人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-12

A Two Microphone-Based Approach for Multiple Speaker Localization on the SIG-2 Humanoid Robot

Ui-Hyun Kim and Hiroshi G. Okuno
Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{euihyun, okuno}@kuis.kyoto-u.ac.jp

Abstract—A system based on the generalized cross-correlation (GCC) method weighted by the phase transform (PHAT) has been developed for multiple speaker localization. In real environments with binaural robot audition, speaker localization is degraded by the interference created when the speech waves arrive at a microphone from two directions around the robot head and by impaired performance when there are multiple speakers. This paper presents a new time difference of arrival (TDOA) factor for the GCC-PHAT method to compensate multipath interference on the assumption of spherical robot head and a multisource speech tracking method consisting of voice activity detection and K-means clustering for multiple speaker situations. The standard K-means clustering algorithm was improved for the purpose of multisource speech tracking by adding two additional steps that increase the number of clusters automatically and eliminate clusters containing incorrect DOA estimations. Experiments conducted on the SIG-2 humanoid robot in a real environment show that our method improved the localization accuracy and can track multiple speakers in real-time with tracking error below 5.3°.

I. INTRODUCTION

'Binaural' literally means having two sound inputs. For a robot, it means having two microphones, one on each side of its head (like ears). Recently, many researchers and engineers have conventionally used lots of microphones for robots to improve the hearing performance [1]–[2]. However, using numerous microphones causes some problems: rising maintenance costs for microphones and computational power, and losing a general-purpose software interface due to the different microphone array configuration for each robot. The cost for a binaural audition device is substantially less than that for a multichannel audition device. Binaural audition hardware and its software can be easily ported to various kinds of robot platforms and embedded in information and communication technology (ICT) devices. Moreover, research on binaural audition can contribute to understanding the human hearing mechanism [3]. For these reasons, binaural audition is particularly important for robots.

Among the various functions required for robot audition, sound source localization (SSL) is one of the most important techniques to achieve more natural and intelligent human-robot interaction (HRI). For example, a robot estimates the directions of sound sources to understand the acoustic scene. Then it faces or tracks the person speaking and signals him/her that it is ready to listen and thereby appear to express its interest in the conversation. 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.). These clues are implicitly included in the head related transfer function (HRTF) [4]. 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 other than front and back. One of the most well-known methods to estimate TDOA for SSL with binaural sound inputs is the generalized cross-correlation (GCC) method with phase transform (PHAT) weighting [5].

The use of a microphone array with many microphones improves localization performance in noisy and reverberant environments. On the other hand, localization performance generally drops as the number of microphones is reduced. Since a binaural robot audition system uses only two microphones embedded on each side of the robot's head, there are difficulties in obtaining a performance as good as that when using the microphone array. The localization performance with only two microphones must be improved to enable robots with a binaural audition system to be deployed in various acoustic environments.

In this paper, we addressed two problems affecting the accuracy of the direction-of-arrival (DOA) estimation based on the GCC-PHAT method in binaural robot audition:

1) *Multipath interference due to diffraction of sound waves caused by shape of robot head:* Sound waves easily bend around the robot's head, resulting in a difference in TDOA between the waves that travel around the front of the head and those that travel around the back of the head.

2) *Correlation between multisource sound sources in real environments:* The accuracy in SSL deteriorates when multiple sound sources are correlated, which is generally the case in real environments, i.e., when the sound sources are speech.

Multipath interference severely degrades localization performance especially for sound sources in the lateral direction (around ±90°) and the correlation between sources limits the number of sound sources that the binaural system can localize to a single source. Our solutions to these two problems are twofold:

1) We incorporate a new TDOA factor for multipath interference into the GCC-PHAT method along with

assuming that the robot's head is spherical.

2) We devised a multisource speech tracking method consisting of voice activity detection (VAD) and K-means clustering in order to eliminate incorrect DOA estimations due to the correlation between multiple sound sources.

Our proposed methods were implemented as a real-time system using the 'HARK' open-source robot audition software [6] and evaluated experimentally in the binaural audition system of the SIG-2 humanoid robot.

The paper is outlined as follows: Section II summarizes the ML-based DOA estimation using the GCC-PHAT method for a multisource situation and describes the two problems related to the SSL accuracy in real environments. Section III gives our solutions to the two problems: a new TDOA factor to compensate for multipath interference and a multisource speech tracking method consisting of VAD and K-means clustering algorithms to eliminate incorrect DOA estimations. Section IV presents the experimental results. Section V concludes the paper.

II. Multisource Direction-Of-Arrival Estimation

In this section, we summarize the ML-based DOA estimation using the GCC-PHAT method for multiple sound sources. Then two problems related to the SSL accuracy with binaural robot audition in real environments are explained.

A. Acoustic Model

This paper employs a *F*-point short-time Fourier transform (STFT) under a far-field assumption [7]. The observed signals from the left and right microphones in a situation with *K* sound sources can be mathematically modeled as

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

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 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 $f \in \{1, \ldots, F\}$ denotes a frequency bin, *F* is the time frame size of the STFT, and *fs* is the sampling frequency; $\alpha_{l,r}$ and $\tau_{l,r}$ are the attenuation factor and time delay from the position of the sound source to each microphone, respectively.

B. ML-Based DOA Estimation for Multiple Sound Sources

The ML-based DOA estimation for multiple sound sources is basically defined by the GCC-PHAT method as follows:

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

where

$$G^{PHAT} = \frac{1}{|X_l[f,n]X_r^*[f,n]|}, \quad (3)$$

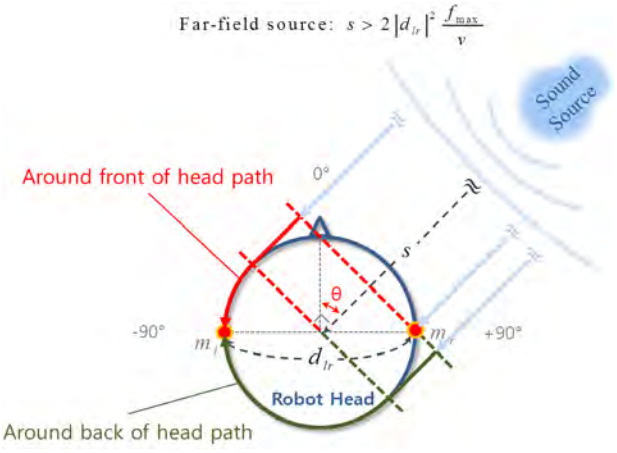

Figure 1. Multipath interference due to diffraction of sound waves with spherical-head assumption.

$$\tau_{lr}(\theta) = \frac{d_{lr}}{v}\sin\left(\frac{\theta}{180}\pi\right). \quad (4)$$

$X_l[f,n]X_r^*[f,n]$, 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 a steering factor for TDOA derived from the free space environment, respectively; $\theta \in \{-90°, \ldots, +90°\}$ is an angle of sound incidence, * is the complex conjugate, d_{lr} is the distance between two microphones, and *v* is the speed of sound (340.5 m/s, at 15 °C, in air). The estimated DOAs θ_{mle_k} of the multiple sound sources in (2)–(4) are obtained by finding several expected angles of sound incidence θ that equally maximizes the sum of the characteristic function obtained from the cross-power spectrum with PHAT weighting in the frequency domain.

C. Problem: Multipath Interference Due to Diffraction of Sound Waves Caused by Shape of Robot Head

Basically, TDOAs are estimated under the assumption that the microphones are located in free space, e.g. in (4). However, this assumption is not applicable to TDOA estimation using two microphones in a robot head because the sound waves easily bend and spread along the shape of the robot head, which creates a difference in TDOA between the waves that travel around the front of the head path and those that travel around the back of the head path. Figure 1 illustrates the two paths created by the diffraction of the sound waves with the assumption that the robot head is spherical. It clearly shows that these two sound-wave paths and multipath interference must be considered if sound source localization in binaural robot audition is to be more accurate.

D. Problem: Correlation between Multiple sound sources in Real Environments

The multisource DOA estimation using (2)–(4) can produce accurate estimates of DOAs in the ideal case that the multiple sound sources S_k are uncorrelated with each other and with additive noise $N_{l,r}$; i.e., $S_1[f,n]S_2[f,n]=0$,

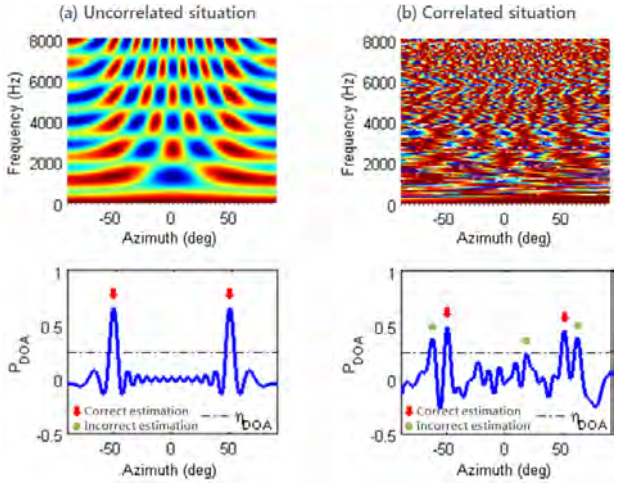

Figure 2. Frequency spectrum and the peak distributions of the ML-based DOA estimation for two sound sources coming from angles of -50° and +50°: (a) A situation with two uncorrelated sources. (b) A situation with two highly correlated sources.

$S_k[f,n]N_{l,r}[f,n]=0$, *and* $N_l[f,n]N_r[f,n]=0$. 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 corrupted by noise and reverberation. For example, if two correlated sound sources are assumed to come from different directions, (1) can be rewritten as:

$$\begin{aligned} X_l[f,n] &= \alpha_{l1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l1}\right) \\ &\quad + \alpha_{l2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l2}\right) + N_l[f,n] \\ X_r[f,n] &= \alpha_{r1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r1}\right) \\ &\quad + \alpha_{r2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r2}\right) + N_r[f,n], \end{aligned} \tag{5}$$

and their cross-power spectrum can be expressed as

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

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 2 shows examples of peak distributions in multisource DOA estimation for two sound signals coming from angles of -50° and +50°. In the uncorrelated situation (a), two sound sources were virtually generated; in the correlated situation (b), two speech signals (one for a male and one for a female) recorded at the same time by the SIG-2 humanoid robot in an experiment room were used. When the sound sources were correlated, the ML-based DOA estimation inaccurately estimated multiple DOAs because of the numerous peaks spread in all directions. In addition, since the intensity of each peak changed over time because of the attenuation factors in (1) applied to each sound source, the ML-based DOA estimation may select peaks in incorrect directions as correct DOAs when the peak intensities in the correct directions are lower than those in the incorrect directions. Furthermore, the ML-based DOA estimation with a threshold η_{DOA} frequently fails to produce the same number of DOAs as sound sources, especially in the absence of information on how many sound sources are active.

These results show that a function is needed to filter out the incorrect DOA estimations in order to get accurate multisource sound localizations.

III. Improved Multisource Direction-Of-Arrival Estimation in Binaural Robot Audition

Our solutions to the two problems in real environments described above are presented here. In binaural DOA estimation, the two problems cause inaccurate and unreliable localizations. We have proposed a new TDOA factor and devised a multisource speech tracking method consisting of voice activity detection (VAD) and K-means clustering. The standard K-means clustering algorithm was extended to enable tracking of an unknown time-varying number of speakers by adding two additional steps that increase the number of clusters automatically and eliminate clusters containing incorrect DOA estimations.

A. New TDOA Factor for Multipath Interference

To solve the problem of multipath interference due to the sound waves traveling along two paths around the robot head, we first apply a simplified formula to these two paths under the assumption that the head is spherical:

$$Path_{front}(\theta) = \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{7}$$

$$Path_{back}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{8}$$

where $Path_{front}$ and $Path_{back}$ are respectively the time delays for the path around the front of the head and that around the back of the head for each sound incidence direction, and *sgn* is a signum function that extracts the sign of θ, i.e., if θ has a negative sign, $sgn(\theta)$ is −1. After the formulas for the two paths are derived, the time delay between them for each sound direction is obtained using

$$Diff_{front-back}(\theta) = Path_{back} - Path_{front} = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right), \tag{9}$$

where $Diff_{front-back}$ is 0 when θ is −90° or +90°. Suppose that the intensity of the multipath interference from $Path_{back}$ for each sound direction complies with that of the ILD ratios

between two microphones located in the robot head and this intensity of the ILD ratios shows the sine function in the ideal condition. We use $Diff_{front\text{-}back}$ multiplied by the absolute sine function with attenuation factor β_{multi} as a factor to compensate for multipath interference:

$$Multi_{front-back}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{multi}\sin\left(\frac{\theta}{180}\pi\right)\right|, \quad (10)$$

where $Multi_{front\text{-}back}$ is the compensation factor for multipath interference in binaural robot audition. The final time delay factor for the binaural DOA estimation can be derived using $Path_{front}$ and $Multi_{front\text{-}back}$:

$$\begin{aligned}\tau_{multi}(\theta) &= Path_{front}(\theta) - Multi_{front-back}(\theta) \\ &= \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right) \\ &\quad - \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{multi}\sin\left(\frac{\theta}{180}\pi\right)\right|.\end{aligned} \quad (11)$$

This new TDOA factor, τ_{multi}, is used instead of τ_{lr} in (4) with the ML-based DOA estimation.

B. Multisource Speech Tracking

Our approach to eliminate incorrect DOAs estimations due to the correlation between multiple sound sources described above is to use data mining in each time frame. For this purpose, we devised a multisource speech tracking module based on two methods:

- *Statistical model-based Voice Activity Detection*

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 to differentiate speech from background noise. We used the statistical model-based VAD algorithm proposed by Sohn et al [8]. This algorithm uses the log likelihood ratio (LLR) between the Gaussian statistical models of speech and background noise for low signal-to-noise (SNR) ratio cases to indicate with high accuracy the presence or absence of speech.

Each time frame is determined to be "speech-present" or "speech-absent" by using a decision procedure based on LLR with a threshold:

$$\begin{aligned}&if\ \hat{P}_{VAD}[n] = \frac{1}{F}\sum_{f=1}^{F}(\gamma[f,n] - \log\gamma[f,n] - 1) > \eta_{VAD} \\ &then\quad n = speech\text{-}present\ frame \\ &else\quad n = speech\text{-}absent\ frame\end{aligned}, \quad (12)$$

$\gamma[f,n]=|X[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* SNR and $\lambda_N[f,n]$ is the estimated variance of $(N_l[f,n]+N_r[f,n])/2$.

- *Improved K-means clustering*

K-means clustering is a commonly used data mining algorithm featuring computational simplicity and high speed. We improved the standard *K*-means clustering algorithm to work well for multisource sound tracking in real situations. If the multiple DOA estimations 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:

$$\Theta_{track_k}^{(i)} = \{\hat{\theta}_{mle_p} : |\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}|^2 \le |\hat{\theta}_{mle_p} - \theta_{track_j}^{(i)}|^2\ \forall 1 \le j \le K\}, \quad (13)$$

where p denotes the index of all estimated DOA in the given time frames, 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_k}.

[Update Step] Calculate the new means to be the centroid of the observations in each cluster:

$$\theta_{track_k}^{(i+1)} = \frac{1}{\left\langle\Theta_{track_k}^{(i)}\right\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\hat{\theta}_{mle_p}, \quad (14)$$

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.

These two problems cause errors in the results of multisource speech tracking. 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:

$$\begin{aligned}&if\ \frac{1}{\left\langle\Theta_{track_k}^{(i)}\right\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\left|\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}\right|^2 > \eta_{C1} \\ &then\ K^{(i+1)} = K^{(i)} + 1\ and\ move\ to\ \text{Assignment Step} \\ &else\qquad\qquad move\ to\ \text{Elimination Step}.\end{aligned} \quad (15)$$

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:

$$\begin{aligned}&if\ \frac{\left\langle\Theta_{track_k}^{(i)}\right\rangle}{\sum_{k=1}^{K}\left\langle\Theta_{track_k}^{(i)}\right\rangle} < \eta_{C2}\quad then\quad \text{eliminate } cluster\ \Theta_{track_k}^{(i)} \\ &else\qquad\qquad \text{keep } cluster\ \Theta_{track_k}^{(i)}\end{aligned}. \quad (16)$$

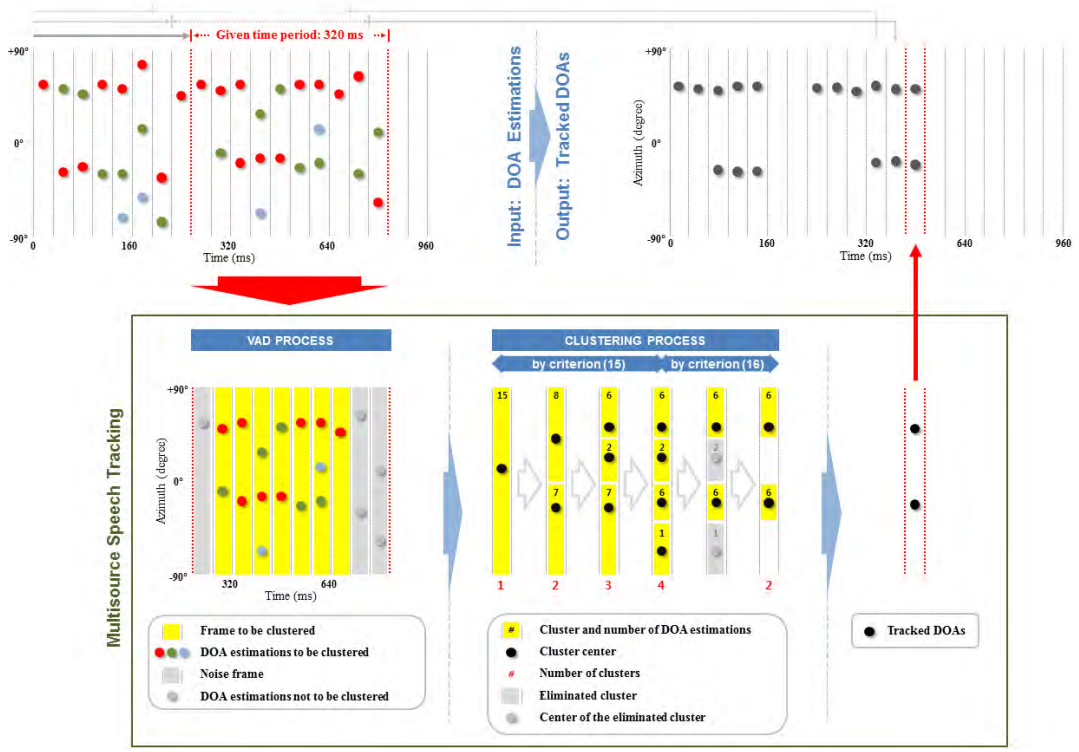

Figure 3. Multisource speech tracking by VAD and *K*-means clustering.

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 (15).

3) All clusters containing incorrect DOA estimations are eliminated on the basis of Criterion (16).

The process of multisource speech tracking with multisource DOA estimations by VAD and K-means clustering is shown in Fig. 3.

IV. Evaluation

We evaluated our ML-based SSL method with the new TDOA factor τ_{multi} to verify that it makes fewer localization errors than with the conventional TDOA factor τ_{lr} in binaural robot audition and tested it with the multisource speech tracking method in a time-varying two or three number of speakers situation. The subject of the experiment was the SIG-2 humanoid robot equipped with two Sennheiser ME 104 omnidirectional microphones and operated by the 'HARK' open-source robot audition software in the real-time. Figure 4 shows the flow of the implemented robot audition system. The tracked DOAs were used to make the robot turn at its neck and waist in order to look in the speaker's directions.

A. Experimental Setup

The experiments were conducted in a room with a reverberation time of about 120 ms and noise from air conditioners and personal computers. To create a noisier environment, background music with lyrics was played as additive noise. The average sound pressure level (SPL) of the background music and the average SNR of the target speech signals were about 70.1 dB and 19.2 dB, respectively. The SIG-2 humanoid robot was placed at the center of the room, and the speakers were located 1.5–2.5 m from the robot. The attenuation factor (β_{multi} in (11)) in the ML-based DOA estimation was set to 0.1 and the values for the thresholds (η_{VAD} in (12), η_{C1} in (15), and η_{C2} in (16)) used in the multisource speech tracking method were set to 50.0, 3.0, and 0.25, respectively. The system recorded the background noise for 2 s before each trial to estimate the noise variance and used the variance as the *a priori* noise variance for the VAD used in the multisource speech tracking method.

To obtain an accurate estimate of the performance improvement with our ML-based DOA estimation with the new TDOA factor in binaural robot audition, we estimated it in a single-speaker situation first. A male and then a female speaker stood at points along the azimuth from −90° to +90° in 10° steps and spoke to the robot five times at each point. Then we evaluated the ML-based DOA estimation with the

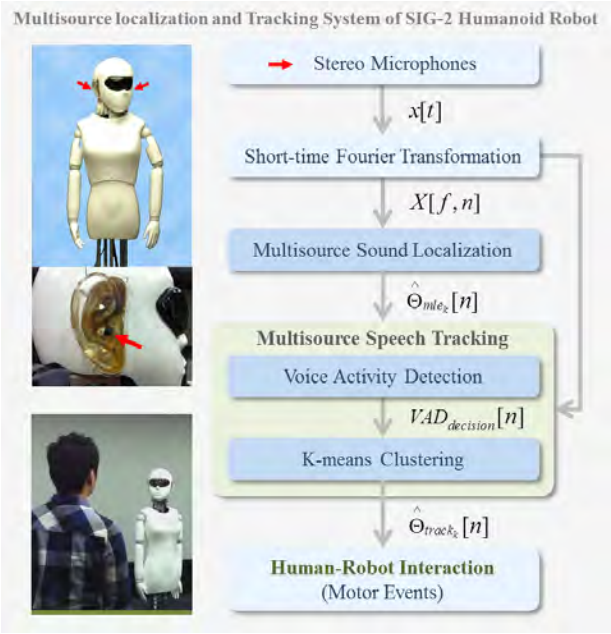

Figure 4. Flowchart of multissource localization and tracking.

multisource speech tracking method in time-varying two- and three-speaker situations for 6 s.

B. *Experimental Results*

Figure 5 shows the root mean square error (RMSE) for the 190 trials (19 points × 5 speech signals × 2 speakers) for the two experimental methods in a single-speaker situation. As shown in the figure, the ML-based SSL methods with the new TDOA factor τ_{multi} had fewer localization errors than with the conventional TDOA factor τ_{lr}. The new TDOA factor τ_{multi} was particularly effective—it reduced the average RMSE by 18.1° and the RMSEs for the side directions by over 37°.

Our tacking method consisting of statistical model-based VAD and improved *K*-means clustering showed good overall performance even though it sometimes failed in tracking with the exact number of directions. Figure 6 shows the experimental results of two- and three-speaker localization and tracking for each 6 s period. Even though the multisource DOA estimation produced many incorrect DOA estimations (shown by (c)), the multisource speech tracking method filtered them out and tracked the direction of each speaker in the running-time domain regardless of changes in the number of speakers over time (shown by (d)). The root mean square error (RMSE) of each tracked DOA for 6 s was less than 5.3° for two- and three-speaker situations.

As a result, despite the use of only two microphones, the robot audition system showed good overall performance for binaural multi-speaker localization in a real environment.

V. Conclusion

We addressed two accuracy problems with the binaural DOA estimation using the GCC-PHAT method in real environments. To solve the problem of multipath interference due to diffraction of the sound waves around the robot head, a

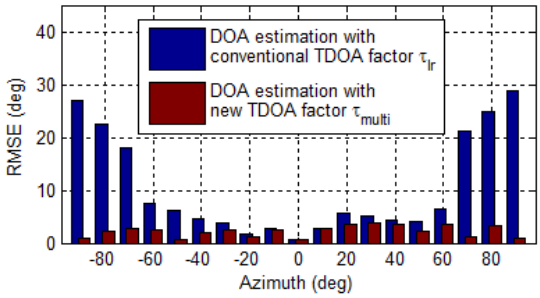

Figure 5. RMSEs for single-speaker localization.

new TDOA factor that takes into account multipath interference is applied to the GCC-PHAT method with the assumption that the robot head is spherical. To overcome the correlation between multiple sound sources, a multisource speech tracking method consisting of statistical model-based VAD and K-means clustering was devised. To make multisource speech tracking more effective, the standard K-means clustering algorithm was improved by adding two additional steps increasing the number of clusters automatically and eliminating clusters containing incorrect direction estimations.

Experimental results demonstrated that taking multipath interference into account when estimating the time delay caused by the diffraction of the sound waves is a key to improving localization performance in binaural robot audition. Doing this with the multisource speech tracking method enabled our real-time binaural robot audition system to correctly track the directions of multiple speakers regardless of the periods during which they spoke and changes in the number of speakers below in tracking error 5.3°.

Future work includes extending our multisource speech tracking method so that it can deal with even more moving speakers. Several problems can occur in a moving-speaker situation, such as incorrect tracking due to ambiguity of speaker identification when moving speakers cross paths or when they are speaking in the same direction. We are planning to implement a blind source separation technique with independent vector analysis [9] in our multisource speech tracking method to handle this problem.

References

[1] U. H. Kim, J. Kim, D. Kim, H. Kim, and B. J. You, "Speaker Localization Using the TDOA-based Feature Matrix for a Humanoid Robot," in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pp. 610-615, Munich, Germany, August 2008.

[2] K. Nakadai, H. Nakajima, G. Ince, and Y. Hasegawa, "Sound Source Separation and Automatic Speech Recognition for Moving Source," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 976-981, Taipei, Taiwan, October 2010.

[3] J. Blauert and J. Braasch, "Binaural Signal Processing," in *Proc. IEEE Int. Conf. on Digital Signal Processing (DSP)*, pp. 1-11, Greece, July 2011.

[4] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Audio Engineering Society*, vol. 49, pp. 231–249, April 2001.

[5] 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.

[6] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System 'HARK' - Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol.24, pp.739-761, 2010.

[7] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*, Cambridge, MA: MIT Press, 1997.

[8] 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.

[9] T. Kim, T. Eltoft, and T. W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components" *International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, LNCS 3889, pp. 165–172, 2006.

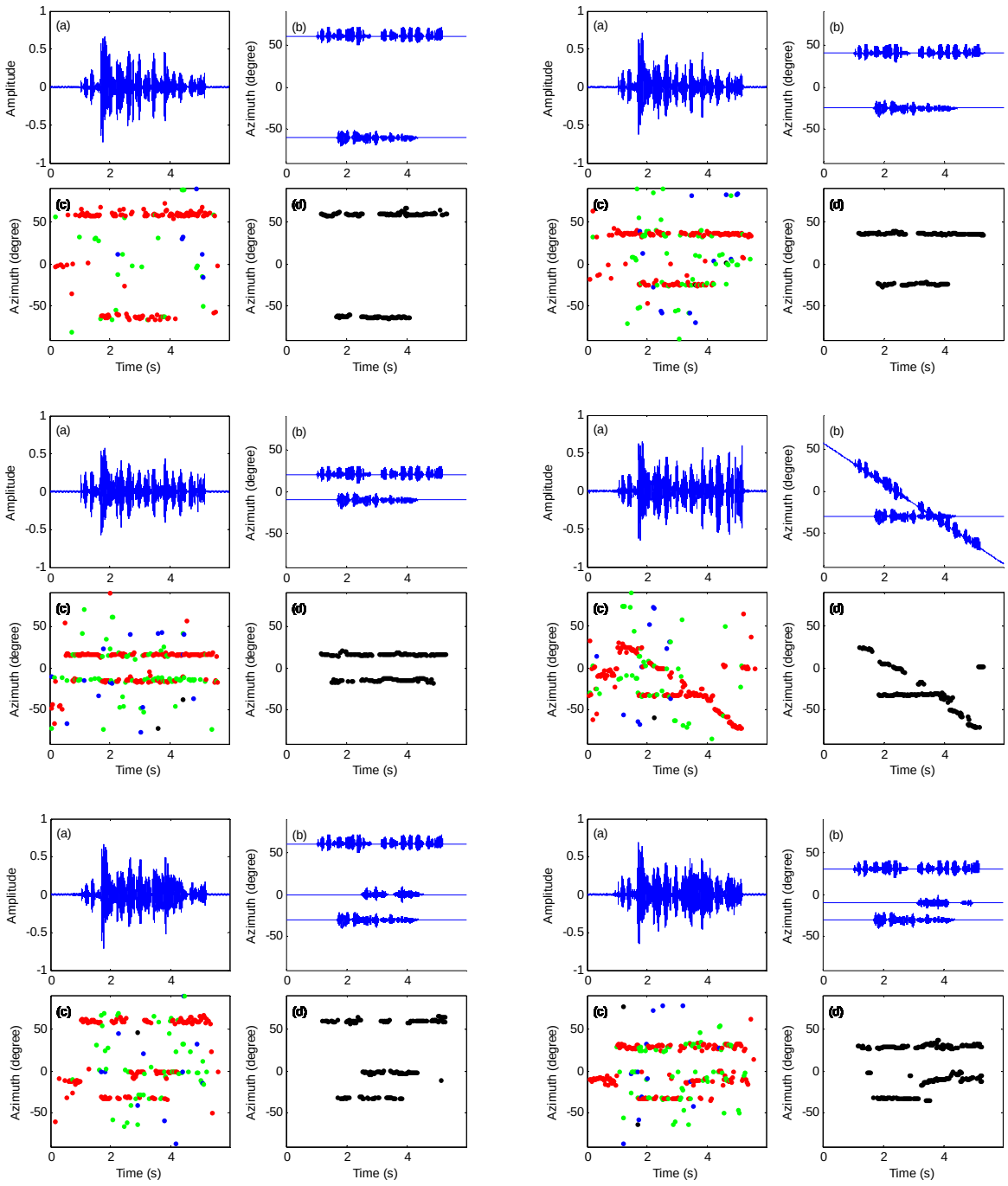

Figure 6. Results of two- and three-speaker tracking with 128-ms time frame, 32-ms time shift, and 320-ms time duration for clustering (10 time frames). (a) Signal input to left microphone consisting of male speech signal and female speech signal. (b) Actual directions and speech durations of two speakers. (c) Results of multisource sound localization, where colors (red, green, and blue) indicate peaks heights in ascending order. (d) Results of multisource speech tracking.