EURASIP Journal on Applied Signal Processing 2003:4, 338–347

The Fusion of Distributed Microphone Arrays for Sound Localization

Parham Aarabi

Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada M5S 3G4
Email: parham@ecf.utoronto.ca

Received 1 November 2001 and in revised form 2 October 2002

This paper presents a general method for the integration of distributed microphone arrays for localization of a sound source. The recently proposed sound localization technique, known as SRP-PHAT, is shown to be a special case of the more general microphone array integration mechanism presented here. The proposed technique utilizes spatial likelihood functions (SLFs) produced by each microphone array and integrates them using a weighted addition of the individual SLFs. This integration strategy accounts for the different levels of access that a microphone array has to different spatial positions, resulting in an intelligent integration strategy that weighs the results of reliable microphone arrays more significantly. Experimental results using 10 2-element microphone arrays show a reduction in the sound localization error from 0.9 m to 0.08 m at a signal-to-noise ratio of 0 dB. The proposed technique also has the advantage of being applicable to multimodal sensor networks.

Keywords and phrases: microphone arrays, sound localization, sensor integration, information fusion, sensor fusion.

1. INTRODUCTION

The localization of sound sources using microphone arrays has been extensively explored in the past [1, 2, 3, 4, 5, 6, 7]. Its applications include, among others, intelligent environments and automatic teleconferencing [8, 9, 10, 11]. In all of these applications, a single microphone array of various sizes and geometries has been used to localize the sound sources using a variety of techniques.

In certain environments, however, multiple microphone arrays may be operating [9, 11, 12, 13]. Integrating the results of these arrays might result in a more robust sound localization system than that obtained by a single array. Furthermore, in large environments such as airports, multiple arrays are required to cover the entire space of interest. In these situations, there will be regions in which multiple arrays overlap in the localization of the sound sources. In these regions, integrating the results of the multiple arrays may yield a more accurate localization than that obtained by the individual arrays.

Another matter that needs to be taken into consideration for large environments is the level of access of each array to different spatial positions. It is clear that as a speaker moves farther away from a microphone array, the array will be less effective in the localization of the speaker due to the attenuation of the sound waves [14]. The manner in which the localization errors increase depends on the background signal-to-noise ratio (SNR) of the environment and the array geometry. Hence, given the same background SNR and geometry for two different arrays, the array closer to the speaker will, on an average, yield more accurate location estimates than the array that is farther away. Consequently, a symmetrical combination of the results of the two arrays may not yield the lowest error since more significance should be placed on the results of the array closer to the speaker. Two questions arise at this point. First, how do we estimate or even define the different levels of access that a microphone array may have to different spatial positions? Second, if we do have a quantitative level-of-access definition, how do we integrate the results of multiple arrays while at the same time accounting for the different levels of access.

In order to accommodate variations in the spatial observability of each sensor, this paper proposes the spatial observability function (SOF), which gives a quantitative indication of how well a microphone array (or a sensor in general) perceives events at different spatial position. Also, each microphone array will have a spatial likelihood function (SLF), which will report the likelihood of a sound source at each spatial position based on the readings of the current microphone array [8, 13, 15]. It is then shown, using simulations and experimental results, that the SOFs and SLFs for different microphone arrays can be combined to result in a robust sound localization system utilizing multiple microphone arrays. The proposed microphone array integration strategy is shown to be equivalent, in the case that all arrays have equal access, to the array integration strategies previously proposed [7, 12].

2. BASIC SOUND LOCALIZATION

Sound localization is accomplished by using differences in the sound signals received at different observation points to estimate the direction and eventually the actual location of the sound source. For example, the human ears, acting as two different sound observation points, enable humans to estimate the direction of arrival of the sound source. Assuming that the sound source is modeled as a point source, two different clues can be utilized in sound localization. The first clue is the interaural level difference (ILD). Emanated sound waves have a loudness that gradually decays as the observation point moves further away from the source [6]. This decay is proportional to the square of the distance between the observation point and the source location.

Knowledge about the ILD at two different observation points can be used to estimate the ratio of the distances between each observation point and the sound source location. Knowing this ratio as well as the locations of the observation points allows us to constrain the sound source location [6]. Another clue that can be utilized for sound localization is the interaural time difference (ITD), more commonly referred to as the time difference of arrival (TDOA). Assuming that the distance between each observation point and the sound source is different, the sound waves produced by the source will arrive at the observation points at different times due to the finite speed of sound.

Knowledge about the TDOA at the different observation points and the velocity of sound in air can be used to estimate the difference in the distances of the observation points to the sound source location. The difference in the distances constrains the sound source location to a hyperbola in two dimensions, or a hyperboloid in three dimensions [8].

By having several sets of observation point pairs, it becomes possible to use both the ILD and the TDOA results in order to accurately localize sound sources. In reality, for speech localization, TDOA-based location estimates are much more accurate and robust than ILD-based location estimates, which are mainly effective for signals with higher frequency components than signals with components at lower frequencies [16]. As a result, most state-of-the-art sound localization systems rely mainly on TDOA results [1, 3, 4, 8, 17].

There are many different algorithms that attempt to estimate the most likely TDOA between a pair of observers [1, 3, 18]. Usually, these algorithms have a heuristic measure that estimates the likelihood of every possible TDOA, and selects the most likely value. There are generally three classes of TDOA estimators, including the general cross-correlation (GCC) approach, the maximum likelihood (ML)approach, and the phase transform (PHAT) or frequency whitening approach [3]. All these approaches attempt to filter the cross-correlation in an optimal or suboptimal manner, and then select the time index of the peak of the result to be the TDOA estimate. A simple model of the signal received by two microphones is shown as [3]

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t), \\ x_2(t) &= h_2(t) * s(t-\tau) + n_2(t). \end{aligned} \tag{1}$$

The two microphones receive a time-delayed version of the source signal $s(t)$, each through channels with possibly different impulse responses $h_1(t)$ and $h_2(t)$, as well as a microphone-dependent noise signal $n_1(t)$ and $n_2(t)$. The main problem is to estimate τ, given the microphone signals $x_1(t)$ and $x_2(t)$. Assuming $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of $x_1(t)$ and $x_2(t)$, respectively, a common solution to this problem is the GCC shown below [3, 7],

$$\hat{\tau} = \arg\max_{\beta} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) \overline{X_2(\omega)} e^{jw\beta} dw, \tag{2}$$

where τ is an estimate of the original source signal delay between the two microphones. The actual choice of the weighing function $W(\omega)$ has been studied at length for general sound and speech sources, and three different choices, the ML [3, 19], the PHAT [3, 17], and the simple cross correlation [6] are shown below,

$$\begin{aligned} W_{\mathrm{ML}}(\omega) &= \frac{|X_1(\omega)||X_2(\omega)|}{|N_1(\omega)|^2 |X_2(\omega)|^2 + |N_2(\omega)|^2 |X_1(\omega)|^2}, \\ W_{\mathrm{PHAT}}(\omega) &= \frac{1}{|X_1(\omega) \cdot \overline{X_2(\omega)}|}, \\ W_{\mathrm{UCC}}(\omega) &= 1, \end{aligned} \tag{3}$$

where $N_1(\omega)$ and $N_2(\omega)$ are the estimated noise spectra for the first and second microphones, respectively.

The ML weights require knowledge about the spectrum of the microphone-dependent noises. The PHAT does not require this knowledge, and hence has been employed more often due to its simplicity. The unfiltered cross correlation (UCC) does not utilize any weighing function.

3. SPATIAL LIKELIHOOD FUNCTIONS

Often, it is beneficial not only to record the most likely TDOA but also the likelihood of other TDOAs [1, 15] in order to contrast the likelihood of a speaker at different spatial positions. The method of producing an array of likelihood parameters that correspond either to the direction or to the position of the sound source can be interpreted as generating a SLF [12, 14, 20]. Each microphone array, consisting of as little as 2 microphones, can produce an SLF for its environment.

An SLF is essentially an approximate (or noisy) measurement of the posterior likelihood $P(\phi(x)|\mathbf{X})$, where $\mathbf{X}$ is a matrix of all the signal samples in a 10–20-ms time segment obtained from a set of microphones and $\phi(x)$ is the event that there is a speaker at position x. Often, the direct computation of $P(\phi(x)|\mathbf{X})$ is not possible (or tractable), and as a result, a variety of methods have been proposed to efficiently measure

$$e(x) = \psi(P(\phi(x)|\mathbf{X})), \tag{4}$$

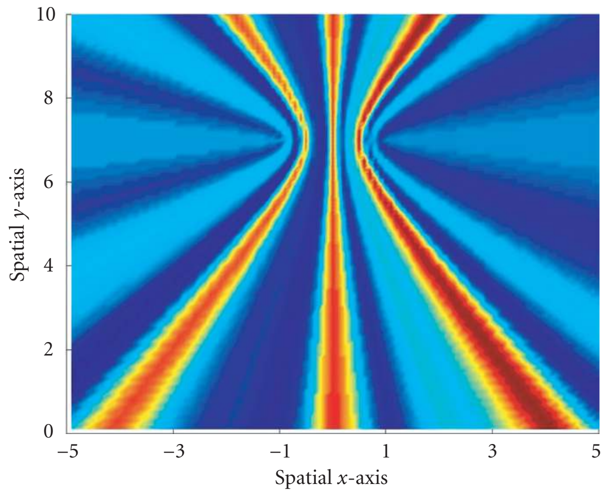

Figure 1: SLF with the dark regions corresponding to a higher likelihood and the light regions corresponding to a lower likelihood.

where $\psi(t)$ is a monotonically nondecreasing function of t. The reason for wanting a monotonically nondecreasing function is that we only care about the relative values (at different spatial locations) of the posterior likelihood and hence any monotonically nondecreasing function of it will suffice for this comparison.

In this paper, whenever we define or refer to an SLF, it is inherently assumed that the SLF is related to the posterior estimate of a speaker at position x, as defined by (4).

The simplest SLF generation method is to use the unfiltered cross correlation between two microphones, as shown in Figure 1. Assuming that $\tau(x)$ is the TDOA between the two microphones for a sound source at position x, we can define the cross-correlation-based SLF as

$$e(x) = \int_{-\infty}^{\infty} X_1(\omega)\overline{X_2(\omega)}e^{jw\tau(x)}dw. \tag{5}$$

The use of the cross correlation for the posterior likelihood estimate merits further discussion. The cross correlation is essentially an observational estimate of $P(\mathbf{X}|\phi(x))$, which is related to the posterior estimate as follows:

$$P(\phi(x)|\mathbf{X}) = \frac{P(\mathbf{X}|\phi(x))P(\phi(x))}{P(\mathbf{X})}. \tag{6}$$

The probability $P(\phi(x))$ is the prior probability of a speaker at position x, which we define as ρ_x. When using the cross correlation (or any other observational estimate) to estimate the posterior probability, we must take into account the "masking" of different positions caused by ρ_x. Note that the $P(\mathbf{X})$ term is not a function of x and hence can be neglected since, for a given signal matrix, it does not change the relative value of the SLF at different positions. In cases where all spatial positions have an equal probability of a speaker (i.e., ρ_x is constant over x), the masking effect is just a constant scaling of the observational estimate, and only in such a case, we do get the posterior estimate of (5).

SLF generation using the unfiltered cross correlation is often referred to as a delay-and-sum beamformer-based energy scans or as steered response power (SRP). Using a simple or filtered cross correlation to obtain the likelihood of different TDOAs and using them as the basis of the SLFs is not the only method for generating SLFs. In fact, for multiple speakers, using a simple cross correlation is one of the least accurate and least robust approaches [4]. Many other methods have generally been employed in multisensor-array SLF generation, including the multiple signal classification (MUSIC) algorithm [21], ML algorithm [22, 23, 24], SRP-PHAT [7], and the iterative spatial probability (ISP) algorithm [1, 15]. There are also several methods developed for wideband source localization, including [25, 26, 27]. Most of these can be classified as wideband extensions of the MUSIC or ML approaches.

The works [1, 15] describe the procedure of obtaining an SLF using TDOA distribution analysis. Basically, for the ith microphone pair, the probability density function (PDF) of the TDOA is estimated from the histogram consisting of the peaks of cross correlations performed on multiple speech segments. Here, it is assumed that the speech source (and hence the TDOA) remains stationary for the duration of time that all speech segments are recorded. Then, each spatial position is assigned a likelihood that is proportional to the probability of its corresponding TDOA. This SLF is scaled so that the maximum value of the SLF is 1 and the minimum value is 0. Higher values here correspond to a higher likelihood of a speaker at those locations.

In [7], SLFs are produced (called SRP-PHATs) for microphone pairs that are generated similarly to [1, 8, 15]. The difference is that, instead of using TDOA distributions, actual filtered cross correlations (using the PHAT cross correlation filter) are used to produce TDOA likelihoods which are then mapped to an SLF, as shown below,

$$e(x) = \sum_k \sum_l \int_{-\infty}^{\infty} \frac{X_k(\omega)\overline{X_l(\omega)}e^{j\omega\tau_{kl}(x)}}{|X_k(\omega)|\,|X_l(\omega)|}d\omega, \tag{7}$$

where $e(x)$ is the SLF, $X_i(\omega)$ is the Fourier transform of the signal received by the ith microphone, and $\tau_{kl}(x)$ is the array steering delay corresponding to the position x and the kth and lth microphones.

In the noiseless situation and in the absence of reverberations, an SLF from a single microphone array will be a representative of the number and the spatial locations of the sound sources in an environment. When there is noise and/or reverberations, the SLF of a single microphone array will be degraded [3, 7, 28]. As a result, in practical situations, it is often necessary to combine the SLFs of multiple microphone arrays in order to result in a more representative overall SLF. Note that in all of the work in [1, 7, 8, 15], SLFs are produced from 2-element microphone arrays and are simply added to produce the overall SLF which, as will be shown, is a special case of the more robust integration mechanism proposed here.

In this paper, we use the notation $e_i(x)$ for the SLF of the ith microphone array over the environment x which can be

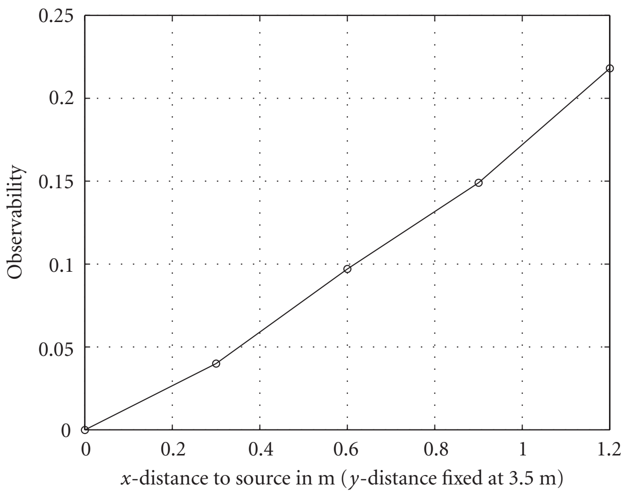

FIGURE 2: Relationship between sensor position and its observability.

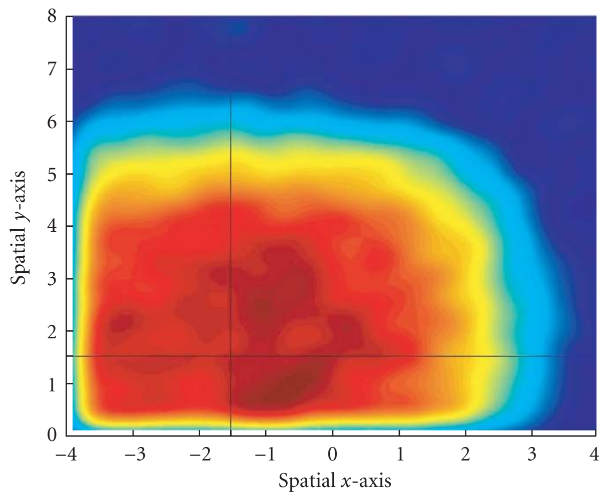

FIGURE 3: A directly estimated SOF for a 2-element microphone array. The darker regions correspond to a lower SOF and the lighter regions correspond to a higher SOF. The location of the array is depicted by the crosshairs.

a 2D or a 3D variable. In the case of 2-element microphone arrays, we also use the notation $e_{kl}(x)$ for the SLF of the microphone pair formed by the kth and lth microphones, also over the environment x.

4. SPATIAL OBSERVABILITY FUNCTIONS

Under normal circumstances, an SLF would be entirely enough to locate all spatial objects and events. However, in some situations, a sensor is not able to make inferences about a specific spatial location (i.e., blocked microphone array) due to the fact that the sensing function provides incorrect information or no information about that position. As a result, the SOF is used as an indication of the accuracy of the SLF. Although several different methods of defining the SOF exist [29, 30], in this paper, the mean square difference between the SLF and the actual probability of an object at a position is used as an indicator of the SOF.

The spatial observability of the ith microphone array corresponding to the position x can thus be expressed as

$$o_i(x) = E\left[\left(e_i(x) - a(x)\right)^2\right], \tag{8}$$

where $o_i(x)$ is the SOF, $e_i(x)$ is the SLF, and $a(x)$ is the actual probability of an object at position x, which can only take a value of 0 or 1. We can relate $a(x)$ to $\phi(x)$ as follows:

$$a(x) = \begin{cases} 1, & \text{if } \phi(x), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The actual probability $a(x)$ is a Bernoulli random variable with parameter ρ_x, the prior probability of an object at position x. This prior probability can be obtained from the nature and geometry of the environment. For example, at spatial locations where an object or a wall prevents the presence of a speaker, ρ_x will be 0 and at other "allowed" spatial regions, ρ_x will take on a constant positive value.

In order to analyze the effects of spatial position of the sound source and the observability of the microphone array, an experiment was conducted with a 2-element microphone array placed at a fixed distance of 3.5 m parallel to the spatial y-axis and a varying x-axis distance to a sound source. The SLF values of the sensor corresponding to the source position were used in conjunction with prior knowledge about the status of the source (i.e., the location of the source was known) in order to estimate the relationship between the observability of the sensor and the x-axis position of the sensor. The results of this experiment, which are shown in Figure 2, suggest that as the distance of the sensor to the source increases, so does the observability.

In practice, the SOF can be directly measured by placing stationary sound sources at known locations in space and comparing it with the array SLF or by modeling the environment and the microphone arrays with a presumed SOF [14]. The modeled SOFs typically are smaller and closer to the microphone array (more accurate localizations) and are larger further away from the array (less accurate localizations) [14]. Clearly, the SOF values will also depend upon the overall noise in the environment. More noise will increase the value of the SOFs (higher localization errors), while less noise will result in lower SOFs (lower localization errors). However, for a given environment with roughly equal noise at most locations, the relative values of the SOF will remain the same, regardless of the noise level. As a result, in practice, we often obtain a distance-to-array-dependent SOF as shown in Figure 3.

5. INTEGRATION OF DISTRIBUTED SENSORS

We will now utilize knowledge about the SLFs and SOFs in order to integrate our microphone arrays. The approach here is analogous to other sensor fusion techniques [12, 14, 20, 31].

Our goal is to find the minimum mean square error (MMSE) estimate of $a(x)$, which can be derived as follows.

Assuming that our estimate is $\tilde{a}(x)$, we can define our mean square error as

$$m(x) = \left(\tilde{a}(x) - a(x)\right)^2. \tag{10}$$

From estimation theory [32], the estimate $\tilde{a}_m(x)$ that minimizes the above mean square error is

$$\tilde{a}_m(x) = E_a\left[a(x) \mid e_0(x), e_1(x), \ldots\right]. \tag{11}$$

Now, if we assume that the SLF has a Gaussian distribution with mean equal to the actual object probability $a(x)$ [14, 20], we can rewrite the MMSE estimate as follows:

$$\begin{aligned} \tilde{a}_m(x) &= 1 \cdot P\left(a(x) = 1 \mid e_0(x), \ldots\right) \\ &\quad + 0 \cdot P\left(a(x) = 0 \mid e_0(x), \ldots\right) \\ &= P\left(a(x) = 1 \mid e_0(x), \ldots\right) \end{aligned} \tag{12}$$

which is exactly equal to (using the assumption that, for a given $a(x)$, all SLFs are independent Gaussians)

$$\tilde{a}_m(x) = \frac{1}{1 + (1-\rho_x)/\rho_x \cdot \exp\left(\sum_i \left(1 - 2e_i(x)/2o_i(x)\right)\right)}, \tag{13}$$

where ρ_x is the prior sound source probability at the location x. It is used to account for known environmental facts such as the location of walls or desks at which a speaker is less likely to be placed. Note that although the Gaussian model for the SLF works well in practice [14], it is not the only model or the best model. Other models have been introduced and analyzed [14, 20].

At this point, it is useful to define the discriminant function V_x as follows:

$$V_x = \sum_i \frac{1 - 2e_i(x)}{2o_i(x)}, \tag{14}$$

and the overall object probability function can be expressed as

$$\tilde{a}_m(x) = \frac{1}{1 + (1-\rho_x) \cdot \exp\left(V_x\right)/\rho_x}. \tag{15}$$

Hence, similar to the approach of [1, 8, 13], additive layers dependent on individual sensors can be summed to result in the overall discriminant. The discriminant is a spatial function indicative of the likelihood of a speaker at different spatial positions, with lower values corresponding to higher probabilities and higher values corresponding to lower probabilities. The discriminant does not take into account the prior sound source probabilities directly and hence a relative comparison of discriminants is only valid for positions with equal prior probabilities.

This decomposition greatly simplifies the integration of the results of multiple sensors. Also, the inclusion of the spatial observabilities allows for a more accurate model of the behavior of the sensors, thereby resulting in greater object localization accuracy. The integration strategy proposed here has been shown to be equivalent to a neural-network-based SLF fusion strategy [31]. Using neural networks often has advantages such as direct influence estimation (obtained from the neural weights) and the existence of strategies for training the network [33].

5.1. *Application to multimedia sensory integration*

The sensor integration strategy here, while focusing on microphone arrays, can be adopted to a wide variety of sensors including cameras and microphones. This work has been explored in [12]. Although observabilities were not used in this work, resulting in a possible nonideal integration of the microphone arrays and cameras, the overall result was impressive. An approximately 50% reduction in the sound localization errors was obtained at all SNRs by utilizing the audiovisual sound localization system compared to the stand-alone acoustic sound localization system. Here, the acoustic sound localization system consisted of a 3-element microphone array and the visual object localization system consisted of a pair of cameras.

5.2. *Equivalence to SRP-PHAT*

In the case when pairs of microphones are integrated without taking the spatial observabilities into account using SLFs obtained using the PHAT technique, the proposed sensor fusion algorithm is equivalent to the SRP-PHAT approach.

Assuming that the SLFs are obtained using the PHAT technique, the SLF for the kth and lth microphones can be written as

$$e_{kl}(x) = \int_{-\infty}^{\infty} \frac{X_k(\omega)\overline{X_l(\omega)}e^{j\omega\tau_{kl}(x)}}{\left|X_k(\omega)\right|\left|X_l(\omega)\right|} d\omega, \tag{16}$$

where $X_k(\omega)$ is the Fourier transform of the signal obtained by the kth microphone, $\overline{X_l(\omega)}$ is the complex conjugate of the Fourier transform of the signal obtained by the lth microphone, and $\tau_{kl}(x)$ is the array steering delay corresponding to the position x and the microphones k and l.

In most applications, we care about the relative likelihoods of objects at different spatial positions. Hence, it suffices to only consider the discriminant function of (14) here. Assuming that the spatial observability of all microphone pairs for all spatial regions is equal, we obtain the following discriminant function:

$$V_x = C_1 - C_2 \sum_i e_i(x), \tag{17}$$

where C_1 and C_2 are positive constants. Since we care only about the relative values of the discriminant, we can reduce (17) to

$$V'_x = \sum_i e_i(x), \tag{18}$$

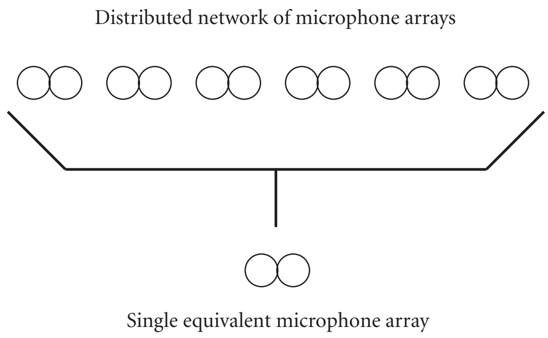

Figure 4: The integration of multiple sensors into a single "super"-sensor.

and we note that while in (17) and (18) higher values of the discriminant were indicative of a lower likelihood of an object, in (18) higher values of the discriminant are now indicative of a higher likelihood of an object. The summation over i is across all the microphone arrays. If we use only microphone pairs and use all available microphones, then we have

$$V'_x = \sum_k \sum_l e_{kl}(x). \tag{19}$$

Utilizing (16), this becomes

$$V'_x = \sum_k \sum_l \int_{-\infty}^{\infty} \frac{X_k(\omega)\overline{X_l(\omega)}e^{j\omega\tau_{kl}(x)}}{|X_k(\omega)|\,|X_l(\omega)|} d\omega \tag{20}$$

which is exactly equal to the SRP-PHAT equation [7].

6. EFFECTIVE SLF AND SOF

After the result of multiple sensors have been integrated, it is useful to get an estimate of the cumulative observability obtained as a result of the integration. This problem is equivalent to finding the SLF and SOF of a single sensor that results in the same overall object probability as that obtained by multiple sensors, as shown in Figure 4.

This can be stated as

$$\begin{aligned} &P(a(x) = 1 | e_0(x), o_0(x), \dots) \\ &\quad = P(a(x) = 1 | \overline{e}(x), \overline{o}(x)), \end{aligned} \tag{21}$$

where $\overline{e}(x)$ is the effective SLF and $\overline{o}(x)$ is the effective SOF of the combined sensors. According to (13), this problem reduces to finding equivalent discriminant functions, one corresponding to the multiple sensors and one corresponding to the effective single sensors. According to (14), this becomes (using the constraint that the effective SLF will also be a Gaussian)

$$\sum_i \frac{1 - 2e_i(x)}{2o_i(x)} = \frac{1 - 2\overline{e}(x)}{2\overline{o}(x)}. \tag{22}$$

Now, we let the effective SOF be the variance of the effective SLF, or in other words, we let the effective SOF be the observability of the effective sensor. We first evaluate the variance of the effective SLF as follows:

$$E\big(\overline{e}(x) - E\overline{e}(x)\big)^2 = \overline{o}(x)^2 E\left(\sum_i \frac{e_i(x) - a(x)}{o_i(x)}\right)^2. \tag{23}$$

The random process $e_i(x) - a(x)$ is a zero-mean Gaussian random process, and the expectation of the square of a sum of an independent set of these random processes is equal to the sum of the expectation of the square of each of these processes [34], as shown below,

$$E\big(\overline{e}(x) - E\overline{e}(x)\big)^2 = \overline{o}(x)^2 \sum_i E\left(\frac{e_i(x) - a(x)}{o_i(x)}\right)^2. \tag{24}$$

This is because all the cross-variances equal zero due to the independency of the sensors and the zero means of the random process. Equation (24) can be simplified to produce

$$E\big(\overline{e}(x) - E\overline{e}(x)\big)^2 = \overline{o}(x)^2 \sum_i E\left(\frac{e_i(x)^2 - a(x)^2}{o_i(x)^2}\right). \tag{25}$$

Now, by setting (25) equal to the effective observability, we obtain

$$\overline{o}(x) = \frac{1}{\sum_i \big(1/o_i(x)^2\big) E\big(e_i(x)^2 - a(x)^2\big)}. \tag{26}$$

Finally, noting that $E(e_i(x)^2 - a(x)^2) = o_i(x)$ according to (8), we obtain

$$\sum_i \frac{1}{o_i(x)} = \frac{1}{\overline{o}(x)}, \tag{27}$$

and the effective SLF then becomes

$$\overline{e}(x) = \frac{1}{2} - \overline{o}(x) \cdot \sum_i \frac{1 - 2e_i(x)}{2o_i(x)} = \overline{o}(x) \cdot \sum_i \frac{e_i(x)}{o_i(x)}. \tag{28}$$

7. SIMULATED AND EXPERIMENTAL RESULTS

Simulations were performed in order to understand the relationship between SNR, sound localization error, and the number of microphone pairs used. Figure 5 illustrates the results of the simulations. The definition of noise in these simulations corresponds to the second speaker (i.e., the interference signal) in the simulations. Hence, SNR in this context really corresponds to the signal-to-interference ratio (SIR).

The results illustrated in Figure 5 were obtained by simulating the presence of a sound source and a noise source at a random location in the environment and observing the sound signals by a pair of microphones. The microphone pair always has an intermicrophone distance of 15 cm but have a random location. In order to get an average over all speaker, noise, and array locations, the simulation was repeated a total of 1000 times.

Figure 5 seems to suggest that accurate and robust sound localization is not possible, because the localization error at low SNRs does not seem to improve when more microphone

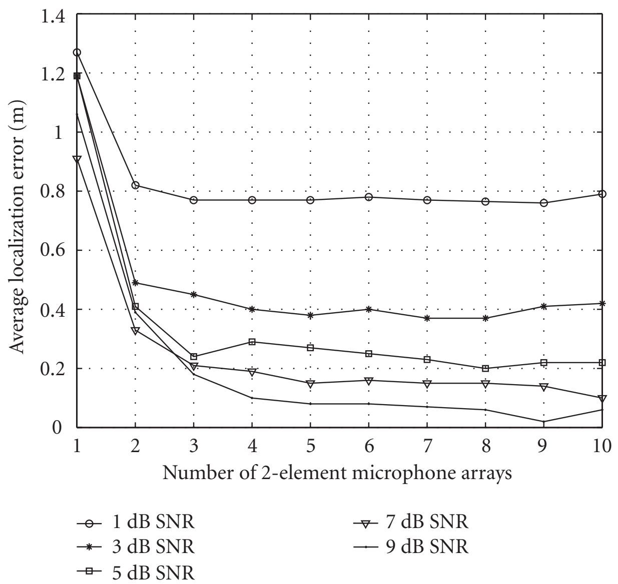

FIGURE 5: Relationship between SNR, simulated sound localization accuracy, and number of binary microphone arrays without taking spatial observabilities into consideration.

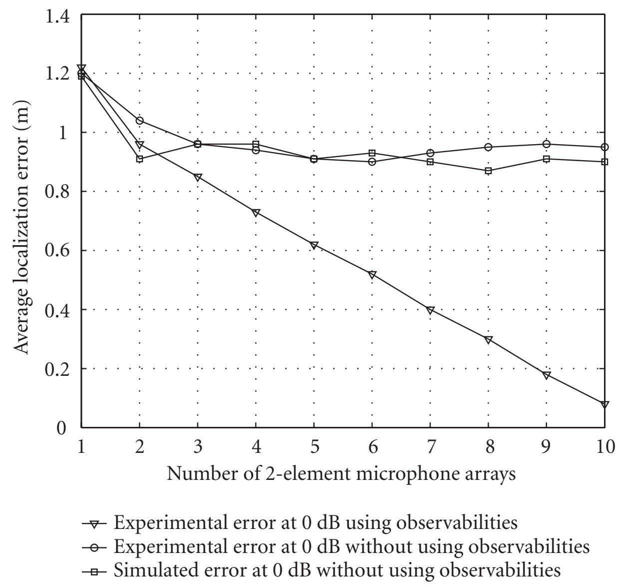

FIGURE 7: Relationship between experimental localization accuracy (at 0 dB) and number of binary microphone arrays both with and without taking spatial observabilities into consideration.

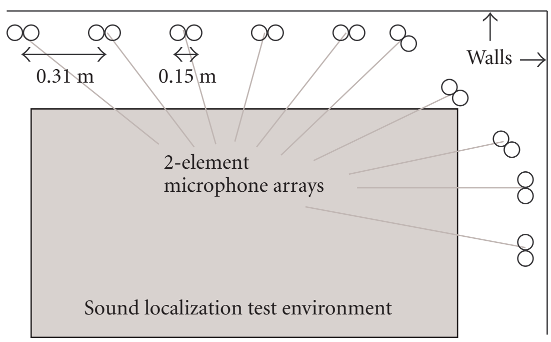

FIGURE 6: The location of the 10 2-element microphone arrays in the test environment.

arrays are added to the environment. On the other hand, at high SNRs, extra microphone arrays do have an impact on the localization error. It should be noted that the results of Figure 5 correspond to an array integration mechanism where all arrays are assumed to have the same observability over all spatial locations. In reality, differences resulting from the spatial orientation of the environment and the attenuation of the source signals usually result in one array to be more observable of a spatial position than another.

An experiment was conducted with 2-element microphone arrays at 10 different spatial positions as shown in Figure 6. Two uncorrelated speakers were placed at random positions in the environment, both with approximately equal vocal intensity that resulted in an overall SNR of 0 dB. The two main peaks of the overall speaker probability estimate were used as speaker location estimates, and for each trial the average localization error in two dimensions was calculated. The trials were repeated approximately 150 times, with the first 50 times used to train the observabilities of each of the microphone arrays by using knowledge about the estimated speaker locations and the actual speaker locations. The localization errors of the remaining 100 trials were averaged to produce the results shown in Figure 7. The localization errors were computed based on the two speaker location estimates and the true location of the speakers. Also, for each trial, the location of the two speech sources was randomly varied in the environment.

As shown in Figure 7, the experimental localization error approximately matches the simulated localization error at 0 dB for the case that all microphone arrays are assumed to equally observe the environment. The error in this case remains close to 1m even as more microphone arrays are used. Figure 7 also shows the localization error for the case that the observabilities obtained from the first 50 trials are used. In this case, the addition of extra arrays significantly reduces the localization error. When the entire set of 10 arrays are integrated, the average localization error for the experimental system is reduced to 8 cm.

The same experiment was conducted with the delay-and-sum beamformer-based SLFs (SRPs with no cross-correlation filtering) instead of the ISP-based SLF generation method. The results are shown in Figure 8.

The localization error of the delay-and-sum beamformer-based SLF generator is reduced by a factor of 2 when observability is taken into account. However, the errors are far greater than the sound localization system that uses the ISP-based SLF generator. When all 10 microphone pairs are taken into account, the localization error is approximately 0.5 m.

Now, we consider an example of the localization of 3 speakers, all speaking with equal vocal intensities. Figure 9

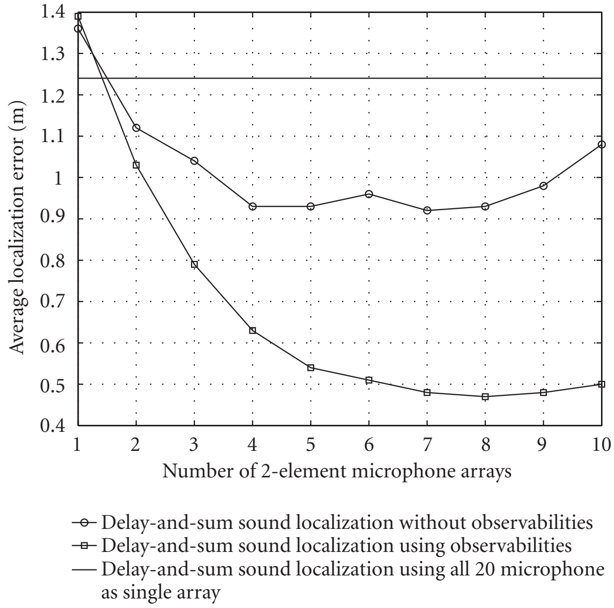

FIGURE 8: Relationship between experimental localization accuracy (at 0 dB) using a delay-and-sum beamformer-based SLFs and number of binary microphone arrays both with and without taking spatial observabilities into consideration.

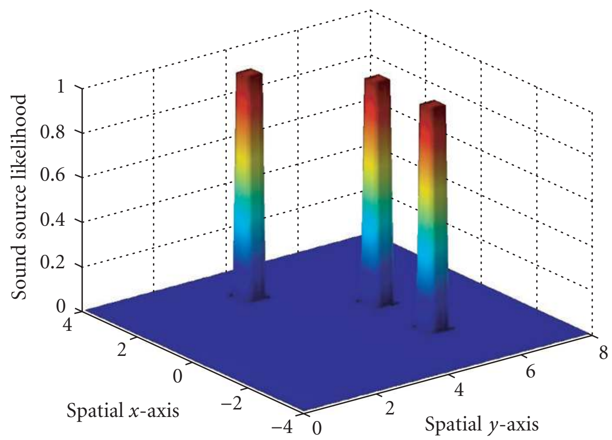

FIGURE 9: The location of 3 speakers in the environment.

illustrates the location of the speakers in a two-dimensional environment. Note that the axis labels of Figures 9, 10, and 11 correspond to 0.31-m steps.

The ISP-based SLF generator, without taking the observability of each microphone pair into account, produces the overall SLF shown in Figure 10.

In Figure 10, it is difficult to determine the true position of the speakers. There is also a third peak that does not correspond to any speaker. Using the same sound signals, an SLF was produced and shown in Figure 11, this time with taking observabilities into account.

This time, the location of the speakers can be clearly determined. Each of the three peaks correspond to the correct location of their corresponding speakers.

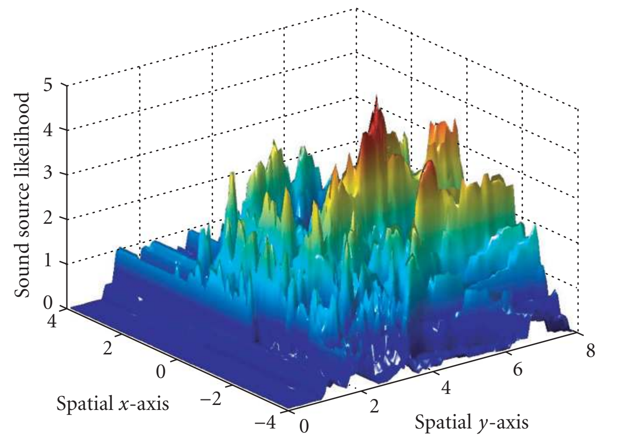

FIGURE 10: Localization of 3 speakers without using observabilities.

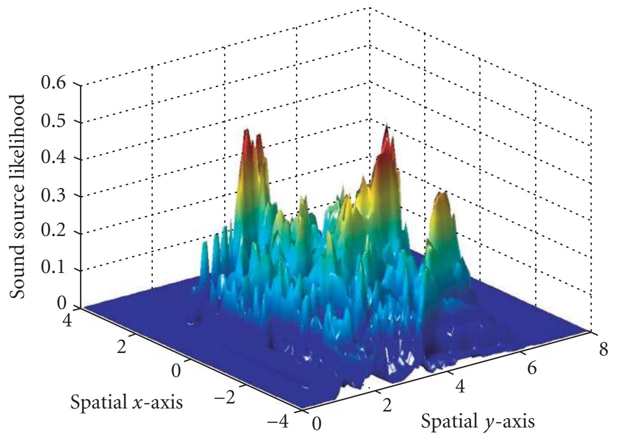

FIGURE 11: Localization of 3 speakers with observabilities.

For the experiments in Figures 10 and 11, the prior probability ρ_x for all spatial positions was assumed to be a constant of 0.3. Furthermore, the SOFs were obtained by experimentally evaluating the SOF function of (8) at several different points (for each microphone pair) and then interpolating the results to obtain an SOF for the entire space. An example of this SOF generation mechanism is the SOF of Figure 3.

The large difference between the results of Figures 10 and 11 merits further discussion. Basically, the main reason for the improvement in Figure 11 is that for locations that are farther away from a microphone pair, the estimates made by that pair are weighted less significantly than microphone pairs that are closer. On the other hand, in Figure 10, the results of all microphone pairs are combined with equal weights. As a result, even if, for every location, there are a few microphone pairs with correct estimates, the integration with the noisy estimates of the other microphone pairs taints the resulting integrated estimate.

8. CONCLUSIONS

This paper introduced the concept of multisensor object localization using different sensor observabilities in order to account for different levels of access to each spatial position. This definition led to the derivation of the minimum mean square error object localization estimates that corresponded to the probability of a speaker at a spatial location given the results of all available sensors. Experimental results using this approach indicate that the average localization error is reduced to 8 cm in a prototype environment with 10 2-element microphone arrays at 0 dB. With prior approaches, the localization error using the exact same network is approximately 0.95 m at 0 dB.

The reason that the proposed approach outperforms its previous counterparts is that, by taking into account which microphone array has better access to each speaker, the effective SNR is increased. Hence, the behaviour and performance of the proposed approach at 0 dB is comparable to that of prior approaches at SNRs greater than 7–10 dB.

Apart from improved performance, the proposed algorithm for the integration of distributed microphone arrays has the advantage of requiring less bandwidth and less computational resources. Less bandwidth is required since each array only reports its SLF, which usually involves far less information than transmitting multiple channels of audio signals. Less computational resources are required since computing an SLF for a single array and then combining the results of multiple microphone arrays by weighted SLF addition (as proposed in this paper) is computationally simpler than producing a single SLF directly from the audio signals of all arrays [14].

One drawback of the proposed technique is the measurement of the SOFs for the arrays. A fruitful direction of future work would be to model the SOF instead of experimentally measuring it, which is a very tedious process. Another area of potential future work is a better model for the speakers in the environment. The proposed model, which assumes that the actual speaker probability is independent of different spatial positions, could be made more realistic by accounting for the spatial dependencies that often exist in practice.

ACKNOWLEDGMENT

Some of the simulation and experimental results presented here have been presented in a less developed manner in [20, 31].

REFERENCES

[1] P. Aarabi and S. Zaky, "Iterative spatial probability based sound localization," in *Proc. 4th World Multi-Conference on Circuits, Systems, Computers, and Communications*, Athens, Greece, July 2000.

[2] P. Aarabi, "The application of spatial likelihood functions to multi-camera object localization," in *Proc. Sensor Fusion: Architectures, Algorithms, and Applications V*, vol. 4385 of *SPIE Proceedings*, pp. 255–265, Orlando, Fla, USA, April 2001.

[3] M. S. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 375–378, Munich, Germany, April 1997.

[4] M. S. Brandstein, *A framework for speech source localization using sensor arrays*, Ph.D. thesis, Brown University, Providence, RI, USA, 1995.

[5] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America*, vol. 78, pp. 1508–1518, November 1985.

[6] K. Guentchev and J. Weng, "Learning-based three dimensional sound localization using a compact non-coplanar array of microphones," in *Proc. AAAI Spring Symposium on Intelligent Environments*, Stanford, Calif, USA, March 1998.

[7] J. DiBiase, H. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., pp. 131–154, Springer Verlag, New York, USA, September 2001.

[8] P. Aarabi, "Multi-sense artificial awareness," M.A.Sc. thesis, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada, 1998.

[9] M. Coen, "Design principles for intelligent environments," in *Proc. 15th National Conference on Artificial Intelligence*, pp. 547–554, Madison, Wis, USA, July 1998.

[10] R. A. Brooks, M. Coen, D. Dang, et al., "The intelligent room project," in *Proc. 2nd International Conference on Cognitive Technology*, Aizu, Japan, August 1997.

[11] A. Pentland, "Smart rooms," *Scientific American*, vol. 274, no. 4, pp. 68–76, 1996.

[12] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 3, no. 2, pp. 209–223, 2001.

[13] P. Aarabi and S. Zaky, "Integrated vision and sound localization," in *Proc. 3rd International Conference on Information Fusion*, Paris, France, July 2000.

[14] P. Aarabi, *The integration and localization of distributed sensor arrays*, Ph.D. thesis, Stanford University, Stanford, Calif, USA, 2001.

[15] P. Aarabi, "Robust multi-source sound localization using temporal power fusion," in *Proc. Sensor Fusion: Architectures, Algorithms, and Applications V*, vol. 4385 of *SPIE Proceedings*, Orlando, Fla, USA, April 2001.

[16] F. L. Wightman and D. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[17] D. Rabinkin, R. J. Ranomeron, A. Dahl, J. French, J. L. Flanagan, and M. H. Bianchi, "A DSP implementation of source location using microphone arrays," in *Proc. 131st Meeting of the Acoustical Society of America*, Indianapolis, Ind, USA, May 1996.

[18] M. S. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech & Language*, vol. 9, no. 2, pp. 153–169, 1995.

[19] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] P. Aarabi, "The integration of distributed microphone arrays," in *Proc. 4th International Conference on Information Fusion*, Montreal, Canada, July 2001.

[21] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[22] H. Watanabe, M. Suzuki, N. Nagai, and N. Miki, "A method for maximum likelihood bearing estimation without nonlinear maximization," *Transactions of the Institute of Electronics, Information and Communication Engineers A*, vol. J72A, no. 8, pp. 303–308, 1989.

[23] H. Watanabe, M. Suzuki, N. Nagai, and N. Miki, "Maximum likelihood bearing estimation by quasi-Newton method using a uniform linear array," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 3325–3328, Toronto, Ontario, Canada, April 1991.
[24] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, no. 10, pp. 1553–1560, 1988.
[25] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.
[26] S. Valaee and P. Kabal, "Wide-band array processing using a two-sided correlation transformation," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 160–172, 1995.
[27] B. Friedlander and A. J. Weiss, "Direction finding for wide-band signals using an interpolated array," *IEEE Trans. Signal Processing*, vol. 41, no. 4, pp. 1618–1634, 1993.
[28] P. Aarabi and A. Mahdavi, "The relation between speech segment selectivity and time-delay estimation accuracy," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, Fla, USA, May 2002.
[29] S. S. Iyengar and D. Thomas, "A distributed sensor network structure with fault tolerant facilities," in *Intelligent Control and Adaptive Systems*, vol. 1196 of *SPIE Proceedings*, Philadelphia, Pa, USA, November 1989.
[30] R. R. Brooks and S. S. Iyengar, *Multi-Sensor Fusion: Fundamentals and Applications with Software*, Prentice Hall, Upper Saddle River, NJ, USA, 1998.
[31] P. Aarabi, "The equivalence of Bayesian multi-sensor information fusion and neural networks," in *Proc. Sensor Fusion: Architectures, Algorithms, and Applications V*, vol. 4385 of *SPIE Proceedings*, Orlando, Fla, USA, April 2001.
[32] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, Addison-Wesley, Reading, Mass, USA, 2nd edition, 1994.
[33] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1985.
[34] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, NY, USA, 2nd edition, 1984.

Parham Aarabi is a Canada Research Chair in Multi-Sensor Information Systems, an Assistant Professor in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto, and the Founder and Director of the Artificial Perception Laboratory. Professor Aarabi received his B.A.S. degree in engineering science (electrical option) in 1998, his M.A.S. degree in electrical and computer engineering in 1999, both from the University of Toronto, and his Ph.D. degree in electrical engineering from Stanford University. In November 2002, he was selected as the Best Computer Engineering Professor of the 2002 fall session. Prior to joining the University of Toronto in June 2001, Professor Aarabi was a Coinstructor at Stanford University as well as a Consultant to various silicon valley companies. His current research interests include sound localization, microphone arrays, speech enhancement, audiovisual signal processing, human-computer interactions, and VLSI implementation of speech processing applications.