Sound Source Localization Using a Profile Fitting Method with Sound Reflectors

Osamu ICHIKAWA†, Tetsuya TAKIGUCHI†, and Masafumi NISHIMURA†, Members

SUMMARY

In a two-microphone approach, interchannel differences in time (ICTD) and interchannel differences in sound level (ICLD) have generally been used for sound source localization. But those cues are not effective for vertical localization in the median plane (direct front). For that purpose, spectral cues based on features of head-related transfer functions (HRTF) have been investigated, but they are not robust enough against signal variations and environmental noise. In this paper, we use a “profile” as a cue while using a combination of reflectors specially designed for vertical localization. The observed sound is converted into a profile containing information about reflections as well as ICTD and ICLD data. The observed profile is decomposed into signal and noise by using template profiles associated with sound source locations. The template minimizing the residual of the decomposition gives the estimated sound source location. Experiments show this method can correctly provide a rough estimate of the vertical location even in a noisy environment.

key words: sound source localization, microphone array, sound reflector, ICTD, ICLD

1. Introduction

In a two-microphone array system, the interchannel cues (ICTD and ICLD) are often referred to for horizontal localization. There have also been several attempts to apply ICTD and ICLD for vertical localization outside of the median plane [1]. In the median plane, ICTD and ICLD do not contribute to vertical localization [2] since they are minimized. To achieve vertical localization in the median plane, it was suggested that a spectral cue model [3], [4] be integrated. However, since the spectral cues depend on the spectrum of the signal source, they are not robust enough against signal variations and environmental noise. Also, it may require special considerations to consolidate the interchannel cues (ICTD and ICLD) and the spectral cues in one localization system [5].

In this paper, we enhance the localization cues for a specific reflection by using reflectors correlated with the location of the sound source. We call this a reflection cue. It can be detected by CSP analysis directly, or it can be observed as a modification of the ICTD, ICLD, or the profile. By using this reflection cue, we believe equi-distant vertical localization in the median plane becomes possible without relying on the spectral cues.

For noise robustness, we introduce the Profile Fitting (PF) method for sound source localization. It was originally proposed for speech enhancement [6], but we show it is also effective for localization in a noisy field because of its noise reduction feature. For the conventional method using ICTD and ICLD, several techniques have been proposed to improve the performance in noisy fields [1], [7], [8]. One of them is to use the onsets (i.e. energy peaks) to get a locally high signal-to-noise ratio (SNR). Another technique is to train the probability density function of the sound location in the actual noise field. However those methods do not have a function to subtract noise, so they depend on the SNR where ICTD and ICLD are trained.

2. Reflector Design

2.1 Reflector Design for Vertical Localization

In the HRTF approach, the pinna shape is a given parameter. In our approach, we deliberately designed the shape of a pinna-like reflector so that the following process can retrieve the localization cues provided by the reflector.

Figure 1 shows the concept of the design. The ellipses are plotted where the two foci for each ellipse are at the microphone location and one of the candidate locations of the sound source. The reflector shape is given by the envelope curve for these ellipses. At the upper part of the reflector, sound waves from a high elevation are reflected to focus on the microphone. At the lower part of the reflector, sound waves from a low elevation are reflected so as to focus on it. Sound waves from unmatched elevations should be diffused by the reflection. Therefore the microphone receives both a direct wave and a reflected wave whose delay time is correlated with the sound source elevation. It should be noted that the actual reflector has a 3D-shape designed as an envelope of the revolutions of the ellipses (spheroids).
ICHIKAWA et al.: SOUND SOURCE LOCALIZATION USING A PROFILE FITTING METHOD WITH SOUND REFLECTORS

2.2 Verifying Prototype Reflector Using CSP Analysis

For our experiment, the reflector was made of gypsum molded from a handmade clay model. We verified the working accuracy by Cross-power Spectrum Phase (CSP) analysis [9] to check that the reflector generated the desired main reflected wave according to the sound source location.

Figure 2 shows the configuration for this test. Human speech in calls for attention (“oh-i”, “moshi-moshi”, etc. in Japanese) of about 5 seconds in length were played back in a soundproof chamber using a loudspeaker located directly in front at a distance of 2 m with elevation angles of $0^\circ$, $15^\circ$, $30^\circ$, $45^\circ$, and $60^\circ$. Two microphones with reflectors recorded the sound signal at a 48 KHz sampling frequency.

As shown in Fig. 3, the output of CSP analysis shows many sub-peaks, so the criteria of the intensity for the acceptable sub-peaks are arbitrary. Here we took the top 3 peaks whose intensities were greater than a tenth of the main peak as valid peaks. Table 1 shows the result of the analysis. The peak in first place is the main peak representing a direct wave. It was observed at position 0. This means the signal source was directly in front. In second and third places, two sub-peaks caused by correlations between the direct wave and the reflected wave should be detected at the designated positions. In these experiments, we observed at least one sub-peak at the designated positions except for $0^\circ$, where the area of the designed surface for the reflection (at the root of the reflector) was zero. The absence of an intense reflection can also be treated as a localization cue.

3. Sound Source Localization

CSP analysis can be used for sound source localization. However, this depends on the assumption that the specific reflected wave is distinct. In a noisy environment, it is difficult for CSP analysis to detect the specific reflected wave, because the sub-peaks associated with the noise sources become dominant. Also, the specific reflected wave can be distinct only when a signal source is located exactly on the designated positions and the working accuracy of the reflectors is precise. Therefore, the conventional method using ICTD and ICLD, and the Profile Fitting method using a profile are investigated in this section. They do not directly utilize the specific reflected wave, but we expect the design method discussed in Sect. 2 will work to make the large modification in the ICTDs, ICLDs, and profiles, so that the localization methods can utilize these reinforced localization cues.

3.1 Conventional Method Using ICTD and ICLD

The probability density function, the likelihood that a source is located at a particular position, can be approximated by the product of the marginal distribution of the ICTD and ICLD at each sub-band frequency [1], [7]. We applied the Gaussian distribution for the likelihood as Equation (1):

$$
\Psi_n = K \cdot \exp \left[ -\frac{1}{2} \sum_\omega \sum_T \left\{ \left( \frac{ICTD_{\omega,T} - \overline{ICTD}_{n,\omega}}{\sigma^2_{ICTD,\omega}} \right)^2 + \left( \frac{ICLD_{\omega,T} - \overline{ICLD}_{n,\omega}}{\sigma^2_{ICLD,\omega}} \right)^2 \right\} \right]
$$

where $\Psi_n$ is the likelihood expected for a signal source at $n$, $\omega$ is the sub-band frequency, $T$ is the time frame number, $\sigma^2_{ICTD,\omega}$ and $\sigma^2_{ICLD,\omega}$ are the variances of the interchannel differences under consideration, and $K$ is a normalizing constant.

We defined the interchannel differences and the variances in Equations (2) to (7):

$$
ICTD_{\omega,T} = \angle \left( R_{\omega,T} \cdot L^*_{\omega,T} \right) \frac{f}{2\pi\omega}
$$

$$
ICLD_{\omega,T} = 10 \log \left( \frac{|R_{\omega,T}|}{|L_{\omega,T}|} \right)
$$

$$
\overline{ICTD}_{n,\omega} = \frac{1}{N_T} \sum_T ICTD_{\omega,T}|_{\text{source=n}}
$$

$$
\overline{ICLD}_{n,\omega} = \frac{1}{N_T} \sum_T ICLD_{\omega,T}|_{\text{source=n}}
$$
\[
\overline{\text{ICLD}}_{n,\omega} = \frac{1}{N_T} \sum_{T} |\text{ICLD}_{n,\omega,T}|_{\text{source}=n} \tag{5}
\]
\[
\sigma^2_{\text{ICTD},\omega} = \frac{1}{N_n N_T} \sum_{n} \sum_{T} (|\text{ICTD}_{n,\omega,T}|_{\text{source}=n} - \overline{\text{ICTD}}_{n,\omega})^2 \tag{6}
\]
\[
\sigma^2_{\text{ICLD},\omega} = \frac{1}{N_n N_T} \sum_{n} \sum_{T} (|\text{ICLD}_{n,\omega,T}|_{\text{source}=n} - \overline{\text{ICLD}}_{n,\omega})^2 \tag{7}
\]

Here, \( R_{n,T} \) and \( L_{n,T} \) are the short-time Fourier transforms of the observations for each of the right and left channels, \( N_T \) is the total number of frames to be examined, and \( N_n \) is the total number of candidate locations. ICLD is measured in dB and ICTD is measured in units of the sampling count.

We selected time frames of 0.2 sec around the onset for the each utterance to be examined.

Before the experiment, \( \overline{\text{ICTD}}_{n,\omega}, \overline{\text{ICLD}}_{n,\omega}, \sigma^2_{\text{ICTD},\omega}, \) and \( \sigma^2_{\text{ICLD},\omega} \) should be trained using a signal from each candidate location \( n \) with or without noise at a specific SNR.

### 3.2 Reflector Effect on ICTD and ICLD

If the left and right reflectors are configured completely symmetrically, ICTD and ICLD still take near-zero values. However, as shown in the CSP output of our prototype (Fig. 3), the desired reflected waves generated by the actual left and right reflectors are not necessarily at the same level. In that case, the ICTD and ICLD values are significantly modified by the reflected waves. Also, it is difficult to predict the actual modification before measurement, because there are many reflected waves and their levels are not balanced. The expectation here is that the reflectors should just cause large modifications at the characteristic positions. For an example, Fig. 4 shows the ICTDs with and without our reflectors. Without reflectors, ICTD plots are similar against variations of signal source elevation. This implies it is difficult to determine the signal source elevation by ICTD without reflectors. With reflectors, we can observe the shape of ICTD plots varies a lot against signal source elevation. As the localization process checks the shapes as a whole, it should not be a problem, even if they are partially similar, under the assumption that the signal is broadband.

### 3.3 Profile Fitting Method

For robustness against noise, we introduce a PF method for sound source localization utilizing the residual of the approximate decomposition of signal and noise. It is based on the concept that the power distribution observed at varying look direction can be approximated by the linear combinations of the template distributions, each associated with a signal source and a noise source. When the assumed location \( n \) is correct, Equation (8) is justified.

\[
X_{\omega}(\theta) \equiv \alpha_{n,\omega} \cdot P_{n,\omega}(\theta) + \beta_{n,\omega} \cdot Q_{\omega}(\theta) \tag{8}
\]

Here, \( X_{\omega}(\theta) \) is the power distribution of the sub-band frequency \( \omega \) observed at the particular look direction \( \theta \) for a delay and sum beamformer. This is called an “observed profile”. \( P_{n,\omega}(\theta) \) is a “template profile” measured by white noise coming from the candidate location \( n \) for the signal source. \( Q_{\omega}(\theta) \) is a “template profile” measured for the noise source. The template profile for the noise source can be measured using a white noise originating from the noise source before the experiment if the location of the noise source is known a priori. Otherwise it should be measured from the actual noise by averaging over noise segments during the experiment.

The PF method determines each of the weight coefficients \( \alpha_{n,\omega} \) and \( \beta_{n,\omega} \) for the template profiles of a signal source and a noise source, so as to minimize the evaluation function \( \Phi_{n,\omega} \) defined by Equation (9):

\[
\Phi_{n,\omega} \equiv \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \left[ X_{\omega}(\theta) - \alpha_{n,\omega} \cdot P_{n,\omega}(\theta) - \beta_{n,\omega} \cdot Q_{\omega}(\theta) \right]^2 d\theta
\tag{9}
\]

We configure the delay and sum beamformer in the time domain, using Equation (10), and the observed profile \( X_{\omega}(\theta) \) is derived by using Equations (11) and (12):
\[ s(t, \theta) = l(t) + r(t + \theta) \]  
\[ S_{\omega,T}(\theta) = DFT [s(t, \theta)] \]  
\[ X_{\omega}(\theta) = \frac{1}{N_T} \sum_{T} S_{\omega,T}(\theta) \cdot S_{\omega,T}(\theta)^* \]

Here, \( l(t) \) and \( r(t) \) are the time domain observations of the left and right channels at the \( t \)th sample, and the look direction \( \theta \) is measured by the delay in the samples. \( T \) is the time frame number and \( N_T \) is the total number of frames. Since the template profile should contain only the directivity information, it is normalized by the power at each sub-band as Equation (13):

\[ P_{n,\omega}(\theta) = \frac{\int_{\min \omega}^{\max \omega} X_{\omega}(\theta) |_{source=n} d\theta}{\int_{\max \omega}^{\min \omega} X_{\omega}(\theta) |_{source=n} d\theta} \]

For speech enhancement, the decomposition using Equation (9) should be done in each time frame, but for sound source localization, it should be done only once. Therefore, \( X_{\omega}(\theta) \) is an averaged observation over a few seconds. As the PF method does not rely on onsets, test data can include non-speech frames before and after the utterances. The coefficients \( \alpha_{n,\omega} \) and \( \beta_{n,\omega} \) can be determined by the Variation Method with non-negative conditions.

Once the coefficients are determined, then the residual \( \Phi_{n,\omega} \) can be determined. With Equation (14), we calculate the normalized residual \( \Phi_n \) as a function of \( n \) by dividing the sub-band power and averaging over the \( \Omega \) sub-bands. Using Equation (15), the location of the signal source is estimated as \( \hat{n} \) so as to minimize the normalized residual.

\[ \Phi_n = \frac{1}{\Omega} \sum_{\omega} \Phi_{n,\omega} \]  
\[ \hat{n} = \arg \min_n (\Phi_n) \]

3.4 Reflector Effect on Profile

A profile contains ICTD information as peak-shifts and ICLD information as a bias. Also, diffusion or reflection of the target signal increases the bias of the profile. Therefore, it should be noted that even though the desired reflected waves generated by the left and right reflectors are completely identical, the bias of the profile still retains the reflection cue, while the peak-shift might be zero in that case.

Figure 5 compares the template profile for an elevation angle of 30° with the one for 0°. At the frequency of 3,375 Hz, the peak-shift and bias are observed in the profile for 30°. They are caused by the reflected waves arriving with their own delays.

4. Experiments and Results

4.1 Preliminary Experiment

In order to verify the PF method with the designed reflectors works correctly, we performed a preliminary experiment using a limited amount of data for vertical localization in a sound-proof chamber.

The recording parameters and the geometry are the same as in Sect. 2.2 for the CSP analysis. In a sound-proof chamber, four utterances about 5 seconds in length were played back from each candidate location for a signal source. As a noise source, white noise was played from a loudspeaker at an azimuth angle of 15°, a distance of 1 m, and an elevation angle of 0° (Fig. 6). The recorded noise was manually mixed with the recorded signal, so that the SNR could be controlled.

Before the experiment, the template profiles for the signal sources and the noise source were individually measured using white noise coming from each sound source location.

Using Equations (16) and (17), a score \( \rho \) is introduced to define the relative degrees of superiority using the second best (smallest) normalized residual as the baseline value. Here, \( n^* \) denotes the correct location. When the correct location has the minimum value, it should be selected by Equation (15) and the score will have a positive value. If the normalized residual is zero, the score becomes 100%. A positive large score means it is estimated with high confidence. If the score decreases close to zero, it means the chances increase that the second best candidate might be incorrectly taken as a result of noise or some other influence. If the correct location does not have the minimum value, then Equation (15) will fail to select the correct location, and the score will have a negative value.
\[ \rho = \frac{\overline{\Phi}_n - \Phi^c_n}{\overline{\Phi}_n} \]  
\[ \overline{n} = \arg\min_{n \neq \overline{n}} (\Phi_n) \]  

On calculating the normalized residual in Equation (14), an averaging operation was performed over the sub-band frequencies from 938 Hz to 7,453 Hz where the reflector effect is most apparent.

Figure 7 shows the experimental results. All elevations maintain large positive scores in spite of SNR degradation. This means the correct signal location was selected from the five candidates without being affected by noise, showing the superiority of the approximate decomposition by the PF method. On the other hand, the reference experiment (marked * in Fig. 7) without using the template profile for the noise source failed in the noisy environment.

4.2 Experiments in a Realistic Environment

In order to evaluate the capability in more realistic conditions, we performed an experiment using more utterances from more locations in a slightly reverberant meeting room with realistic noise.

As shown in Fig. 8, 21 locations were defined as a signal source location. They are also candidate locations for the localization. They have 5 horizontal steps from −30° to +30°, and 5 vertical steps from 0° to 60°. As a noise source, cafeteria noise in stereo was played from two loudspeakers at azimuth angles of 30° and −30°, a distance of 2 m, and an elevation angle of 0°. The recorded noise was manually mixed with the recorded signal, so that the SNR could be controlled. The recording was done in our meeting room whose reverberant time is about 0.22 sec.

Per location, a total of 108 utterances of personal names spoken by 6 male and 6 female speakers were played back. In order to evaluate the robustness, we projected an imaginary grid around each candidate location as shown in Fig. 9, and played back almost same numbers of utterances from each grid point. Here, we categorize the utterances by the offset error from the candidate location. Category A is for the utterances from the exact candidate location. Category B is for the utterances whose azimuth angle and elevation angle are correct but whose distance contains about ±10% error. Category C is for the utterances whose azimuth angle is correct but whose elevation angle contains about ±4° error or whose distance contains about ±10% error. Category D is for all the utterances that contain at least one of the errors in azimuth of about ±4°, elevation of about ±4°, or distance of about ±10%. It should be noted Category B, Category C and Category D do not include Category A, and therefore the Categories other than A involve offset errors in one or multiple dimensions.

The sizes of the offset errors should not be too large with reference to the design points and the neighboring candidate locations. Here, the offset errors in azimuth and elevation are about a quarter of the angles between the candidate locations. The offset error in distance is chosen as a simple fraction of the distance between the microphone and the candidate locations, so that it will be near to the actual length of the offset errors in azimuth and elevation.

For the PF method, the template profile for the noise source was measured from the actual noise for 1 sec just before each utterance. It should be noted that the template does not contain any spectral information, but just records the directivity information as it is normalized by a power at each sub-band.

![Fig. 8](image-url)  
Testing configuration for the experiment in a realistic environment.

![Fig. 9](image-url)  
Category by the offset from the location.
Both for the PF method and the conventional method, the sub-bands to be examined were selected from 938 Hz to 7,453 Hz where the reflector effect is most apparent.

Figure 10 shows the success rates for the localization of 5 signal source locations in the median plane out of 21 candidate locations. The SNR was 11 dB. Both the PF method and the conventional method (trained by the utterances in Category A) showed high success rates for the utterances in Category A that have very little offset error from the candidate locations. On the other hand, the success rates are significantly decreased for the utterances in Category D that have much larger offset errors. Figure 10 also shows the result of the conventional method trained using the utterances in Category D. This improved the success rate for the utterances in Category D. In that case, the probability density
functions have broad distributions, as they are trained with large offset errors associated with Category D. Therefore, that causes a significant loss of accuracy for the utterances in Category A.

In order to evaluate the dependency on SNR, we also tried this localization without adding noise. The SNR was 28 dB. Figure 11 shows the resulting success rate. It also shows the result of the conventional method that was trained in a noisy environment (11 dB). In that case, the SNR was unmatched between the training and the localization. The success rate of this unmatched case was worse than the matched cases shown in Fig. 10 (at 11 dB) and Fig. 11 (at 28 dB). We conclude the conventional method is dependent on the SNR when it is trained. Also, there is concern that the conventional method is dependent on the noise color as well as the SNR, because the probability density functions are trained for each sub-band. On the other hand, the PF method is less dependent on them, because it does not require any training in advance.

Using not only the 5 signal source locations in the median plane, but also using all of the 21 signal source locations, Fig. 12 shows the success rates resulting for the localization out of 21 candidate locations. We see the PF method outperformed the conventional method in all categories. It should be noted that the conventional method checked the utterances only around onsets where the SNR was locally high, both for training and localization. The PF method did not use this technique and still had an advantage in the experimental results.

Figure 13 shows maps of the signal source locations and the estimated locations for the utterances from the 5 signal source locations in the median plane in Categories A, B, and C. In the error cases, the locations estimated by the PF method were closer to the correct locations than the ones using the conventional method. This trend was still observed when the SNR was reduced to 11 dB. In both methods, the azimuth estimation was very accurate.

5. Concluding Remarks

We have proposed a design method for a reflector that generates reflection cues for vertical localization. We have also proposed a framework for sound source localization using the PF method. This can reduce the effect of noise by exploiting the approximate decomposition of signal and noise. In the PF method combined with reflectors, the process for horizontal localization and the process for vertical localization can be consolidated into a single process. Experiments showed this method can correctly provide a rough estimate of the vertical location in the median plane even in a noisy environment. The PF method showed more robustness against SNR variations than the conventional method using ICTD and ICLD.

References


Osamu Ichikawa received B.E. and M.E. degrees in Aeronautical Engineering from Tokyo University in 1986 and 1988, respectively. In 1988 he joined IBM Japan, Ltd. From April 1999 to March 2001, he was a funded researcher at the Institute of Space and Astronautical Science. He is now working on signal processing for automatic speech recognition at the Tokyo Research Laboratory. His research interests include microphone arrays, robust speech recognition, and computational fluid dynamics. He is a member of the Acoustical Society of Japan and the Japan Society of Mechanical Engineers.

Tetsuya Takiguchi received the B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and the M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1996 to 1999, he was a Research Assistant at Nara Institute of Science and Technology. He is currently a research staff member at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. His research interests include robust speech recognition, auditory scene analysis, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE and the Acoustical Society of Japan.
Masafumi Nishimura received his B.E. and M.E. degrees in Biophysical Engineering from Osaka University in 1981 and 1983, and his Dr. Eng. degree from Toyohashi University of Technology in 1998. In 1983 he joined the Tokyo Research Laboratory of IBM Japan, working on speech and language processing. He is currently a group leader of the speech technology group of TRL. He received the SIG Research Award from the Information Processing Society of Japan in 1998, and the Outstanding Technological Development in Acoustics Prize from the Acoustical Society of Japan in 1999. He is a member of the IPSJ and the ASJ.