SOUND SOURCE LOCALIZATION USING A PINNA-BASED PROFILE FITTING METHOD

Osamu ICHIKAWA, Tetsuya TAKIGUCHI, and Masafumi NISHIMURA

Tokyo Research Laboratory, IBM Japan Ltd.
Yamato-shi, 242-8502, JAPAN
E-mail: ichikaw@jp.ibm.com

ABSTRACT

In a two-microphone approach, interaural differences in time (ITD) and interaural differences in sound intensity (IID) 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 pinna specially designed for vertical localization. The observed sound is converted into a profile containing information about reflections as well as ITD and IID 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.

1. INTRODUCTION

In a binaural system, ITD and IID are often cited for horizontal localization and vertical localization. But many of the attempts avoid vertical localization in the median plane, where ITD and IID are minimized [1]. To rectify this problem, it is suggested that a spectral cue model [2] be integrated. However, since the spectral cues depend on a spectrum with a locally constant slope, it is not always applicable in certain noisy environments.

In this paper, we enhance the localization cue for a specific reflection by using pinna 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 ITD, IID, or the profile. By using this reflection cue vertical localization in the median plane becomes possible without relying on the spectral cue.

For noise robustness, we introduce the Profile Fitting (PF) method for sound source localization. It was originally proposed for speech enhancement [3], but we show it is also effective for localization in a noisy field because of its noise reduction feature. For the conventional method using ITD and IID, several methods have been proposed to improve the performance in noisy fields. Martin [1] used onsets (i.e. energy peaks) to get a locally high signal-to-noise ratio (SNR). Nix et al. [4] trained 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 ITD and IID are trained.

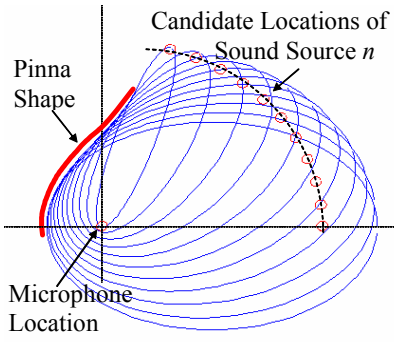

Fig. 1. Concept of pinna design.

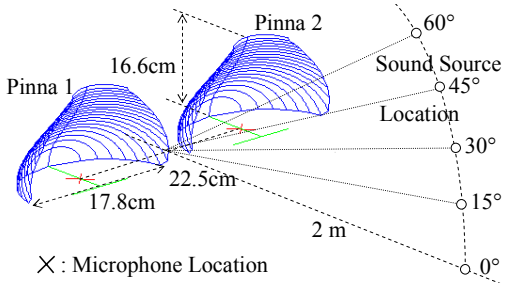

Fig. 2. Testing configuration.

2. PINNA DESIGN

2.1. Pinna Design for Vertical Localization

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

Fig. 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 pinna shape is given by the envelope curve for these ellipses. At the upper part of the pinna, sound waves from a high elevation are reflected to focus on the microphone. At the lower part of the pinna, 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 pinna has a 3D-shape designed as an envelope of the revolutions of the ellipses (spheroids).

From a practical point of view, sound sources are not necessarily at the designated locations, and the working accuracy of pinna is not always sufficient for the detection of the reflected wave. But this design method still works as a guideline to cause large changes in the ITDs, IIDs, and profiles for the sound sources around the designated locations.

2.2. Verifying Prototype Pinna using CSP Analysis

For our experiment, the pinna was made of gypsum molded from a handmade clay model. We verified the working accuracy by Cross-power Spectrum Phase (CSP) analysis [5] to check that the pinna generated the desired main reflective wave according to the sound source location.

Fig. 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 using a loudspeaker located directly in front at a distance of 2 m with elevation angles of 0°, 15°, 30°, 45°, and 60°. Two microphones with pinna 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°, where the area of the designed surface for the reflection (at the root of the pinna) was zero. The absence of an intense reflection can also be treated as a localization cue.

CSP analysis can be used for sound source localization. But, in a noisy environment, it is difficult for CSP analysis to detect a specific reflective wave, because the sub-peaks associated with the noise sources become dominant. Therefore, conventional methods using ITD and IID, and the Profile Fitting method using a profile are investigated in the later sections.

3. SOUND SOURCE LOCALIZATION

3.1. Conventional Method using ITD and IID

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 ITD and IID at each sub-band frequency. Nix et al. [4] defined the marginal distribution by a histogram of the training data. As our following experiment is using a limited amount of speech data, we applied the Gaussian distribution as Martin did [1] for the likelihood as Equation (1):

$$\Psi_n = K \cdot \exp\left[-\frac{1}{2} \sum_{\omega} \sum_{T} \left\{ \frac{\left(ITD_{\omega,T} - \overline{ITD}_{n,\omega} \right)^2}{\sigma^2_{ITD,\omega}} + \frac{\left(IID_{\omega,T} - \overline{IID}_{n,\omega} \right)^2}{\sigma^2_{IID,\omega}} \right\} \right] \quad (1)$$

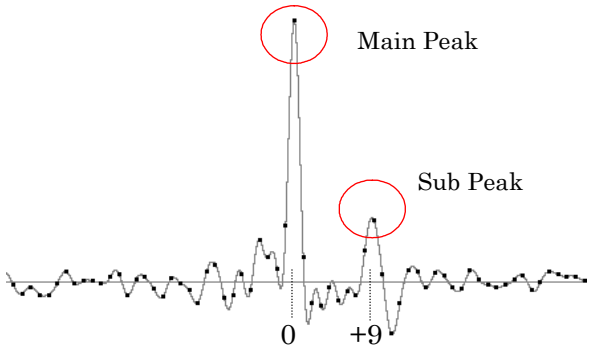

Fig. 3. Output of CSP analysis with pinna for a signal source at an elevation angle of 30°.

Table 1. Peak locations detected by CSP analysis

Elevation angle of sound source	0°	15°	30°	45°	60°
Peak in 1st place	**0**	**0**	**0**	**0**	**0**
Peak in 2nd place	N/A	**10**	**9**	**6**	**2**
Peak in 3rd place	N/A	N/A	N/A	**-6**	-10
Design point	±14	±12	±9	±5.5	±2.5

where Ψ_n is the likelihood expected for a signal source at n, ω is the sub-band frequency, T is the time frame number, $\sigma^2_{ITD,\omega}$ and $\sigma^2_{IID,\omega}$ are the variances of the interaural differences under consideration, and K is a normalizing constant. IID is measured in dB and ITD is measured in units of the sampling count.

In the training step, $\overline{ITD}_{n,\omega}$, $\overline{IID}_{n,\omega}$, $\sigma^2_{ITD,\omega}$, and $\sigma^2_{IID,\omega}$ were measured by listening to a signal from each candidate location n with or without noise at a specific SNR.

3.2. Pinna Effect on ITD and IID

If the left and right pinnas are configured completely symmetrically, ITD and IID still take near-zero values. But, as shown in the CSP output of our prototype (Fig.3), the desired reflection waves generated by the actual left and right pinnas are not necessarily at the same level. In that case, the ITD and IID are significantly modified by the reflection waves. For an example, Fig. 4 shows the ITDs with and without our pinnas for a signal source elevation angle of 30°.

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. The PF method determines the weight coefficient $\alpha_{n,\omega}$ for the template profile of a signal source and the weight coefficient $\beta_{n,\omega}$ for the template profile of a noise source, so as to minimize the evaluation function $\Phi_{n,\omega}$ defined by Equation (2):

$$\Phi_{n,\omega} = \int_{\min\_\theta}^{\max\_\theta} \left\{ X_\omega(\theta) - \alpha_{n,\omega} \cdot P_{n,\omega}(\theta) - \beta_{n,\omega} \cdot Q_\omega(\theta) \right\}^2 d\theta \quad (2)$$

Here, $X_\omega(\theta)$ is the power distribution of the sub-band frequency ω observed at the particular look direction θ 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 using a white noise originating from the noise source. We configure the delay and sum beamformer in the time domain, using Equation (3), and the observed profile $X_\omega(\theta)$ is derived by using Equations (4) and (5):

$$s(t,\theta) = l(t) + r(t+\theta) \tag{3}$$

$$S_{\omega,T}(\theta) = DFT[s(t,\theta)] \tag{4}$$

$$X_\omega(\theta) = \frac{1}{N_T}\sum_T S_{\omega,T}(\theta)\cdot S_{\omega,T}(\theta)^* \tag{5}$$

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 θ 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 in Equation (6):

$$P_{n,\omega}(\theta) = \frac{X_\omega(\theta)|_{source=n}}{\int_{\min\_\theta}^{\max\_\theta} X_\omega(\theta)|_{source=n}\, d\theta} \tag{6}$$

The template profile for the signal source should be measured before the experiment without any noise. The template profile for the noise source can be measured before the experiment if the location of the noise source is known, or it can be measured from the noise itself by averaging over long intervals during the experiment.

For speech enhancement, the decomposition using Equation (2) 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. 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 (7), we calculate the normalized residual $\bar{\Phi}_n$ as a function of n by dividing the sub-band power and averaging over the Ω sub-bands. Using Equation (8), the location of the signal source is estimated as $\hat{n}$ so as to minimize the normalized residual.

$$\bar{\Phi}_n = \frac{1}{\Omega}\sum_\omega \frac{\Phi_{n,\omega}}{\int_{\min\_\theta}^{\max\_\theta}\{X_\omega(\theta)\}^2 d\theta} \tag{7}$$

$$\hat{n} = \underset{n}{\operatorname{argmin}}\left(\bar{\Phi}_n\right) \tag{8}$$

3.4. Pinna Effect on Profile

A profile contains ITD information as peak-shifts and IID 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 reflection waves generated by the left and right pinnas are completely identical, the bias of the profile still retains the reflection cue, while the peak-shift might be zero in that case.

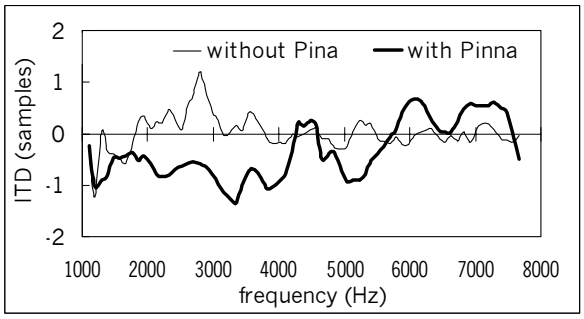

Fig. 4. ITD plots with and without pinna for a signal source at an elevation angle of 30°. The plots are smoothed over 8 sub-bands (=375 Hz).

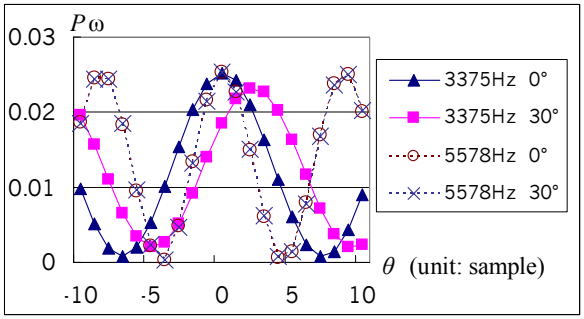

Fig. 5. Template profiles for a signal source at elevation angles of 0° and 30° measured with pinna.

Fig. 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 caused by the reflected waves arriving with about 0.19 ms of delay are observed in the profile for 30°. On the other hand, the two profiles are almost identical at the frequency of 5,578 Hz, where the reflective wave is in phase with the direct wave with one cycle of delay.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The recording parameters and the geometry are the same as in Section 2.2 for the CSP analysis. In a soundproof chamber, four utterances about 5 seconds in length were played back from the 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 speech, so that the SNR could be controlled.

4.2. Conventional Method using ITD and IID

The likelihood parameters $\overline{ITD}_{n,\omega}$, $\overline{IID}_{n,\omega}$, $\sigma^2_{ITD,\omega}$, and $\sigma^2_{IID,\omega}$ were trained with the actual utterances mixed with the noise at a SNR of 16dB. We used the same parameters obtained in the training step for all the SNRs for calculating the likelihood in the experiments, because the SNR in the actual situation is considered to be unknown before the utterances.

In Equation (1), the summation was performed over the sub-band frequencies from 938 Hz to 7,453 Hz where the pinna effect is most apparent. Fig. 7 plots the estimated locations

where the likelihood was maximized. The conventional method was only successful in a limited range of SNRs.

We also tried to train the parameters with white noise coming from the signal source locations instead of the actual utterances. The SNR was same as previous trial. However, the localization performance was severely degraded. It suggested that the performance is dependent on the spectral conformance between the training and the actual localization

4.3. Profile Fitting Method

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

Using Equations (9) and (10), a score ρ is introduced to define the relative degrees of superiority using the second best (smallest) normalized residual as the base value. Here, n° denotes the correct location. When the correct location has the minimum value, it should be selected by Equation (8) and the score has a positive value. If the normalized residual is zero, the score becomes 100%. However when the correct location does not have the minimum value and Equation (8) fails to select the correct location, the score has a negative value.

$$\rho = \frac{\overline{\Phi}_{\overline{n}} - \overline{\Phi}_{n^\circ}}{\overline{\Phi}_{\overline{n}}} \tag{9}$$

$$\overline{n} = \underset{n \neq n^\circ}{\operatorname{argmin}} \left(\overline{\Phi}_n \right) \tag{10}$$

On calculating the normalized residual in Equation (7), an averaging operation was performed over the sub-band frequencies from 938 Hz to 7,453 Hz.

Fig. 8 shows the experimental results. In all cases, the correct signal location was selected from the five candidates without being affected by noise, showing the superiority of the approximate decomposition of the PF method. On the other hand, the reference experiment (marked * in Fig. 8) without using the template profile for the noise source failed in the noisy environment.

5. CONCLUDING REMARKS

We have proposed a design method for a pinna that generates reflection cues for vertical localization. We have also proposed a framework for sound source localization using the PF method. It can reduce the effect of noise by exploiting the approximate decomposition of signal and noise. In the pinna-based PF method, the process for horizontal localization and the process for vertical localization using reflection cues can be consolidated into a single process. Experiments in a soundproof chamber showed this method can correctly provide a rough estimate of the vertical location in the median plane even in a noisy environment. We will verify the performance in a reverberant environment in the near future.

6. REFERENCES

[1] Keith D. Martin, "Estimating azimuth and elevation from interaural differences," *IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio*, Oct. 1995.

[2] P. Zakarauskas and M.S. Cynader, "A computational theory of spectral cue localization", *J. Acoust. Soc. Amer.*, Vol.94, pp. 1323-1331, 1993.

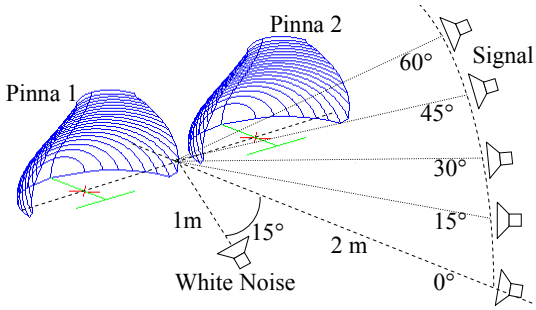

Fig. 6. Sound source geometries for the experiment.

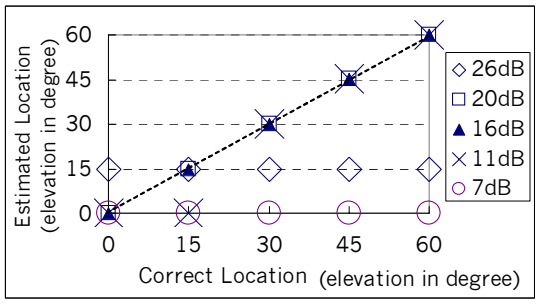

Fig. 7. Estimated locations by conventional method plotted for the utterances from each location at 5 different SNRs.

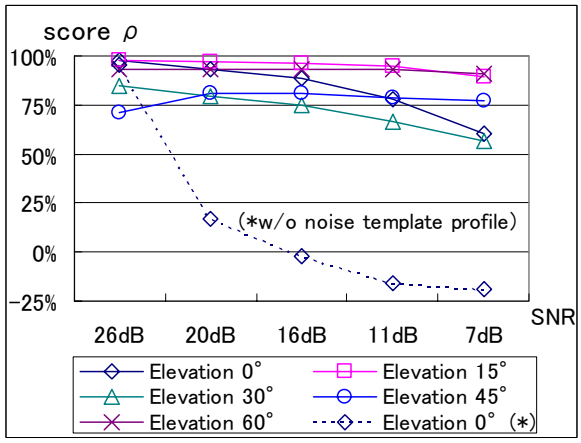

Fig. 8. Resulting score for sound source localization by PF method. (*) denotes a reference trial without using the template profile for the noise source.

[3] O. Ichikawa, T. Takiguchi, and M. Nishimura, "Speech enhancement by profile fitting method," *IEICE Transactions on Info. and Sys.*, Vol. E86-D, No. 3, pp. 514-521, Mar. 2003.

[4] Johannes Nix and V. Hohmann, "Enhancing sound sources by use of binaural spatial cues," *Consistent & Reliable Acoustic Cues for sound analysis One-day workshop*, Sep. 2001.

[5] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. ICASSP94*, pp. 273-276, 1994.