Localization of Simultaneous Moving Sound Sources for Mobile Robot Using a Frequency-Domain Steered Beamformer Approach

Jean-Marc Valin, François Michaud, Brahim Hadjou, Jean Rouat
LABORIUS – Research Laboratory on Mobile Robotics and Intelligent Systems
Department of Electrical Engineering and Computer Engineering
Université de Sherbrooke, Sherbrooke (Quebec) CANADA, J1K 2R1
{Jean-Marc.Valin, Francois.Michaud, Brahim.Hadjou, Jean.Rouat}@USherbrooke.ca

***Abstract*— Mobile robots in real-life settings would benefit from being able to localize sound sources. Such a capability can nicely complement vision to help localize a person or an interesting event in the environment, and also to provide enhanced processing for other capabilities such as speech recognition. In this paper we present a robust sound source localization method in three-dimensional space using an array of 8 microphones. The method is based on a frequency-domain implementation of a steered beamformer along with a probabilistic post-processor. Results show that a mobile robot can localize in real time multiple moving sources of different types over a range of 5 meters with a delay of 200 ms.**

I. Introduction

The sense of hearing is quite important in providing information in a real life environment: it can draw attention to particular and discriminate events in the world that can be further analyzed using other senses such as vision, or it allows to exchange information through language. For those who do not have hearing impairments, it is hard to imagine going a day without being able to hear, especially given the fact that we are moving in many different environments (indoor and outdoor).

Signal processing research that address artificial audition is often geared toward specific tasks such as speaker tracking for videoconferencing. However, artificial hearing for mobile robots is still in its infancy. The SAIL robot uses one microphone to develop online audio-driven behaviors [1]. The robot ROBITA uses two microphones to follow a conversation between two people [2]. SIG, a humanoid robot uses two pairs of microphones; one pair is installed on both sides of the head, while the other pair is placed inside the head to record internal sounds (such as motor noise) for noise cancellation [3], [4]. Like humans, these last two robots use binaural localization, i.e. the ability to locate the source of sound in three dimensional space.

It is difficult to localize sounds with only two input sources. The human auditory system accounts for the acoustic shadow of the head and the ridges of the outer ear. Without this ability, only localization in two dimensions is possible without the possibility to distinguish if the sounds come from the front or the back. Also, it may be difficult to obtain high-precision readings when the sound source is in the same axis of the pair of microphones.

Robots are not inherently limited to two microphones; we decided to use more microphones to better approach the localization abilities of the human auditory system. This way, increased resolution can be obtained in three-dimensional space. This also means increased robustness, since multiple signals greatly helps reduce the effects of noise (instead of trying to isolate the noise source by putting sensors inside the robot's head, as with SIG) and discriminate multiple sound sources. There are already robots available with more than two microphones; the Sony SRD-4X has seven.

An artificial audition system can be used for three things: 1) localizing sound sources, 2) separating sound sources in order to process only signals that are relevant to a particular event in the environment, and 3) processing sound sources to extract useful information from the environment (like speech recognition for instance). This paper focuses on sound source localization. In previous work [5], we presented a method based on time delay of arrival (TDOA) estimation. The method works for far-field and near-field sound sources and was validated using a Pioneer 2 mobile robotic platform.

In this paper, we present an approch with the same objective, but is based on a frequency-domain beamformer that is steered in all possible directions to detect sources. Instead of measuring TDOAs and then converting to a position, the search is performed in a single step. This makes the system more robust, especially in the case where an obstacle prevents one or more microphones from properly receiving the signals. The results are then enhanced by probability-based post-processing which prevents false detection of sources. This makes the system sensitive enough for simultaneous localization of multiple moving sound sources.

The paper is organized as follows. Section II presents a brief overview of the system and Section III describes our frequency-domain implementation of a steered beamformer. Section IV explains how we enhance the results from the beamformer using a probabilistic post-processor, followed by experimental results in Section V.

II. System overview

The proposed localization system as shown in Figure 1 is composed of three parts:

- A microphone array;
- A memoryless localization algorithm based on a steered beamformer;
- A probability-based post-processor.

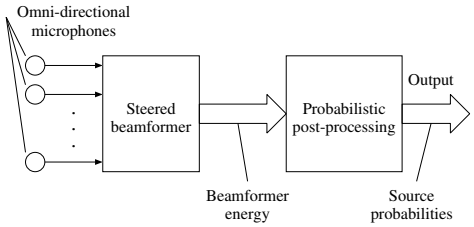

Fig. 1. Overview of the system

The microphone array is composed of a number of omni-directional elements mounted on the robot. The signals are used by a beamformer that is steered in all possible directions in order to maximize the output power. The initial localization performed by the steered beamformer is then used as the input of a post-processing stage that uses Bayesian probability rules to compute the probability of source presence for every directions.

The output of the localization can be used to direct the robot attention to the source. It can also be used as part of a source separation algorithm to isolate the sound coming from a single source [6].

III. Localization by steered beamformer

The basic idea behind the steered beamformer approach to source localization is to direct a beamformer in all possible directions and look for maximal output. For this task, we try to maximize the output power of a simple delay-and-sum beamformer.

A. Delay-and-sum beamformer

The output of an M-microphone delay-and-sum beamformer is defined as:

$$y(n) = \sum_{m=0}^{M-1} x_m\left(n - \tau_m\right) \tag{1}$$

where $x_m\left(n\right)$ is the signal from the m^{th} microphone and τ_m is the delay of arrival for that microphone. The output energy of the beamformer over a frame of length L is thus given by:

$$\begin{aligned} E &= \sum_{n=0}^{L-1} \left[y(n)\right]^2 \\ &= \sum_{n=0}^{L-1} \left[x_0\left(n-\tau_0\right) + \ldots + x_{M-1}\left(n-\tau_{M-1}\right)\right]^2 \end{aligned} \tag{2}$$

Assuming that one sound source is present, we can see that E will be maximal when the delays τ_m are such that the microphone signals are in phase (and therefore add constructively).

There is, however, a problem with that technique in that energy peaks are very wide [7], which means that the resolution is poor. Moreover, in the case of multiple sources, it makes it more likely to have sources responses overlap.

One way to narrow the peaks is to whiten the microphone signals prior to computing the energy [8]. Unfortunately, the coarse-fine search methods as proposed in [7] cannot be used because the narrow peaks can be missed during the coarse search. Therefore, a fine search is necessary, which requires increased computing power. It is however possible to reduce the amount of computation by calculating the beamformer energy in the frequency domain. This also has the advantage of making the whitening of the signal easier.

We first notice that the beamformer output energy in Equation 2 can be expanded as:

$$\begin{aligned} E &= \sum_{m=0}^{M-1}\sum_{n=0}^{L-1} x_m^2\left(n-\tau_m\right) \\ &+ 2\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{m_1-1}\sum_{n=0}^{L-1} x_{m_1}\left(n-\tau_{m_1}\right) x_{m_2}\left(n-\tau_{m_2}\right) \end{aligned} \tag{3}$$

which in turn can be rewritten in terms of cross-correlations:

$$E = K + 2\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{m_1-1} R_{x_{m_1},x_{m_2}}\left(\tau_{m_1} - \tau_{m_2}\right) \tag{4}$$

where $K = \sum_{m=0}^{M-1}\sum_{n=0}^{L-1} x_m^2\left(n-\tau_m\right)$ is nearly constant with respect to the τ_m delays and can thus be ignored when maximizing E. The cross-correlation function can be approximated in the frequency domain as:

$$R_{ij}(\tau) \approx \sum_{k=0}^{L-1} X_i(k) X_j(k)^* e^{\jmath 2\pi k\tau/L} \tag{5}$$

where $X_i(k)$ is the discrete Fourier transform of $x_i[n]$, $X_i(k)X_j(k)^*$ is the cross-spectrum of $x_i[n]$ and $x_j[n]$ and $(\cdot)^*$ denotes the complex conjugate. The power spectra and cross-power spectra are computed on overlapping windows (50% overlap) of $L = 1024$ samples at 48 kHz. Once the $R_{ij}(\tau)$ are precomputed, it is possible to compute E using only $N(N-1)/2$ lookup and accumulation operations.

Because of the reduced complexity, it is possible to use two different source detectors; a short-term one for percussive noise like dropped objects or handclaps and a medium-term one for speech and other continuous sounds.

For each estimator, the $R_{ij}(\tau)$ are computed by averaging the cross-power spectra $X_i(k)X_j(k)^*$ over two different time periods. In our implementation, the short- and medium-term estimators average the cross-power spectra over 4 frames (40 ms) and 20 frames (200 ms), respectively.

B. Spectral weighting

As stated in the previous subsection, we chose to whiten the signal prior to computing the beamformer energy. In the frequency domain, the whitened cross-correlation is thus computed as:

$$R_{ij}^{(w)}(\tau) \approx \sum_{k=0}^{L-1} \frac{X_i(k)X_j(k)^*}{|X_i(k)|\,|X_j(k)|} e^{j2\pi k\tau/L} \tag{6}$$

While it produces much sharper cross-correlation peaks, the whitened cross-correlation has a drawback. Each frequency bin of the spectrum contributes the same amount to the final correlation, even if the signal at that frequency is dominated by noise. This makes the system less robust to noise, while making detection of voice (which has a narrow bandwidth) more difficult.

In order to resolve the problem, we developed a weighting function for the spectrum. This function gives more weight to regions in the spectrum where the local signal-to-noise ratio (SNR) is the highest. Let $Y(k)$ be the mean power spectral density for all the microphones at a given time and $Y_N(k)$ be a noise estimate based on the time average of previous $Y(k)$. We define a noise masking weight by:

$$w(k) = \begin{cases} 1 & , Y(k) \leq Y_N(k) \\ \left(\frac{Y(k)}{Y_N(k)}\right)^{\gamma} & , Y(k) > Y_N(k) \end{cases} \tag{7}$$

where the exponent $0 < \gamma < 1$ gives more weight to regions where the signal is much higher than the noise. For our system, we empirically set γ to 0.1. The resulting enhanced cross-correlation is defined as:

$$R_{ij}^{(e)}(\tau) = \sum_{k=0}^{L-1} \frac{w^2(k)X_i(k)X_j(k)^*}{|X_i(k)|\,|X_j(k)|} e^{j2\pi k\tau/L} \tag{8}$$

C. Direction search on spherical grid

In order to reduce the computation required and to make the system isotropic, we define a uniform triangular grid for the surface of a sphere. In order to create the grid, we start from an initial icosahedral grid [9]. Each triangle in the initial 20-element grid is then recursively subdivided into four smaller triangles as shown in Figure 2. The resulting grid is composed of 5120 triangles and 2562 points. The beamformer energy is then computed for the hexagonal region associated with each of these points. Each of the 2562 regions covers a radius of about $2.5°$ around its center, which means that it introduces at most an error of $2.5°$.

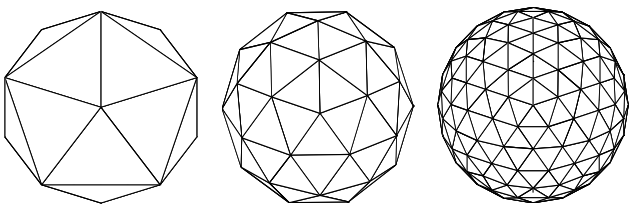

Fig. 2. Recursive subdivision (2 levels) of a triangular element

Once the cross-correlations $R_{ij}^{(e)}(\tau)$ are computed, the search for the best direction on the grid is performed as described by Algorithm 1.

Algorithm 1 Steered beamformer direction search

```
for all grid index d do
    E_d ← 0
    for all microphone pair ij do
        τ ← lookup(d, ij)
        E_d ← E_d + R_ij^(e)(τ)
    end for
end for
direction of source ← argmax_d (E_d)
```

In Algorithm 1, *lookup* is a precomputed table of the time delay of arrival (TDOA) for each microphone pair and each direction on the sphere. By making the far-field assumption [5], the TDOA in samples is computed as:

$$\tau_{ij} = \frac{F_s}{c}\left(\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\right) \cdot \vec{\mathbf{u}} \tag{9}$$

where $\vec{\mathbf{x}}_i$ is the position of microphone i, $\vec{\mathbf{u}}$ is a unit-vector that points in the direction of the source, c is the speed of sound and F_s is the sampling rate.

For an array of M microphones and a K-element grid, the algorithm requires $M(M-1)K$ table memory accesses and $M(M-1)K/2$ additions. In the proposed configuration ($K = 2562$, $M = 8$), the accesses data can be made to fit in a modern processor's L2 cache.

The algorithm described above is able to find the loudest source present by maximizing the energy of a steered beamformer. In order to localize other sources that may be present, we remove the contribution of the first source to the cross-correlations. The process is then repeated, which leads to Algorithm 2.

Algorithm 2 Localization of multiple sources

```
for k = 1 to desired number of sources do
    D_k ← Steered beamformer direction search
    for all microphone pair ij do
        τ ← lookup(D_k, ij)
        R_ij^(e)(τ) = 0
    end for
end for
```

The number of desired sources is a constant for each estimator. We consider that the short-term estimator is able to locate at most two sources at the same time, while the medium-term estimator is able to detect four sources. When less sources are present this leads to false detection of a source. That problem is handled by the probabilistic post-processing described next.

IV. Probabilistic post-processing

In order to prevent false detection of sources and keep the system sensitive enough to weak sources, we introduce

a post-processing step that provides some smoothing in time, while combining the results of the short- and medium-term estimators. Using the same quantized sphere as in the previous section, we associate a probability of source presence to each region of the grid (we omit the grid region index for clarity). We note H_1^n the hypothesis of source presence at discrete time n and H_0^n the hypothesis of no source being present at that time. Also, the steered beamformer observation for time n is denoted o_n, with $\mathbf{O}_n = (o_1, o_2, \ldots, o_n)$ the set of all observations up to time n.

We first introduce an instantaneous probability estimation that uses the results of the steered beamformer of Section III. The idea is that the higher the output energy of the beamformer, the more likely that a source is present. We thus approximate the instantaneous probability of a source being present as:

$$P\left(H_1^n \left| o_n \right.\right) = \max\left[1 - \exp\left(1 - \frac{E}{E_{min}}\right), p_{min}\right] \quad (10)$$

where E is the energy at the output of the beamformer, E_{min} is an energy threshold corresponding to the value when no source is present, and p_{min} is the minimal probability we want to assign for a source that is detected by the steered beamformer (with $p_{min} = 0.1$). In the case where there is no source detected by the beamformer at a certain point, we assign a floor probability $p_{floor} = 0.005$ that accounts for the possibility that the beamformer doesn't detect anything even though a sound source is present.

A. Temporal integration

At time N, we use Bayes' rule to express the probability of source presence given all observations as:

$$P\left(H_1^n \left| \mathbf{O}_n \right.\right) = \frac{P\left(\mathbf{O}_n \left| H_1^n \right.\right) P\left(H_1^n\right)}{P\left(\mathbf{O}_n\right)} \quad (11)$$

Because the energy of the steered beamformer is computed on non-overlapping segments, we assume conditional independence of the observations with respect to the presence or absence of a source. We can thus rewrite Equation 11 as:

$$\begin{aligned} P\left(H_1^N \left| \mathbf{O}_n \right.\right) &= \frac{P\left(\mathbf{O}_{n-1} \left| H_1^n \right.\right) P\left(o_n \left| H_1^n \right.\right) p_1}{P\left(\mathbf{O}_n\right)} \\ &= \frac{P\left(H_1^n \left| \mathbf{O}_{n-1} \right.\right) P\left(H_1^n \left| o_n \right.\right)}{p_1} \\ &\cdot \frac{P\left(\mathbf{O}_{n-1}\right) P\left(o_n\right)}{P\left(\mathbf{O}_n\right)} \end{aligned} \quad (12)$$

where $p_1 = P\left(H_1^n\right) = P\left(H_1\right)$ is the constant *a priori* probability of source presence. Similarly, it follows that the complementary probability is given by:

$$\begin{aligned} P\left(H_0^n \left| \mathbf{O}_N \right.\right) &= \frac{\left[1 - P\left(H_1^n \left| \mathbf{O}_{n-1} \right.\right)\right]\left[1 - P\left(H_1^n \left| o_n \right.\right)\right]}{(1 - p_1)} \\ &\cdot \frac{P\left(\mathbf{O}_{n-1}\right) P\left(o_n\right)}{P\left(\mathbf{O}_n\right)} \end{aligned} \quad (13)$$

We assume that the transitions between H_0 and H_1 can be modeled as a first order Markov process with transition probabilities $\alpha_{ij} = P\left(H_j^n \middle| H_i^{n-1}\right)$. This leads to:

$$\begin{aligned} P\left(H_1^n \left| \mathbf{O}_{n-1} \right.\right) &= \alpha_{01}\left[1 - P\left(H_1^{n-1} \left| \mathbf{O}_{n-1} \right.\right)\right] \\ &+ \alpha_{11} P\left(H_1^{n-1} \left| \mathbf{O}_{n-1} \right.\right) \end{aligned} \quad (14)$$

For this work, we use $\alpha_{01} = 0.00004$, $\alpha_{11} = 0.992$ for the short-term estimator and $\alpha_{01} = 0.0002$, $\alpha_{11} = 0.96$ for the medium-term estimator. The reason for the differences in values is that the medium-term estimator is updated less often.

In order to avoid computing $P\left(\mathbf{O}_n\right)$, $P\left(\mathbf{O}_{n-1}\right)$ and $P\left(o_n\right)$ terms that do not depend on H_0 or H_1, we introduce the unnormalized probabilities $\pi\left(H_1^N \left| \mathbf{O}_n \right.\right)$ and $\pi\left(H_0^N \left| \mathbf{O}_n \right.\right)$ that omit these terms. For example, from Equation 12, we have the unnormalized probability:

$$\pi\left(H_1^N \left| \mathbf{O}_n \right.\right) = \frac{P\left(H_1^n \left| \mathbf{O}_{n-1} \right.\right) P\left(H_1^n \left| o_n \right.\right)}{p_1} \quad (15)$$

From there, it is easy to compute $P\left(H_1^N \left| \mathbf{O}_n \right.\right)$ as:

$$\begin{aligned} P\left(H_1^N \left| \mathbf{O}_N \right.\right) &= \frac{\pi\left(H_1^N \left| \mathbf{O}_N \right.\right)}{\pi\left(H_1^N \left| \mathbf{O}_N \right.\right) + \pi\left(H_0^N \left| \mathbf{O}_N \right.\right)} \\ &= \frac{1}{1 + \frac{\pi\left(H_0^N \left| \mathbf{O}_N \right.\right)}{\pi\left(H_1^N \left| \mathbf{O}_N \right.\right)}} \end{aligned} \quad (16)$$

B. Combination of estimator probabilities

After using the temporal integration method to derive the short-term and medium-term estimators, the last step consists of combining these probabilities to infer a unique probability of source presence. Let $\mathbf{O}^s, \mathbf{O}^m$ respectively be all the observations made by the short- and medium-term estimators up to a certain time, we can first write using Bayes' rule:

$$P\left(H_1 \left| \mathbf{O}^s, \mathbf{O}^m \right.\right) = \frac{P\left(\mathbf{O}^s, \mathbf{O}^m \left| H_1 \right.\right) P\left(H_1\right)}{P\left(\mathbf{O}^s, \mathbf{O}^m\right)} \quad (17)$$

Unfortunately, we cannot assume that $\mathbf{O}^s$ and $\mathbf{O}^m$ are conditionally independent. To represent that, we approximate the combined probability as a weighted geometric average of two hypotheses: 1) complete independence of $\mathbf{O}^s$ and $\mathbf{O}^m$ and 2) equivalence of $\mathbf{O}^s$ and $\mathbf{O}^m$.

If we consider the hypothesis of complete conditional independence of the different estimators, we have:

$$\begin{aligned} P_i\left(H_1 \left| \mathbf{O}^s, \mathbf{O}^m \right.\right) &= \frac{P\left(\mathbf{O}^s \left| H_1 \right.\right) P\left(\mathbf{O}^m \left| H_1 \right.\right) p_1}{P\left(\mathbf{O}^s, \mathbf{O}^m\right)} \\ &= \frac{P\left(H_1 \left| \mathbf{O}^s \right.\right) P\left(H_1 \left| \mathbf{O}^m \right.\right)}{p_1} \\ &\cdot \frac{P\left(\mathbf{O}^s\right) P\left(\mathbf{O}^m\right)}{P\left(\mathbf{O}^s, \mathbf{O}^m\right)} \end{aligned} \quad (18)$$

The complementary probability $P_i\left(H_0 \left| \mathbf{O}^s, \mathbf{O}^m \right.\right)$ can be estimated similarly.

In addition to the complete conditional independence hypothesis, we consider the case where $\mathbf{O}^s, \mathbf{O}^m$ bring exactly the same information about source presence or absence. In

that case, all probabilities should be equal, so we rewrite the probability as:

$$P_d\left(H_1|\mathbf{O}^s,\mathbf{O}^m\right) = \sqrt{P\left(H_1|\mathbf{O}^s\right)P\left(H_1|\mathbf{O}^m\right)} \quad (19)$$

The reality lies in between the situation described by Equations 18 and 19. We express the combined probability estimation as:

$$\begin{aligned} P\left(H_1|\,\mathbf{O}^s,\mathbf{O}^m\right) &\approx \left[P_d\left(H_1|\,\mathbf{O}^s,\mathbf{O}^m\right)\right]^{\beta} \\ &\cdot \left[P_i\left(H_1|\,\mathbf{O}^s,\mathbf{O}^m\right)\right]^{1-\beta} \end{aligned} \quad (20)$$

where $0 \leq \beta \leq 1$ expresses the degree of dependence between the observations ($\beta = 0$ is complete independence), P_i is the probability assuming complete independence and P_d is the probability assuming equivalence of $\mathbf{O}^s$ and $\mathbf{O}^m$. For this paper, we use $\beta = 0.7$.

Using the same unnormalized probabilities defined in the Section IV-A, we have:

$$P\left(H_1|\,\mathbf{O}^s,\mathbf{O}^m\right) \approx \frac{1}{1+\frac{\pi(H_0|\mathbf{O}^s,\mathbf{O}^m)}{\pi(H_1|\mathbf{O}^s,\mathbf{O}^m)}} \quad (21)$$

where:

$$\pi(H_1|\mathbf{O}^s,\mathbf{O}^m) = \frac{\left[P(H_1|\mathbf{O}^s)P(H_1|\mathbf{O}^m)\right]^{1-\frac{\beta}{2}}}{p_1^{1-\beta}} \quad (22)$$

$$\pi(H_0|\mathbf{O}^s,\mathbf{O}^m) = \frac{\left[(1-P(H_1|\mathbf{O}^s))(1-P(H_1|\mathbf{O}^m))\right]^{1-\frac{\beta}{2}}}{(1-p_1)^{1-\beta}} \quad (23)$$

Our choice of the geometric mean is based on the fact that the probabilities can have a very wide dynamic range that is not suitable for the arithmetic mean.

V. Results

The array used for experimentation is composed of eight microphones arranged on the summits of a rectangular prism. The array is mounted on an ActivMedia Pioneer 2 robot, as shown in Figure 3. However, due to processor and space limitations (the acquisition is performed using an 8-channel PCI soundcard that cannot be installed on the robot), the signal acquisition and processing is performed on a desktop computer (Athlon XP 2000+). The algorithm currently requires 30% CPU to work in real-time, but this amount could be reduced by lowering the grid resolution or by using approximations in computing the source probabilities. It is worth mentioning that the CPU time does not increase with the number of sources.

For all results presented in this paper, we used real multi-channel recordings in a noisy environment with moderate reverberation. The system is tested under different conditions. First, we measure the maximum distance at which the system is able to detect different sound sources. During the test, the sound source is produced 50 times with the robot placed in different positions. The source detection rates are shown in Table I. We note that the system is able to reliably detect sources at distances up to 5 meters. Also, while the system is able to detect bursts of white noise reliably at great distance, it is mostly unable to detect pure tones. This behavior is explained by the fact that sinusoids occupy only a very small region of the spectrum and thus have a very small contribution to the cross-correlations, even with the proposed weighting. It must be noted that tones tend to be difficult to localize even for the human auditory system.

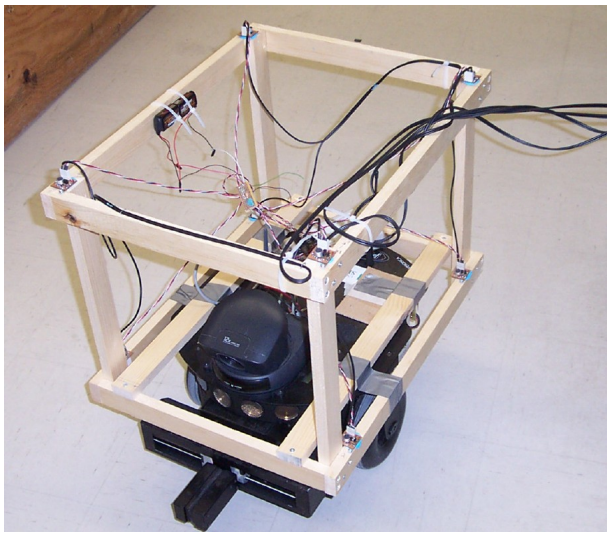

Fig. 3. Pioneer 2 robot with an array of eight microphones

TABLE I
Detection rate as a function of distance for different sounds

Sound source	3 m	5 m	7 m
Hands clapping	92%	94%	84%
Speech (“test”)	100%	90%	42%
Noise burst (250 ms)	100%	100%	100%

The second task for which the system is evaluated is speaker tracking. In this experiment, several people talk to the robot simultaneously and in two of the three cases presented, the speakers are moving while they talk. In Figure 4, we plot the regions where the probability of source presence is at least 0.6. Only azimuth is shown, since the sources are all located in the same elevation range. From the Figure, it can be observed that the system has no difficulty tracking up to 4 moving speakers. With 7 speakers, the system becomes unable to detect all speakers simultaneously, but nonetheless succeeds in localizing them all at over a period of time.

A third test is performed with two stationary speakers, and a moving robot. Figure 5 shows how the robot localizes the speakers as it moves. This demonstrates that the system is able to function despite the noise caused by its motors. The two sources that are sometimes detected at 0° and 90° elevation are respectively a computer fan located at 1.5 meter and a ceiling ventilation trap.

A last experiment was conducted in which we verified that the system still works when the microphone array is not completely open. Even when some sides of the array are filled and some microphones no longer have a line of sight with the source, the system’s reliability is not significantly affected.

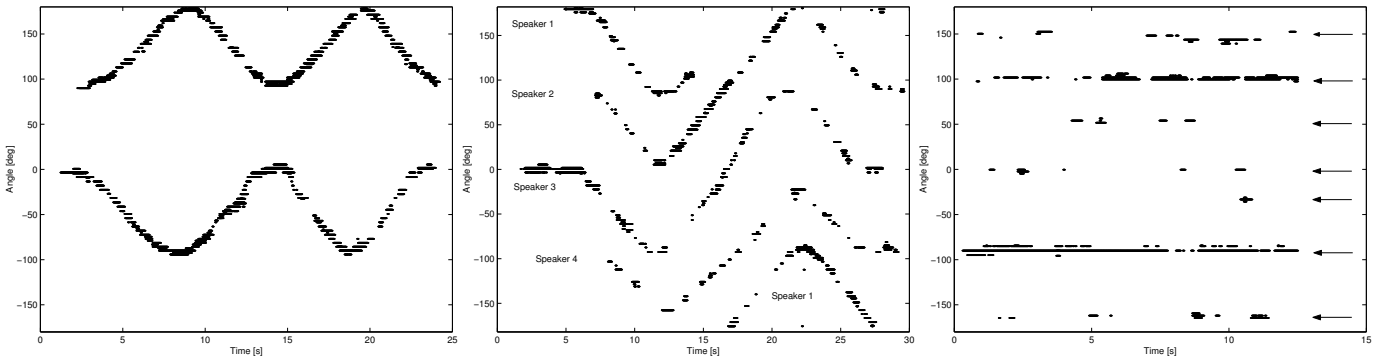

Fig. 4. Tracking of speech at a distance of 1-2 meters. a) 2 moving speakers b) 4 moving speakers c) 7 stationary speakers (positions denoted by arrows).

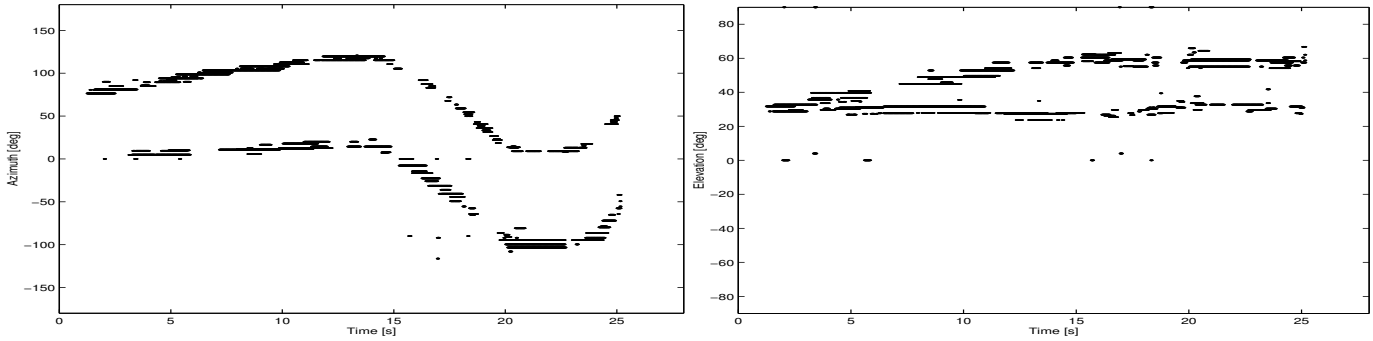

Fig. 5. Two stationary speakers with robot moving and rotating a) Azimuth of sources b) Elevation of sources

VI. Conclusion

Using an array of 8 microphones, we have implemented a system that is able to reliably localize sounds up to five meters away, even in the presence of noise. It is also possible to detect and track simultaneous and moving sound sources. Moreover, our system is adapted to both short-duration sounds like handclaps and longer duration sounds like speech.

In the proposed system, localization is performed in two steps. The first step consists of a beamformer that is steered in all possible directions, trying to maximize output power. The second step uses Bayesian probability combinations to enhance the results produced by the steered beamformer and output probabilities of source presence.

In its current form, the localization system is very sensitive and is sometimes able to detect weak sounds like computer fans located within 2-3 meters. While this may in some cases be desirable, it may be desirable in the future to design an algorithm capable of ranking sound sources in terms of potential interest to the robot.

Acknowledgment

François Michaud holds the Canada Research Chair (CRC) in Mobile Robotics and Autonomous Intelligent Systems. This research is supported financially by the CRC Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Foundation for Innovation (CFI). Special thanks to Dominic Létourneau, Serge Caron, Nicolas Bégin, Mathieu Lemay, Pierre Lepage and Nathan Sharfi for their help in this work.

References

[1] Y. Zhang and J. Weng, "Grounded auditory development by a developmental robot," in *Proceedings INNS/IEEE International Joint Conference on Neural Networks*, 2001, pp. 1059–1064.

[2] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface - a robot who communicate with multi-user," in *Proceedings EUROSPEECH*, 1999, pp. 1723–1726.

[3] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *Proceedings IEEE International Conference on Spoken Language Processing*, 2002, pp. 193–196.

[4] H. G. Okuno, K. Nakadai, and H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *Proceedings of Eighteenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2002, pp. 725–735.

[5] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings International Conference on Intelligent Robots and Systems*, 2003.

[6] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Submitted to ICASSP 2004*.

[7] R. Duraiswami, D. Zotkin, and L. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[8] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. II–273–II–276.

[9] F. Giraldo, "Lagrange-galerkin methods on spherical geodesic grids," *Journal of Computational Physics*, vol. 136, pp. 197–213, 1997.