# Robust Recognition of Simultaneous Speech By a Mobile Robot

Jean-Marc Valin*◇, Shun'ichi Yamamoto†, Jean Rouat◇,
François Michaud◇, Kazuhiro Nakadai‡, Hiroshi G. Okuno†
*CSIRO ICT Centre, Sydney, Australia
◇Department of Electrical Engineering and Computer Engineering
Université de Sherbrooke, 2500 boul. Université, Sherbrooke, Quebec, J1K 2R1, Canada
†Graduate School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan
‡HONDA Research Institute Japan Co., Ltd, 8-1 Honcho, Wako, Saitama, 351-0114, Japan
Jean-Marc.Valin@csiro.au
{shunichi, okuno}@kuis.kyoto-u.ac.jp
{Jean.Rouat, Francois.Michaud}@USherbrooke.ca
nakadai@jp.honda-ri.com

***Abstract*— This paper describes a system that gives a mobile robot the ability to perform automatic speech recognition with simultaneous speakers. A microphone array is used along with a real-time implementation of Geometric Source Separation and a post-filter that gives a further reduction of interference from other sources. The post-filter is also used to estimate the reliability of spectral features and compute a missing feature mask. The mask is used in a missing feature theory-based speech recognition system to recognize the speech from simultaneous Japanese speakers in the context of a humanoid robot. Recognition rates are presented for three simultaneous speakers located at 2 meters from the robot. The system was evaluated on a 200-word vocabulary at different azimuths between sources, ranging from 10 to 90 degrees. Compared to the use of the microphone array source separation alone, we demonstrate an average reduction in relative recognition error rate of 24% with the post-filter and of 42% when the missing features approach is combined with the post-filter. We demonstrate the effectiveness of our multi-source microphone array post-filter and the improvement it provides when used in conjunction with the missing features theory.**

***Index Terms*— Robot audition, microphone array, geometric source separation, missing feature theory, speech recognition, cocktail party**

## I. Introduction

The human hearing sense is very good at focusing on a single source of interest and following a conversation even when several people are speaking at the same time. This ability is known as the *cocktail party effect* [1]. To operate in human and natural settings, autonomous mobile robots should be able to do the same. This means that a mobile robot should be able to separate and recognize all sound sources present in the environment at any moment. This requires the robots not only to detect sounds, but also to locate their origin, separate the different sound sources (since sounds may occur simultaneously), and process all of this data to be able to extract useful information about the world from these sound sources.

*Corresponding author: Jean-Marc Valin (jean-marc.valin@csiro.au)

Recently, studies on robot audition have become increasingly active [2], [3], [4], [5], [6], [7], [8]. Most studies focus on sound source localization and separation. Recognition of separated sounds has not been addressed as much, because it requires integration of sound source separation capability with automatic speech recognition, which is not trivial. Robust speech recognition usually assumes source separation and/or noise removal from the feature vectors. When several people speak at the same time, each separated speech signal is severely distorted in spectrum from its original signal. This kind of interference is more difficult to counter than background noise because it is non-stationary and similar to the signal of interest. Therefore, conventional noise reduction techniques such as spectral subtraction [9], used as a front-end of an automatic speech recognizer, usually do not work well in practice.

We propose the use of a microphone array and a sound source localization system integrated with an automatic speech recognizer using the missing feature theory [10], [11] to improve robustness against non-stationary noise. In previous work [5], missing feature theory was demonstrated using a mask computed from clean (non-mixed) speech. The system we now propose can be used in a real environment by computing the missing feature mask only from the data available to the robot. To do so, a microphone array is used and a missing feature mask is generated based only on the signals available from the array post-filtering module.

This paper focuses on the integration of speech/signal processing and speech recognition techniques into a complete system operating in a real (non-simulated) environment, demonstrating that such an approach is functional and can operate in real-time. The novelty of this approach lies in the way we estimate the missing feature mask in the speech recognizer and in the tight integration of the different modules.

More specifically, we propose an original way of computing the missing feature mask for the speech recognizer that relies on a measure of frequency bins quality, estimated by our proposed postfilter. In opposition to most missing feature techniques, our approach does not need estimation of prior characteristics of the corrupting sources or noise. This leads to new capabilities in robot speech recognition with simultaneous speakers. As an example, for three simultaneous speakers, our system can allow at least three speech recognizers running simultaneously on the three separated speaker signals.

It is one of the first systems that runs in real-time on real robots while performing simultaneous speech recognition. The real-time constraints guided us in the integration of signal and speech processing techniques that are sufficiently fast and efficient. We therefore had to reject signal processing techniques that are too complex, even if potentially yielding better performance.

The paper is organized as follows. Section II discusses the state of the art and limitations of speech enhancement and missing feature-based speech recognition. Section III gives an overview of the system. Section IV presents the linear separation algorithm and Section V describes the proposed post-filter. Speech recognition integration and computation of the missing feature mask are shown in Section VI. Results are presented in Section VII, followed by the conclusion.

## II. Audition in Mobile Robotics

Artificial hearing for robots is a research topic still in its infancy, at least when compared to the work already done on artificial vision in robotics. However, the field of artificial audition has been the subject of much research in recent years. In 2004, the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) included for the first time a special session on robot audition. Initial work on sound localisation by Irie [12] for the Cog [13] and Kismet robots can be found as early as 1995. The capabilities implemented were however very limited, partly because of the necessity to overcome hardware limitations.

The SIG robot[1] and its successor SIG2[2], both developed at Kyoto University, have integrated increasing auditory capabilities [14], [15], [16], [17], [18], [19], [20] over the years (from 2000 to now). Both robots are based on binaural audition, which is still the most common form of artificial audition on mobile robots. Original work by Nakadai *et al.* [14], [15] on active audition has made it possible to locate sound sources in the horizontal plane using binaural audition and active behaviour to disambiguate front from rear. Later work has focused more on sound source separation [18], [19] and speech recognition [5], [6].

The ROBITA robot, designed at Waseda University, uses two microphones to follow a conversation between two people, originally requiring each participant to wear a headset [21], although a more recent version uses binaural audition [22].

A completely different approach is used by Zhang and Weng [23] in the SAIL robot with the goal of making a robot develop auditory capabilities autonomously. In this case, the *Q-learning* unsupervised learning algorithm is used instead of supervised learning, which is most commonly used in the field of speech recognition. The approach is validated by making the robot learn simple voice commands. Although current speech recognition accuracy using conventional methods is usually higher than the results obtained, the advantage is that the robot learns words autonomously.

More recently, robots have started taking advantage of using more than two microphones. This is the case of the Sony QRIO SDR-4XII robot [24] that features seven microphones. Unfortunately, little information is available regarding the processing done with those microphones. A service robot by Choi *et al.* [25] uses eight microphones organised in a circular array to perform speech enhancement and recognition. The enhancement is provided by an adaptive beamforming algorithm. Work by Asano, Asoh, *et al.* [2], [26], [27] also uses a circular array composed of eight microphones on a mobile robot to perform both localisation and separation of sound sources. In more recent work [28], particle filtering is used to integrate vision and audition in order to track sound sources.

In general, human-robot interface is a popular area of audition-related research in robotics. Works on robot audition for human-robot interface has also been done by Prodanov *et al.* [29] and Theobalt *et al.* [30], based on a single microphone near the speaker. Even though human-robot interface is the most common goal of robot audition research, there is research being conducted for other goals. Huang *et al.* [31] use binaural audition to help robots navigate in their environment, allowing a mobile robot to move toward sound-emitting objects without colliding with those object. The approach even works when those objects are not visible (i.e., not in line of sight), which is an advantage over vision.

## III. System overview

One goal of the proposed system is to integrate the different steps of source separation, speech enhancement and speech recognition as closely as possible to maximize recognition accuracy by using as much of the available information as possible and with a strong real-time constraint. We use a microphone array composed of omni-directional elements mounted on the robot. The missing feature mask is generated in the time-frequency plane since the separation module and the post-filter already use this signal representation. We assume that all sources are detected and localized by an algorithm such as [32], [33], although our approach is not specific to any localization algorithm. The estimated location of the sources is used by a linear separation algorithm. The separation algorithm we use is a modified version of the Geometric Source Separation (GSS) approach proposed by Parra and Alvino [34], designed to suit our needs for real-time and real-life applications. We show that it is possible to implement the separation with relatively low complexity that grows linearly with the number of microphones. The method is interesting for use in the mobile robotics context because it makes it easy to dynamically add or remove sound sources as they appear or disappear. The output of the GSS still contains residual background noise and interference, that we further attenuate through a multi-channel post-filter. The

[1]http://winnie.kuis.kyoto-u.ac.jp/SIG/oldsig/
[2]http://winnie.kuis.kyoto-u.ac.jp/SIG/

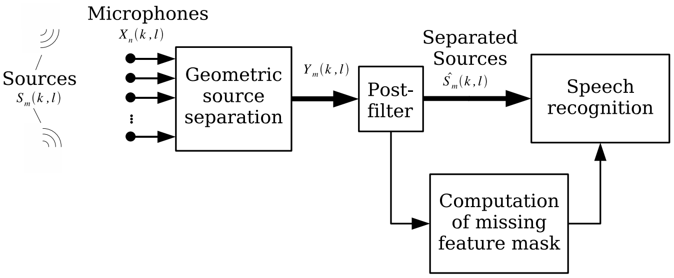


Fig. 1. Overview of the separation system with the post-filter being used both to improve the audio quality and to estimate the missing feature mask.

novel aspect of this post-filter is that, for each source of interest, the noise estimate is decomposed into stationary and transient components assumed to be due to leakage between the output channels of the initial separation stage. In the results, the performance of that post-filter is shown to be superior to those obtained when considering each separated source independently.

The post-filter we use can not only reduce the amount of noise and interference, but its behavior provides useful information that is used to evaluate the reliability of different regions of the time-frequency plane for the separated signals. Based also on the ability of the post-filter to model independently background noise and interference, we propose a novel way to estimate the missing feature mask to further improve speech recognition accuracy. This also has the advantage that acoustic models trained on clean data can be used and that no multi-condition training is required.

The structure of the proposed system is shown in Fig. 1 and its four main parts are:

1) Linear separation of the sources, implemented as a variant of the Geometric Source Separation (GSS) algorithm;
2) Multi-channel post-filtering of the separated output;
3) Computation of the missing feature mask from the post-filter output;
4) Speech recognition using the separated audio and the missing feature mask.

## IV. Geometric Source Separation

Although the work we present can be adapted to systems with any linear source separation algorithm, we propose to use the Geometric Source Separation (GSS) algorithm because it is simple and well suited to a mobile robotics application. More specificly, the approach has the advantage that it can make use of the location of the sources. In this work, we only make use of the direction information, which can be obtained with a high degree of accuracy using the method described in [3]. It was shown in [32] that distance can be estimated as well. The use of location information is important when new sources are observed. In that situation, the system can still provide acceptable separation performance (at least equivalent to the delay-and-sum beamformer), even if the adaptation has not yet taken place.

The method operates in the frequency domain using a frame length of 21 ms (1024 samples at 48 kHz). Let $S_m(k,\ell)$ be the real (unknown) sound source $m$ at time frame $\ell$ and for discrete frequency $k$. We denote as $\mathbf{s}(k,\ell)$ the vector of the sources $S_m(k,\ell)$ and matrix $\mathbf{A}(k)$ as the transfer function from the sources to the microphones. The signal received at the microphones is thus given by:

$$\mathbf{x}(k,\ell) = \mathbf{A}(k)\mathbf{s}(k,\ell) + \mathbf{n}(k,\ell) \tag{1}$$

where $\mathbf{n}(k,\ell)$ is the non-coherent background noise received at the microphones. The matrix $\mathbf{A}(k)$ can be estimated using the result of a sound localization algorithm by assuming that all transfer functions have unity gain and that no diffraction occurs. The elements of $\mathbf{A}(k)$ are thus expressed as:

$$a_{ij}(k) = e^{-\jmath 2\pi k \delta_{ij}} \tag{2}$$

where $\delta_{ij}$ is the time delay (in samples) to reach microphone $i$ from source $j$.

The separation result is then defined as $\mathbf{y}(k,\ell) = \mathbf{W}(k,\ell)\mathbf{x}(k,\ell)$, where $\mathbf{W}(k,\ell)$ is the separation matrix that must be estimated. This is done by providing two constraints (the index $\ell$ is omitted for the sake of clarity):

1) Decorrelation of the separation algorithm outputs (second order statistics are sufficient for non-stationary sources), expressed as $\mathbf{R_{yy}}(k) - \mathrm{diag}\left[\mathbf{R_{yy}}(k)\right] = \mathbf{0}$.
2) The geometric constraint $\mathbf{W}(k)\mathbf{A}(k) = \mathbf{I}$, which ensures unity gain in the direction of the source of interest and places zeros in the direction of interferences.

In theory, constraint 2) could be used alone for separation (the method is referred to as LS-C2 [34]), but this is insufficient in practice, as the method does not take into account reverberation or errors in localization. It is also subject to instability if $\mathbf{A}(k)$ is not invertible at a specific frequency. When used together, constraints 1) and 2) are too strong. For this reason, we use a "soft" constraint (refereed to as GSS-C2 in [34]) combining 1) and 2) in the context of a gradient descent algorithm.

Two cost functions are created by computing the square of the error associated with constraints 1) and 2). These cost functions are defined as, respectively:

$$J_1(\mathbf{W}(k)) = \left\|\mathbf{R_{yy}}(k) - \mathrm{diag}\left[\mathbf{R_{yy}}(k)\right]\right\|^2 \tag{3}$$

$$J_2(\mathbf{W}(k)) = \left\|\mathbf{W}(k)\mathbf{A}(k) - \mathbf{I}\right\|^2 \tag{4}$$

where the matrix norm is defined as $\|\mathbf{M}\|^2 = \mathrm{trace}\left[\mathbf{M}\mathbf{M}^H\right]$ and is equal to the sum of the square of all elements in the matrix. The gradient of the cost functions with respect to $\mathbf{W}(k)$ is equal to [34]:

$$\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 4\mathbf{E}(k)\mathbf{W}(k)\mathbf{R_{xx}}(k) \tag{5}$$

$$\frac{\partial J_2(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 2\left[\mathbf{W}(k)\mathbf{A}(k) - \mathbf{I}\right]\mathbf{A}(k) \tag{6}$$

where $\mathbf{E}(k) = \mathbf{R_{yy}}(k) - \mathrm{diag}\left[\mathbf{R_{yy}}(k)\right]$.

The separation matrix $\mathbf{W}(k)$ is then updated as follows:

$$\mathbf{W}^{n+1}(k) = \mathbf{W}^n(k) - \mu\left[\alpha(k)\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} + \frac{\partial J_2(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)}\right] \tag{7}$$

where $\alpha(f)$ is an energy normalisation factor equal to $\|\mathbf{R_{xx}}(k)\|^{-2}$ and $\mu$ is the adaptation rate.

The difference between our implementation and the original GSS algorithm described in [34] lies in the way the correlation matrices $\mathbf{R_{xx}}(k)$ and $\mathbf{R_{yy}}(k)$ are computed. Instead of using several seconds of data, our approach uses instantaneous estimates, as used in the stochastic gradient adaptation of the Least Mean Square (LMS) adaptive filter [35]. We thus assume that:

$$\mathbf{R_{xx}}(k) = \mathbf{x}(k)\mathbf{x}(k)^H \tag{8}$$
$$\mathbf{R_{yy}}(k) = \mathbf{y}(k)\mathbf{y}(k)^H \tag{9}$$

It is then possible to rewrite (5) as:

$$\frac{\partial J_1(\mathbf{W}(k))}{\partial \mathbf{W}^*(k)} = 4\left[\mathbf{E}(k)\mathbf{W}(k)\mathbf{x}(k)\right]\mathbf{x}(k)^H \tag{10}$$

which only requires matrix-by-vector products, greatly reducing the complexity of the algorithm. Similarly, the normalisation factor $\alpha(k)$ can also be simplified as $\left[\|\mathbf{x}(k)\|^2\right]^{-2}$. With a small update rate, it means that the time averaging is performed implicitly. In early experiments, the instantaneous estimate of the correlation was found to have no significant impact on the performance of the separation, but is necessary for real-time implementation.

The weight initialization we use corresponds to a delay-and-sum beamformer, referred to as the I1 (or C1) initialization method in [34]. Such initialization ensures that prior to adaptation, the performances are at worst equivalent to a delay-and-sum beamformer. In fact, if only a single source is present, our algorithm is strictly equivalent to a delay-and-sum beamformer implemented in the frequency domain.

## V. MULTI-CHANNEL POST-FILTER

To enhance the output of the GSS algorithm presented in Section IV, we derive a frequency-domain post-filter that is based on the optimal estimator originally proposed by Ephraim and Malah [36], [37]. Several approaches to microphone array post-filtering have been proposed in the past. Most of these post-filters address reduction of stationary background noise [38], [39]. Recently, a multi-channel post-filter taking into account non-stationary interferences was proposed by Cohen [40]. The novelty of our post-filter resides in the fact that, for a given channel output of the GSS, the transient components of the corrupting sources are assumed to be due to leakage from the other channels during the GSS process. Furthermore, for a given channel, the stationary and the transient components are combined into a single noise estimator used for noise suppression, as shown in Fig. 2. In addition, we explore different suppression criteria ($\alpha$ values) for optimizing speech recognition instead of perceptual quality. Again, when only one source is present, this post-filter is strictly equivalent to standard single-channel noise suppresion techniques.

### A. Noise Estimation

This section describes the estimation of noise variances that are used to compute the weighting function $G_m(k,\ell)$ by which the outputs $Y_m(k,\ell)$ of the GSS is multiplied to generate a

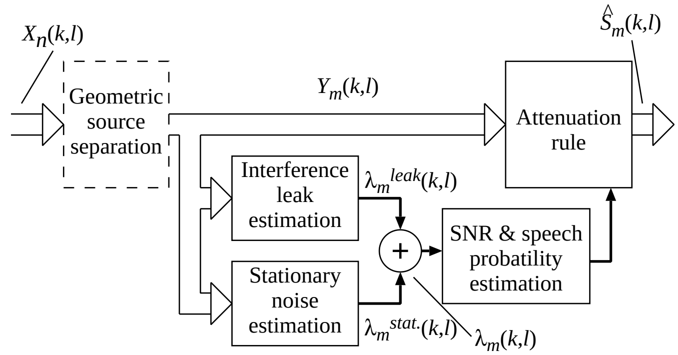


Fig. 2. Overview of the post-filter.
$X_n(k,\ell), n = 0 \ldots N-1$: Microphone inputs, $Y_m(k,\ell)$, $m = 0 \ldots M-1$: Inputs to the post-filter, $\hat{S}_m(k,\ell) = G_m(k,\ell)Y_m(k,\ell)$, $m = 0 \ldots M-1$: Post-filter outputs.

cleaned signal whose spectrum is denoted $\hat{S}_m(k,\ell)$. The noise variance estimation $\lambda_m(k,\ell)$ is expressed as:

$$\lambda_m(k,\ell) = \lambda_m^{stat.}(k,\ell) + \lambda_m^{leak}(k,\ell) \tag{11}$$

where $\lambda_m^{stat.}(k,\ell)$ is the estimate of the stationary component of the noise for source $m$ at frame $\ell$ for frequency $k$, and $\lambda_m^{leak}(k,\ell)$ is the estimate of source leakage.

We compute the stationary noise estimate $\lambda_m^{stat.}(k,\ell)$ using the Minima Controlled Recursive Average (MCRA) technique proposed by Cohen [41].

To estimate $\lambda_m^{leak}$ we assume that the interference from other sources has been reduced by a factor $\eta$ (typically $-10\,\mathrm{dB} \leq \eta \leq -3\,\mathrm{dB}$) by the separation algorithm (GSS). The leakage estimate is thus expressed as:

$$\lambda_m^{leak}(k,\ell) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,\ell) \tag{12}$$

where $Z_m(k,\ell)$ is the smoothed spectrum of the $m^{th}$ source, $Y_m(k,\ell)$, and is recursively defined (with $\alpha_s = 0.7$) as:

$$Z_m(k,\ell) = \alpha_s Z_m(k,\ell-1) + (1-\alpha_s)\left|Y_m(k,\ell)\right|^2 \tag{13}$$

It is worth noting that if $\eta = 0$ or $M = 1$, then the noise estimate becomes $\lambda_m(k,\ell) = \lambda_m^{stat.}(k,\ell)$ and our multi-source post-filter is reduced to a single-source post-filter.

### B. Suppression Rule

From here on, unless otherwise stated, the $m$ index and the $\ell$ arguments are omitted for clarity and the equations are given for each $m$ and for each $\ell$. The proposed noise suppression rule is based on minimum mean-square error (MMSE) estimation of the spectral amplitude in the ($|X(k)|^\alpha$) domain. The power coefficient $\alpha$ is chosen to maximize the recognition results.

Assuming that speech is present, the spectral amplitude estimator is defined by:

$$\hat{A}(k) = \left(E\left[|S(k)|^\alpha \left|Y(k)\right.\right]\right)^{\frac{1}{\alpha}} = G_{H_1}(k)\left|Y(k)\right| \tag{14}$$

where $G_{H_1}(k)$ is the spectral gain assuming that speech is present.

The spectral gain for arbitrary $\alpha$ is derived from Equation 13 in [37]:

$$G_{H_1}(k)=\frac{\sqrt{\upsilon(k)}}{\gamma(k)}\left[\Gamma\left(1+\frac{\alpha}{2}\right)M\left(-\frac{\alpha}{2};1;-\upsilon(k)\right)\right]^{\frac{1}{\alpha}} \quad (15)$$

where $M(a;c;x)$ is the confluent hypergeometric function, $\gamma(k)\triangleq|Y(k)|^2/\lambda(k)$ and $\xi(k)\triangleq E\left[|S(k)|^2\right]/\lambda(k)$ are respectively the *a posteriori* SNR and the *a priori* SNR. We also have $\upsilon(k)\triangleq\gamma(k)\xi(k)/\left(\xi(k)+1\right)$ [36].

The *a priori* SNR $\xi(k)$ is estimated recursively as [36]:

$$\begin{aligned}\hat{\xi}(k,\ell)=&\alpha_p G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1)\\&+(1-\alpha_p)\max\{\gamma(k,\ell)-1,0\}\end{aligned} \quad (16)$$

When taking into account the probability of speech presence, we obtain the modified spectral gain:

$$G(k)=p^{1/\alpha}(k)G_{H_1}(k) \quad (17)$$

where $p(k)$ is the probability that speech is present in the frequency band $k$ and given by:

$$p(k)=\left\{1+\frac{\hat{q}(k)}{1-\hat{q}(k)}\left(1+\xi(k)\right)\exp\left(-\upsilon(k)\right)\right\}^{-1} \quad (18)$$

The *a priori* probability of speech presence $\hat{q}(k)$ is computed as in [41] using speech measurements on the current frame for a local frequency window, a larger frequency and for the whole frame.

## VI. Integration With Speech Recognition

Robustness against noise in conventional[3] automatic speech recognition (ASR) is being extensively studied, in particular, in the AURORA project [42], [43]. To realize noise-robust speech recognition, *multi-condition training* (training on a mixture of clean speech and noises) has been studied [44], [45]. This is currently the most common method for vehicle and telephone applications. Because an acoustic model obtained by multi-condition training reflects all expected noises in specific conditions, recognizer's use of the acoustic model is effective as long as the noise is stationary. This assumption holds for example with background noise in a vehicle and on a telephone. However, multi-condition training is not effective for mobile robots, since those usually work in dynamically changing noisy environments and furthermore multi-condition training requires an important amount of data to learn from.

Source separation and speech enhancement algorithms for robust recognition are another potential alternative for automatic speech recognition on mobile robots. However, their common use is to maximize the perceptual quality of the resulting signal. This is not always effective since most preprocessing source separation and speech enhancement techniques distort the spectrum and consequently degrade features, reducing the recognition rate (even if the signal is perceived to be cleaner by naïve listeners [46]). For example, the work of Seltzer *et al.* [47] on microphone arrays addresses the problem of optimizing the array processing specifically for speech recognition (and not for a better perception). Recently, Araki *et al.* [48] have applied ICA to the separation of three sources using only two microphones. Aarabi and Shi [49] have shown speech enhancement feasibility, for speech recognition, using only the phase of the signals from an array of microphones.

[3]We use conventional in the sense of speech recognition for applications where a single microphone is used in a static environment such as a vehicle or an office.

### A. Missing Features Theory and Speech Recognition

Research of confident islands in the time-frequency plane representation has been shown to be effective in various applications and can be implemented with different strategies. One of the most effective is the missing feature strategy. Cooke *et al.* [50], [51] propose a probabilistic estimation of a mask in regions of the time-frequency plane where the information is not reliable. Then, after masking, the parameters for speech recognition are generated and can be used in conventional speech recognition systems. They obtain a significant increase in recognition rates without any explicit modeling of the noise [52]. In this scheme, the mask is essentially based on the dominance speech/interference criteria and a probabilistic estimation of the mask is used.

Conventional missing feature theory based ASR is a Hidden Markov Model (HMM) based recognizer, which output probability (emission probability) is modified to keep only the reliable feature distributions. According to the work by Cooke *et al.* [51], HMMs are trained on clean data. Density in each state $S_i$ is modelled using mixtures of $M$ Gaussians with diagonal-only covariance.

Let $f(\mathbf{x}|S)$ be the output probability density of feature vector $\mathbf{x}$ in state $S_i$, and $P(j|S_i)$ represent the mixture coefficients expressed as a probability. The output probability density is defined by:

$$f(\mathbf{x}|S_i)=\sum_{j=1}^{M}P(j|S_i)f(\mathbf{x}|j,S_i) \quad (19)$$

Cooke *et al.* [51] propose to transform (19) to take into consideration the only reliable features $x_r$ from $\mathbf{x}$ and to remove the unreliable features. This is equivalent to using the marginalization probability density functions $f(x_r|j,S_i)$ instead of $f(\mathbf{x}|j,S_i)$ by simply implementing a binary mask. Consequently, only reliable features are used in the probability calculation, and the recognizer can avoid undesirable effects due to unreliable features.

Hugo van Hamme [53] formulates the missing feature approach for speech recognizers using conventional parameters such as mel frequency cepstral coefficients (MFCC). He uses data imputation according to Cooke [51] and proposes a suitable transformation to be used with MFCC for missing features. The acoustic model evaluation of the unreliable features is modified to express that their clean values are unknown or confined within bounds. In a more recent paper, Hugo van Hamme [54] presents speech recognition results by integrating harmonicity in the signal to noise ratio for noise estimation. He uses only static MFCC as, according to his observations, dynamic MFCC do not increase sufficiently the speech recognition rate when used in the context of missing features framework. The need to estimate pitch and voiced regions in the time-space representation is a limit to this approach. In a similar approach, Raj, Seltzer and Stern [55] propose to modify the spectral representation to derive cepstral vectors. They present two missing feature algorithms that reconstruct spectrograms from incomplete noisy spectral representations (*masked* representations). Cepstral vectors can be derived from the reconstructed spectrograms for missing feature recognition. Seltzer *et al.* [56] propose the use of a Bayesian classifier to determine the reliability of spectrographic elements. Ming, Jancovic and Smith [57], [58] propose the *probabilistic union model* as an alternative to the missing feature framework. According to the authors, methods based on the missing feature framework usually require the identification of the noisy bands. This identification can be difficult for noise with unknown, time-varying band characteristics. They designed an approach for speech recognition involving partial, unknown corrupted frequency-bands. In their approach, they combine the local frequency-band information based on the union of random events, to reduce the dependence of the model on information about the noise. Cho and Oh [59] apply the *union model* to improve robust speech recognition based on frequency bands selection. From this selection, they generate "channel-attentive" mel frequency cepstral coefficients. Even if the use of missing features for robust recognition is relatively recent, many applications have already been designed.

To avoid the use of multi-condition training, we propose to merge a multi-microphone source separation and speech enhancement system with the missing feature approach. Very little work has been done with arrays of microphones in the context of missing feature theory. To our knowledge, only McCowan *et al.* [60] apply the missing feature framework to microphone arrays. Their approach defines a missing feature mask based on the input-to-output ratio of a post-filter but is however only validated on stationary noise.

Some missing feature mask techniques can also require the estimation of prior characteristics of the corrupting sources or noise. They usually assume that the noise or interference characteristics vary slowly with time. This is not possible in the context of a mobile robot. We propose to estimate quasi-instantaneously the mask (without preliminary training) by exploiting the post-filter outputs along with the local gains (in the time-frequency plane representation) of the post-filter. These local gains are used to generate the missing feature mask. Thus, the speech recognizer with clean acoustic models can adapt to the distorted sounds by consulting the post-filter feature missing masks. This approach is also a solution to the automatic generation of simultaneous missing feature masks (one for each speaker). It allows the use of simultaneous speech recognizers (one for each separated sound source) with their own mask.

### B. Reliability estimation

The post-filter uses adaptive spectral estimation of background noise and interfering sources to enhance the signal produced during the initial separation. The main idea lies in the fact that, for each source of interest, the noise estimate is decomposed into stationary and transient components assumed to be due to leakage between the output channels of the initial separation stage. It also provides useful information concerning the amount of noise present at a certain time, for each particular frequency. Hence, we use the post-filter to estimate a missing feature mask that indicates how reliable each spectral feature is when performing recognition.

### C. Computation of Missing Feature Masks

The missing feature mask is a matrix representing the reliability of each feature in the time-frequency plane. More specifically, this reliability is computed for each frame and for each mel-frequency band. This reliability can be either a continuous value from 0 to 1, or a discrete value of 0 or 1. In this paper, discrete masks are used. It is worth mentioning that computing the mask in the mel-frequency band domain means that it is not possible to use MFCC features, since the effect of the DCT cannot be applied to the missing feature mask.

For each mel-frequency band, the feature is considered reliable if the ratio of the post-filter output energy over the input energy is greater than a threshold $T$. The reason for this choice is that it is assumed that the more noise is present in a certain frequency band, the lower the post-filter gain will be for that band.

One of the dangers of computing missing feature masks based on a signal-to-noise measure is that there is a tendency to consider all silent periods as non-reliable, because they are dominated by noise. This leads to large time-frequency areas where no information is available to the ASR, preventing it from correctly identifying silence (we made this observation from practice). For this reason, it is desirable to consider as reliable at least some of the silence, especially when there is no non-stationary interference.

The missing feature mask is computed in two steps: for each frame $\ell$ and for each mel frequency band $i$:

1) We compute a continuous mask $m_\ell(i)$ that reflects the reliability of the band:

$$m_\ell(i) = \frac{S_\ell^{out}(i) + N_\ell(i)}{S_\ell^{in}(i)} \tag{20}$$

where $S_\ell^{in}(i)$ and $S_\ell^{out}(i)$ are respectively the post-filter input and output energy for frame $\ell$ at mel-frequency band $i$, and $N_\ell(i)$ is the background noise estimate. The values $S_\ell^{in}(i)$, $S_\ell^{out}(i)$ and $N_\ell(i)$ are computed using a mel-scale filterbank with triangular bandpass filters, based on linear-frequency post-filter data.

2) We deduce a binary mask $M_\ell(i)$. This mask will be used to remove the unreliable mel frequency bands at frame $\ell$:

$$M_\ell(i) = \begin{cases} 1, & m_\ell(i) > T \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $T$ is the mask threshold. We use the value $T = 0.25$, which produces the best results over a range of experiments. In practice the algorithm is not very sensitive to $T$ and all values in the $[0.15, 0.30]$ interval generally produce equivalent results.

In comparison to McCowan *et al.* [60], the use of the multi-source post-filter allows a better reliability estimation by distinguishing between interference and background noise. We include the background noise estimate $N_\ell(i)$ in the numerator of (20) to ensure that the missing feature mask equals 1 when no speech source is present (as long as there is no interference). Using a more conventional post-filter as proposed by McCowan *et al.* [60] and Cohen *et al.* [40] would not allow the mask to preserve silence features, which is known to degrade ASR accuracy. The distinction between background noise and interference also reflects the fact that background noise cancellation is generally more efficient than interference cancellation.

An example of a computed missing feature mask is shown in Fig. 3. It is observed that the mask indeed preserves the silent periods and considers unreliable the regions of the spectrum dominated by other sources. The missing feature mask for delta-features is computed using the mask for the static features. The dynamic mask $\Delta M_\ell(i)$ is computed as:

$$\Delta M_\ell(i) = \prod_{k=-2}^{2} M_{\ell-k}(i) \tag{22}$$

and is non-zero only when all the mel features used to compute the delta-cepstrum are deemed reliable.

### D. Speech Analysis for Missing Feature Masks

Since MFCC cannot be easily used directly with a missing feature mask and as the post-filter gains are expressed in the time–frequency plane, we use spectral features that are derived from MFCC features with the Inverse Discrete Cosine Transform (IDCT). The detailed steps for feature generation are as follows:

1) [FFT] The speech signal sampled at 16 kHz is analyzed using a 400-sample FFT with a 160-sample frame shift.
2) [Mel] The spectrum is analyzed by a $24^{\text{th}}$ order mel-scale filter bank.
3) [Log] The mel-scale spectrum of the $24^{\text{th}}$ order is converted to log-energies.
4) [DCT] The log mel-scale spectrum is converted by Discrete Cosine Transform to the Cepstrum.
5) [Lifter] Cepstral features 0 and 13-23 are set to zero so as to make the spectrum smoother.
6) [CMS] Convolutive effects are removed using Cepstral Mean Subtraction.
7) [IDCT] The normalized Cepstrum is transformed back to the log mel-scale spectral through an Inverse DCT.
8) [Differentiation] The features are differentiated in the time, producing 24 delta features in addition to the static features.

The [CMS] step is necessary to remove the effect of convolutive noise, such as reverberation and microphone frequency response.

The same features are used for training and evaluation. Training is performed on clean speech, without any effect from the post-filter. In practice, this means that the acoustic model does not need to be adapted in any way to our method. During

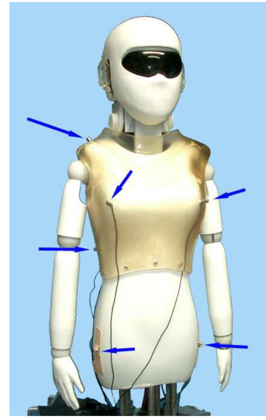

Fig. 4. SIG 2 robot with eight microphones (two are occluded).

evaluation, the only difference with a conventional ASR is the use of the missing feature mask as represented in (19).

### E. The Missing Feature based Automatic Speech Recognizer

Let $f(x|s)$ be the output probability density of feature vector $x$ in state $S$. The output probability density is defined by (19), page 5 and becomes:

$$f(x|S) = \sum_{k=1}^{M} P(k|S) f(x_r|k,S), \tag{23}$$

where $M$ is the dimensionality of the Gaussian mixture, and $x_r$ are the reliable features in $x$. This means that only reliable features are used in probability calculation, and thus the recognizer can avoid undesirable effects due to unreliable features. We used two speech recognizers. The first one is based on the CASA Tool Kit (CTK) [52] hosted at Sheffield University, U.K.[4] and the second on is the Julius open-source Japanese ASR [61] that we extended to support the above decoding process[5]. According to our preliminary experiments with these two recognizers, CTK provides slightly better recognition accuracy, while Julius runs much faster.

## VII. Results

Our system is evaluated on the SIG2 humanoid robot, on which eight omni-directional (for the system to work in all directions) microphones are installed as shown in Fig. 4. The microphone positions are constrained by the geometry of the robot because the system is designed to be fitted on any robot. All microphones are enclosed within a 22 cm $\times$ 17 cm $\times$ 47 cm bounding box. To test the system, three Japanese speakers (two males, one female) are recorded simultaneously: one in front, one on the left, and one on the right. In nine different experiments, the angle between the center speaker and the side speakers is varied from 10 degrees to 90 degrees. The speakers are placed two meters away from the robot, as shown in Fig. 5. The distance between the speakers and the robot was not found to have a significant impact on

[4] http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/ctk/
[5] http://julius.sourceforge.jp/

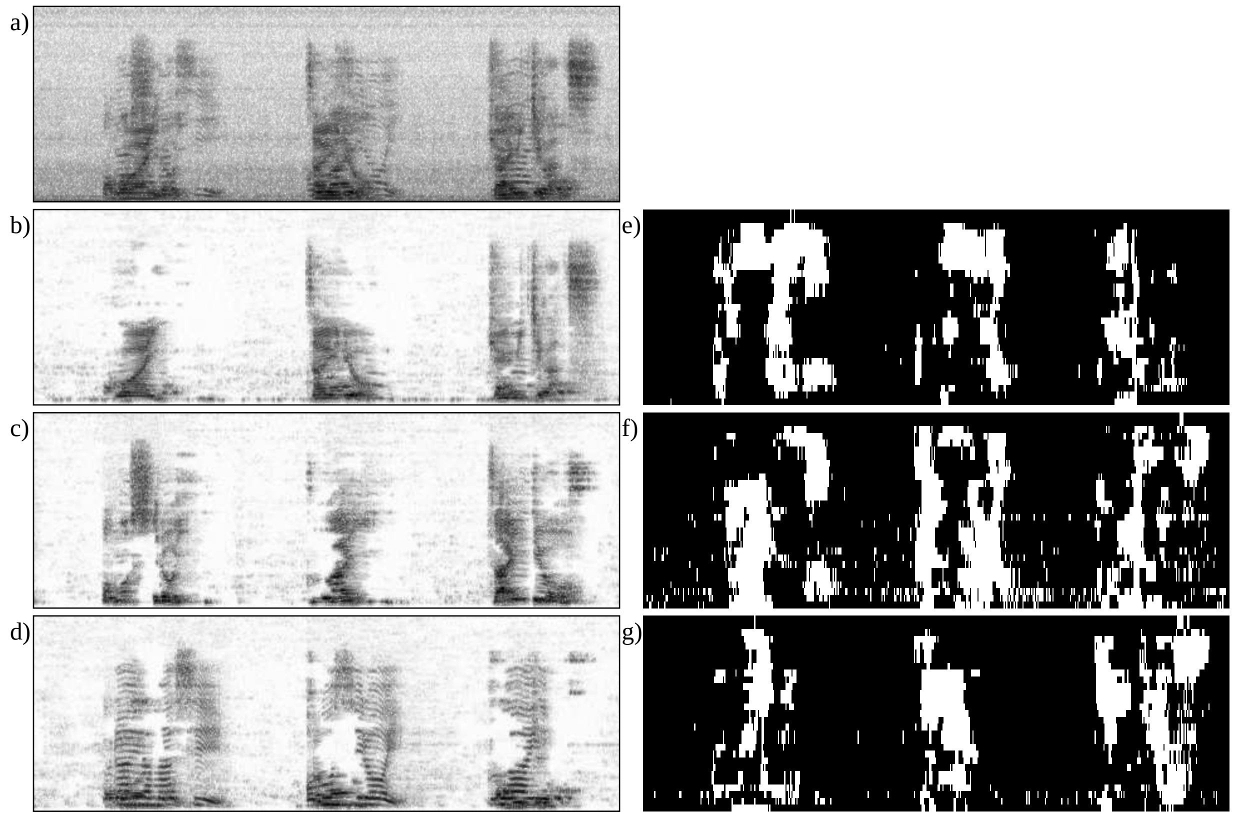


Fig. 3. Spectrograms for separation of three speakers, $90^\circ$ apart with post-filter. a) signal as captured at microphone #1 b) separated right speaker c) separated center speaker d) separated left speaker. e) - g) corresponding mel-frequency missing feature mask for static features with **reliable features ($M_\ell(i) = 1$) shown in black**. Time is represented on the *x-axis* and frequency (0-8 kHz) on the *y-axis*.

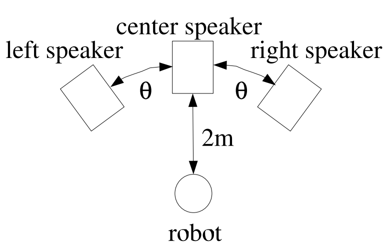


Fig. 5. Position of the speakers relative to the robot in the experimental setup.

the performance of the system. The only exception is for short distances (<50 cm) where performance decreases due to the far-field assumption we make in this particular work. The position of the speakers used for the GSS algorithm is computed automatically using the algorithm described in [3].

The room in which the experiment took place is 5 m $\times$ 4 m and has a reverberation time ($-60\,\mathrm{dB}$) of approximately $0.3$ seconds. The post-filter parameter $\alpha = 1$ (corresponding to a short-term spectral amplitude (STSA) MMSE estimator) is used since it was found to maximize speech recognition accuracy[6]. When combined together, the GSS, post-filter and missing feature mask computation require 25% of a 1.6 GHz Pentium-M to run in real-time when three sources are present[7]. Speech recognition complexity is not reported as it usually varies greatly between different engine and settings.

### *A. Separated Signals*

Spectrograms showing separation of the three speakers[8] are shown in Fig. 3, along with the corresponding mask for static features. Even though the task involves non-stationary interference with the same frequency content as the signal of interest, we observe that our post-filter is able to remove most of the interference. Informal subjective evaluation has confirmed that the post-filter has a positive impact on both quality and intelligibility of the speech. This is confirmed by improved recognition results.

### *B. Speech Recognition Accuracy*

We report speech recognition experiments obtained using the CTK toolkit. Isolated word recognition on Japanese words is performed using a triphone acoustic model. We use a speaker-independent 3-state model trained on 22 speakers (10 males, 12 females), not present in the test set. The test

[6]The difference between $\alpha = 1$ and $\alpha = 2$ on a subset of the test set was less than one percent in recognition rate

[7]Source code for part of the proposed system will be available at http://manyears.sourceforge.net/

[8]Audio signals and spectrograms for all three sources are available at: `http://www.gel.usherbrooke.ca/laborius/projects/Audible/sap/`

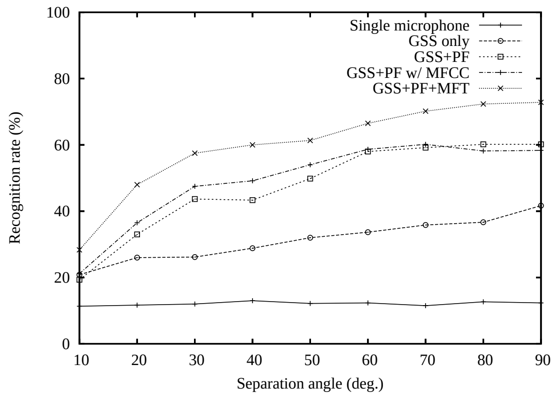


Fig. 6. Speech recognition accuracy results for intervals ranging from 10° to 90° averaged over the three speakers.

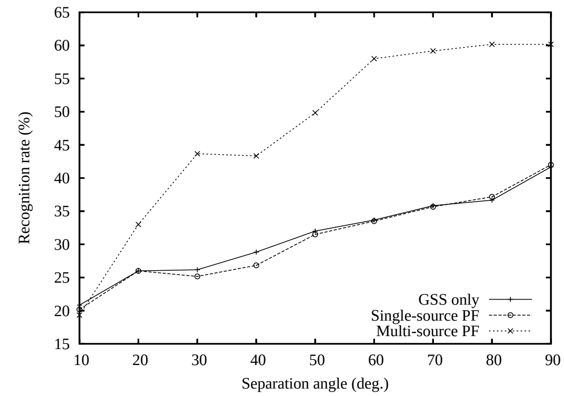


Fig. 7. Effect of the multi-source post-filter on speech recognition accuracy.

set includes 200 different ATR phonetically-balanced isolated Japanese words (300 seconds) for each of the three speakers and is used on a 200-word vocabulary (each word spoken once). Speech recognition accuracy on the clean data (no interference, no noise) varies between 94% and 99%.

Speech recognition accuracy results are presented for five different conditions:

1) Single-microphone recording
2) Geometric Source Separation (GSS) only;
3) GSS with post-filter (GSS+PF);
4) GSS with post-filter using MFCC features (GSS+PF w/ MFCC)
5) GSS with post-filter and missing feature mask (GSS+PF+MFT).

Results are shown in Fig. 6 as a function of the angle between sources and averaged over the three simultaneous speakers. As expected, the separation problem becomes more difficult as sources are located closer to each other because the difference in the transfer functions becomes smaller. We find that the proposed system (GSS+PF+MFT) provides a reduction in relative error rate compared to GSS alone that ranges from 10% to 55%, with an average of 42%. The post-filter provides an average of 24% relative error rate reduction over use of GSS alone. The relative error rate reduction is computed as the difference in errors divided by the number of errors in the reference setup. The results of the post-filter with MFCC features (4) are included to show that the use of mel spectral features only has a small effect on the ASR accuracy.

While they seem poor, the results with GSS only can be explained by the highly non-stationary interference coming from the two other speakers (especially when the speakers are close to each other) and the fact that the microphones' placement is constrained by the robot dimensions. The single microphone results are provided only as a baseline. The results are very low because a single omni-directional microphone does not have any acoustic directivity.

In Fig. 7 we compare the accuracy of the multi-source post-filter to that of a "classic" (single-source) post-filter that removes background noise but does not take interference from other sources into account ($\eta = 0$). Because the level of background noise is very low, the single-source post-filter has almost no effect and most of the accuracy improvement is due to the multi-source version of the post-filter, which can effectively remove part of the interference from the other sources. The proposed multi-source post-filter was also shown in [62] to be more effective for multiple sources than multi-channel approach in [40].

## VIII. Conclusion

In this paper we demonstrate a complete multi-microphone speech recognition system capable of performing speech recognition on three simultaneous speakers. The system closely integrates all stages of source separation and missing features recognition so as to maximize accuracy in the context of simultaneous speakers. We use a linear source separator based on a simplification of the geometric source separation algorithm. The non-linear post-filter that follows the initial separation step is a short-term spectral amplitude MMSE estimator. It uses a background noise estimate as well as information from all other sources obtained from the geometric source separation algorithm.

In addition to removing part of the background noise and interference from other sources, the post-filter is used to compute a missing feature mask representing the reliability of mel spectral features. The mask is designed so that only spectral regions dominated by interference are marked as unreliable. When compared to the GSS alone, the post-filter contributes to a 24% (relative) reduction in the word error rate while the use of the missing feature theory-based modules yields a reduction of 42% (also when compared to GSS alone). The approach is specifically designed for recognition on multiple sources and we did not attempt to improve speech recognition of a single source with background noise. In fact, for a single sound source, the proposed work is strictly equivalent to commonly used single-source techniques.

We have shown that robust simultaneous speakers speech recognition is possible when combining the missing feature framework with speech enhancement and source separation with an array of eight microphones. To our knowledge, there is no work reporting multi-speaker speech recognition using missing feature theory. This is why this paper is meant more as a proof of concept for a complete auditory system than a comparison between algorithms for performing specific signal processing tasks. Indeed, the main challenge here is the adaptation and integration of the algorithms on a mobile robot so that the system can work in a real environment (moderate reverberation) and that real-time speech recognition with simultaneous speakers be possible.

In future work, we plan to perform the speech recognition with moving speakers and adapt the post-filter to work even in highly reverberant environments, in the hope of developing new capabilities for natural communication between robots and humans. Also, we have shown that the cepstral-domain speech recognition usually performs slightly better, so it would be desirable for the technique to be generalized to the use of cepstral features instead of spectral features.

## ACKNOWLEDGEMENT

Jean-Marc Valin was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Quebec Fonds de recherche sur la nature et les technologies and the JSPS short-term exchange student scholarship. Jean Rouat is supported by NSERC. François Michaud holds the Canada Research Chair (CRC) in Mobile Robotics and Autonomous Intelligent Systems. This research is supported financially by the CRC Program, NSERC and the Canadian Foundation for Innovation (CFI). This research was also partially supported by MEXT and JSPS, Grant-in-Aid for Scientific Research (A) No.15200015 and Informatics No.16016251, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan).

## REFERENCES


[1] E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustic Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time source localization and separation system and its application to automatic speech recognition," in *Proceedings Eurospeech*, 2001, pp. 1013–1016.

[3] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proceedings IEEE International Conference on Robotics and Automation*, vol. 1, 2004, pp. 1033–1038.

[4] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 2123–2128.

[5] S. Yamamoto, K. Nakadai, H. Tsujino, and H. Okuno, "Assessment of general applicability of robot audition system by recognizing three simultaneous speeches." in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 2111–2116.

[6] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory." in *Proceedings IEEE International Conference on Robotics and Automation*, 2004, pp. 1517–1523.

[7] Q. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion (Special Issue on Robust Speech Processing)*, vol. 5, no. 2, pp. 131–140, 2004.

[8] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Transactions on Systems, Man, and Cybernetics Part B*, vol. 34, no. 3, pp. 1526–1540, 2004.

[9] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, pp. 200–203.

[10] P. Renevey, R. Vetter, and J. Kraus, "Robust speech recognition using missing feature theory and vector quantization," in *Proceedings Eurospeech*, 2001, pp. 1107–1110.

[11] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proceedings IEEE International Conference on Spoken Language Processing*, vol. I, 2000, pp. 373–376.

[12] R. Irie, "Robust sound localization: An application of an auditory perception system for a humanoid robot," Master's thesis, MIT Department of Electrical Engineering and Computer Science, 1995.

[13] R. Brooks, C. Breazeal, M. Marjanovie, B. Scassellati, and M. Williamson, "The Cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, C. Nehaniv, Ed. Spriver-Verlag, 1999, pp. 52–87.

[14] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proceedings National Conference on Artificial Intelligence*, 2000, pp. 832–839.

[15] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2000, pp. 1453–1461.

[16] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-time multiple speaker tracking by multi-modal integration for mobile robots," in *Proceedings Eurospeech*, 2001, pp. 1193–1196.

[17] H. G. Okuno, K. Nakadai, K.-I. Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2001, pp. 1402–1409.

[18] 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.

[19] ——, "Exploiting auditory fovea in humanoid-human interaction," in *Proceedings National Conference on Artificial Intelligence*, 2002, pp. 431–438.

[20] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 1147–1152.

[21] 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.

[22] Y. Matsusaka, S. Fujie, and T. Kobayashi, "Modeling of conversational strategy for the robot participating in the group conversation," in *Proceedings Eurospeech*, 2001.

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

[24] M. Fujita, Y. Kuroki, T. Ishida, and T. Doi, "Autonomous behavior control architecture of entertainment humanoid robot SDR-4X," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 960–967.

[25] C. Choi, D. Kong, J. Kim, and S. Bang, "Speech enhancement and recognition using circular microphone array for service robots," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 3516–3521.

[26] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui, "Socially embedded learning of the office-conversant mobile robot *jijo-2*," in *Proceedings International Joint Conference on Artificial Intelligence*, vol. 1, 1997, pp. 880–885.

[27] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot "Jijo-2"," in *Proceedings International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 1999, pp. 243–248.

[28] H. Asoh, F. Asano, K. Yamamoto, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, and J. Ogata, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Proceedings International Conference on Information Fusion*, 2004, pp. 805–812.

[29] P. J. Prodanov, A. Drygajlo, G. Ramel, M. Meisser, and R. Siegwart, "Voice enabled interface for interactive tour-guided robots," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002, pp. 1332–1337.

[30] C. Theobalt, J. Bos, T. Chapman, A. Espinosa-Romero, M. Fraser, G. Hayes, E. Klein, T. Oka, and R. Reeve, "Talking to Godot: Dialogue

with a mobile robot," in *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002, pp. 1338–1343.

[31] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robots and Autonomous Systems*, vol. 27, no. 4, pp. 199–209, 1999.

[32] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[33] ——, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems (Elsevier)*, vol. 55, no. 3, pp. 216–228, 2007.

[34] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[35] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall, 2002.

[36] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[37] ——, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, 1985.

[38] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1988, pp. 2578–2581.

[39] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. 905–908.

[40] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 901–904.

[41] ——, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 2, pp. 2403–2418, 2001.

[42] AURORA, "http://www.elda.fr/proj/aurora1.html" "http://www.elda.fr/proj/aurora2.html."

[43] D. Pearce, "Developing the ETSI Aurora advanced distributed speech recognition front-end & what next," in *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, 2001.

[44] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-styletraining for robust isolated-word speech recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987, pp. 705–708.

[45] M. Blanchet, J. Boudy, and P. Lockwood, "Environment adaptation for speech recognition in noise," in *Proceedings European Signal Processing Conference*, vol. VI, 1992, pp. 391–394.

[46] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1272–1305, Sept. 2003.

[47] M. L. Seltzer and R. M. Stern, "Subband parameter optimization of microphone arrays for speech recognition in reverberant environments," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. 408–411.

[48] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ica," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 881–884.

[49] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 34, no. 4, part B: Cybernetics, pp. 1763–1773, 2004.

[50] M. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Proceedings IEEE International Conference on Spoken Language Processing*, 1994, p. Paper 26.20.

[51] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[52] J. Barker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proceedings Eurospeech*, 2001, pp. 213–216.

[53] H. van Hamme, "Robust speech recognition using missing feature theory in the cepstral or LDA domain," in *Proceedings Eurospeech*, 2003, pp. 1973–1976.

[54] ——, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 213–216.

[55] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[56] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian framework for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, Sept 2004.

[57] J. Ming, P. Jancovic, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 403–414, 2002.

[58] J. Ming and F. J. Smith, "Speech recognition with unknown partial feature corruption – a review of the union model," *Computer Speech and Language*, vol. 17, pp. 287–305, 2003.

[59] H.-Y. Cho and Y.-H. Oh, "On the use of channel-attentive MFCC for robust recognition of partially corrupted speech," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 581–584, 2004.

[60] I. McCowan, A. Morris, and H. Bourlard, "Improved speech recognition performance of small microphone arrays using missing data techniques," in *Proceedings IEEE International Conference on Spoken Language Processing*, 2002, pp. 2181–2184.

[61] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open-source real-time large vocabulary recognition engine," in *Proceedings Eurospeech*, 2001, pp. 1691–1694.

[62] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

**Jean-Marc Valin** (S'03-M'05) holds B.Eng. ('99), M.A.Sc. ('01) and Ph.D. ('05) degrees in electrical engineering from the University of Sherbrooke. His Ph.D. research focused on bringing auditory capabilities to a mobile robotics platform, including sound source localization and separation. Prior to his PhD, Jean-Marc has also worked in the field of automatic speech recognition. In 2002, he wrote the Speex open source speech codec, which he keeps maintaining to this date. Since 2005, he is a postdoctoral fellow at the CSIRO ICT Centre in Sydney, Australia. His research topics include acoustic echo cancellation and microphone array processing. He is a member of the IEEE signal processing society.

**Shun'ichi Yamamoto** received the B.S. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2003 and 2005, respectively. He is currently pursuing the Ph.D degree in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. His research interests include automatic speech recognition, sound source separation, sound source localization for robot audition. Mr. Yamamoto is a member of IEEE. He has received several awards including the IEEE Robotics and Automation Society Japan Chapter Young Award.

**Jean Rouat** holds a master degree in Physics from Univ. de Bretagne, France (1981), an E. & E. master degree in speech coding and speech recognition from Université de Sherbrooke (1984) and an E. & E. Ph.D. in cognitive and statistical speech recognition jointly with Université de Sherbrooke and McGill University (1988). From 1988 to 2001 he was with Université du Québec à Chicoutimi (UQAC). In 1995 and 1996, he was on a sabbatical leave with the Medical Research Council, Applied Psychological Unit, Cambridge, UK and the Institute of Physiology, Lausanne, Switzerland. In 1990 he founded the ERMETIS, Microelectronics and Signal Processing Research Group from UQAC. He is now with Université de Sherbrooke where he founded the Computational Neuroscience and Intelligent Signal Processing Research group. He regularly acts as a reviewer for speech, neural networks and signal processing journals. He is an active member of scienti?c associations (Acoustical Society of America, Int. Speech Communication, IEEE, Int. Neural Networks Society, Association for Research in Otolaryngology, etc.). He is a senior member of the IEEE and was on the IEEE technical committee on Machine Learning for Signal Processing from 2001 to 2005.

**François Michaud** (M'90) received his bachelor's degree ('92), Master's degree ('93) and Ph.D. degree ('96) in electrical engineering from the Université de Sherbrooke, Québec Canada. After completing postdoctoral work at Brandeis University, Waltham MA ('97), he became a faculty member in the Department of Electrical Engineering and Computer Engineering of the Université de Sherbrooke, and founded LABORIUS, a research laboratory working on designing intelligent autonomous systems that can assist humans in living environments. His research interests are in architectural methodologies for intelligent decision-making, design of autonomous mobile robotics, social robotics, robot for children with autism, robot learning and intelligent systems. Prof. Michaud is the Canada Research Chairholder in Autonomous Mobile Robots and Intelligent Systems. He is a member of IEEE, AAAI and OIQ (Ordre des ingénieurs du Québec). In 2003 he received the Young Engineer Achievement Award from the Canadian Council of Professional Engineers.

**Kazuhiro Nakadai** received B.E. in electrical engineering in 1993, M.E. in information engineering in 1995 and Ph.D. in electrical engineering in 2003, from the University of Tokyo. He worked for Nippon Telegraph and Telephone and NTT Comware Corporation from 1995 to 1999. He was a researcher for Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corporation from 1999 to 2003. He is currently a senior researcher in Honda Research Institute Japan, Co., Ltd. Since 2006, he is also Visiting Associate Professor at Tokyo Institute of Technology. His research interests include signal and speech processing, AI and robotics, and he is currently engaged in computational auditory scene analysis, multi-modal integration and robot audition. He received the best paper award at IEA/AIE-2001 from the International Society for Applied Intelligence, and the best paper finalist of International Conference on Intellignet Robots and Systems (IROS 2001). He is a member of RSJ, JSAI, ASJ, and IEEE.

**Hiroshi G. Okuno** Hiroshi G. Okuno received the B.A. and Ph.D degrees from the University of Tokyo, Tokyo, Japan, in 1972 and 1996, respectively. He worked for Nippon Telegraph and Telephone, Kitano Symbiotic Systems Project, and Tokyo University of Science. He is currently a Professor in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan. He was a Visiting Scholar at Stanford University, Stanford, CA, and Visiting Associate Professor at the University of Tokyo. He has done research in programming languages, parallel processing, and reasoning mechanisms in AI, and is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition. He edited (with D. Rosenthal) Computational Auditory Scene Analysis (Princeton, NJ: Lawrence Erlbaum, 1998) and (with T. Yuasa) Advanced Lisp Technology (London, U.K.: Taylor &Francis, 2002). Dr. Okuno has received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001 and 2005, and IEEE/RSJ Nakamura Award for IROS-2001 Best Paper Nomination Finalist. He was also awarded 2003 Funai Information Science Achievement Award. He is a member of the IPSJ, JSAI, JSSST, JSCS, RSJ, ACM, AAAI, ASA, and ISCA.