Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory

Shun'ichi Yamamoto†, Jean-Marc Valin†,‡, Kazuhiro Nakadai*,
Jean Rouat‡, François Michaud‡, Tetsuya Ogata†, and Hiroshi G. Okuno†
† Graduate School of Informatics, Kyoto University, Kyoto, 606-8501 Japan
‡ LABORIUS, Department of Electrical Engineering and Computer Engineering
Université de Sherbrooke, Sherbrooke, Quebec, CANADA
∗ Honda Research Institute Japan, Co. Ltd., Wako, Saitama, 351-0114 Japan
{shunichi, jmvalin, nakadai, ogata, okuno}@kuis.kyoto-u.ac.jp

***Abstract*— A humanoid robot under real-world environments usually hears mixtures of sounds, and thus three capabilities are essential for robot audition; sound source localization, separation, and recognition of separated sounds. While the first two are frequently addressed, the last one has not been studied so much. We present a system that gives a humanoid robot the ability to localize, separate and recognize simultaneous sound sources. A microphone array is used along with a real-time dedicated implementation of Geometric Source Separation (GSS) and a multi-channel post-filter that gives us a further reduction of interferences from other sources. An automatic speech recognizer (ASR) based on the Missing Feature Theory (MFT) recognizes separated sounds in real-time by generating missing feature masks automatically from the post-filtering step. The main advantage of this approach for humanoid robots resides in the fact that the ASR with a clean acoustic model can adapt the distortion of separated sound by consulting the post-filter feature masks. Recognition rates are presented for three simultaneous speakers located at 2m from the robot. Use of both the post-filter and the missing feature mask results in an average reduction in error rate of 42% (relative).**

I. Introduction

Due to increasing demands for symbiosis of humans and robots, humanoid robots are increasingly expected to possess perceptual capabilities similar to humans. In particular, hearing capabilities are essential for social interaction, because spoken communication is very important for normal-hearing people. Unfortunately, current speech recognition technology, which usually assumes a single sound source is present, is not sufficiently accurate for real-world environments. When confronted with a mixture of sounds, three main capabilities are essential for robot audition; *sound source localization*, *separation*, and *recognition of the separated sounds*. While the first two are often addressed, the last one has not been studied as much.

A conventional approach used in human-robot interaction is to use microphones near the speaker's mouth to collect only the desired speech. Kismet of MIT has a pair of microphones with pinnae, but a human partner still uses a microphone close to the speaker's mouth [1]. A group communication robot, Robita of Waseda University, assumes that each human participant uses a headset microphone [2].

Recently, studies on robot audition has become increasingly active. IROS-2004 had, for the first time, organized sessions on Robot Audition where ten papers were presented. Most studies, however, focus on sound source localization and separation. Recognition of separated sounds has not been addressed as much, because it needs integration of sound source separation capability with automatic speech recognition.

The improvement of robustness against noises in automatic speech recognition (ASR) has been studied, in particular, in the AURORA project [3], [4]. In order to realize noise-robust speech recognition, *multi-condition training* (training on a mixture of clean speech and noises) has been studied [5], [6]. This is currently the most common method for car and telephone applications. Because an acoustic model obtained by multi-condition training reflects all expected noises in specific conditions, ASR's use of the acoustic model is effective as long as the noise is stationary. This assumption holds well for background noises in a car and on a telephone. However, multi-condition training may not be effective for robots, since they usually work under dynamically changing noisy environment. Under such conditions, missing feature theory is often used as an alternative method [7], [8].

In this paper, we use the missing feature theory to improve the robustness against non-stationary noise. In previous work [9], the idea of missing feature theory was demonstrated using a feature mask computed from the clean (non-mixed) speech. In this paper, we bring the idea one step further by computing the missing feature mask only from the data available to the robot in a real environment. In order to do so, a microphone array is used and the missing feature mask is computed using only the signals generated during the array post-filtering step [10].

The remainder of this paper is organized as follows. Section II exposes the basis of speech recognition using the missing feature theory. Section III provides an overview of the proposed recognition system. Section IV details the post-filter used and Section V explains how the missing

feature mask is computed. Results are provided in Section VI with a discussion in Section VII.

II. Missing Feature Theory

When several people speak at the same time, each separated speech is severely distorted in spectrum from its original signal. These kinds of interfering sounds, such as simultaneous speakers, are more complicated than static background noises and reverberation. Therefore, conventional noise reduction techniques such as spectral subtraction [11] will not work well, which is usually used as a front-end of an automatic speech recognizer (ASR). In this paper, we use a new ASR based on the missing feature theory [7], [8], [9], [12], [13].

The idea of the missing feature theory is that the ASR masks acoustic features according to a missing feature mask during the decoding process. The two main topics in ASR technolgy based on Missing Feature Theory (MFT) are as follows:

1) The features used in the decoding process of ASR,
2) Automatic generation of missing feature mask.

Conventional ASR usually uses MFCC (Mel Frequency Cepstral Coefficients) that capture the characteristics of voiced speech well. However, the missing feature mask can only be computed in the spectral domain and it is not possible to convert it to the cepstral domain. For that reason, we use Mel-scale spectral features in the decoding phase of ASR.

A. Missing Feature Based Speech Recognition

Since MFCC is not appropriate for recognizing separated sounds from simultaneous speeches, we use spectral features that are obtained by applying Inverse Discrete Cosine Transform (DCT) to the MFCC features. The detailed flow of calculation is as follows:

1) [FFT] 16 bit acoustic signals sampled by 16kHz are analyzed by FFT with 400 points of window and 160 frame shift to obtain spectrum.
2) [Mel] Spectrum is analyzed by Mel-scale filter bank to obtain Mel-scale spectrum of 24th order.
3) [Log] Mel-scale spectrum of 24th 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 Substraction.
7) [IDCT] The normalized Cepstrum is transformed back to the log Mel-scale spectral domain by means of an Inverse DCT.
8) [Differentiation] The features are differentiated in the time domain. Thus, we obtain 24 log spectral features as well as their first-order time derivatives.

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

B. Missing Feature Mask

The *a priori* mask has been used successfully in MFT-based ASR applications with MFCC or spectral features such as comb-filter banks. An *a priori* mask is a missing feature mask generated by comparing MFCC or spectral features of the separated speech with those of the corresponding clean speech. This kind of missing feature mask is easily generated and ASR using an *a priori* mask demonstrates quite high recognition rates. Therefore, the recognition rate by ASR with a priori mask indicates the upper limit of the performance of MFT-based ASR [12], [9]. In other words, *a priori* mask is an ideal missing feature mask.

Automatic generation of missing feature mask needs information about which spectral parts of a separated sound are distorted. This kind of information may be obtained by a sound source separation system. We use the post-filter gains as reference data to generate the missing feature mask automatically. Since we use a feature vector of 48 spectral features, the missing feature mask is a vector containing the 48 corresponding values. The value may be binary (1, reliable, or 0, unreliable) or continuous between 0 and 1.

C. Speech Recognition Based on Missing Feature Theory

Missing Feature Theory based ASR is a Hidden Markov Model (HMM) based recognizer, which is commonly used by most ASRs nowadays. Their differences reside only in the decoding process. In conventional ASR systems, estimation of a path with maximum likelihood is based on state transition probabilities and output probability in Viterbi algorithm. In case of missing feature based recognition, estimation of the output probability is different from conventional ASR systems.

Let $f(\mathbf{x}|S)$ be the output probability of feature vector $\mathbf{x}$ in state S. The output probability is defined by

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

where M is the dimensionality of the Gaussian mixture, and x_r are the reliable features in $\mathbf{x}$.

This means that only reliable features are used in probability calculation, and thus the recognizer can avoid undesirable effects due to unreliable features.

III. System overview

The speech recognition system, as shown in Figure 1, is composed of four parts:

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.

The microphone array used is composed of a number of omni-directional elements mounted on the robot. We

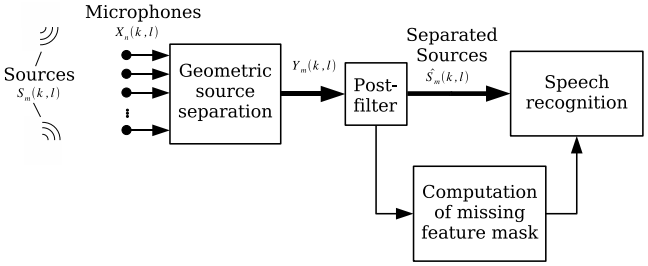

Fig. 1. Overview of the system

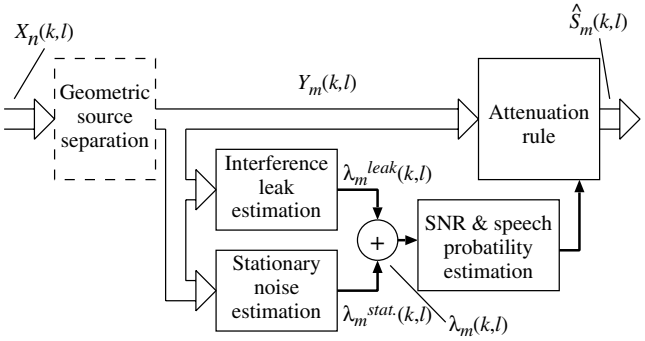

Fig. 2. Overview of the post-filter.
$X_n(k,\ell), n = 0 \dots N-1$: Microphone inputs, $Y_m(k,\ell), m = 0 \dots M-1$: Inputs to the post-filter, $\hat{S}_m(k,\ell) = G_m(k,\ell)Y_m(k,\ell)$, $m = 0 \dots M-1$: Post-filter outputs.

assume that these sources are detected and localized by an algorithm such as [14] (our approach is not specific to any localization algorithm).

A. Source Separation

The source separation stage consists of a linear separation based on the Geometric Source Separation (GSS) approach proposed by Parra and Alvino [15]. It is modified so as to provide faster adaptation using stochastic gradient estimation and shorter time frames estimations [10].

B. Multi-channel post-filter

The initial separation using GSS is followed by a multi-channel post-filter that is based on a generalization of beamformer post-filtering [16], [17] for multiple sources. This post-filter uses adaptive spectral estimation of background noise and interfering sources to enhance the signal produced during the initial separation. The main idea resides 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.

C. Missing Feature Mask Computation

The multi-channel post-filter is not only useful for reducing the amount of interference in the separated sounds. It also provides useful information concerning the amount of noise present at a certain time, at a 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.

D. Recognition

For speech recognition, we use the CASA Tool Kit (CTK) [7], which is based on the missing feature theory. The toolkit uses triphone acoustic models, and a search algorithm with a beam. Since CTK does not yet support statistical language models, we use isolated word recognition only.

IV. Multi-channel post-filter

In order to enhance the output of the GSS algorithm, we derive a frequency-domain post-filter that is based on the optimal estimator originally proposed by Ephraim and Malah [18], [19]. Several approaches to microphone array post-filtering have been proposed in the past. Most of these post-filters address reduction of stationary background noise [20], [21]. Recently, a multi-channel post-filter taking into account non-stationary interferences was proposed by Cohen [16]. The novelty of our approach resides in the fact that, for a given channel output of the GSS, the transient components of the corrupting sources is 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 Figure 2.

For this post-filter, we consider that all interferences (except the background noise) are localized (detected by the localization algorithm) sources and we assume that the leakage between channels is constant. This leakage is due to reverberation, localization error, differences in microphone frequency responses, near-field effects, etc.

Section IV-A describes the estimation of noise variances that are used to compute the weighting function G_m by which the outputs Y_m of the GSS is multiplied to generate a cleaned signal whose spectrum is denoted $\hat{S}_m$.

A. Noise estimation

The noise variance estimation $\lambda_m(k,\ell)$ is expressed as:

$$\lambda_m(k,\ell) = \lambda_m^{stat.}(k,\ell) + \lambda_m^{leak}(k,\ell) \quad (1)$$

where $\lambda_m^{stat.}(k,\ell)$ is the estimate of the stationary component of the noise for source m at frame ℓ 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 [22].

To estimate λ_m^{leak} we assume that the interference from other sources is reduced by a factor η (typically $-10\,\mathrm{dB} \leq \eta \leq -5\mathrm{dB}$) by the separation algorithm (GSS). The leakage estimate is thus expressed as:

$$\lambda_m^{leak}(k,\ell) = \eta \sum_{i=0, i\neq m}^{M-1} Z_i(k,\ell) \quad (2)$$

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:

$$Z_m(k,\ell) = \alpha_s Z_m(k,\ell-1) + (1-\alpha_s)\left|Y_m(k,\ell)\right|^2 \quad (3)$$

B. Suppression rule in the presence of speech

We now derive the suppression rule under H_1, the hypothesis that speech is present. From here on, unless otherwise stated, the m index and the ℓ arguments are omitted for clarity and the equations are given for each m and for each ℓ.

The proposed noise suppression rule is based on minimum mean-square error (MMSE) estimation of the spectral amplitude in the loudness domain, $|X(k)|^{1/2}$. The choice of the loudness domain over the spectral amplitude [18] or log-spectral amplitude [19] is motivated by better results obtained using this technique, mostly when dealing with speech presence uncertainty (Section IV-C).

The loudness-domain amplitude estimator is defined by:

$$\hat{A}(k) = \left(E\left[|S(k)|^{\alpha}\,|Y(k)|\right]\right)^{\frac{1}{\alpha}} = G_{H_1}(k)\,|Y(k)| \qquad (4)$$

where $\alpha = 1/2$ for the loudness domain and $G_{H_1}(k)$ is the spectral gain assuming that speech is present.

The spectral gain for arbitrary α is derived from Equation 13 in [19]:

$$G_{H_1}(k) = \frac{\sqrt{\upsilon(k)}}{\gamma(k)}\left[\Gamma\left(1+\frac{\alpha}{2}\right) M\left(-\frac{\alpha}{2}; 1; -\upsilon(k)\right)\right]^{\frac{1}{\alpha}} \qquad (5)$$

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 $\upsilon(k) \triangleq \gamma(k)\xi(k)/\left(\xi(k)+1\right)$ [18].

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

$$\begin{aligned}\hat{\xi}(k,l) &= \alpha_p G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) \\ &+ (1-\alpha_p)\max\left\{\gamma(k,\ell)-1, 0\right\} \qquad (6)\end{aligned}$$

using the modifications proposed in [22] to take into account speech presence uncertainty.

C. Optimal gain modification under speech presence uncertainty

In order to take into account the probability of speech presence, we derive the estimator for the loudness domain:

$$\hat{A}(k) = \left(E\left[A^{\alpha}(k)|\,Y(k)\right]\right)^{\frac{1}{\alpha}} \qquad (7)$$

Considering H_1, the hypothesis of speech presence for source m, and H_0, the hypothesis of speech absence, we obtain:

$$\begin{aligned}E[A^{\alpha}(k)|Y(k)] &= p(k)E\left[A^{\alpha}(k)|\,H_1, Y(k)\right] \\ &+ [1-p(k)]E[A^{\alpha}(k)|H_0, Y(k)] \qquad (8)\end{aligned}$$

where $p(k)$ is the probability of speech at frequency k.

The optimally modified gain is thus given by:

$$G(k) = \left[p(k)G_{H_1}^{\alpha}(k) + (1-p(k))G_{min}^{\alpha}\right]^{\frac{1}{\alpha}} \qquad (9)$$

where $G_{H_1}(k)$ is defined in (5), and G_{min} is the minimum gain allowed when speech is absent. Unlike the log-amplitude case, it is possible to set $G_{min} = 0$ without running into problems. For $\alpha = 1/2$, this leads to:

$$G(k) = p^2(k)G_{H_1}(k) \qquad (10)$$

Setting $G_{min} = 0$ means that there is no arbitrary limit on attenuation. Therefore, when the signal is certain to be non-speech, the gain can tend toward zero. This is especially important when the interference is also speech since, unlike stationary noise, residual babble noise always results in musical noise.

The probability of speech presence is computed as:

$$p(k) = \left\{1 + \frac{\hat{q}(k)}{1-\hat{q}(k)}\left(1+\xi(k)\right)\exp\left(-\upsilon(k)\right)\right\}^{-1} \qquad (11)$$

where $\hat{q}(k)$ is the *a priori* probability of speech presence for frequency k and is defined as:

$$\hat{q}(k) = 1 - P_{local}(k)P_{global}(k)P_{frame} \qquad (12)$$

where $P_{local}(k)$, $P_{global}(k)$ and P_{frame} are defined in [22] and correspond respectively to a speech measurement on the current frame for a local frequency window, a larger frequency and for the whole frame.

V. Computation of Missing Feature Mask

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 bank 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.

We compute the missing feature mask by comparing the input and the output of the multi-channel post-filter presented in Section IV. For each Mel-frequency band, the feature is considered reliable if the ratio of the 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 present in a certain frequency band, the lower the post-filter gain will be for that band. The continuous missing feature mask $m_k(i)$ is thus computed as:

$$m_k(i) = \frac{S_k^{out}(i) + N_k(i)}{S_k^{in}(i)} \qquad (13)$$

where $S_k^{in}(i)$ and $S_k^{out}(i)$ are respectively the post-filter input and output energy for frame k, at Mel-frequency band i and $N_k(i)$ is the background noise estimate for that band. The main reason for including the noise estimate $N_k(i)$ in the numerator of equation 13 is that it ensures that the missing feature mask equals 1 when no speech source is present. This allows the silence model to work properly. From the continuous mask $m_k(i)$, we derive a binary mask $M_k(i)$ as:

$$M_k(i) = \begin{cases} 1, & m_k(i) > T \\ 0, & \text{otherwise} \end{cases}$$

where T is an arbitrary threshold (we use $T = 0.3$). An example computation of the mask is shown in Figure 3.

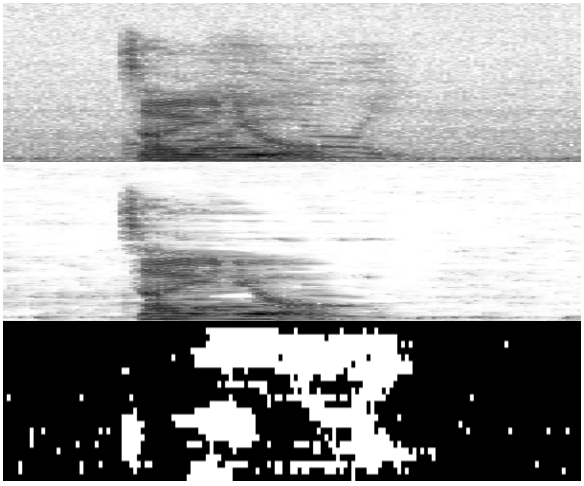

Fig. 3. Missing feature mask computation a) Spectrogram at post-filter input b) Spectrogram at post-filter output c) Mel-frequency missing feature mask with reliable features (value of 1) shown in black

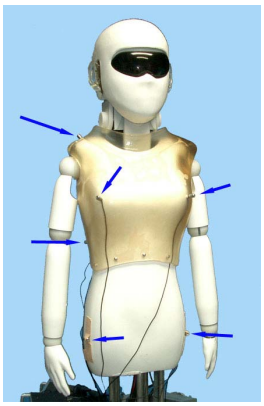

Fig. 4. SIG 2 robot with eight microphones (two are occluded).

The missing feature mask for delta-features is computed using the mask for the static features. Let $M_k(i)$ be the static mask for frame i and band k, the dynamic mask $\Delta M_k(i)$ is computed as:

$$\Delta M_k(i) = M_{k-2}(i)M_{k-1}(i)M_{k+1}(i)M_{k+2}(i)$$

VI. Results

The proposed system is evaluated on the SIG2 humanoid robot, on which an array of eight microphones is installed (Fig. 4). In order to test the system, three voices are recorded simultaneously from loudspeakers placed two meters away from the robot. The room size is $5\,m \times 4\,m$, with a reverberation time of 0.3 – 0.4 sec. We use combinations of three different words selected from a set of 216 phonemically-balanced Japanese words.

Two experiments are performed. In the first, the loudspeakers are placed at $90°$ interval ($-90°$, $0°$ and $90°$) relative to the robot. In the second experiment, we use a $60°$ interval ($-60°$, $0°$ and $60°$). Recognition is performed using vocabulary sizes of 10, 50, 100 and 200 words.

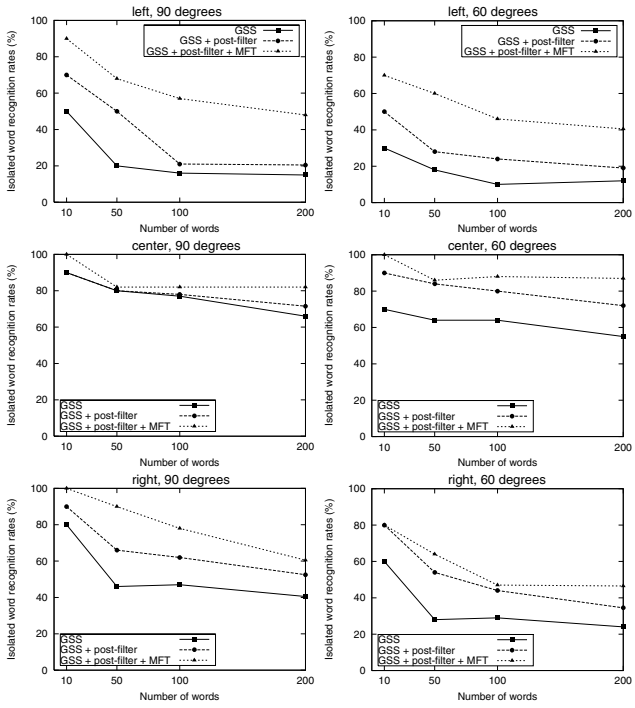

Fig. 5. Speech recognition accuracy results for $60°$ and $90°$ interval (left, center and right).

A. Acoustic Model for Speech Recognition

Even though many direction- and speaker-dependent acoustic models have been used in the past, we use only one triphone-based acoustic model for this system. The acoustic model is based on Hidden Markov Models (HMM) and is trained on clean speech. The training data includes utterance sets from 25 male and female speakers. Each utterance set consists of 216 phonemically-balanced Japanese words. The acoustic model uses 3 states and 8 Gaussians per mixture.

B. Recognition accuracy

We present speech recognition accuracy results obtained in three different conditions:

1) GSS separation only;
2) GSS separation plus post-filter;
3) GSS separation plus most-filter and missing feature mask.

Results are presented in Figure 5 for all vocabulary sizes. On a vocabulary of 200 words, the post-filter alone brings an average reduction in error rate of 17% (relative). When the post-filter is combined with missing feature theory, the improvement becomes 42%. It is worth noting that the large differences in recognition accuracy as a function of direction (left, center, right) are mainly due to differences in playback level, resulting in differences in SNR levels after GSS.

VII. Discussion And Conclusion

In order for humanoid and mobile robots to evolve in real-world environments communicate with human by means of spoken languages, robot audition capabilities are required. Since in such environments, robots hear a mixture

of sounds, three capabilities are required: sound source localization, separation and recognition.

In this paper, we focus on the last capability by integrating GSS and post-filtering sound source separation and missing feature theory based automatic speech recognition. When assuming that the acoustic environment does not change much, ASR with multi-condition trained acoustic model tends to work well. However, robots that interact with multiple people are requested under dynamically changing environments. Therefore, ASR should work with a single acoustic model by adapting it to a current environment. That is why we adopt a missing feature theory based ASR.

The system consists of geometric source separation, post-filtering, computation of missing feature mask and missing feature theory based ASR. Recognition experiments were conducted, during which three simultaneous Japanese word utterances were played from loudspeakers placed at a distance of 2 meters from the robot. On average, combination of the post-filter and missing feature theory provides a reduction of 42% (relative) in error rate, while the post-filter alone contributes to a 17% reduction in error rate.

For robot audition, these preliminary results are promising in two ways:

- The whole system runs in real-time.
- Automatic generation of missing feature mask is achieved.

We believe that the latter is the first successful report as far as we know.

The future work includes extensive verification of the performance for different directions and distance, improvement of peripheral speakers, and application to group interactions. Also, the theory that validates the methodology of this paper should be established so as to apply it to a wider area of applications.

ACKNOWLEDGMENT

This research was 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). Jean-Marc Valin was supported by the JSPS short-term exchange student scholarship and the Quebec *Fonds de recherche sur la nature et les technologies*.

REFERENCES

[1] C. Breazeal, "Emotive qualities in robot speech," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)*. IEEE, 2001, pp. 1389–1394.

[2] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," in *Proc. of Eurospeech-1999*, 1999, pp. 1723–1726.

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

[4] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & what next," in *Proc. of Eurospeech-2001*. ESCA, 2001.

[5] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-styletraining for robust isolated-word speech recognition," in *Proc. of ICASSP-87*. IEEE, 1987, pp. 705–708.

[6] M. Blanchet, J. Boudy, and P. Lockwood, "Environmentadaptation for speech recognition in noise," in *Proc. of EUSIPCO-92*, vol. VI, 1992, pp. 391–394.

[7] 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 *Proc. of Eurospeech-2001*. ESCA, 2001, pp. 213–216.

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

[9] S. Yamamoto, K. Nakadai, H. Tsujino, and H. Okuno, "Assessment of general applicability of robot audition system by recognizing three simultaneous speeches." in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004)*. IEEE and RSJ, 2001, p. to appear.

[10] 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.

[11] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proceedings of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*. IEEE, 1979, pp. 200–203.

[12] 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 *Proc. of IEEE-RAS International Conference on Robots and Automation (ICRA-2004)*. IEEE, 2001, pp. 1517–1523.

[13] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. of 6th International Conference on Spoken Language Processing (ICSLP-2000)*, vol. I, 2000, pp. 373–376.

[14] 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*, 2004.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] ——, "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.

[20] 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.

[21] 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.

[22] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 2, pp. 2403–2418, 2001.