

Time Delay Estimation Based On Hearing Characteristic

Zhaoli Yan^{I,II}, Limin Du^{II}, Jianqiang Wei^{II}, Hui Zeng^{II}

Institute of Physics, Chinese Academy of Sciences^I
Institute of Acoustics, Chinese Academy of Sciences^{II}
yanzl@iis.ac.cn, dulm@iis.ac.cn

Abstract

This paper proposes a new time delay estimation model, Summary Cross-correlation Function (SCCF). It is based on a hearing model of the human ear, which comes from a pitch perception model. The inherent relation between some time delay estimation (TDE) and pitch perception method is mentioned, and propose an idea - some pitch perception models' pre-processing can be used for references in TDE model and vice versa. The new TDE model is proposed based on this viewpoint. Then SCCF is analyzed further, and compares its performance with Phase Transform (PHTA) and Modified Cross-power Spectrum (M-CPSP). The simulated experiments show that the new model is more robust to noise than PHAT and M-CPSP.

1. Introduction

Time-Delay Estimation (TDE) is a basic step of localization of acoustic sources and beamforming, which can be used in many acoustic systems such as video-conference systems, hands-free systems, etc. While in the real environment, background noise and room reverberation attenuates the accuracy of TDE, even if the difference between a pair of microphones is cancelled. Many methods like Generalized Cross-Correlation [1] and adaptive algorithm [2] were brought forward to TDE. The Maximum Likelihood (ML) time delay estimation is derived from signal-to-noise ratio (SNR)-weighted version of the General Cross-Correlation (GCC) function [1]. The whitening of the cross-correlation spectrum deduces Phase Transform (PHTA) [1] method. This approach has been attended recently as the base of speech source localization. A pitch-based approach is proposed in [3]. According to PHTA method, which is called Cross-power Spectrum Phase (CPSP) in [6], Daniel V. Rabinkin proposed a modified CPSP. However their performance is still susceptible to noise and reverberation, their robustness needs to be improved further. Further more, the computation cost of adaptive method is considerable.

This paper proposes a thought way for TDE. As we know, cross-correlation is used in many TDE methods. At the same time, auto-correlation is employed to pitch perception models - L. R. Rabiner's model [4] and Meddis-O'Mard's model [5], for example. Using correlation (auto-correlation for pitch perception and cross-correlation for TDE) is their common characteristic. So some pitch perception models' pre-processing can be used for references in TDE model and vice versa. According to those discussed above, a TDE model - Summary Cross-correlation Function (SCCF) employing a hearing model is proposed. The hearing model comes from Karjalainen and Tolonen's pitch perception model [7][8], which is a simplified version of Meddis-O'Mard unitary model.

This paper is organized as follows. The GCC and Modified Cross-power Spectrum Phase (M-CPSP) are mentioned in Section II. The SCCF model is proposed and analyzed in Section III. Then the experiments are discussed and the conclusion is given at the end of this paper.

2. The Generalized Cross-Correlation (GCC)

For a given signal source $s(t)$ propagating in a noisy environment, the signal acquired by microphone i can be modeled as follows:

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \quad (1)$$

where α_i is an attenuation factor due to propagation effects, τ_i is the propagation time and $n_i(t)$ includes all the noises which are assumed to be uncorrelated with signal $s(t)$.

The Generalized Cross-Correlation function between $x_i(t)$ and $x_j(t)$ is defined as:

$$R_{x_i x_j}(\tau) = \int_{-\infty}^{+\infty} \psi(f) X_i(f) X_j^*(f) e^{2\pi f \tau} df \quad (2)$$

where $\psi(f)$ is a general frequency weighting filter, $X_i(f)$ is the Fourier transform of $x_i(t)$, $(\cdot)^*$ is the complex conjugate operator. As to ML time delay estimation, $\psi(f)$ is determined according to SNR of the contaminated signal $x_i(t)$. $\psi(f)$ is given a larger value if $X_i(f)$ has a higher SNR.

A way to sharpen the cross-correlation peak is to "whiten" the cross-power spectrum in Equ.2, which leads to the so-called Phase Transform (PHTA) or Cross-power Spectrum Phase (CPSP). $\psi(f)$ is determined as:

$$\psi(f) = \frac{1}{|X_i(f) X_j^*(f)|} \quad (3)$$

The whitening of the cross-correlation spectrum is under the assumption that the statistical behavior of both signal and noise is uniform all over the entire spectrum, namely, the SNR of each frequency is stationary. In fact, the SNR varies with frequency. Like the ML method's idea, Daniel V. Rabinkin proposed a modified CPSP [6] considering SNR:

$$CPSP = IDFT \left(\frac{\mathbf{X}_i \mathbf{X}_j^*}{(\|\mathbf{X}_i\| \|\mathbf{X}_j\|)^r} \right) \quad (4)$$

$$\tau_{ij} = k : \max_k CPSP(k) \quad (5)$$

where $0 < r \leq 1$, \mathbf{X}_i and \mathbf{X}_j are the Discrete Fourier Transform of microphone signal x_i and x_j respectively, and τ_{ij} is the delay estimation.

At the frequency where the SNR level is higher, \mathbf{X}_i should have greater power. As a result, It's reasonable to set r between 0 and 1, to weigh the cross-correlation part in Equ.4 at the frequency where \mathbf{X}_i , \mathbf{X}_j have greater magnitude. An optimal value for r was determined to be 0.75 according to experiments. For more details, see [6]. The final experiment shows that its performance is better than that of CPSP.

3. The Proposed Time Delay Estimation Model

In Meddis-O'Mard unitary pitch perception model, the signal is split into channels such as ERB (equivalent rectangular bandwidth) channels and each channel is half-wave rectified and low-pass filtered (at 1 kHz) in order to simulate the hearing neural transduction. The auto-correlation function (ACF) of each channel is calculated, and summed together to produce a summary auto-correlation (SACF). Fig.1 is Karjalainen and Tolonen's pitch perception model, a simplified version of Meddis-O'Mard unitary model. It condenses the primary several tens of channels into two, therefore, the computational complexity is reduced largely.

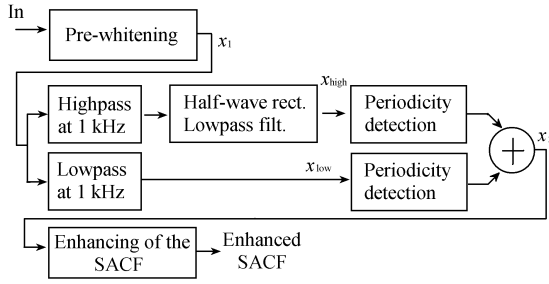


Figure 1: Karjalainen and Tolonen's pitch perception

Referring to these two pitch perception models discussed above, a TDE model is proposed and illustrated in Fig. 2.

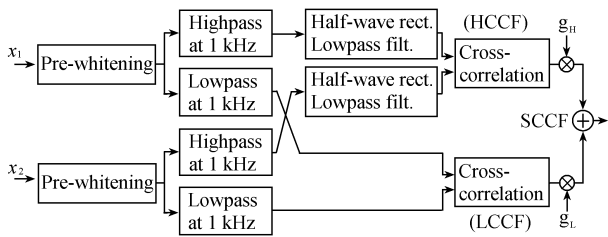


Figure 2: The proposed time delay estimation model – SCCF

Obviously, its pre-processing is similar to the pitch perception model's in Fig.1. Pre-whitening filter is to obtain the exciter signal of speech using LPC. The outputs of pre-whitening filter are divided into two sub-channels respectively. The portions with frequency below 1 kHz are directly used to calculate the low-frequency cross-correlation function (LCCF), while another portions with frequency

above 1 kHz are half-wave rectified and filtered by a 1 kHz low-pass filter, and then the filter results are used to calculate the high-frequency cross-correlation function (HCCF). The two cross-correlation functions are added with different weights to get the summary cross-correlation function (SCCF).

Some details about the proposed model need to be clarified. In Meddis-O'Mard model, although the band-pass filter is set to be a gammatone filterbank to simulate the frequency selectivity of the cochlea, a fourth-order FIR filter is recommended in this dual channel TDE model. The cross-correlation algorithm is the same as Equ.4. The optimal value of r is chosen to be 0.6 according to the experiments. If x_1 and x_2 are substituted by the same signal x , the auto-correlation function will be produced:

$$acor(\tau) = IDFT[\mathbf{X}^{2(1-r)}] \quad (6)$$

where \mathbf{X} is the Discrete Fourier Transform of signal x . The ACF in [7][8] is $acor(\tau) = IDFT[\mathbf{X}^{2/3}]$. If $r = 2/3$, the ACF will be equivalent to that in [7][8]. So it can be said that setting r to be 0.6 in this paper is consistent with that in the pitch perception model. The weights g_L and g_H can be calculated through SNRs of the two channels:

$$g_L = SNR_L / (SNR_L + SNR_H) \quad (7)$$

$$g_H = SNR_H / (SNR_L + SNR_H) \quad (8)$$

where SNR_L and SNR_H are SNR of the high-pass and low-pass channels respectively. The SNR is

$$SNR = \frac{E[x^2] - E[n^2]}{E[n^2]} \quad (9)$$

where x is one of the channel's signal, n is the noise estimation. The noise level can be obtained during speech pause for SNR estimation. If it's hard to get SNR, g_L and g_H are assumed to be 1. Like ML and M-CPSP, SCCF method is also based on the assumption that the SNR level is variable with the frequency of the signal. The sub-channels with higher SNR levels will get a better result, which is tenable in practice.

In order to analyze SCCF, The examples of cross-correlation calculated by CPSP, M-CPSP and SCCF are given in Fig.3.

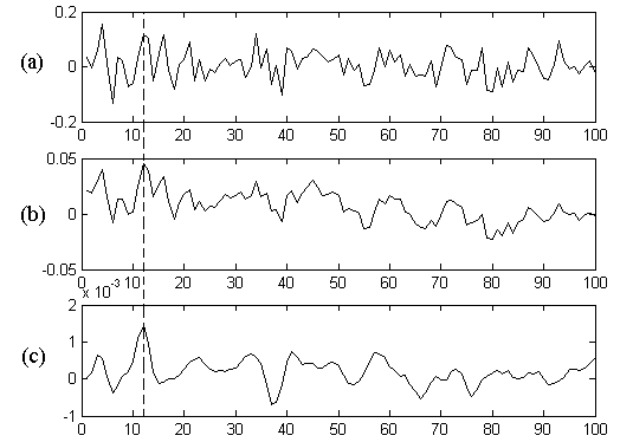


Figure 3: (a) CPSP of the signals. (b) M-CPSP of the signals. (c) SCCF of the signals.

The input signal is recorded in a real environment of office, being contaminated by the noises from many computers fans and air-conditioners. The dash-line marks the correct position. The figure shows that CPSP gives an error result; SCCF has a steeper peak than M-CPSP dose. The proposed model is more robust to noise.

4. Experiments

4.1. The simulated experiment

The performances of SCCF, CPSP and M-CPSP are compared with various SNRs and walls' reflection coefficients in a simulated rectangle room (7m x 4m x 2.75m). The white Gaussian noises were mixed with the signals to get different SNRs ranging from 0dB to 30dB with 10dB steps. The reflection coefficients of the six reflection surfaces are equivalent and frequency-independent. The reflection coefficients were obtained using Eyring's formula:

$$\beta = \exp\{-13.82/[(1/L_x + 1/L_y + 1/L_z)cT]\} \quad (9)$$

where L_x , L_y , L_z are the room size, c is the sound spreading speed and T is the reverberation time which is ranging from 0s to 0.2s in this experiment. The acoustical room transfer functions were produced via image method [9]. The speech signal with 22kHz sampling rate was convolved with the room transfer function.

The receivers and sound source were placed as Fig.4. The positions of two receivers are (3.0m, 2.0m, 1.5m) and (3.3m, 2.0m, 1.5m). D is equal to 30cm. The sound source has three positions: A ($r = 1.5m$, $\alpha = 75^\circ$), B ($r = 1.5m$, $\alpha = 35^\circ$) and C ($r = 2.5m$, $\alpha = 35^\circ$).

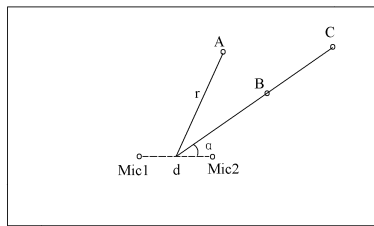


Figure 4: The simulated experiment environment. Mic1 and Mic2 are microphones' position. A, B and C are sound source positions.

The experiments were carried out under the condition of different positions, SNRs and reverberation times. The results of sound source B are illustrated in Fig.5. The delay estimations within ± 2 sample points' error are regarded as valid results in the correct rate statistic. Fig.5.a1, b1 and c1 are correct rate illustration of the estimation, and the variances are shown in Fig.5.a2, b2 and c2.

4.2. The experiment in the real environment

In order to test the performance of the proposed model, the experiment in real office environment has been finished. The microphones are Panasonic WM-52BP; The distance between microphones is 28cm; The noises come from the computer fans, air-conditioners all around and a sound box beside the microphone array. The sound box plays the noises recorded in office in advance. It is used to control the background noise.

The azimuth angle of speaker changes around the array system with the variance of SNR, and the distance to the array system is 70cm. The sampling rate is 32kHz. The correct rate statistic is the same as experiment 1.

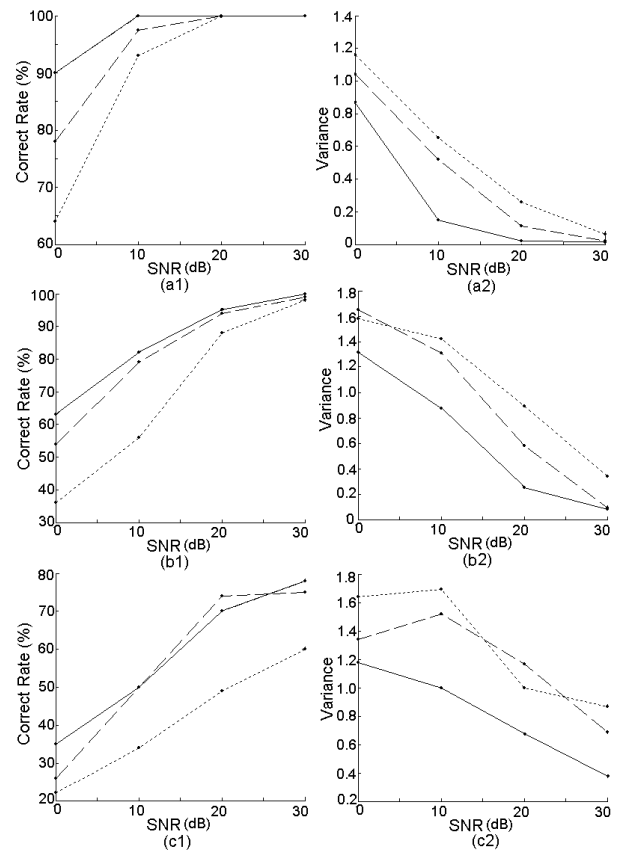


Figure 5: The simulated TDE experiment results. (a1) is the correct rate, (a2) is the variance of TDE at reverberation time $T = 0s$; (b1) and (b2) have the same meaning at $T = 0.1s$. (c1) and (c2) have the same meaning at $T = 0.2s$. The solid-lines are the results of SCCF. The dash-lines are the results of M-CPSP. The dot-lines are the results of CPSP.

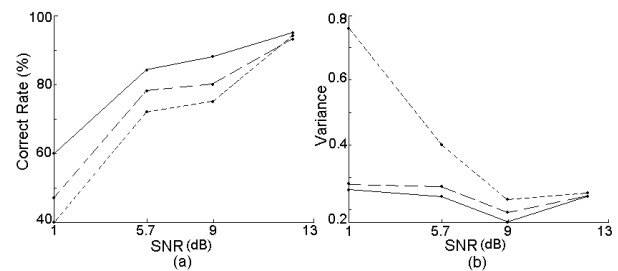


Figure 6: The TDE results in real environment. (a) is the correct rate, (b) is the variance of the results. The solid-lines are the results of SCCF. The dash-lines are the results of M-CPSP. The dot-lines are the results of CPSP.

The figures show that the proposed method has better performance than CPSP and M-CPSP do. In experiment 2, due to the non-uniform of the experiments in real environment, the variance may not remain monotony with the variance of SNR. But the sequence of the performances of the three methods is not changed in the same experiment.

5. Conclusion

A TDE model SCCF has been proposed in this paper, which has the similar pre-processing as Karjalainen and Tolonen's pitch perception model. The cross-correlation of sub-channels are calculated and added to estimate the time delay. At high SNR level, both the correct rate and variance tend to be uniform. While the proposed SCCF model has a better performance than M-CPSP and CPSP at low SNR level. This has been confirmed by the simulated experiments and real experiments. Further more, we can try to use more other pitch perception models based on auto-correlation as references in our TDE research in the future.

6. References

- [1] C. H. Knapp and G. Clifford Carter, "The Generalized Correlation Method for Estimation of Time Delay" in *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP – 24, No. 4, August 1976, pp. 320-327.
- [2] D.H. Youn, N. Ahmed, G. C. Carter, "On Using the LMS Algorithm for Time Delay Estimation" in *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 30(5), 1982, pp. 798-801.
- [3] Michael S. Brandstein, "A Pitch-Based Approach to Time_Delay Estimation of Reverberant Speech" in *IEEE 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, October 19-22 1997.
- [4] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection" in *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-25, No 1, February 1977, pp. 24-33.
- [5] R. Meddis and L. O'Mard, "A unitary model for pitch perception" in *J. Acoust. Soc. Am.*, Vol. 89, June 1991, pp. 2866-2822.
- [6] Daniel V. Rabinkin, "Optimum Sensor Placement for Microphone Arrays", *Doctoral Thesis*, 1998, pp. 47.
- [7] Matti Karjalainen and Tero Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'99)*, Vol. 2, March 1999, pp. 929-932.
- [8] Matti Karjalainen and Tero Tolonen, "Separation of speech signals using iterative multi-pitch analysis and prediction" in *Eurospeech '99*, Vol. 6, September 1999, pp. 2187-2190.
- [9] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small room acoustics" in *J. Acoust. Soc. Am.*, Vol. 65, No. 4, April 1979, pp. 943-950.