A linear phase unwrapping method for binaural sound source localization on a robot

Danfeng Li, Stephen E. Levinson

Abstract — A robust linear phase unwrapping method is proposed to solve the 2π discontinuities in the phase of the cross power spectrum from the binaural inputs using two omnidirectional microphones. The relative incident angle of the interested sound is then estimated according to the time difference of arrival (TDOA) which is obtained from the unwrapped phase of the cross power spectrum. The frequency components associated with the high power are clustered into groups by the phase and frequency distance, and the dominant group is then used to obtain the initial slope estimation. The phase is unwrapped by checking the difference between the actual and the predicted phase by the estimated slope. The re-estimation is proposed to solve the high power are clustered into groups by the phase according to the spectrum. The frequency components associated with the power field. The re-estimation is then performed by the unwrapped phase. The algorithm is tested under different incident angles and signal to noise ratio (SNR) using real speech signal and white Gaussian noise. The simulation results show the high accuracy and the robustness. This method is also implemented to control a robot to adaptively adjust itself to the position facing the sound source directly. The satisfactory result was achieved in an open house demonstration.

Keywords — Phase unwrapping, Binaural, TDOA, Cross power spectrum, Robot.

I. INTRODUCTION

Sound source localization has been an active research area for years. It is often used in many diverse fields of science and engineering, such as in radar, sonar, astronomy, and seismology, etc. [1] Along with the development of automatic speech recognition (ASR), sound source localization becomes increasingly important in many applications in our daily lives. [2] For example, in video-conference, the camera may track the speaker based on the dynamic speech source localization. In hand-free ASR applications, background noise can be reduced by beam-forming the microphones to the dominant speaker.

Currently existing sound source localization techniques can be loosely divided into three groups. The first group of methods is based on time difference of arrivals (TDOA), where the sound location is estimated by the time delay between the microphones. [2] [3] The second group is the steered-beamformer approach, which is based on steering the array to various positions in space and looking for the maximum power field. [4] The last group is the eigendecomposition-based technique, which takes the subspace approach for multiple signal classification. [5]

When applying TDOA methods for sound source localization, phase unwrapping is the crucial step. Many existing algorithms are aimed to solve the general phase unwrapping problem and computationally complicated. According to the linear characteristic of the phase from the cross power spectrum in the TDOA estimation problem, we proposed a simple and robust linear phase unwrapping algorithm.

This paper is organized as follows. Section II describes the TDOA algorithm. In Section III the new phase unwrapping algorithm is proposed. Section IV and Section V are the experiments and the conclusion.

II. TDOA ALGORITHM

Consider a pair of microphones m1 and m2 are placed symmetrically against the origin on the x axis, the sound source s is located in the x-y plane, as shown in Fig 1. Our goal is to estimate the angle θ between so and the y axis. The distance between the two microphones is denoted by d, and from the source to the origin by r. If r is relatively large compared to d, the incident angles to the two microphones can be treated as the same, which are equal to θ. In this case, θ can be calculated as

$$\theta = \sin^{-1}\left(\frac{C \tau_c}{d}\right),$$  \hspace{1cm} (1)

where C is the speed of sound, τ is the time difference of arrival (TDOA). Therefore, the problem is converted to estimate the τ.

Assuming the relative signal attenuation between the microphones due to propagation distance and source size and orientation are negligible. The received signals at the two microphones can be expressed as

$$Y_1[t] = X[t] + N_1[t]$$  \hspace{1cm} (2)
$$Y_2[t] = X[t - \tau] + N_2[t].$$  \hspace{1cm} (3)

where Y1, Y2, X, N1 and N2 are all random sequences. X is the signal generated by the source. N1 and N2 are assumed to be zero mean uncorrelated noises, mainly due to the acquisition and the quantization noise. τ
Fig. 1. Approximation calculation of the incident angle in 2-D.

Here is the time difference of arrival in sample units. If the sampling frequency is $f_s$, the TDOA $\tau = \tau / f_s$.

One way to estimate $\tau$ is by the cross power spectrum. Assuming $X$ is wide sense stationary (WSS). Even though some real signals are not WSS, because the frame size is small, within each frame, the WSS assumption usually holds. The cross correlation of $Y_1$ and $Y_2$ can be expressed as

$$
R_{Y_1Y_2}[k] = E[Y_1[l + k]Y_2[l]]
$$

$$
= E[(X[l + k] + N_1[l + k])(X[l - \tau] + N_2[l])]
$$

$$
= E[X[l + k]X[l - \tau]]
$$

$$
= R_{XX}[k + \tau]. \quad (4)
$$

Then the cross power spectrum is obtained by taking the DFT of $R_{Y_1Y_2}[k]$

$$
S_{Y_1Y_2}(\omega) = DFT(R_{XX}[k + \tau]) = S_{XX}e^{j\omega \tau}. \quad (5)
$$

Note $S_{XX}(\omega)$ is real because $R_{XX}[k]$ is symmetric. Therefore the phase of cross power spectrum

$$
\phi = \omega \tau, \quad (6)
$$

which is a linear function of TDOA.

In practice, DFT need to be calculated by FFT, therefore, only discrete samples are obtained. Equation (6) can rewritten as

$$
\phi = \omega \tau, \quad (7)
$$

where $\phi$ and $\omega$ are vectors. There is normally no solution for this equation, unless $\phi \in R(\omega)$, where $R(\cdot)$ denotes range. However, there is a unique least square solution

$$
\hat{\tau}_{LS} = \arg \min_\tau \| \omega \tau - \phi \|^2_2
$$

$$
\hat{\tau}_{LS} = (\omega^H \omega)^{-1} \omega^H \phi
$$

$$
= \sum_i \omega_i \phi_i \sum_i \omega_i^2
$$

It can be shown this is also the best linear unbiased estimator. [6]

### III. Linear Phase Unwrapping

Even though the equation (6) gives the linear relationship between the phase of the cross power spectrum and the TDOA, In practice, the phase will be wrapped between $-\pi$ and $\pi$. The equation (6) will become

$$
\phi_i = \omega_i \tau + 2\pi k_i, \quad (9)
$$

where $k_i$, the wrapping factor, is an integer that makes $-\pi \leq \phi_i \leq \pi$. In order to estimate the TDOA, which is the slope of the unwrapped phase, the correct $k_i$ must be found first.

There are two important facts of the phase from the cross power spectrum for localization problem. The first one is that the phase is linear, and the second one is that the phase always starts from zero at zero frequency. A straightforward method of phase unwrapping is to use the low frequency part, which is very unlikely to be wrapped, to estimate the slope and then predict the phases in higher frequencies, unwrap the phase if necessary by checking the difference between the predicted and the actual phase, and then re-estimate the slope. [7] This method works fine when the SNR is high in the low frequency part. When this condition does not hold, the false estimation will be resulted.

Figure 2 shows the magnitude and the phase from a cross power spectrum. While a human inspector can easily identify the slope, the method described above, which using the low frequency to make the prediction, will not work at all. Given the knowledge that the phase is linear, a human inspector will first identify the part of spectrum that will be the most useful to estimate the slope, which is relatively smooth comparing to the other part of the spectrum in term of the phase, and then separates this clean part into segments according to the presence of the phase jump (resulting no phase wrapping within each segment), and then estimates the slope from those segments. Inspired by this procedure, we proposed a robust phase unwrapping method, which borrows the clustering idea from pattern recognition.
This method consists of three steps: key segment identification, initial slope estimation, and phase unwrapping and final estimation. There are two assumptions. The spectrum of the signal is continuous, and the signal is dominant at some high power frequencies. Those assumptions usually hold in practice.

In order to find the key segment to be used for the initial estimation of the slope, the frequency components are first ranked by power. The $M$ most significant components are then picked as the potential candidates for the key segment. These components are clustered into groups according to the distances in both frequency and phase direction. The group which contains the most components is assigned as the key segment. The clustering procedure is as follows:

1. Select the $M$ most dominant frequency components by power.
2. Group and label the components by the adjacent frequency distance and phase distance, respectively.
3. Add the two sets of label together for each component. The components with the same label are grouped together.
4. Pick the largest group as the key segment.

Figure 3 shows an illustration of the key segment identification, where the components are first clustered by frequency distance into two groups (group 1 and 2), and then by the phase distance into four groups (group 3 to 6). There are totally four groups generated after combining the results. The largest group is group 5 in this illustration. The "X" denotes the component which is unused in clustering. Figure 2 shows a real example, the group with "*" is the key segment.

The second step is to estimate the initial slope from the key segment. Let $\theta = 2\pi k$. Because there is no phase discontinuity, the following equation holds within the key segment.

$$\phi_i = \omega iT + \theta.$$  

Consider a merit function

$$\chi^2(\tau, \theta) = \sum_{i=1}^{N} \frac{(\phi_i - \omega_i T - \theta)^2}{\sigma_i^2},$$  

where $\sigma_i$ is the uncertainty associated with each measurement $\phi_i$ and $N$ is the total number of components within this segment. The least-square solution of $\tau$ and $\theta$ can be obtained by minimizing the equation (11). At its minimum, derivatives of $\chi^2(\tau, \theta)$ with respect to $\tau$, $\theta$ vanish.

$$0 = \frac{\partial \chi^2}{\partial \tau} = -2 \sum_{i=1}^{N} \frac{\omega_i (\phi_i - \omega_i T - \theta)}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial \theta} = -2 \sum_{i=1}^{N} \frac{\phi_i - \omega_i T - \theta}{\sigma_i^2}.$$  

For the simplicity of the notation, define

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad S_{\omega} \equiv \sum_{i=1}^{N} \frac{\omega_i}{\sigma_i^2}, \quad S_{\phi} \equiv \sum_{i=1}^{N} \frac{\phi_i}{\sigma_i^2},$$

$$S_{\omega \omega} \equiv \sum_{i=1}^{N} \frac{\omega_i^2}{\sigma_i^2}, \quad S_{\omega \phi} \equiv \sum_{i=1}^{N} \frac{\omega_i \phi_i}{\sigma_i^2}.$$  

Then the equation (12) and (13) become

$$S_{\omega \omega} \tau + S_{\omega} \theta = S_{\omega \phi}$$

$$S_{\omega} \tau + S_{\phi} \theta = S_{\phi}$$

The solution of these two equations will be

$$\hat{\tau} = S_{\omega \omega} S_{\phi} - S_{\omega} S_{\phi}$$

$$\hat{\theta} = S_{\omega \phi} S_{\phi} - S_{\phi}^2.$$

Fig. 3. An illustration of key segment identification.
However, the \( \hat{\theta} \) has to be equal to \( 2\pi k \), therefore
\[
\hat{\theta} = 2\pi k, \quad \hat{k} = \arg\min_k |\hat{\theta} - 2\pi k|.
\] (18)

Then the initial estimate of \( \tau \) can be obtained by
\[
\hat{\tau}_{\text{ini}} = \frac{S_{\phi} - S_{\hat{\theta}}}{S_{\omega}}
\] (19)

The uncertainty factor \( \sigma_i \) is unknown in practice, the power is used instead. The assumption is that the higher the power, the lower the uncertainty. The relation between them is
\[
\frac{1}{\sigma_i^2} = W_i
\] (20)

where \( W_i \) is the power of the \( i \)-th component. The final estimation formulas for \( \hat{\theta} \) and \( \hat{\tau}_{\text{ini}} \) become
\[
\hat{\theta} = \frac{\sum_{i=1}^{N} W_i \omega_i^2 \sum_{i=1}^{N} W_i \phi_i - \sum_{i=1}^{N} W_i \omega_i \sum_{i=1}^{N} W_i \phi_i}{\sum_{i=1}^{N} W_i \omega_i^2} - (\sum_{i=1}^{N} W_i \omega_i)^2
\]
\[
\hat{\tau}_{\text{ini}} = \frac{\sum_{i=1}^{N} W_i \omega_i (\phi_i - \hat{\theta})}{\sum_{i=1}^{N} W_i \omega_i^2}
\] (21) (22)

Note that the equation (22) is consistent with equation (8) except the offset term \( \hat{\theta} \). The \( \hat{\theta} \) is obtained by equation (18) according to \( \hat{k} \). It is also an unbiased estimation given the wrapping factor \( k \) is correct.

The final step is the phase unwrapping and the final estimation of the slope. Given the phase is linear and start from zero at the zero frequency, the phase at each frequency can be predicted by the initial slope. Compared with the predicted phase, the actual phase from the spectrum can be unwrapped. Then the slope can be re-estimated by Equation (22) using all the points. Figure 4 shows a phase unwrapping result.

IV. EXPERIMENT RESULTS

Simulations are performed on real speech signal from TIDIGITS database. The digit "0" is used for testing. The sampling frequency is 20kHz. A hanning window is applied to each frame before the cross power spectrum is calculated to eliminate the effect of the end points. The frame size and the window length is 512, which corresponds to 25.6ms. 1024 point FFT is performed in calculating the cross power spectrum. The test is performed under different SNR, and different sample delay. Each situation is repeated for 100 times. Table I shows the mean and variance of the results. Instead of using the incident angle, the sample delay between the two microphones is used to show the results. For example, under 5dB, when the actually delay is 10 samples, the mean of 100 simulation is 9.9585 and the variance is 0.1091. That means in most of the case the error is within one sample, which reflects high accuracy.

<table>
<thead>
<tr>
<th>Mean</th>
<th>delay \ SNR</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>14.9714</td>
<td>14.9681</td>
<td>15.0309</td>
<td>15.0791</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.9881</td>
<td>4.9900</td>
<td>5.0287</td>
<td>4.9671</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.0044</td>
<td>0.9858</td>
<td>0.9999</td>
<td>1.1578</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.0016</td>
<td>0.0026</td>
<td>0.0243</td>
<td>0.0480</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance</th>
<th>delay \ SNR</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>0.0025</td>
<td>0.0225</td>
<td>0.0744</td>
<td>0.4861</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0021</td>
<td>0.0260</td>
<td>0.1091</td>
<td>0.6039</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0024</td>
<td>0.0251</td>
<td>0.0720</td>
<td>0.5813</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0020</td>
<td>0.0237</td>
<td>0.0753</td>
<td>0.6644</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.0020</td>
<td>0.0265</td>
<td>0.0956</td>
<td>0.5528</td>
<td></td>
</tr>
</tbody>
</table>

When the assumption of the phase unwrapping algorithm holds: the spectrum of the signal is continuous and the signal power is dominant at some high power frequencies of the corrupt signal, the algorithm performed relatively well even under -5dB SNR. This is because the algorithm can pick up the key components in a very noisy spectrum, and uses those key components to make the estimation, which is very robust against the noise.

The algorithm is also tested on a robot in a real life
environment. Two omnidirectional microphones are mounted on a robot. When people speak to the robot, it will turn itself towards the speaker. This was a part of the demonstration at 2001 Beckman Open House, University of Illinois at Urbana-Champaign, March, 2001. A person with a food can is walking around and call the robot, the robot will turn to the person and locate the food can by the onboard camera, then run into the food can and grab it by its hands. The demonstration area is about 4m×4m, the localization accuracy is about ±10°, which is good enough for the robot to see the object.

V. CONCLUSION AND FUTURE WORKS

A new and robust linear phase unwrapping algorithm is proposed for the sound source localization using two microphones. Because no searching or re-iteration are involved, the algorithm is fast and computationally effective. In simulations the algorithm performed relatively well even under −5dB SNR. It has also shown the efficiency in a real, noisy environment.

The possible future works will include multiple sources localization and object tracking.

ACKNOWLEDGMENTS

The authors would like to thank Kevin M. Squire for helping to setup the robot and also like to thank Matthew D. Kleffner for his audio acquisition program.

REFERENCES