BINAURAL SOUND SOURCE LOCALIZATION BASED ON UPWARD ZERO-CROSSING POINTS DETECTION IN REVERBERANT FIELDS
Xiangyang Zeng, Huaizhen Cai
School of Marine Science and Technology, Northwestern Polytechnical University, 710072 Xi’an, China
e-mail: zenggxy@nwpu.edu.cn

This paper mainly deals with the problem of sound source localization based on upward zero-crossing points detection with two microphones. Gammatone filter bank are used to divide the signal into different frequencies, SNR can be obtained by the statistical properties of the time difference samples, and then reliable time difference samples can be selected according to the SNR limitation. Finally, the practical time difference is achieved by the analysis of the time difference samples. Consequently, the time differences are transformed into azimuths by use of geometrical relationship. Experiments are conducted with different sounds, and sounds with wider spectrum seem to perform more precisely; Evidence also shows that the method is robust to sound signals with different SNR, and experiments in different rooms reveal that reverberation apparently degrades the localization accuracy.

1. Introduction
The development of audio-visual technology as well as the approaching of intelligent era, both of which facilitates the research on auto-sound source localization in rooms contain many sources and large reverberation, which plays an important role in fields like picture phone meeting, auto REC, intelligent monitoring system, robot design and so on.

The Microphone arrays are widely used in sound source localization. However, arrays’ advantage in accuracy is at the cost of larger computation and higher cost, which works no more when cost and space are preferentially considered in practical applications. The human auditory system is known to have an excellent ability in sound localization, which is able to distinguish a specific sound source among multiple sound sources even in reverberant environments. The mechanism is based on the difference between sounds reaching to two ears. Two primarily used cues are ITD (interaural time difference) and ILD (interaural level difference).

It is widely known that ITDs are the main cues used at low frequency less than 1.5 kHz while IIDs are used in high frequency scale. Estimations of sound source direction based on ITDs has smaller variation, while estimations based on ILDs has larger variation. ILDs are more sensitive to reverberation. In this paper, a Gammatone filter bank is applied to divide the signal into different frequencies, which is a simulation to human auditory, the sound source direction is identified by the estimation of time difference (TD) between two microphones, and a more reliable localization method which is robust to environment has been used. And we investigate the upward zero-crossing points based method for a possible solution of these problems.

The paper is organized as follows. In section 2, we presents the estimation of TD samples based on upward zero-crossing points detection and the relationship between SNRs to TD samples,
2. Zero-crossing points based method

2.1 Time-domain Gammatone filters

Since the method used in this paper is based on two microphones, which can take the advantage of human auditory system. A Gammatone filter is a simulation to the frequency selective characteristic of human’s basilar membrane, which has the impulse responses of the form:

$$g(t) = at^{n-1}e^{-2\pi br} \cos(2\pi f_c + \phi)$$  (1)

Where $f_c$ is the central frequency of the filter, $\phi$ is the phase which is usually set to be 0. The constant $a$ controls the gain, and $n$ represents the order of the filter, usually set to be no larger than 4. $b$ is the decay factor which is related to the central frequency $f_c$ and is given by:

$$b = 1.019 \times 24.7 \times (4.37 \times f_c / 1000 + 1)$$  (2)

A series of Gammatone filter with different $f_c$ form a Gammatone filter bank, which is used to divide signals into various frequencies. As a simulation of human auditory behaviour, the central frequencies of the filter bank are usually equally distributed on the Bark scale. The amplitude-frequency response of the Gammatone filter bank used in this paper is showed in figure 1, which includes 33 filters from 0 to 5000Hz.

![Amplitude-frequency response of the gammatone filter bank](image)

Figure 1. Amplitude-frequency response of the gammatone filter bank

2.2 Estimation of time difference (TD) samples

In this section, we will describe the estimation for TD samples using upward zero-crossing points detection. Firstly, the dual channel signals (distinguished by L, R) was filtered by a series of band-pass Gammatone filter bank. The outputs of $i$th filter are represented by $x^L_i(t)$ and $x^R_i(t)$, the corresponding amount of upward zero-crossing points are $N$ and $M$ for $x^L_i(t)$ and $x^R_i(t)$, respectively. The upward zero-crossing time delay for the left and right channel are $t^L_n,n = 1...N$ and $t^R_m, m = 1...M$, respectively. Here, assuming that the magnitude of the signal at zero-crossing time satisfies $x^L_i(t^L_n) = 0, x^R_i(t^R_m) = 0$. We now describe the principle of determining TD samples using zero-crossing points. To eliminate the possible phase ambiguity, the time delay of left channel ($t^L_n$)
is used as a reference, and \( k \) is an index, we estimate three types of TD samples between two nearest left and right zero-crossings at the \( i \)th channel as:

The nearest zero-crossing points TD:

\[
\Delta t_i(n) = \min \| t^L_n - t^R_k \| \quad (3)
\]

The nearest zero-crossing points TD in the positive direction:

\[
\Delta t_2(n) = \min \| t^L_n - t^R_k \| \quad t^L_n > t^R_k \quad (4)
\]

The nearest zero-crossing points TD in the negative direction:

\[
\Delta t_3(n) = \min \| t^L_n - t^R_k \| \quad t^L_n < t^R_k \quad (5)
\]

In other words, we want to get the minimum time difference from \( t^L_n \) in the left channel to the nearest zero-crossing time \( t^R_k \) in the right channel. The TD samples corresponding to the nearest zero-crossing points in the positive and negative directions are required to resolve a possible phase ambiguity in estimating TD samples when zero-crossing time differences (usually between \(-800\) and \(+800\) \(\mu s\)) are larger than the half of the channel period defined by the inverse of the center frequency of the filter bank channel\(^8\).

Additionally, according to the distance between two microphones, denoted by \( D \), we can estimate the reasonable limitation of TD:

\[
\max \| TD \| < D / c \quad (6)
\]

Where \( c \) represents sound speed in air.

### 2.3 Effective TD samples selection based on SNRs

Since the outputs of band-pass filter bank not only contain sinusoid signal but some other components, such as noise due to environment and measurement error. This can be described by the following equations:

\[
\begin{align*}
    t^L_n &= t^L_i + e^L_n \\
    t^R_n &= t^R_i + e^R_n
\end{align*}
\]

Where \( e^L_n, e^R_n \) are random deviations describing the perturbations of zero-crossing points due to noise in the left and right channel signals, respectively. Here, we assume that both \( e^L_n, e^R_n \) are independent and identically distributed with mean zero, and \( e^L_i, e^R_i \) are also independent of each other. Then the mean and variance of TD samples are:

\[
\begin{align*}
    E(\Delta t(n)) &= E[(t^L_n - t^R_k) + (e^L_n - e^R_k)] \\
    &= E[\Delta t + (e^L_n - e^R_k)] = \Delta t \\
    Var(\Delta t(n)) &= E[(\Delta t(n) - \Delta t)^2] \\
    &= E[(e^L_n - e^R_k)^2] = Var(e^L_n) + Var(e^R_n)
\end{align*}
\]

Since we assume that \( e^L_n, e^R_n \) are independent of each other and have zero mean. We can concluded from equation(9) that the mean value of TD samples is an unbiased estimation. According to reference 9, the variance of TD samples in equation(10) can be achieved by:

\[
Var(\Delta t_i) = Var(e^L_n) + Var(e^R_n) = \frac{1}{\omega_i^2} \left( \frac{1}{2 \times 10^{\text{SNR}^L/10}} + \frac{1}{2 \times 10^{\text{SNR}^R/10}} \right) \quad (11)
\]

When the distance between two microphones is much less than the source distance, the sound intensity difference is negligible, then we can get:

\[
\begin{align*}
    \text{SNR}^L_i &= \text{SNR}^R_i = \text{SNR}_i \\
    \text{SNR}_i &= 10 \log_{10} \frac{1}{\omega_i^2 Var(\Delta t_i)}
\end{align*}
\]

Then the relationship between TD samples and SNR can be used to select reliable TD samples, a reasonable SNR limitation is needed before that.

### 2.4 Detail description of the method
Step1: Zero-crossing points detection and measurement of TD samples: the signals of both channel are passed through a series of band-pass filter bank, respectively, and detect upward zero-crossing points for each frequency band, estimate the TD samples using equation (3-5).

Step2: Selection of reliable TD samples: apply a window to TD samples within a frequency band, compute the mean and variance across all TD samples in the window, and the SNR can be estimated by use of equation (12), select the reliable TD samples when the estimated SNR is greater than the limitation value.

Step3: Estimation of TD values: analyze all the reliable TD samples, and the corresponding bin of the statistical graph is plotted, the TD value corresponding to the dominant peak of the graph is the TD value we need.

Step4: The geometric relationship between TD and the azimuth is shown in figure 2, equation (13) represents how the azimuth is estimated in this paper. The definition of azimuth is the angle included by sound source and the median surface of two microphones. A negative azimuth means the source in the right and vice versa.

\[ \theta = \arccos\left(\frac{d}{D}\right) - \pi / 2 \]

\[ d = c \cdot TD \]

Figure 2. Geometric relationship between TD and the azimuth

3. Experimental results

The signal in the experiment used are amplified by BK2716 and the microphone we used are BSWA MZ211. Four kinds of sound sources are selected: 1kHz pure tone, male and female speech and tap-tap, the duration are all 1 second, the frequency spectrum are shown in figure 3, the distance between sound source and the midpoint of the microphones is 1.9 meter, the distance between the microphones is 19 centimeter, 9 azimuth are selected uniformly spaced from -60° to 60° as figure 4 showed. The experiment is conduct in semi-anechoic room, ordinary room and reverberant room. The Gammatone filter bank with 33 pass bands is used for signal frequency division; the SNR limitation is 0dB, above which the TD samples used to compute SNR are regarded as reliable TD samples. The window length used in step 2 is 14, figure 5 shows how the window is applied, in which numbers represent number of TD samples. The results showed in this paper have transfer TD into azimuth by use of equation (13).

Firstly, the influence of different sound sources have on the localization method is verified in a semi-anechoic room. The results are showed in figure 6, the standard deviations of four sound source localization in 9 positions are 9.7°, 7.6°, 8.1°, 5.8°, respectively. It seems that the localization accuracy increased by the source bandwidth.

Additionally, considering that the zero-crossing points based method selects reliable TD samples by use of SNR, it is indispensable to compare the localization results of signals with different SNR. The tap-tap source located in 30° was played in this experiment, and white noise with SNR(0dB, 5dB, 10 dB) was added to the signal, the localization results are showed in figure 7 (value in the box represents the estimated azimuth). We can concluded that the method is robust
to different SNR, since the errors for three SNRs are 1.8°, 0.5°, 2.1°, respectively.

Figure 3. The normalized frequency spectrum of four sound sources

Figure 4. The position of two microphones and 9 sound source locations (‘>’, ‘*’ represent the location of microphone and sound sources, respectively)

Figure 5. Diagram for applying window
Finally, as a main points in this paper, the influence of different reverberation time on the method is studied. Experiments were conducted in rooms with different reverberation times $T_{60} = 1.26s, 4.11s$, the tap-tap sound source was played in this experiment. The localization results in three azimuths (30°, 0°, -30°) are showed in figure 8-9.

It can be concluded from figure 7-9 that a higher reverberation leads to a less sharp peak;
Besides, the localization deviations in figure 8 and figure 9 are 13° and 23°, respectively, which is a demonstration for the speculation that reverberation has a negative impact on the zero-crossing points based method.

**Figure 8.** Statistical graph of TD samples when the tap-tap sound was played in a reverberant room at azimuth 30°, 0°, -30° from top to bottom, respectively ($T_{60} = 1.26s$)

**Figure 9.** Statistical graph of TD samples when the tap-tap sound was played in a reverberant room at azimuth 30°, 0°, -30° from top to bottom, respectively ($T_{60} = 4.11s$)
4. Conclusion
This paper mainly studied the localization method based on zero-crossing points detection, the signals was filtered by 33 Gammatone filters for different frequency. The TD between two microphones was estimated by use of statistical property of TD samples graph. The method has been proved to cost less computation, and performed well in low reverberation, and seems to perform better with broader band signals, the data in this paper showed an increase of 3.9° in accuracy. It also verified to be robust to different SNRs, and was negatively influenced by reverberation. The further study will study the de-reverberation of this method.

REFERENCES