from the University of Edinburgh with an MSc in acoustics and music technology, specialiszing in room acoustics simulation. Alice is a member of the AES.

●

Josh Reiss is a senior lecturer with the Centre for Digital Music at Queen Mary University of London. He received his Ph.D. in physics from Georgia Tech, specializing in analysis of nonlinear systems. He made the transition to audio and musical signal processing through his work on sigma delta modulators, which led to patents and a nomination for a best paper award from the IEEE. He has investigated music retrieval systems, time scaling and pitch shifting techniques, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering. Dr. Reiss has published over 150 scientific papers and serves on several steering and technical committees. He is a member of the AES Board of Governors and co-chair of the Technical Committee on High-resolution Audio. As coordinator of the EASAIER project, he led an international consortium working to improve access to sound archives in museums, libraries, and cultural heritage institutions. He is co-founder of the start-up company Mix Genius, providing intelligent tools for audio production.