The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Missing-Feature-Theory-based Robust Simultaneous Speech Recognition System with Non-clean Speech Acoustic Model

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

***Abstract*— A humanoid robot must recognize a target speech signal while people around the robot chat with them in real-world. To recognize the target speech signal, robot has to separate the target speech signal among other speech signals and recognize the separated speech signal. As separated signal includes distortion, automatic speech recognition (ASR) performance degrades. To avoid the degradation, we trained an acoustic model from non-clean speech signals to adapt acoustic feature of distorted signal and adding white noise to separated speech signal before extracting acoustic feature. The issues are (1) To determine optimal noise level to add the training speech signals, and (2) To determine optimal noise level to add the separated signal.**

In this paper, we investigate how much noises should be added to clean speech data for training and how speech recognition performance improves for different positions of three talkers with soft masking. Experimental results show that the best performance is obtained by adding white noises of 30 dB. The ASR with the acoustic model outperforms with ASR with the clean acoustic model by 4 points.

I. INTRODUCTION

Robust automatic speech recognition (ASR) is important for efficient and friendly human computer interaction (HCI). There are many aspects for accurate ASR. We focus on speech robustness, that is, robustness against interfering non-target speech signals. The reason of focusing the issue is a robot must recognize the target speech signal while people around the robot chat with them. Several talkers may simultaneously speak to a robot capable of audition. If it is possible for a robot to recognize all speech signals simultaneously, the robot can understand who is speaking to the robot and who is NOT speaking to the robot. In [1] , it is shown that human can understand two or three simultaneous speech signals. The robot capable of audition can provide clues for the dialogue manager, which can determine a target speech signal from the recognition results and smoothly maintain each talker's utterance in the dialogue history. Otherwise, the dialogue manager has to determine which talker is the target talker by means of other ways such as using dialogue history. Then, it determines the target talker and separates his/her utterance from mixed speech signals. The problem with simultaneous listening in robot audition is to recognize speech against other interfering speech signals. Such speech interference is stronger when a robot interacts with multiple talkers.

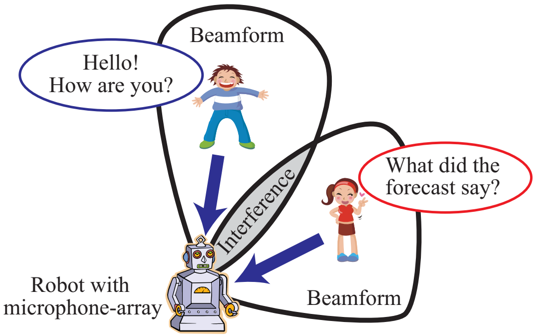

Fig. 1. Simultaneous speech recognition and speech inteference

We have developed the HRI-JP Audition for Robots with Kyoto University (HARK) [2], simultaneous speech recognition system based on the missing-feature theory (MFT) [3], [4]. ASR system based on HARK can recognize simultaneous speech signals. HARK can separate mixed speech signal into each speech signal because components of mixed speech are received from different directions. It is possible by steering the directivity of microphone array to the localized sound direction. In this process, each speech signal is distorted as a side-effect of sound source separation. As distorted acoustic features are mismatched with clean acoustic model, strongly distorted time-frequency position is estimated and input speech feature is masked at the position. Simple masking is a hard mask, that is, either 0 or 1 for estimating the reliability of each acoustic feature of separated speech signals. We attained further ASR improvement by using soft masks. The soft mask is represented as a continuous value between 0 and 1 [5]. Additional improvement was attained by using the acoustic model of which parameters were trained by adding white noises at the signal-to-noise ratio (SNR) of 30-40 dB to the Japanese Newspaper Article Sentences (JNAS) clean speech corpus attains speech robustness.

In this paper we investigated how much white noises should be added to clean speech data for acoustic model training to improve word correct rate when there are three talkers. The rest of the paper is organized as follows. In section 2, we describe our simultaneous speech recognition system. In section 3, the experiment setup and results are shown. In section 4, we discuss the experimental results, and finally conclude the paper.

II. SIMULTANEOUS SPEECH RECOGNITION SYSTEM

The system is developed to recognize simultaneous speech signals by multi-talker. The general architecture in Fig. 2

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

K. Nakadai is with Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan nakadai@jp.honda-ri.com

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

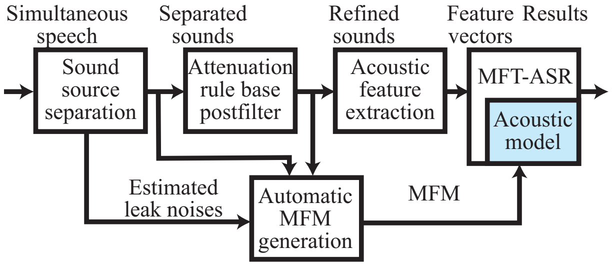

Fig. 2. MFM-ASR system overview.

consists of five components:

1. Sound Source Separation,
2. Attenuation rule base postfilter,
3. Acoustic feature extraction,
4. MFT-ASR, and
5. Automatic missing-feature mask (MFM) generation.

A. *Multi-talker interference*

To capture simultaneous speech by multi-talker, our system uses an 8-channel microphone array. After source localization, geometric source separation is applied to emphasize each source. In this process, one source is interfered with others. When talkers close each other, interference increases. There is distortion in separated speech feature shown in Fig. 3 and 4. P1 and P9 conditions correspond to three talker aligned 10 and 90 degree apart, respectively. Detailed condition describes in experiment section. These features contain separated distortion compared to clean speech feature in Fig. 5. P1 separated distortion level is higher than P9 separated distortion level because distance between talkers in P1 condition are closer than distance between talkers in P9 condition. The separated distortion is multi-talker interference. Multi-talker interference causes temporal variation of acoustic feature. Recognition results are affected by the temporal variation. Strongly affected acoustic features are unreliable. This is because automatic generated MFM is introduced after sound source separation. By masking unreliable acoustic feature, it is possible to avoid getting high likelihood of the input feature with unreliable feature by chance.

MFMs are automatically generated from estimated leak noises, separated sounds, and refined sounds which are filtered by attenuation rule base postfilter. Acoustic feature vector is extracted from refined sounds. Additional spectral feature distortion is added to speech by postfiltering. This is because noise added acoustic model is used in ASR. The acoustic feature vector is recognized by hidden Marcov model base recognizer.

B. *Acoustic model mismatch*

A problem is mismatch between separated speech feature and acoustic model, when likelihood from separated speech feature and acoustic model is calculated. Input speech feature is distorted by sound source separation. It cannot be matched with clear speech model. Separated speech feature is matched with non-clean speech model. In addition, reliable feature is more matched with clearer speech than unreliable feature. If noise level of acoustic model is determined by reliable feature, contribution of unreliable feature to total likelihood decreases.

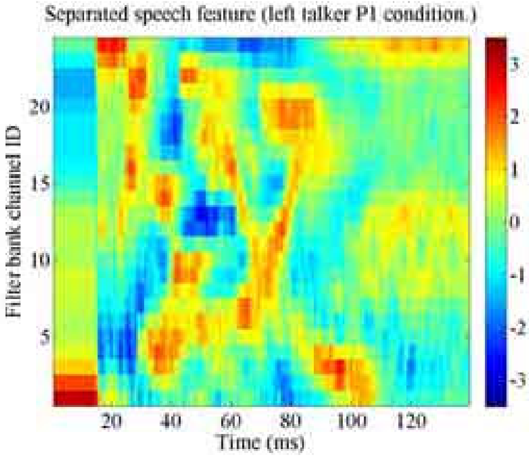

Fig. 3. Separated speech feature (P1 condition).

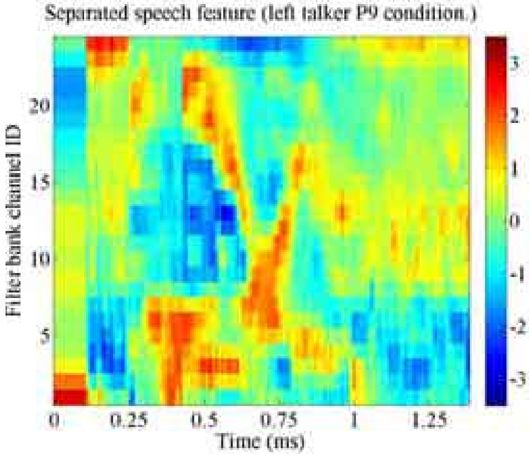

Fig. 4. Separated speech feature (P9 condition).

We investigated that the robustness of ASR is improved by using an acoustic model trained by speeches with added white noises. It is possible to defuse mismatch between acoustic feature of separated sounds and acoustic model. It is reported that the addition of a colored noises is effective for noise-robust ASR [6]. We tried to add white noises because characteristics of temporal variation are unknown in advance. MFT-ASR results using acoustic models trained from various SNR speeches and MFMs are compared in this paper.

C. *MFT-ASR*

In conventional ASR systems, estimation of a path with maximum likelihood is based on state transition and output probabilities in the hidden Markov model (HMM). An output probability estimation process is modified in the MFT-ASR system as follows: let $M = [M(1), \cdots M(F)]$ be an MFM vector and $M(f)$ represent the reliability of the f-th acoustic feature. The output probability $b_j(x)$ is given by

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{\sum_{f=1}^{F} M(f) \log g(x(f)|l, S_j)\right\}, \tag{1}$$

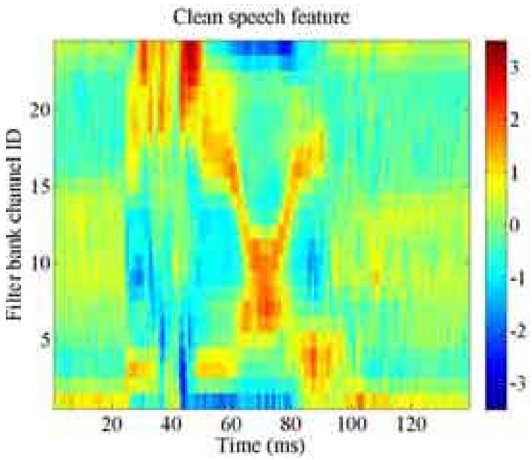

Fig. 5. Clean speech feature.

where $P(\cdot)$ is a probability operator, $\boldsymbol{x} = [x(1), \cdots, x(F)]$ is an acoustic feature vector, F is the size of the acoustic feature vector, S_j is the j-th state, and $g(x(f)|S_j)$ is a mixture of Gaussian distribution in j-th state. This definition is natural extension of output probability because the equation of output probability is equivalent to the equation of conventional output probability when all mask values are one. If reliability is not available, all mask values are one. Note that this output probability definition is formed when all off-diagonal elements of covariance matrix in the output probability is zero.

III. Acoustic feature and MFM

A word correct rate of ASR is improved by using a MFM. That kind of ASR is called MFT-ASR. The mask corresponds to a reliability of acoustic feature. MFT-ASR can cover feature mismatch between input acoustic feature and an acoustic model with the masks. When the reliability of acoustic features is known in advance, the only reliable acoustic features without the unreliable acoustic features are used to recognize. If the unreliable acoustic features are regarded as reliable, those features cause degradation of a word correct rate.

An acoustic feature vector is calculated in the MFTASR component before being recognized by the MFT-ASR system. An acoustic feature vector consists of static and dynamic features. The static acoustic feature is extracted from separated speech and includes separation noises. We used Mel-scale logarithmic spectrum (MSLS) [9] as an acoustic feature although the Mel-frequency ceptral coefficient (MFCC) is commonly used because spectral distortion in band limited frequency is confined to a limited order of MSLS coefficients. For the MFCC, the distortion spreads over all MFCC coefficients by the discrete cosine transform of the spectral parameter.

MSLS coefficient vector based on N channel filter banks is defined as

$$\boldsymbol{p}(t) \quad = \quad [p(1,t), p(2,t), ..., p(N,t)]. \tag{2}$$

This also represents static acoustic feature. Dynamic acoustic feature is defined as

$$\begin{aligned} \delta\boldsymbol{p}(t) &= [\delta p(1,t), \delta p(2,t), ..., \delta p(N,t)], \\ &= \frac{\sum_{k=-2}^{2} k\boldsymbol{p}(t+k)}{\sum_{k=-2}^{2} k^2}. \end{aligned} \tag{3}$$

The acoustic feature vector is defined as

$$\begin{aligned} \boldsymbol{x}(t) = \quad & [p(1,t), p(2,t), ..., p(N,t), \\ & \delta p(1,t), \delta p(2,t), ..., \delta p(N,t)]. \end{aligned} \tag{4}$$

$$\tag{5}$$

MFM vector corresponds to acoustic feature vector.

$$\begin{aligned} \boldsymbol{M}(t) = \quad & [M(1,t), M(2,t), ..., M(N,t), \\ & \delta M(1,t), \delta M(2,t), ..., \delta M(N,t)]. \end{aligned} \tag{6}$$

$$\tag{7}$$

Hard [7] and soft mask [5], [8] generation methods were developed to generate such reliabilities. Hard mask represents binary status as 0 and 1. Soft mask represents values between 0 and 1. Reliability of acoustic feature is represented as a continuous value. We revealed that the MFT based ASR with soft masking outperforms with hard masking (0 or 1) by 5 points. In this paper, the soft mask is used for MFT-ASR. The hard mask is calculated from the reliability r using a mapping function.

$$M(f,t) \quad = \quad \begin{cases} 1 & r > \theta_{hard} \\ 0 & r \leq \theta_{hard} \end{cases}, \tag{8}$$

The soft mask generation is described in [5]. We briefly described the soft mask in this subsection. The soft mask is calculated from the reliability r using a mapping function. The function is defined as a sigmoid function which has three tunable parameter, i.e., weight w, tilt k, and threshold θ_{soft}.

$$\begin{aligned} & M(f,t) = \\ & g(r(f,t)|w, k, \theta_{soft}) = \\ & \begin{cases} \dfrac{w}{1+\exp(-k(r(f,t)-\theta))}, & r(f,t) > \theta_{soft} \\ 0, & r(f,t) \leq \theta_{soft} \end{cases} \end{aligned} ,(9)$$

where $0.0 \leq r \leq 1.0$. f and t are frequency and time, respectively. The reliability r is determined in a postfilter processing.

$$r(f,t) \quad = \quad \frac{\hat{S}_m(f,t) + B(f,t)}{Y_m(f,t)}, \tag{10}$$

A block diagram of GSS with postfiltering is shown in Fig 6. $y_m(f,t)$, $\hat{s}_m(f,t)$, and $b(f,t)$ are the input, the output, and the estimated background noise. There parameters are calculated from the multi-channel input speech with object related transfer function (ORTF). The variables filtered by the Mel filter bank are $Y_m(f,t)$, $\hat{S}_m(f,t)$, and $B(f,t)$, respectively. The key idea of determining the reliability is that frequency bands where spectral shape is refined by the postfilter are unreliable and frequency bands where spectral shape is not refined by the postfilter are reliable.

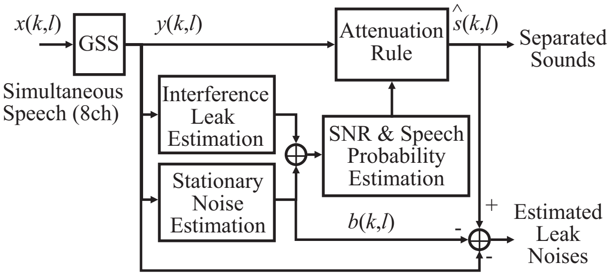
x(k,l)
GSS
y(k,l)
Attenuation Rule
ŝ(k,l)
Separated Sounds
Simultaneous Speech (8ch)
Interference Leak Estimation
SNR & Speech Probability Estimation
Stationary Noise Estimation
b(k,l)
+
-
-
Estimated Leak Noises

Fig. 6. GSS with postfiltering.

TABLE I
NINE LOUD SPEAKER LAYOUT PATTERNS

Pattern	f101	m101	m102
P1.	-10 degree	0 degree	10 degree
P2.	-20 degree	0 degree	20 degree
P3.	-30 degree	0 degree	30 degree
P4.	-40 degree	0 degree	40 degree
P5.	-50 degree	0 degree	50 degree
P6.	-60 degree	0 degree	60 degree
P7.	-70 degree	0 degree	70 degree
P8.	-80 degree	0 degree	80 degree
P9.	-90 degree	0 degree	90 degree

IV. EXPERIMENT

To determine the best level adding white noise to training speeches for acoustic model, six types of acoustic models were compared. An evaluation of simultaneous speech recognition was conducted. Word correct rates were calculated using HMM recognition system with soft MFMs based on Julius [10], [11].

A. Experimental setup

Six types of acoustic models based on the HMM were trained from phonetically balanced speeches of JNAS which is sampled at 16 kHz. Each acoustic model was trained from clean and noisy speeches. Noise models were trained from speeches with added white noise at 0 dB, 10 dB, 20 dB, 30 dB, and 40 dB levels. We call these models C, N_0, N_{10}, N_{20}, N_{30}, and N_{40} Parameters of three states of a triphone HMM were trained from the speeches. The total number of states was about 2000 by state sharing.

Instead of talking three talkers at once to a robot, recorded speeches are played through three loud speakers. Three talkers are two male and one female signified as “m101”, “m102”, and “f101”. We installed our system into a robot (SIG2) is placed in the center of a circle. SIG2 has eight microphone on its body as is shown in Fig. 7. Loud speakers were placed in a circle with a radius of 200 centimeters. One of the talkers was placed in front of our system, and the others were placed on both sides of the robot. Figure 8 and 9 shows the layout of our system and the loud speakers. The word correct rates for nine loud speaker layout patterns (P1,P2,...,P9) were compared. All patterns are detailed in table I.

We used Mel-scale logarithmic spectrum (MSLS) [9] base acoustic feature. The acoustic feature vector is composed of 48 spectral-related acoustic features, i.e., mean normalized MSLS 24 spectral features and 24 differential features. Analysis frame length and frame shift length were 25 ms and 10 ms, respectively.

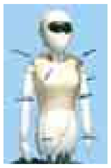

Fig. 7. Eight microphones on the robot SIG2.

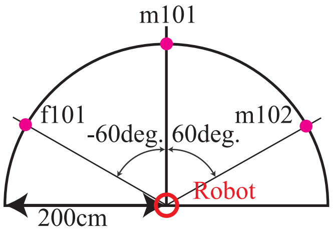

Fig. 8. Pettern 6 (P6) : A position of robot and loud speakers .

The soft masks were generated using developed method described in [5]. The tunable parameters $\{w, k, \theta_{soft}\} = \{$ 0.6, 140, 0.3 $\}$ in Eq. (9) for static feature part of acoustic feature vector, i.e., elements of feature vector from 1st to 24th, were used. The parameters $\{w, k, \theta_{soft}\} = \{$ 1.0, 140, 0.3 $\}$ for differential feature part of acoustic feature vector, i.e., elements of feature vector from 25st to 48th, were used. These parameters are numerically optimized. In preliminarily experiments, appropriate w, k, and, θ_{soft} are from 0.3 to 0.6, around 140, and from 0.2 to 0.4.

For speech spectral features, static feature variance is generally wider than dynamic feature variance. For average vector, likelihood for a model with wider variance is smaller than one for a model with tighter variance. Therefore static features are weighted to equalize contribution for total likelihood.

B. Experimental results

First, relationship between model noise level and input speech noise level is shown. C, N_0, N_{10}, N_{20}, N_{30}, and N_{40} for data set T_c. All models were evaluated using six data sets named T_c, T_0, T_{10}, T_{20}, T_{30}, and T_{40}. The data set T_c consisted fo clean speeches. The others consisted of speeches which white nose were added into. Subscript represents input noise level in dB which added to clean speeches.

The average word correct rate contour maps is shown in Fig. 10. Vertical axis shows that input noise level added to clean speech in dB. Horizontal axis shows that model noise level added to clean speeches for training speech database. Word correct rate is high for one noise level model when input noise level is 10 dB higher than model noise level.

Second, all models were evaluated using two data sets named T_c and T_n. The former consisted of clean speeches and the latter consisted of speeches separated from mixed speech based on geometric source separation (GSS) [12]. These speeches were constructed from phonetically balanced words in Advanced Telecommunications Research

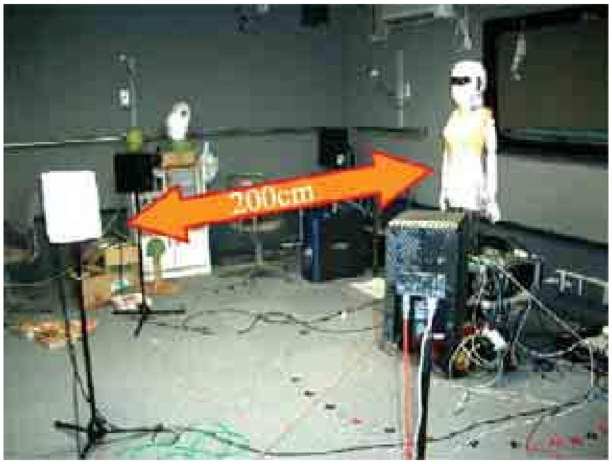

Fig. 9. Humanoid robot SIG2 and loud speaker location.

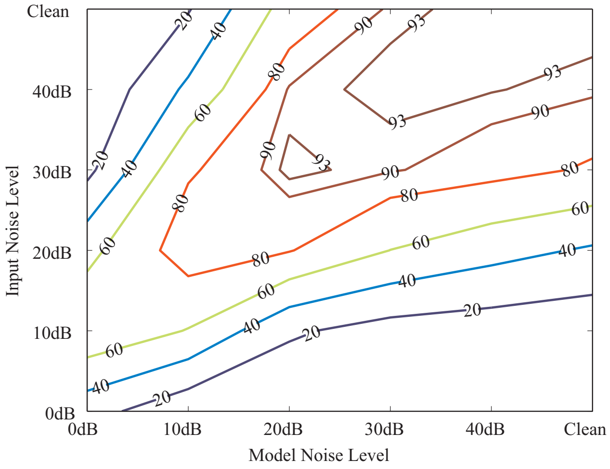

Fig. 10. Word correct rate for noisy models.

Institute International (ATR) speech database. T_c included 200 isolated words from 25 talkers (male: 12, female: 13). Tn included 200 isolated words from 3 talker's (male: 2, female:1).

Figure 11, 12, 13 show word correct rates with N_0, N_{10}, N_{20}, N_{30},N_{40} and C for data set T_n. A soft MFM was applied [5] to calculate the word correct rate. The horizontal and vertical axes show model IDs and word correct rates, respectively.

V. DISCUSSION

A. Word correct rate of cleen speech

Figure 10 shows that the word correct rate decreases when the model noise level is higher and the word correct rate decreases when the input speech noise level is higher. These mean that the training speech has to be clean if the test speech is clean. As the test speech is not clean in real-world recognition, acoustic model parameters should not be trained from clean speeches. The reason of the decrease in the word correct rate shown in fig. 10 is the mismatch of acoustic features between training and testing speeches. The test speech is non-clean speech and the type of its distortion is unknown. If the type of distortion is known in advance, the model should be trained from speeches with distortions. We added white noise into training speeches.

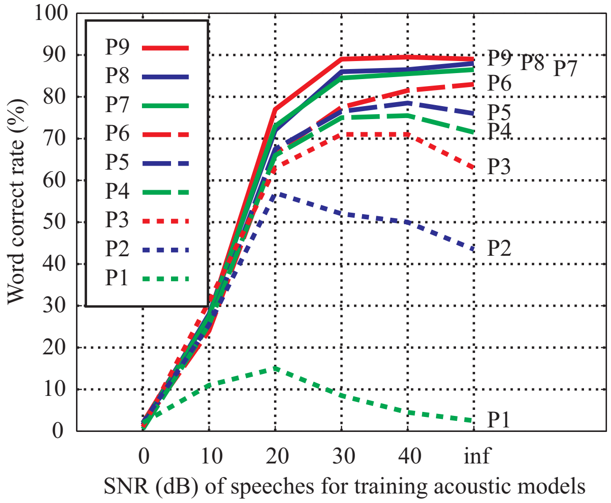

Fig. 11. Separated speech WCR based on MFT-ASR for center talker

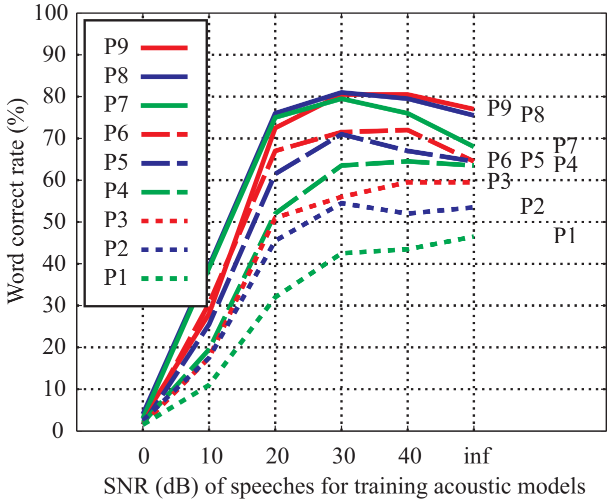

Fig. 12. Separated speech WCR based on MFT-ASR for right talker

B. Word correct rate of separated speech

In Figs. 11, 12, 13, there are peaks of the word correct rate between SNR 20–40dB. In general, speech separation based on geometric source separation is difficult when the distance between talkers is short because space sparseness is assumed to be separated.

Since smaller angle means that interference between talkers has increased, it becomes more difficult to separate the target speech from a mixed speech signal. This trend clearly appears in the center and right talkers. As the acoustic feature of a separated speech signal is much different from a clean speech signal, separated speech is non-clean speech. Therefore a peak of the separated speech correct rate appears speech around N_{30} and N_{40}. To obtain a better separated correct rate, an acoustic model should be trained from non-clean speeches.

When the distance is great (like P7, P8, P9 conditions), the word correct rates are relatively high because the speech interference between talkers is smaller. In these conditions, the acoustic feature of separated speech is similar to that of

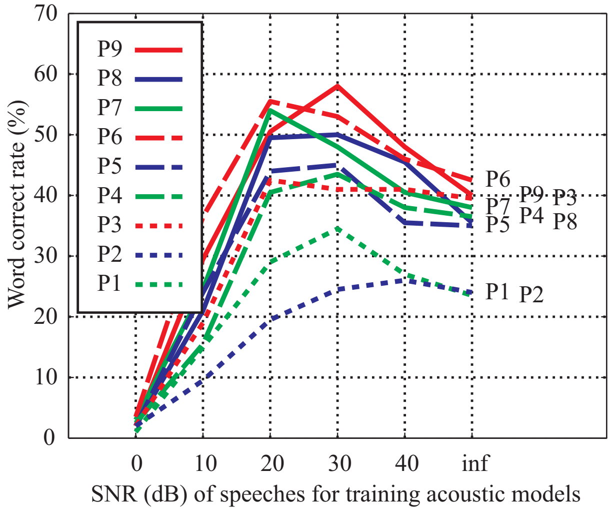

Fig. 13. Separated speech WCR rate based on MFT-ASR for left talker

N_{40} or C speech because separation of mixed speech works well. When the distance is short (like P1, P2, P3 conditions), the word correct rates are relatively low. In these conditions, the acoustic feature of separated speech is similar to that of N_{20} or N_{30} speech as it is difficult to separate. From these results, we conclude that it is possible to choose the best acoustic model among N_{20}, N_{30} and N_{40} according to the talkers position, assuming unknown spectral distortion caused by separation to be the added white noise. Talker position is obtained from the GSS process.

To compare the word correct rate of acoustic model trained with clean speech, the word correct rates are improved in P1,P2,....,P5 (more than 10 % in P1, P2), as shown in Fig. 11. The word correct rate was relatively low in P1 because interference of non-target speech from both sides was strong. When the angles between the target talker and non-target talkers were over 20 degrees, the interference softened. In P6, P7, P8, and P9, the acoustic feature of separated speech was enough to be similar to the clean acoustic feature because the distance between talkers is great and SIG2 has high separability to the center talker.

In Fig. 12, the word correct rates improved in P6, P7, P8, and P9 (more than 20 % in P8). In contrast, the word correct rates of the right and left talkers improved at higher SNR. The acoustic feature of separated speech was different from the clean acoustic feature because SIG2 has middle separability to peripheral talkers although distance between talkers was great.

For the center talker, the acoustic models should be selected according to talker angles. When the angle is small and large, the acoustic model is trained from SNR 20-dB and SNR 40-dB speeches, respectively. For the peripheral talkers, we can always obtain high recognition performance by using an acoustic model trained from SNR 30-dB speeches. Average word correct rate with acoustic model trained from SNR 30-dB speeches is improved by 4 points compared to acoustic model trained from clean speeches.

VI. Conclusion

We conducted simultaneous speech recognition experiments using six types of acoustic models. Each model is trained from speeches with added white noises of different SNR. A soft MFM is also applied based on the MFT. From the experiments, we found that

(1) for the center talker, acoustic models trained from SNR 20-dB, 30-dB, and 40-dB, and speech should be used if the angle between the center talker and peripheral talkers is 10-20, 30-40, or 50 degrees.

(2) for peripheral talkers, acoustic model trained from SNR 30-dB speech should be used.

(3) Average word correct rate with acoustic model trained from SNR 30-dB is improved by 4 points compared to acoustic model trained from clean.

We used single condition acoustic model, but it also can be improve using multi-condition acoustic model. It is necessary to improve the word correct rate for selecting a dialogue strategy. In future work, we will compare our model to multi-condition model and develop another soft MFM generation method. Since it is difficult to prepare various SNR acoustic models with training from speech database, we will also develop a method for converting clean acoustic models into arbitrary SNR acoustic models.

VII. Acknowledgments

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

References

[1] M. Kashino, *et al.*, "One, two, many - judging the number of concurrent talkers," *Journal of Acoustic Society of America*, vol. 99, no. 4, pp.Pt2, 2569, 1996, ASA.

[2] *http://winnie.kuis.kyoto-u.ac.jp/HARK/*

[3] B. Raj, *et al.*, "Missing-Feature Approaches in Speech Recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp.101–116, 2005, IEEE.

[4] S. Yamamoto, *et al.*, "Design Automatic Speech Recognition and Understanding," *Proc. of ASRU 2007*, pp.111–116, 2007, IEEE.

[5] T. Takahashi, *et al.*, "Soft Missing-Feature Mask Generation for Simultaneous Speech Recognition System in Robots," *Proc. of Interspeech 2008*, pp.992–995, 2008, ISCA.

[6] S. Yamada, *et al.*, "Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments," *Proc. of Eurospeech 2003*, pp.1493–1496, 2003, ISCA.

[7] S. Yamamoto, *et al.*, "Genetic Algorithm-Based Improvement of Robot Hearing Capabilities in Separating and Recognizing Simultaneous Speech Signals," *Proc. of IEA/AIE 2006 / LNSA 4031*, pp.207–217, 2006, AAAI.

[8] M. L. Seltzer, *et al.*, "A Bayesian framework for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp.379–393, 2004, ISCA.

[9] S. Yamamoto, *et al.*, "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory," *Proc. of ICRA 2005*, pp.1489–1494, 2005, IEEE.

[10] Y. Nishimura, *et al.*, "Noise-robust speech recognition using multiband spectral features," *Proc. of 148th ASA Meetings*, no. 1aSC7, 2004, ASA.

[11] T. Kawahara, *et al.*, "Free software toolkit for Japanese large vocabulary continuous speech recognition," *Proc. of ICSLP 2000*, vol. 4, pp.476–479, 2000, ISCA.

[12] L. C. Parra, *et al.*, "Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp.352–362, 2002, IEEE.