2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Improvement in Listening Capability for Humanoid Robot HRP-2

Toru Takahashi, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno.

Abstract— **This paper describes improvement of sound source separation for a simultaneous automatic speech recognition (ASR) system of a humanoid robot. A recognition error in the system is caused by a separation error and interferences of other sources. In separability, an original geometric source separation (GSS) is improved. Our GSS uses a measured robot's head related transfer function (HRTF) to estimate a separation matrix. As an original GSS uses a simulated HRTF calculated based on a distance between microphone and sound source, there is a large mismatch between the simulated and the measured transfer functions. The mismatch causes a severe degradation of recognition performance.**

Faster convergence speed of separation matrix reduces separation error. Our approach gives a nearer initial separation matrix based on a measured transfer function from an optimal separation matrix than a simulated one. As a result, we expect that our GSS improves the convergence speed. Our GSS is also able to handle an adaptive step-size parameter.

These new features are added into open source robot audition software (OSS) called "HARK" which is newly updated as version 1.0.0. The HARK has been installed on a HRP-2 humanoid with an 8-element microphone array. The listening capability of HRP-2 is evaluated by recognizing a target speech signal which is separated from a simultaneous speech signal by three talkers. The word correct rate (WCR) of ASR improves by 5 points under normal acoustic environments and by 10 points under noisy environments. Experimental results show that HARK 1.0.0 improves the robustness against noises.

I. INTRODUCTION

Automatic speech recognition (ASR) is essential to a humanoid robot which interacts with humans. In a dairy life, it is required for a humanoid robot to have the listening capability of recognizing simultaneous speech signals. The capability enables a humanoid robot to work where multiple sound sources exist besides target speech sources and noise sources radiated from robot's own motors. A typical speech recognition system assumes a single target speech source. Such a system avoided multiple speech recognition problems by means of making a user wear a headset microphone [1].

Separation error and interference of other sources cause a recognition error in a simultaneous speech recognition system. The error and interference contaminate acoustic feature, which is extracted from the separated signal. Thus, the acoustic feature mismatches an acoustic model of an ASR system. Although an acoustic model adaptation technique is available for reducing the mismatch, we have to know about the error and interference in advance. The separated speech signals are required as an adaptive training data set. Such signals are hard to correct in general.

We improve an original geometric source separation (GSS) in separability. As microphones are installed on a robot head of a humanoid robot, each microphones are receives a direct-wave and an indirect-wave, such as a reflected wave from the robot head. Thus, our GSS uses a measured robot's head related transfer function (HRTF) instead of a simulated one to estimate a proper separation matrix. Separation results based on a proper separation matrix enables the robots to improve the robustness against noises. An original GSS uses a simulated robot's HRTF calculated based on a distance between positions of microphones and sound sources to estimate a separation matrix. It's assumed that microphones are located in an acoustical free field condition. Thus, it's assumed that each microphone only receives a direct-wave component of source signal. The assumption is unsatisfied in an ASR system of a humanoid robot.

Our simultaneous speech recognition system consists of capturing sounds with a microphone array, localizing sound sources, separating each sound source, and recognizing each separated source by an ASR system. It is based on "Robot Audition", which can handle recognition of noisy speech such as simultaneous speakers by using robot-embedded microphones, that is, the ears of a robot, was proposed in [2]. It has been studied actively for recent years [3], [4], [5], [6], [7], [8], [9], [10]. We provides the platform as an open source robot audition software (OSS) called HARK stands for Honda Research Institute Japan Audition for Robots with Kyoto University, which has a meaning of "listen" in old English. It is available at http://winnie.kuis.kyoto-u.ac.jp/HARK/, which is newly updated as version 1.0.0. Our new GSS is also included.

The rest of the chapter is organized as follows: Section II introduces sound source separation. Section III describes issues and approach. Section IV explains the implementation of HARK. How to use HARK to construct robot audition systems. Section V describes the evaluation of the system, and the last section concludes the paper.

II. SOUND SOURCE SEPARATION

First we describe two algorithms of sound source separation in a previous version HARK 0.1.7, that is a Delay-and-Sum (DS) beamforming and a Geometric Source Separation (GSS) [11]. A GSS is available by providing a patch for SeparGSS which changes I/O IF to be able to use as a HARK module.

T. Takahashi K. Komatani, T. Ogata and H. G. Okuno are with the Department of Intelligence andScience and Technology. Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan {tall, Komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

K. Nakadai is with Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0114, JAPAN, and also with Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8552, JAPAN nakadai@jp.honda-ri.com

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

A DS beamforming separates sound sources by using sound source tracking results. It is easy to control beamformer parameters, and it shows high robustness for environmental noises. A large number of microphones are necessary to get high separation performance. A GSS is a kind of hybrid algorithm of a Blind Source Separation (BSS) and a beamforming.

As a GSS shows higher separation performance than a DS beamforming, we improve a GSS. A current implementation has four problems, thus we re-implemented GSS module and improve the four problems.

A. Formulation of GSS

Suppose that there are M sources and N ($\geq M$) microphones. A spectrum vector of M sources at frequency ω, $\boldsymbol{s}(\omega)$, is denoted as $[s_1(\omega)s_2(\omega)\ldots s_M(\omega)]^T$, and a spectrum vector of signals captured by the N microphones at frequency ω, $\boldsymbol{x}(\omega)$, is denoted as $[x_1(\omega)x_2(\omega)\ldots x_N(\omega)]^T$, where T represents a transpose operator. $\boldsymbol{x}(\omega)$ is, then, calculated as

$$\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega), \tag{1}$$

where $\boldsymbol{H}(\omega)$ is a transfer function matrix. Each component H_{nm} of the transfer function matrix represents the transfer function from the m-th source to the n-th microphone. The source separation is generally formulated as

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega), \tag{2}$$

where $\boldsymbol{W}(\omega)$ is called a *separation matrix*. The separation is defined as finding $\boldsymbol{W}(\omega)$ which satisfies the condition that output signal $\boldsymbol{y}(\omega)$ is the same as $\boldsymbol{s}(\omega)$. In order to estimate $\boldsymbol{W}(\omega)$, GSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}) defined by

$$J_{SS}(\boldsymbol{W}) = \|E[\boldsymbol{y}\boldsymbol{y}^H - \mathrm{diag}[\boldsymbol{y}\boldsymbol{y}^H]]\|^2, \tag{3}$$

$$J_{GC}(\boldsymbol{W}) = \|\mathrm{diag}[\boldsymbol{W}\boldsymbol{D} - \boldsymbol{I}]\|^2, \tag{4}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, diag$[\cdot]$ is the diagonal operator, $E[\cdot]$ represents the expectation operator and H represents the conjugate transpose operator. $\boldsymbol{D}$ shows a transfer function matrix based on a direct sound path between a sound source and each microphone. The total cost function $J(\boldsymbol{W})$ is represented as

$$J(\boldsymbol{W}) = \alpha_S J_{SS}(\boldsymbol{W}) + J_{GC}(\boldsymbol{W}), \tag{5}$$

where α_S means the weight parameter that controls the weight between the separation cost and the cost of the geometric constraint. This parameter is usually set to $\|\boldsymbol{x}^H\boldsymbol{x}\|^{-2}$ according to [12]. In an online version of GSS, $\boldsymbol{W}$ is updated by minimizing $J(\boldsymbol{W})$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu\boldsymbol{J}'(\boldsymbol{W}_t), \tag{6}$$

where $\boldsymbol{W}_t$ denotes $\boldsymbol{W}$ at the current time step t, $\boldsymbol{J}'(\boldsymbol{W})$ is defined as an update direction of $\boldsymbol{W}$, and μ means a step-size parameter.

III. ISSUES AND APPROACHES FOR HARK

When we constructed a robot audition system based on HARK 0.1.7, we found problems both in separation and in ASR.

A. Issues in Sound Source Separation

GSS has high separation performance originating from BSS (Eq. (3)), and also relaxes BSS's limitations such as permutation and scaling problems by introducing "geometric constraints" obtained from the locations of microphones and sound sources (Eq. (4)). Therefore, GSS has better performance than delay-and-sum beamforming with a small number of microphones. However, current implementation has the following problems, thus we re-implemented a GSS module to solve these problems.

1) The transfer function $\boldsymbol{D}$ is calculated from the relationship between microphone and source locations. This means that the effect of a robot head was not considered to get $\boldsymbol{D}$, and the calculated $\boldsymbol{D}$ has large errors. This lead to low separation performance and slow convergence of $\boldsymbol{W}$.
2) Usually, a robot has fans and actuators that generate stationary noise. This kind of noise always should be removed in separation. However, the number of sound sources is decided by thresholding a spatial spectrum estimated in sound source localization. Sometimes it fails in detecting a robot's stationary noise with a high threshold, or too many erroneous noise sources are detected with a low threshold.
3) $\boldsymbol{W}$ is initialized at the beginning of each utterance, and the initial value of $\boldsymbol{W}$ ' is calculated from $\boldsymbol{D}$. However, this initial value includes a lot of errors, and thus the convergence of $\boldsymbol{W}$ is slow.
4) Moving sources were not considered. An update of $\boldsymbol{W}$ (Eq. (6)) is based on sound source direction. This means that $\boldsymbol{W}$ is successfully updated only when a sound source is stationary.

B. Approaches in Sound Source Separation

For four problems, we take following approaches to solve the problem.

1) To use more realistic transfer function $\boldsymbol{D}$ than calculated transfer function $\boldsymbol{D}$ from the relationship between microphone and source location, our new GSS implementation can support measured transfer functions A tool which converts measured impulse responses into a transfer function matrix file for GSS is also provided. The measurement based transfer function is expected to have better performance in separation.
2) To deal with robot's noises such as fans and actuators, we implemented our new GSS module so that we can specify a fixed direction of noise source. When this is specified, this module always removes the corresponding sound source as a robot's noise in spite of sound source localization results.
3) To provide faster convergence of the separation matrix, we add a new function to our new GSS module which

can import initial $\boldsymbol{W}$ from a separation matrix file on initialization. If we can prepare a good separation matrix in advance, the matrix can be given as initial $\boldsymbol{W}$. We also add a separation matrix export function to generate the separation matrix file. When we have a converged separation matrix as the separation matrix file, the error of initial $\boldsymbol{W}$ will be smaller.

4) To separate moving sound sources, the criteria and timing of the separation matrix update are controllable in our new implementation. We can select either direction-based initialization or speaker-ID-based initialization. As other techniques, we are trying to add two new features to GSS, that is, adaptive step-size control that provides faster convergence of the separation matrix [13] and Optima Controlled Recursive Average [14] that controls window size adaptively for better separation. We are testing these features and have some promising results [15]. They will be included in a future HARK release.

C. Issues in ASR

We have a problem regarding acoustic feature extraction, and we have another room to improve ASR performance in terms of acoustic model.

1) We use a Mel-Scale Log-Spectrum (MSLS) feature as an acoustic feature. We showed that this acoustic feature is more noise-robust than a commonly-used MFCC feature when we used it with sound source separation. Our acoustic feature consists of a 24-dim MSLS feature, and a 24-dim Δ MSLS feature. The total dimension of the acoustic model is 48. This may be too many because a MFCC-based acoustic feature usually has 25-27 dimensions. In addition, it is well-known that Δ power feature improves noise-robustness, but we did not use it.
2) We used only a clean acoustic model so far, while ASR basically has better performance with a noise-adapted acoustic model.

D. Approaches in ASR

1) We propose a new 27-dim acoustic feature which consists of 13 MSLS, 13 Δ MSLS and Δ power. To realize this, we add new HARK modules called FeatureRemover and DeltaPowerMask.
2) We trained a noise-adapted acoustic model for ASR, and try a combination of separation, MFT and ASR with the noise-adapted acoustic model.

IV. IMPLEMENTATION OF HARK

HARK works on middle ware named Flowdesigner which is OSS. Flowdesigner is a data flow oriented development environment. It can be used to build an application such as robot audition system by combining small, reusable building blocks, named modules. An application is described by some modules and arcs for connecting between two modules.

TABLE I
MODULES PROVIDED BY HARK 1.0.0

Category Name	Module Name
Multi-channel Audio I/O	AudioStreamFromMic AudioStreamFromWave SaveRawPCM
Sound Source Localization and Tracking	LocalizeMUSIC ConstantLocalization SourceTracker DisplayLocalization SaveSourceLocation LoadSourceLocation SourceIntervalExtender
Sound Source Separation	DSBeamformer GSS Postfilter BGNEstimator
Acoustic Feature Extraction	MelFilterBank MFCCExtraction MSLSExtraction SpectralMeanNormalization Delta FeatureRemover PreEmphasis SaveFeatures
Automatic Missing Feature Mask Generation	MFMGeneration DeltaMask DeltaPowerMask
ASR Interface	SpeechRecognitionClient SpeechRecognitionSMNClient
MFT-ASR	Multiband Julius/Julian (non-FlowDesigner module)
Data Conversion and Operation	MultiFFT Synthesize WhiteNoiseAdder ChannelSelector SourceSelectorByDirection SourceSelectorByID MatrixToMap PowerCalcForMap PowerCalcForMatrix

Figure 1 displays an overview of a robot audition system using HARK. The system consists of six part named, Multi-Channel Sound Input, Sound Source Localization & Tracking, Sound Source Separation, Acoustic Feature Extraction, Missing Feature Mask Generation, and ASR Interface. Part of HARK is implemented as modules. HARK modules consists of eight categories. Non-module part of HARK is ASR subsystem. As we can see, it is easy to construct a robot system using HARK by connecting some modules. A new modules can be developed by a user. Some modules may be combined to compose a specific function. Table I shows the module list provided in HARK 1.0.0. Main part of improvement of HARK 1.0.0 is sound source separation modules.

Each modules has custumizable property. Property values can be changed if you need. Figure 3 shows property of LocalizeMUSIC module. For example, A_MATRIX property represents a file name of transfer function between microphones and sound sources. MIN_DEG and MAX_DEG properties represent direction range of localization.

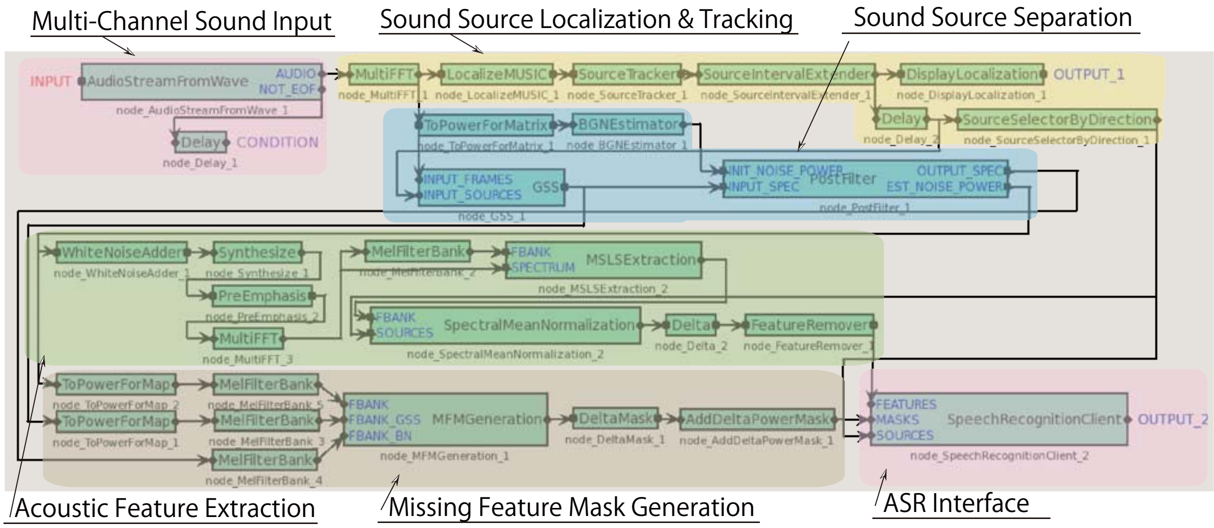


Fig. 1. An example of a robot audition system using HARK.

V. EVALUATION

We conducted two experiments for comparing new features of HARK 1.0.0 with features of HARK 0.1.7. For the fist problem, we conducted an experiment for GSS with mesured transfer function (HARK 1.0.0) and with calculated transfer function from locations of microphones and sound sources, that is conventional GSS (HARK 0.1.7).

A. Experiment 1

An evaluation task is a simultaneous speech recognition experiment by three males, signified as “m101”, “m102”, and “m103”. Simultaneous speech signal by three talkers is highly interfered. Therefore a separation matrix has to be estimated accurately to achieve high separation performance. We considered that a separation matrix was estimated more accurate based on a measured transfer function than based on calculated transfer function.

Figure 4 shows a robot and three talkers in virtual space. Instead of talking three talkers at once to a robot, each speech signals were convoluted impulse response corresponds to eight microphones and sound source. Then the mixed speech signals composed of eight tracks were localized, separated, and recognized. Distance between each sound sources and the robot is 100 centimeters. In HARK 0.1.7, impulse response was calculated from microphone locations. In HARK 1.0.0, impulse response was measured by a humanoid robot HRP-2 whose body shows in Figure 5. HRP-2 has eight microphones in his head as shown in Figure 6.

We used Mel-scale logarithmic spectrum (MSLS) base acoustic feature. The acoustic feature vector is composed of 27 spectral-related acoustic features, i.e., mean normalized MSLS 13 spectral features and 13 differential features, and delta logarithmic power. Analysis frame length and frame shift length were 25 ms and 10 ms.

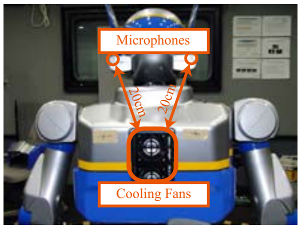


Fig. 2. Cooling fans on HRP-2 and the nearest microphones from them.

TABLE II

ACOUSTIC MODEL

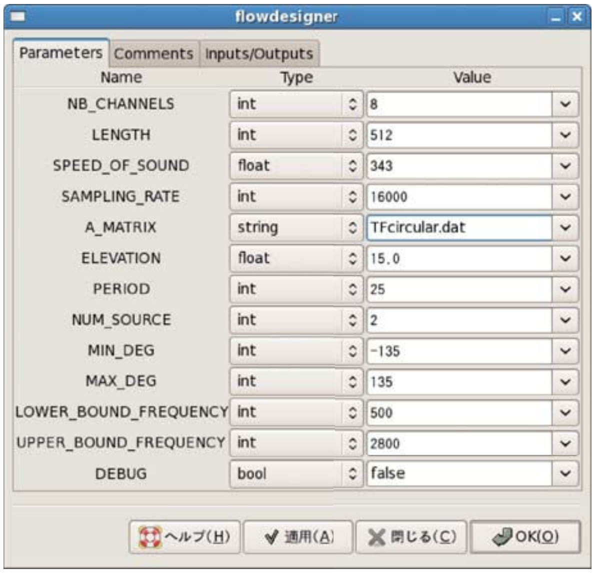
HMM Type	Triphone HMM
Num. of mixture	4
States	3 stats left-to-right model
Num. of states	2000
Training data	Phonetically balanced speech signals 15,370 sentences (Japanese News Article Speech database)
Data format	16 kHz (sampling rate) 16 bit Linear PCM

Hidden Markov Model base ASR is used. A training condition of the HMM is detailed in Table II. ASR is based on missing-feature theory. When acoustic logarithmic likelihood is caluculated, unreliable acoustic feature is masked generated from MFMGeneration.

Test speech signals were constructed from phonetically balanced words in Advanced Telecommunications Research Institute International (ATR). Test set includes 200 isolated words from 3 talkers.

Separated speech was recognized based on Julius 3.5, which is a HMM base speech recognition engine. It is also OSS. We modified it to support a missing-feature theory base recognition.

Figure 7 shows the word correct rates (WCR). Solid, dotted and dashed lines show WCR of center, left and right talkers. Red and blue lines show WCR using HARK 1.0.0 and HARK 0.1.7, that is using measured transfer function and calculated transfer function which is calculated from



Name	Type	Value
NB_CHANNELS	int	8
LENGTH	int	512
SPEED_OF_SOUND	float	343
SAMPLING_RATE	int	16000
A_MATRIX	string	TFcircular.dat
ELEVATION	float	15.0
PERIOD	int	25
NUM_SOURCE	int	2
MIN_DEG	int	-135
MAX_DEG	int	135
LOWER_BOUND_FREQUENCY	int	500
UPPER_BOUND_FREQUENCY	int	2800
DEBUG	bool	false

Fig. 3. A window for property setting in LocalizeMUSIC.

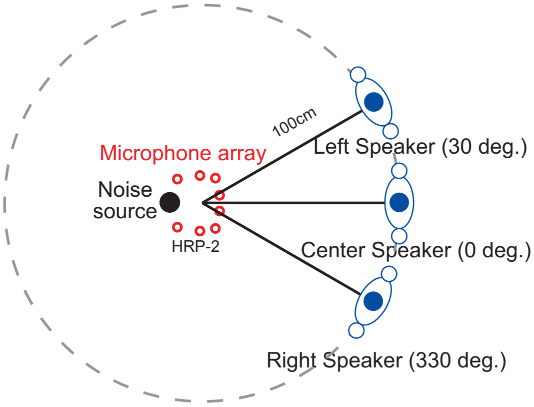


Fig. 4. HRP-2 and three talkers.

microphone positions. The horizontal and vertical axes show angles between talkers and WCR. These results show that measured transfer function improved WCR. For center talker, WCR is improved about 5 points. For peripheral talkers, WCR is improved about 10 to 30 points. As focusing on center talker, angles between talkers using HARK 1.0.0 base system to achieve the same WCR is narrower than using HARK 0.1.7.

B. Experiment 2

We evaluated effectiveness of noise source removing. An evaluation task is a simultaneous speech recognition experiment by three males when a noise source from a fixed direction exists. HRP-2 has fans on his back. Figure 2 shows two fans on HRP-2. The nearest microphones from a hole for air flows is apart from 20 cm. This fan noise achieves at 50–60 dBA. Therefore HARK 0.1.7 localizes the fan noise source as well as other sound sources. GSS in HARK 1.0.0 can be ignore the noise source among other sources. Figure 3 shows virtual noise source as black point.

Fig. 5. HRP-2 Humanoid.

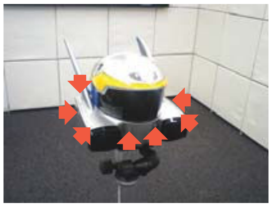


Fig. 6. HRP-2 head mock-up.

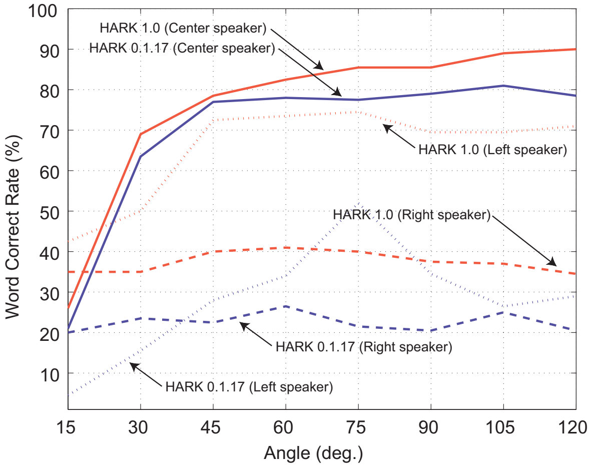


Fig. 7. Comparison simultaneous speech recognition system between old HARK 1.0.0 HARK 0.1.7.

For this experiment, an acoustic model was trained. Multi-condition training was applied to train the acoustic model. The model is more robust than an acoustic model trained from clean speech database. Other experimental conditions were the same as the experiment 1. First, clean model was trained. Second, some parameters in a robot audition system mare tuned to maximise word correct rate. A clean model is used to recognize. Third, speech source separation for all speech signals in JNAS by single talker was applied. Finally, HMM was trained from clean and separated speech signal. By a multi-condition training, a robust acoustic model for separation distortion is obtained.

Figure 8 shows experimental result. Almost all WCR of HARK 1.0.0 outperform that of HARK 0.1.7. Center talker's WCR of HARK 1.0.0 increases for angle between talkers. In contrast, that of HARK 0.1.7 shows a dip around 45 degree. Around 75 degree, both WCRs of HARK 1.0.0 and 0.1.7 forms dip. This may cause by head shape of HRP-2, that is two fins.

VI. CONCLUSIONS AND FUTURE WORK

We developed a new GSS which incorporates a measured robot's HRTF, a fixed-noise-removing-function, and an adaptive-step size. When a robot's HRTF has strong resonance for indirect-wave components, a measured HRTF base GSS is more effective than the original GSS. For a fixed noise source, such as cooling fans, a fixed-noise-removing-function enables the robot to neglect the noise. By neglecting

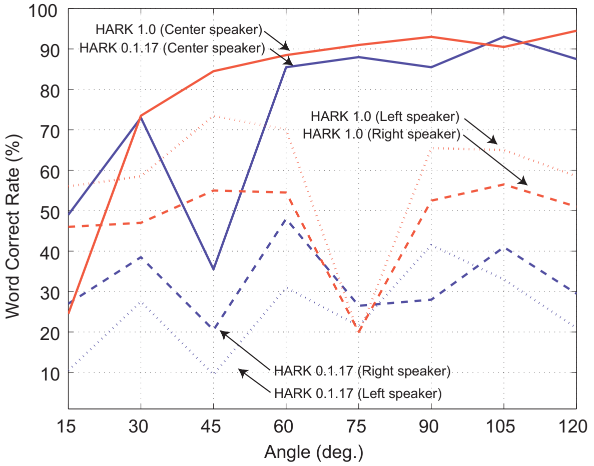


Fig. 8. Comparison simultaneous speech recognition system between old HARK and new HARK with directional noise generated from HRP-2 cooling fans.

it, error that the robot localizes the fixed noise source never has been happened.

Experimental results prove that a WCR of separated speech improves by using our GSS which incorporates a measured robot's HRTF instead of the original GSS. For a center talker, a WCR improves about 5 points. For peripheral talkers, WCRs improve about 10 to 30 points. By a fixed-noise-removing-function, WCR of separated speech improves. Almost all WCRs of HARK 1.0.0 outperform those of HARK 0.1.7. For a center talker, WCRs of HARK 1.0.0 increase for angles between talkers. In contrast, that of HARK 0.1.7 shows a dip around 45 degrees. Around 75 degrees, both WCRs of HARK 1.0.0 and 0.1.7 form dips. Two fins installed on a HRP-2 may cause the dips.

We have released a new version of robot audition software HARK. One year has been passed since we released a first version. For one year, we developed many modules and improved some modules for the new version. This paper describes some of the modules are evaluated. A robot audition system is possible to construct from only HARK. This means that all robot audition researchers who want to use HAKR have to do is to sign up a HARK license. An old version of HARK depends on other modules distributed from a different developer.

We are going to develop an automatic system turning method to optimize many parameters in a robot audition system. The parameters relate each other, and a WCR is nonlinear with respect to one specific parameter. Thus, the optimization is based on empirical knowledge. By providing a turning method, HARK will be rapid prototyping system for designing a robot audition system.

VII. ACKNOWLEDGMENTS

Our research is partially supported by the Grant-in-Aid for Scientific Research and Global COE Program.

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] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*. AAAI, 2000, pp. 832–839.

[3] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid (hrp-2)," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*. IEEE, 2004, pp. 2404–2410.

[4] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Communication*, vol. 44, pp. 97–112, 2004.

[5] S. Yamamoto, J.-M. Valin, K. Nakadai, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 1489–1494.

[6] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2004, pp. 2133–2128.

[7] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems Journal*, vol. 55, no. 3, pp. 216–228, 2007.

[8] F. Michaud, C. Côté, D. Létourneau, J.-M. Valin, E. Beaudry, C. Räievsky, A. Ponchon, P. Moisan, P. Lepage, Y. Morin, F. Gagnon, P. Giguére, M.-A. Roux, S. Caron, P. Frenette, and F. Kabanza, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 742–752, 2007.

[9] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H. G. Okuno, "Real-time robot audition system that recognizes simultaneous speech in the real world," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*. IEEE, 2006, pp. 5333–5338.

[10] H.-D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Human tracking system integrating sound and face localization using em algorithm in real environments," *Advanced Robotics*, vol. 23, no. 6, pp. 629–653, 2007.

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

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

[13] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Adaptive step-size parameter control for real-world blind source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 149–152.

[14] ——, "High performance sound source separation adaptable to environmental changes for robot audition," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2008)*, 2008, pp. 2165–2171.

[15] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, 2009, pp. 3685–3688.