POSTERIOR PROBABILISTIC MODELING FOR INTER-CHANNEL PHASE AND TIME DIFFERENCE ESTIMATION IN AUDIO SIGNALS

Chao-Wen Li, and Yi-Wen Liu

Dept. Electrical Engineering, National Tsing Hua University, Hsinchu, 30013, Taiwan

ABSTRACT

A method is proposed for the estimation of the time difference of arrival (TDOA) from a sound source to a pair of microphones. Given noisy observations of the source, the magnitude spectrum of the source is first estimated via smoothing across time frames. Then, a probability density function (PDF) for the phase conditioned upon the magnitude is constructed based on the signal-to-noise ratio (SNR) at every frequency. Subsequently, the conditional PDF for the inter-channel phase difference (IPD) is calculated, and the computation can be accelerated via Gaussian curve fitting. Finally, by combining the information from all frequencies, the TDOA can be estimated in the *maximum a posteriori* (MAP) sense. Results from "anechoic" simulation showed that, at various SNRs (0 to 40 dB), the proposed method consistently produced more accurate estimation than the well known GCC-PHAT [1] and a recent method that was also based on IPD modeling [2]. Experiments conducted in an office environment are also reported, using speech and footsteps as test materials.

Index Terms— spectral estimation, time difference of arrival, multichannel audio processing

1. INTRODUCTION

With the advent of Internet of Things, there has been much discussion on connecting microphones to the networks for sound detection, enhancement, and localization. Sound localization in particular, could be achieved via estimation of the time differences of arrival (TDOA) from sources to microphones [3]. Many methods for TDOA estimation are based on peak finding from generalized cross-correlation (GCC) functions [1]. GCC, when correctly weighted, leads to maximum-likelihood (ML) estimation of TDOA [4], but its effectiveness for audio signals might be compromised — the ML implementation of GCC requires *a priori* and complete knowledge of the cross spectrum, but audio signals are generally non-stationary. Alternatively, some methods achieve sound localization via decomposition of multi-channel (array) signals into signal and noise subspaces. *Multiple signal classification* (MUSIC)[5], a well-known and powerful method of this kind, has enabled simultaneous direction-of-arrival estimation for multiple sources. However, the cost involved in high dimensional eigenspace decomposition might hinder its deployment to sensor networks when computing resource is limited.

Therefore, interest in conducting TDOA estimation for microphone pairs (e.g., [2, 6, 7, 8, 9]) has resurfaced with a set of new constraints and goals. First, the computation needs to be efficient; secondly, information extracted from microphone pairs pertaining the TDOA should be easy to integrate; finally, the method needs to adapt to the environment as signal and noise statistics are constantly changing. Toward these goals we aimed to "re-invent" TDOA estimation by calculating, at least approximately, the probability density function (PDF) of the time difference variable given the observed signals. We envision that such PDFs, collected from distributed microphone pairs in a network, can be combined for network-based sound enhancement and sound localization in a robust manner.

The proposed algorithm turns out to be most akin to the method in [2], which uses power-weighted histogram to infer the TDOA without explicitly deriving the PDF. Detailed comparison of the performance is reported, and the rest of this paper is organized as follows: Section 2 describes the methods, Section 3 reports on the results, and discussions and conclusion follow in Section 4.

2. METHODS

Described in this section is a TDOA estimation method which extracts information from the phase difference between the right and the left channel. The proposed algorithm is based upon a probabilistic model for noise. The noise model induces a probability distribution of TDOA, whose Gaussian approximation enables fast implementation.

2.1. Probabilistic modeling of TDOA

Considering audio signal interfered by additive white Gaussian noise in the time domain, the signal $y[n]$ received by a microphone can be described as follows,

$$y[n] = x[n] + u[n], \tag{1}$$

where $x[n]$ stands for the original signal and $u[n] \sim \mathcal{N}(0, \sigma^2)$ denotes the white Gaussian noise with zero mean and a variance of σ^2.

For a single frame of length N, denote the discrete Fourier transform (DFT) of $y[n]$, $x[n]$, and $u[n]$ as $Y[k]$, $X[k]$ and $U[k]$, respectively. In particular, we have

$$U[k] = \sum_{n=0}^{N-1} u[n] e^{-jk\frac{2\pi}{N}n} = U_r[k] + jU_i[k], \qquad (2)$$

where U_r and U_i denote the real and the imaginary parts of U, respectively. Since $u[n]$ is assumed to be Gaussian and independent and identically distributed (i.i.d.), it is straightforward to show that $U_r[k]$ and $U_i[k]$ are Gaussian and uncorrelated, and their variance is

$$\text{Var}(U_i[k]) = \text{Var}(U_r[k]) = \frac{N\sigma^2}{2}, \forall k = 1, 2, ..., N-1.$$

Therefore, the joint PDF of $U_r[k]$ and $U_i[k]$ is given as follows,

$$\begin{aligned} p(U_r[k], U_i[k]) &= p_{U_r}(U_r[k]) \cdot p_{U_i}(U_i[k]) \\ &= \frac{1}{\pi N\sigma^2} \exp\left\{-\frac{U_r^2[k] + U_i^2[k]}{N\sigma^2}\right\}. \end{aligned}$$

Denote the magnitude and phase of $X[k]$ as $R_X = |X|$ and $\theta_X = \angle X$, respectively; similarly, define $\theta_Y = \angle Y$ (hereafter, the frequency index k is omitted to simplify the presentation). To change the coordinates between (U_r, U_i) and (R_X, θ_X), we have

$$\begin{bmatrix} U_r \\ U_i \end{bmatrix} = \begin{bmatrix} |Y|\cos\theta_Y - R_X\cos\theta_X \\ |Y|\sin\theta_Y - R_X\sin\theta_X \end{bmatrix}, \qquad (3)$$

where $|Y|$ and θ_Y are known but R_X and θ_X are treated as random variables. Thus, the Jacobian matrix for the transformation is

$$J = \begin{bmatrix} \frac{\partial U_r}{\partial R_X} & \frac{\partial U_r}{\partial \theta_X} \\ \frac{\partial U_i}{\partial R_X} & \frac{\partial U_i}{\partial \theta_X} \end{bmatrix} = \begin{bmatrix} -\cos\theta_X & R_X\sin\theta_X \\ -\sin\theta_X & -R_X\cos\theta_X \end{bmatrix}. \qquad (4)$$

Then, the joint PDF of (R_X, θ_X) is

$$\begin{aligned} p_X(R_X, \theta_X) &= |\det(J)| \cdot p(U_r, U_i) \\ &= \frac{R_X}{\pi N\sigma^2} \exp\left\{-\frac{U_r^2 + U_i^2}{N\sigma^2}\right\} \\ &= \frac{R_X}{\pi N\sigma^2} \cdot \\ &\quad \exp\left\{-\frac{|Y|^2 + R_X^2 - 2|Y|R_X\cos(\Delta\theta)}{N\sigma^2}\right\}, \end{aligned}$$

where $\Delta\theta = \theta_Y - \theta_X$. To model the phase θ_X, the following estimation for R_X is made first,

$$\hat{R}_{X,m}[k] = \frac{1}{M}\sum_{i=0}^{M-1} |Y_{m-i}[k]|, \qquad (5)$$

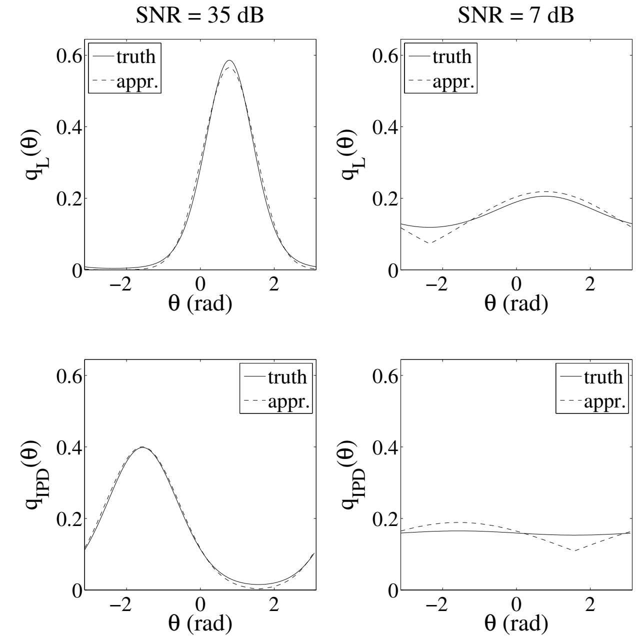


Fig. 1. Comparison of $q(\theta)$ and $q_{\text{IPD}}(\theta)$ and their Gaussian approximations. The upper panels show examples of $q(\theta)$, the PDF of the channel phase at a frequency, and the lower panels show examples of $q_{\text{IPD}}(\theta)$ when information from both channels is combined. The solid lines mark the "accurate" value of the PDFs as defined by Eqs. (6) and (7) and the dashed lines mark the Gaussian approximation. Two panels on the left are simulated at 35 dB SNR and the two on the right at 7 dB.

where m denote the index for the present frame, so Y_{m-i} denotes the DFT of $y[n]$ for the ith frame preceding the present frame. Eq. (5) is based on the assumption that $x[n]$ is quasi-stationary so $|X_m[k]|$ does not vary much for M successive frames. Therefore, if the SNR is sufficiently high, $|X[k]|$ can be estimated by averaging $|Y[k]|$. Now, replacing R_X in $p_X(R_X, \theta_X)$ by $\hat{R}_{X,m}$ in Eq. (5), we obtain the following conditional PDF,

$$\begin{aligned} q(\theta_X) \triangleq p(\theta_X | R_X = \hat{R}_{X,m}) &= \frac{p(\hat{R}_{X,m}, \theta_X)}{\int_{-\pi}^{\pi} p(\hat{R}_{X,m}, \theta_X) d\theta_X} \\ &= \frac{1}{C} \exp\left\{\frac{2|Y|\hat{R}_{X,m}\cos(\theta_Y - \theta_X)}{N\sigma^2}\right\}, \qquad (6) \end{aligned}$$

where C is a constant so that $\int_{-\pi}^{\pi} q(\theta)d\theta = 1$.

Next, the conditional PDF in Eq. (6) can be obtained for both the left and the right channels, given their respective estimates of R_X using Eq. (5) and their respective noise level σ^2. Denoting the results as $q_L(\theta_X^{(L)})$ and $q_R(\theta_X^{(R)})$ respectively, and define the inter-channel phase difference (IPD) as $\phi = \theta_X^{(L)} - \theta_X^{(R)}$. Assuming that the noise signals received by both channels are independent, the PDF for ϕ conditioned upon the estimates of R_X for both channels can be calculated as follows,

$$q_{\text{IPD}}(\phi) = \int_0^{2\pi} q_L(\theta) \cdot q_R(\theta - \phi) d\theta. \tag{7}$$

Hence, the PDF for the TDOA, conditioned upon the estimates of $\hat{R}_X$ for both channels and across all frequencies, can be written as follows,

$$p_{\text{TDOA}}(\tau) = \prod_{k=1}^{N/2-1} q_{\text{IPD},k}\big(2\pi k f_s \tau / N \bmod 2\pi\big). \tag{8}$$

where the notation $q_{\text{IPD},k}$ emphasizes that it varies against the frequency index k, and f_s denotes the sampling frequency. If the distance d_{mic} between two microphones is known, the TDOA should fall inside $[-\frac{d_{\text{mic}}}{c}, \frac{d_{\text{mic}}}{c}]$, where c denotes the speed of sound. Hence, an estimate of the TDOA can be obtained by maximizing Eq. (8); that is,

$$\hat{\tau} \triangleq \arg\max_{\tau \in [-\frac{d_{\text{mic}}}{c}, \frac{d_{\text{mic}}}{c}]} p_{\text{TDOA}}(\tau). \tag{9}$$

2.2. Simplification via Gaussian approximation

The estimate of TDOA can in principle be calculated as described previously, but the computation cost is high because the integral in Eq. (7) does not have a closed form. In this section, the computation is simplified via Gaussian approximation. First, note that in Eq. (6), the peak always occurs at $\theta_X = \theta_Y$, and the following equation also always holds,

$$\frac{q(\theta_Y)}{q(\theta_Y - \frac{\pi}{2})} = \exp\Big(\frac{2|Y|\hat{R}_{X,m}}{N\sigma^2}\Big). \tag{10}$$

Therefore, we can fit a Gaussian curve to approximate $q(\theta_X)$; the curve would represent the PDF of a Gaussian random variable $\tilde{\theta}_X \sim \mathcal{N}(\theta_Y, \sigma^2_{\text{appr}})$, where

$$\sigma_{\text{appr}} = \frac{\pi}{2}\sqrt{\frac{N\sigma^2}{2|Y|\hat{R}_{X,m}}}. \tag{11}$$

Thus, two Gaussian random variables $\tilde{\theta}_X^{(L)}$ and $\tilde{\theta}_X^{(R)}$ can be constructed for both channels respectively. Consequently, $\tilde{\phi} = \tilde{\theta}_X^{(L)} - \tilde{\theta}_X^{(R)}$ would also be Gaussian, with mean $\mu = \theta_Y^{(L)} - \theta_Y^{(R)}$ and a variance of $\sigma^2_{\text{tot}} = \sigma^2_{\text{appr},L} + \sigma^2_{\text{appr},R}$. Therefore, an approximation to Eq. (7) is obtained,

$$q_{\text{IPD}}(\phi) \approx \frac{1}{\sqrt{2\pi\sigma^2_{\text{tot}}}} \exp\left\{-\frac{(\phi-\mu)^2}{2\sigma^2_{\text{tot}}}\right\}. \tag{12}$$

Eq. (12) can be substituted into Eq. (8), so an estimation of τ is obtained with a reduced computational load.

Figure 1 shows a few examples of $q(\theta)$, $q_{\text{IPD}}(\theta)$, and their Gaussian approximation at various SNRs for comparison.

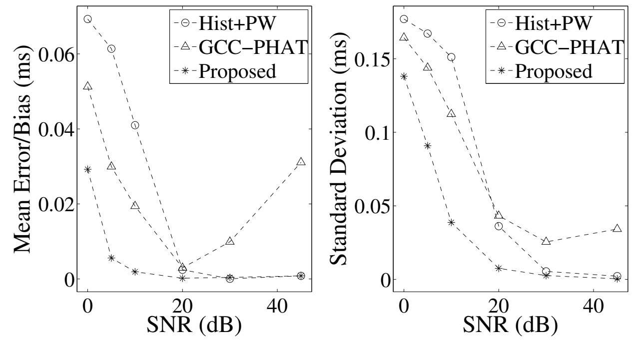


Fig. 2. Bias and standard deviation of three TDOA estimators at different levels of SNR. Statistics were obtained by averaging across 1100 frames.

3. SIMULATION AND EXPERIMENT

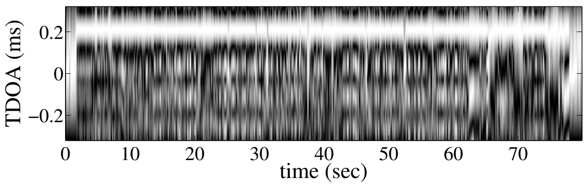
The performance of the proposed method was evaluated and compared against Fujii *et al*'s histogram-based method [2] and the well-known GCC-PHAT [1]. The following parameters were used for simulation and experiments: $d_{\text{mic}} = 11$ cm, $c = 348$ m/s, the length of FFT $= 2048$, and $f_s = 44.1$ kHz. The number of frames for smoothing in Eq. (5) was $M = 10$. The test materials included (a) an *adagio* music played by a string ensemble and a female vocal, (b) news-reporting speech produced by a female native speaker of English, and (c) footsteps recorded in the authors' laboratory. Electret microphones (Horn Audio, Shenzhen, China), unidirectional $(0 - 180$ degree) and with a sensitivity of -42 ± 3 dB (relative to 1V/Pa), were used in the experiments.

3.1. Simulation

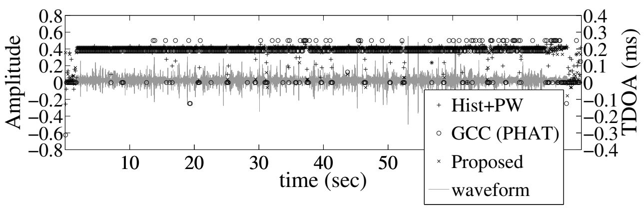
In a simulation, the adagio music was used as the source, and the TDOA was set at $= -3$ samples or -0.068 ms. White Gaussian noise was injected to the signal in a controlled manner at various SNRs. Figure 2 compares the estimation performance achieved by the three methods. Though not always significantly, the proposed method outperformed both the power-weighted histogram method (Hist+PW) [2] and the GCC-PHAT at all SNRs.

3.2. Experiments

A pair of microphones were set up to record sounds in an office environment. In one experiment, a loudspeaker was placed in front of the microphones at distances of 6.25 cm and 6.95 cm, respectively. The female speech signal was played back from the loudspeaker, so the true TDOA between the microphones should have been approximately 0.7cm/348(m/s) = 0.20 ms. The signals received by the microphones were amplified by a custom-made circuit using

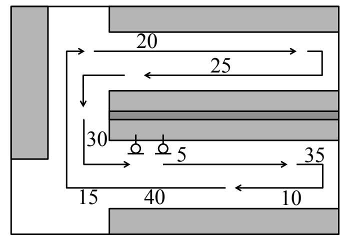


(a) $p_{\text{TDOA}}(\tau)$ for a non-moving source

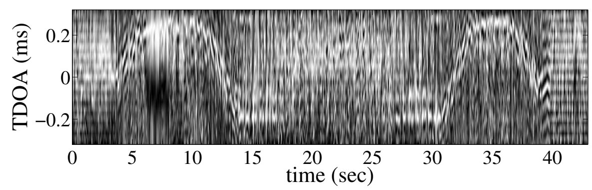


(b) TDOA estimated by three methods

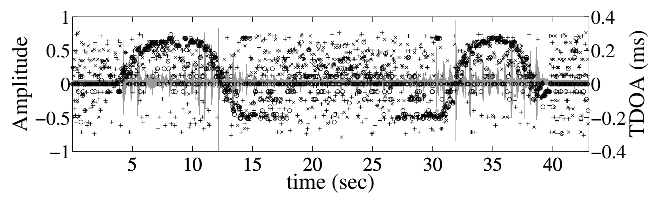
Fig. 3. TDOA estimation for a non-moving speech source.



(a) Recording room and the path of footsteps



(b) $p_{\text{TDOA}}(\tau)$ for the footsteps



(c) TDOA estimated by three methods.

Fig. 4. TDOA estimation for footsteps. Same legends as in Fig. 3(b).

the LM386 IC [10] before digitally sampled. A short period of silence was left at the beginning of the source signal for the purpose of estimating the noise level σ^2. Fig. 3 shows the results of TDOA estimation. In Fig. 3(a), the brightness represents the value of $p_{\text{TDOA}}(\tau)$ calculated in Eq. (8). Possibly because the loudspeaker was placed pretty close to the microphones and hence the SNR was sufficiently high, the proposed method performed well throughout the duration of the test. For this experiment, the results of TDOA estimation by all three estimators are shown in Fig. 3(b).

In another experiment, we used real footsteps in the lab as the source signal to test the TDOA estimators. Fig. 4(a) depicts the floor-plan of the space; the lab was partitioned in the middle, and the microphones were placed on a desk near the middle of the room. The path the person took while walking around the office is marked with the numbers indicating the time (in seconds) when the person reached a particular location. The size of the space was about 6.7×7.0 m^2.

Fig. 4(b) and (c) show the results of TDOA estimation. The estimated TDOA clearly correlates to the direction of sound arrival, except during $t = 17$ to 27 sec when the direct paths from the footstep to the microphones were blocked by the partition. Nevertheless, when direct paths were cleared, the proposed algorithm appeared to perform better in the sense that its results had fewer outliers than GCC (marked with $\circ$) and Hist+PW (marked with $+$). However, many of the outliers might be due to the footsteps being intermittent; sometimes, the source signal was just silent.

4. CONCLUSION

In this work, we proposed a new method for estimating the TDOA from signals received by a pair of microphones. Simulation showed that, the proposed method outperforms an existing IPD estimation method [2] and the well-known GCC-PHAT at all levels of SNR. The performance is improved largely due to the present method's capability to adjust the probabilistic model based on dynamic estimation of the signal level and the noise floor. However, this relies on the assumption that the amplitude of the signal does not change in a short period of time, and the appropriate time course may be difficult to determine. It surely depends on the characteristics of the signals of interest. The effectiveness can thus be improved if a more accurate signal model is available.

For indoor applications, reverberation is also an important issue to address in the future. In our "footstep" experiment, the reverberation characteristics might have been just like noise since the office was filled with a lot of furnitures. The reverberation did not affect our experiment much, but should be investigated further to make the method more robust.

Finally, the value of the present method is not only to provide an estimate of the TDOA, but also an estimate of its PDF. Under a networked configuration, such PDFs can be transformed and combined across sensor nodes, in a manner similar to [7], to achieve sound source localization in a collaborative manner. Future work along this direction is warranted.

5. REFERENCES

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] F. Fujii, N. Hogaki, and Y. Watanabe, "A simple and robust binaural sound source localization system using interaural time difference as a cue," in *Proc. IEEE Int. Conf. Mechatronics and Automation*, Takamatsu, Japan, Aug. 2013, pp. 1095–1101.

[3] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 2, pp. 439–443, 2013.

[4] E. J. Hannan and P. J. Thompson, "The estimation of coherence and group delay," *Biometrika*, vol. 58, pp. 469–481, 1971.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[6] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, vol. 5, pp. V–341–V–344.

[7] H.-K. Hao, H.-M. Liang, and Y.-W. Liu, "Particle methods for real-time sound source localization based on the Multiple Signal Classification algorithm," in *Proc. IEEE Int. Conf. Intelligent Green Building and Smart Grid*, Taipei, Taiwan, Apr. 2014, pp. 1–5.

[8] Z. Zohny and J. Chambers, "Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL," in *Proc. 19th International Conference on Digital Signal Processing*, Hong Kong, China, Aug. 2014, pp. 59–62.

[9] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013.

[10] National Semiconductor, "LM386 Low Voltage Audio Power Amplifier," Tech. Rep., 2000.