ROBUST LOCALIZATION AND TRACKING OF MULTIPLE SPEAKERS IN REAL ENVIRONMENTS FOR BINAURAL ROBOT AUDITION

Ui-Hyun Kim and Hiroshi G. Okuno

Dept. of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University,
Kyoto, Japan

ABSTRACT

This paper presents a multisource sound localization method based on the generalized cross-correlation (GCC) method weighted by the phase transform (PHAT) and a novel multisource speech tracking method consisting of voice activity detection (VAD) and K-means clustering algorithm for binaural robot audition. The standard K-means clustering algorithm was improved for the purpose of multisource speech tracking by adding two additional steps. Experiments conducted on the SIG-2 humanoid robot in a real environment show that our method can track multiple speakers in real-time with tracking error below 4.35°.

1. INTRODUCTION

'Binaural' literally means having two sound inputs. For a robot, it means having two microphones, one on each side of its head like human ears. Among the various functions required for binaural robot audition, sound source localization (SSL) is one of the most important techniques to achieve more natural and intelligent human-robot interaction (HRI). SSL has been extensively studied by a number of researchers and the primary clues have revealed. They include the interaural level difference (ILD), the interaural time difference (ITD), and the spectral modifications caused by parts of the body (the pinna, head, shoulders, etc.). The ITD, more commonly referred to as the time difference of arrival (TDOA), plays an important role in SSL; the sound signals arrive at each microphone at different times for directions. One of the most widely used SSL methods based on the TDOA between binaural sound inputs is the generalized cross-correlation (GCC) method with phase transform (PHAT) weighting [1]. Since SSL performance generally drops as the number of microphones is reduced, a significant problem to be overcome for binaural SSL is the difficulty with multisource speech localization. The number of speech sources that the binaural robot system can localize in real environments has been limited to a single speaker [2].

In this paper, our approach to realizing multisource speech localization for binaural robot audition is twofold: 1) we analyzed the difficulties with multisource speech localization in real environments for the maximum-likelihood (ML)-based direction-of-arrival (DOA) estimation using the GCC-PHAT method and 2) devised a multisource speech tracking method consisting of voice activity detection (VAD) and improved K-means clustering.

Our multisource speech localization and tracking method was implemented as a real-time system and evaluated experimentally in the binaural audition system of the SIG-2 humanoid robot.

The paper is outlined as follows: Section 2 describes the ML-based DOA estimation using the GCC-PHAT method and addresses its difficulties in estimating correct multiple DOAs. Section 3 gives our solution to the difficulties with multisource localization: a multisource speech tracking method consisting of VAD and improved K-means clustering algorithms. Section 4 presents the experimental results. Section 5 concludes the paper.

2. MULTISOURCE SOUND LOCALIZATION

In this section, we summarize the ML-based DOA estimation using the GCC-PHAT method. Then we address the difficulties with multisource DOA estimation for speech signals in real environments.

2.1. Acoustic Model

This paper employs a F-point short-time Fourier transform (STFT) under a far-field assumption [3]. The observed signals from the left and right microphones in a situation with K sound sources can be mathematically modeled as

$$X_l[f,n] = \sum_{k=1}^{K} \alpha_{lk}[f] \left|S_k[f,n]\right| \exp\left(-j2\pi \frac{f}{F} fs\tau_{lk}\right) + N_l[f,n] \quad (1)$$
$$X_r[f,n] = \sum_{k=1}^{K} \alpha_{rk}[f] \left|S_k[f,n]\right| \exp\left(-j2\pi \frac{f}{F} fs\tau_{rk}\right) + N_r[f,n],$$

where $X_{l/r}[f,n]$, $S_k[f,n]$, and $N_{l/r}[f,n]$ are the f-th elements of the STFT of the measured signals from each of the two microphones (l and r), the sound sources (k denotes the index of each sound source), and uncorrelated additive noise, respectively, on the n-th time frame index; the $f \in \{1,$

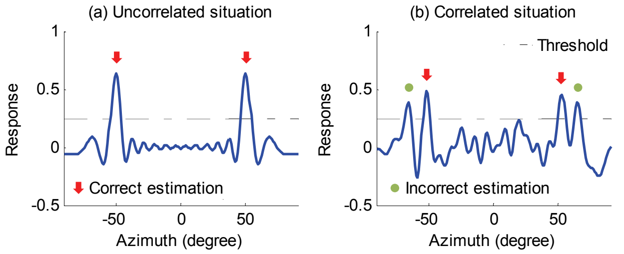

Fig. 1. Peak distributions of the ML-based DOA estimation for two sound sources coming from angles of -50° and +50°: (a) two uncorrelated sources (two sources were vertually generated). (b) two correlated sources (two speech signals were used).

..., F} denotes a frequency bin, F is the time frame size of the STFT, and fs is the sampling frequency; $\alpha_{lk/rk}$ and $\tau_{lk/rk}$ are the attenuation factor and time delay from the position of the k-th sound source to each microphone, respectively.

2.2. ML-Based DOA Estimation Using the GCC-PHAT Method for Binaural Robot Audition

The ML-based DOA estimation for binaural robot audition is basically defined by the GCC-PHAT method and the compensation factor for diffraction of sound waves with multipath interference around the binaural robot head [2] as follows:

$$\hat{\theta}_{mle}[n] = \arg\max_{\theta} \frac{1}{F}\sum_{f=1}^{F} G^{PHAT} X_l[f,n]X_r^*[f,n]\exp\left(j2\pi\frac{f}{F} fs\tau_{lr}(\theta)\right), \tag{2}$$

where

$$G^{PHAT} = \frac{1}{\left|X_l[f,n]X_r^*[f,n]\right|}, \tag{3}$$

$$\tau_{lr}(\theta) = \frac{d_{lr}}{2v}(\theta+\sin(\theta)) - \frac{d_{lr}}{2v}(\mathrm{sgn}(\theta)\pi - 2\theta)\cdot\left|\beta\sin(\theta)\right|; \tag{4}$$

$X_l[f,n]X_r^*[f,n]$, G^{PHAT}, and τ_{lr} are the cross-power spectrum, the PHAT weighting that preserves only the phase information in the cross-power spectrum, and the TDOA compensation factor for multipath interference, respectively; θ_{mle} is the estimated DOA, $\theta \in \{-\pi/2, ..., +\pi/2\}$ is an steering angle of sound incidence, * is the complex conjugate, d_{lr} is the distance between two microphones, v is the speed of sound (340.5 m/s, at 15 °C, in air), $0 < \beta < 0.15$ is the constant attenuation factor, and *sgn* is the signum function that extracts the sign of θ, i.e., ., if θ has the negative sign, then $sgn(\theta)$ will be -1.

2.2. Difficulties with Multisource DOA Estimation for Speech Signals in Real Environments

For the ideal case, the multiple sound sources S_k are uncorrelated with each other and with additive noise $N_{l/r}$; i.e., $E\{S_1[f,n]S_2[f,n]\}=0$, $E\{S_k[f,n]N_{l/r}[f,n]\}=0$, and $E\{N_l[f,n]N_r[f,n]\}=0$. Accordingly, the cross-power spectrum for multiple sound sources is expressed from (1) as

$$X_l[f,n]X_r^*[f,n] = \sum_{k=1}^{K} \alpha_{lk}[f]\alpha_{rk}[f]\left|S_k[f,n]\right|^2 \exp\left(j2\pi\frac{f}{F} fs(\tau_{rk}-\tau_{lk})\right), \tag{5}$$

where TDOAs $(\tau_{rk}-\tau_{lk})$ of K sound sources exist independently and the K directions corresponding to them are represented as a set of K maximum values in (2). However, the accuracy deteriorates when multiple sound sources are correlated, which is generally the case in real environments, i.e., when the sound sources are speech. For example, if two correlated sound sources are assumed to come from different directions, their cross-power spectrum can be expressed as

$$\begin{aligned} X_l[f,n]X_r^*[f,n] &= \alpha_{l1}[f]\alpha_{r1}[f]\left|S_1[f,n]\right|^2 \exp\left(j2\pi\frac{f}{F} fs(\tau_{r1}-\tau_{l1})\right) \\ &+ \alpha_{l2}[f]\alpha_{r2}[f]\left|S_2[f,n]\right|^2 \exp\left(j2\pi\frac{f}{F} fs(\tau_{r2}-\tau_{l2})\right) \\ &+ \alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F} fs(\tau_{r2}-\tau_{l1})\right) \\ &+ \alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F} fs(\tau_{r1}-\tau_{l2})\right). \end{aligned} \tag{6}$$

We can verify that there are two more incorrect TDOAs produced by the correlation between two sound sources in (6). Moreover, if we assume a situation in which there are more than two sources or in which additive noise and reverberation are correlated with other sound sources, the number of incorrect TDOAs will increase geometrically. This phenomenon causes ambiguity in multisource DOA estimation because there will be many peaks in incorrect directions as well in correct ones. Figure 1 shows examples of peak distributions in multisource DOA estimation for two sound signals coming from angles of -50° and +50°. When the sound sources were correlated, the ML-based DOA estimation inaccurately estimated multiple DOAs because of the numerous peaks spread in all directions.

3. MULTISOURCE SPEECH TRACKING

This section describes our solution to the difficulties with multisource DOA estimation described above. Our approach is to use data mining to eliminate incorrect DOAs estimations in each time frame in order to get accurate multisource sound localizations. For this purpose, we devised a multisource speech tracking module based on two methods: A statistical model-based VAD and Improved K-means clustering.

3.1. Statistical Model-Based VAD Algorithm

If the target sound sources are localized speech in noisy environments, all DOA estimations during the noisy periods can be eliminated by using the VAD method. We used the statistical model-based VAD algorithm [4]. Each time frame is determined to be "speech-present" or "speech-absent" by using a decision procedure with a threshold:

$$\begin{aligned} & if\ \hat{P}_{VAD}[n] = \frac{1}{F}\sum_{f=1}^{F}(\gamma[f,n] - \log\gamma[f,n] - 1) > \eta_{VAD} \\ & then \quad n = speech\text{-}present\ frame \\ & else \quad n = speech\text{-}absent\ frame \end{aligned} \tag{7}$$

where $\gamma[f,n]=|X_l[f,n]X_r^*[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* signal-to-noise (SNR) and $\lambda_N[f,n]$ is the estimated variance of $|N_l[f,n]N_r^*[f,n]|$.

3.2. Improved K-means Clustering Algorithm

We improved the standard K-means clustering algorithm to work well for multisource sound tracking in real situations. If the DOAs estimated using (2)–(4) in the given time frames are the observations to be clustered and if their cluster centers represent the tracked DOAs for a specific time frame, i.e., given the initial sets of observations $(\theta_{mle\_1}, \theta_{mle\_2}, \ldots, \theta_{mle\_p})$ and K-clusters $(\Theta_{track\_1}, \Theta_{track\_2}, \ldots, \Theta_{track\_k})$ with their center means $(\theta_{track\_1}, \theta_{track\_2}, \ldots, \theta_{track\_k})$, the standard K-means algorithm proceeds by alternating between two steps:

• Assignment Step. Assign each observation to the cluster with the closest mean:

$$\Theta_{track_k}^{(i)} = \{\hat{\theta}_{mle_p} : \ |\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}|^2 \le |\hat{\theta}_{mle_p} - \theta_{track_j}^{(i)}|^2 \ \forall 1 \le j \le K\}, \tag{8}$$

where p denotes the index of all estimated DOA in the given time frames and i denotes the iteration number. Each initial center mean is randomly assigned and each DOA estimation θ_{mle_p} goes into exactly one cluster Θ_{track}.

• Update Step. Calculate the new means to be the centroid of the observations in each cluster:

$$\theta_{track_k}^{(i+1)} = \frac{1}{\langle\Theta_{track_k}^{(i)}\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\hat{\theta}_{mle_p}, \tag{9}$$

where $<\Theta_{track\_k}>$ is the number of estimated DOAs belonging to cluster Θ_{track_k}. These two steps are repeated until the assignments no longer change.

There are two problems with the standard K-means clustering when it is to be used for multisource speech tracking: 1) Fixed number of clusters. The number of clusters is fixed from the beginning to the end of the standard K-means clustering calculations. This means that the number of speech sources needs to be known in advance for exact clustering. Furthermore, the number of clusters cannot be automatically changed in the observation period for clustering even though speech signals independently appear and disappear over time. 2) Absence of a function for filtering out incorrect DOA estimations. In the standard K-means clustering, the tracked directions of the speech signals are not correct because even incorrect direction estimations are used for calculating the center of each cluster.

For accurate multisource speech tracking, we improved the standard K-means clustering by including two additional steps with new criteria:

• Increase Step. Increase the number of clusters automatically:

$$\begin{aligned} & if \ \frac{1}{\langle\Theta_{track_k}^{(i)}\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\left|\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}\right|^2 > \eta_{C1} \\ & then \ K^{(i+1)} = K^{(i)} + 1 \ and \ move\ to\ \text{Assignment Step} \\ & else \qquad\qquad move\ to\ \text{Elimination Step}. \end{aligned} \tag{10}$$

The K-means clustering algorithm begins with one cluster $(K=1)$. After executing the assignment step and the update step, it adds another cluster $(K=K+1)$ if the variance of observations in each cluster is more than a given threshold η_{C1}.

• Elimination Step. Eliminate clusters containing incorrect direction estimations:

$$\begin{aligned} & if \ \frac{\langle\Theta_{track_k}^{(i)}\rangle}{\sum_{k=1}^{K}\langle\Theta_{track_k}^{(i)}\rangle} < \eta_{C2} \quad then \quad \text{eliminate } cluster \ \Theta_{track_k}^{(i)} \\ & else \qquad\qquad\qquad \text{keep } cluster \ \Theta_{track_k}^{(i)} \end{aligned} \tag{11}$$

The increase step maximizes the number of clusters by using the variance of DOA estimations in each cluster. In this case, some clusters will likely contain few DOA estimations that are all incorrect. The elimination step filters out the clusters containing incorrect direction estimations by checking the ratio between the number of DOA estimations in each cluster and the number of all DOA estimations in the given time frames with a given threshold η_{C2}.

The process of the improved K-means clustering algorithm for multisource speech tracking is thus as follows:

1) The standard K-means algorithm (the assignment step and the update step) is executed with $K=1$.

2) The standard K-means algorithm is repeated with $K=K+1$ on the basis of Criterion (10).

3) All clusters containing incorrect DOA estimations are eliminated on the basis of Criterion (11).

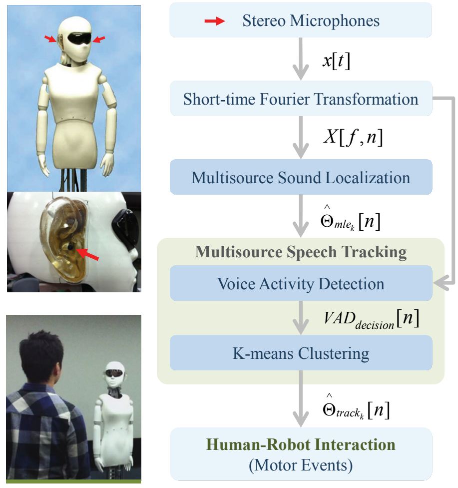

Fig. 2. Flowchart of multisource localization and tracking.

Table 1. RMSEs for two- and three-speaker tracking.

Two speakers positions (°)	Two speakers RMSEs (°)	Three speakers positions (°)	Three speakers RMSEs (°)
(-90, +90)	(1.47, 3.07)	(-60, +30, +70)	(1.03, 1.79, 1.73)
(-60, +60)	(1.59, 2.15)	(-50, -10, +30)	(1.25, 4.10, 2.33)
(-50, +30)	(2.09, 3.36)	(-50, +30, +60)	(0.82, 1.56, 1.83)
(-20, +40)	(2.39, 2.62)	(-30, +10, +30)	(0.44, 4.35, 2.11)
(-10, +20)	(3.32, 4.34)	(-30, 0, +60)	(1.37, 2.21, 3.68)
(-10, +30)	(2.93, 3.05)	(-10, +10, +30)	(1.39, 3.71, 1.88)

4. EVALUATION

We evaluated our methods for multisource sound localization and speech tracking in a two- or three-speaker situation. The subject of the experiment was the SIG-2 humanoid robot. Figure 2 shows the flow of the implemented robot audition system. The tracked DOAs were used to make the robot turn at its neck and waist to look in the directions of the speakers as an act of HRI. The experimental were conducted in a room with a reverberation time of about 120 ms and noise from air conditioners, personal computers, and background music. The SIG-2 humanoid robot was placed at the center of the room and two male and one female speaker stood at predetermined points along the azimuth in twelve different conditions of speakers' positions. They spoke to the robot randomly and their speech signals were captured continuously. The average sound pressure level (SPL) of background noise was about 67.7 dB and those of target speech signals were about 87.3 dB. The multisource speech localization and tracking method was evaluated for each frame.

Table 1 shows the root mean square error (RMSE) for two- and three-speaker tracking in the twelve different conditions of speakers' positions individually, where each result of RMSEs corresponds to each position of speakers. The RMSEs for each tracked directions were less than 4.35° for both the two- and three-speaker situations. Figure 3 shows two of their graphical results for 6 s. Even though the ML-based DOA estimation using (2)–(4), produced many incorrect direction estimations (shown by (d)), the multisource speech tracking method filtered them out and tracked the direction of each speaker regardless of changes in the number of speakers over time (shown by (e)).

As a result, despite the use of only two microphones, the binaural robot audition system with our multisource speech localization and tacking method showed good overall performance in a real environment even though it sometimes tracked directions incorrectly.

5. CONCLUSION

We presented the ML-based DOA estimation using the GCC-PHAT method for multisource sound localization and described the difficulties with multisource DOA estimation using two microphones in a practical situation. We addressed the difficulties by devising a multisource speech tracking method consisting of VAD and K-means clustering. For effective multisource speech tracking, we improved the standard K-means clustering algorithm by applying two new steps that increase the number of clusters automatically and eliminate clusters including incorrect DOA estimations. Experimental results demonstrated that our method can track directions of multiple speakers over time with tracking error below 4.35°

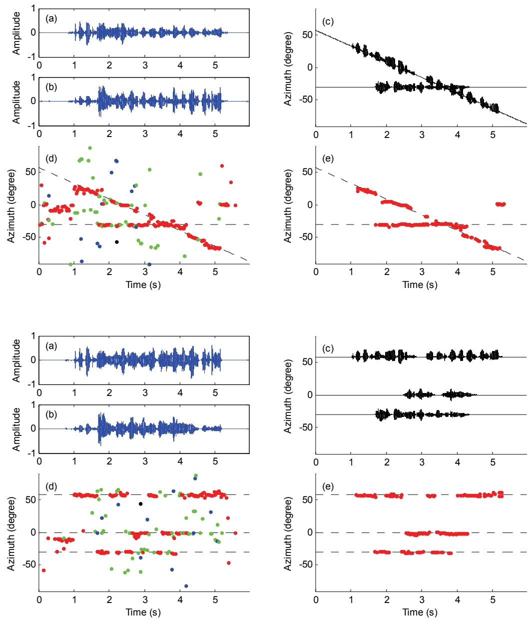

Fig. 3. Results of localization and tracking of multiple speakers. (a)(b) Signal inputs to right and left microphones. (c) Actual directions and speech durations of multiple speakers. (d) Results of multisource localization using (2)–(4). (e) Results of multisource speech tracking using (7)–(11), where the dotted lines show the actual directions.

6. REFERENCES

[1] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.
[2] U. H. Kim, T. Mizumoto, T. Ogata, and H. G. Okuno, "Improvement of Speaker Localization by Considering Multipath Interference of Sound Wave for Binaural Robot Audition," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2910-2915, San Francisco, USA, September 2011.
[3] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*, Cambridge, MA: MIT Press, 1997.
[4] J. Sohn, N. S. Kim, and W. Sung, "A Statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, January 1999.