PAPER *Special Section on Statistical Modeling for Speech Processing*

ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles

Shigeki MATSUDA[†a)], **Takatoshi JITSUHIRO**[†], ***Members***, **Konstantin MARKOV**[†], ***Nonmember***, ***and*** **Satoshi NAKAMURA**[†], ***Member***

SUMMARY In this paper, we describe a parallel decoding-based ASR system developed of ATR that is robust to noise type, SNR and speaking style. It is difficult to recognize speech affected by various factors, especially when an ASR system contains only a single acoustic model. One solution is to employ multiple acoustic models, one model for each different condition. Even though the robustness of each acoustic model is limited, the whole ASR system can handle various conditions appropriately. In our system, there are two recognition sub-systems which use different features such as MFCC and Differential MFCC (DMFCC). Each sub-system has several acoustic models depending on SNR, speaker gender and speaking style, and during recognition each acoustic model is adapted by fast noise adaptation. From each sub-system, one hypothesis is selected based on posterior probability. The final recognition result is obtained by combining the best hypotheses from the two sub-systems. On the AURORA-2J task used widely for the evaluation of noise robustness, our system achieved higher recognition performance than a system which contains only a single model. Also, our system was tested using normal and hyper-articulated speech contaminated by several background noises, and exhibited high robustness to noise and speaking styles.

key words: *automatic speech recognition, parallel decoding, multiple acoustic models, fast noise adaptation, speaking style, hyper-articulated speech*

1. Introduction

In a real environment, there is a wide variety of noises such as engine noise from automobiles, babble noise in convention halls, street traffic noise, etc. Moreover, natural speech exhibits various speaking styles such as fast and slow utterances, hyper-articulation and whispering. Therefore, it is important to have a system that can handle such a wide variety of noises and speaking styles.

To date, many techniques have been proposed that improve noise robustness [1]. In the field of speech enhancement, Spectrum Subtraction (SS) [2] and two-stage mel-scaled Wiener-filtering [3] have been proposed. RASTA processing [4] and Cepstrum Mean Normalization (CMN) have also been developed as noise-robust feature extraction techniques. In addition, Parallel Model Combination (PMC) [5] and Maximum Likelihood Linear Regression (MLLR) [6] have been proposed to adapt models to a particular noise environment. Multi-condition training is widely used for generating a model robust to noise type and level. To deal with variations in speaking style, there are some techniques for robust recognition of speech distorted by the Lombard effect [7], for hyper-articulated speech [8] and fast spontaneous speech [9], [10].

These techniques can be classified into three typical categories. Acoustic modeling techniques for improving robustness are classified into the first category as shown in Fig. 1 (1). The multi-condition training belongs to the first category, where a single acoustic model is estimated with a large database which contains several environments including different noise types, noise levels and speaking styles. An acoustic model estimated by this method can robustly recognize speech uttered in an environment which is included in the training data. However, an acoustic model trained with data collected in a specific environment can accurately recognize speech uttered in the same environment (matched condition), and achieves higher recognition performance than a model estimated by the multi-condition training. Therefore, for a single model, there is a trade-off between robustness and accuracy. As principly shown in Fig. 2, the variety of speaking environments for which a single model can have high performance is limited.

The speech enhancement and the noise-robust feature extraction methods belong to the second category, where noisy speech is mapped into a specific environment such as clean speech. Figure 1 (2) conceptually illustrates mappings from a noisy environment to a clean one. If the specific environment is modeled completely by an acoustic model, the model can accurately recognize the speech from that environment. SS and Wiener-filtering are used to reduce background noise. The types of environment which can be mapped depend on individual techniques applied in this category.

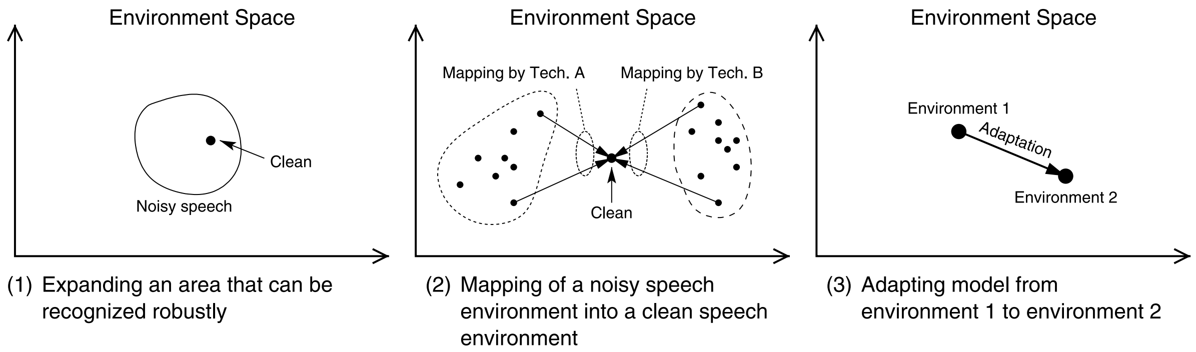
The model adaptation methods and techniques for dealing with variable speaking styles belong to the third category. They are used to transform the model to a different speaking environment. Figure 1 (3) depicts conversion from an acoustic model for environment 1 to environment 2. The area of the speaking environment space that can be transformed generally depends on the acoustic model. One approach to extend the area that can be adapted is to prepare multiple acoustic models depending on different environments, and then to perform model adaptation of the model that is close to the input speech.

In this paper, we describe an ASR system developed by ATR which is based on parallel decoding with improved robustness to noise, SNR and speaking styles. Our sys-

Manuscript received July 11, 2005.
Manuscript revised September 30, 2005.

[†]The authors are with ATR Spoken Language Communication Research Laboratories, Kyoto-fu, 619–0288 Japan.
a) E-mail: shigeki.matsuda@atr.jp

DOI: 10.1093/ietisy/e89–d.3.989



(1) Expanding an area that can be recognized robustly

(2) Mapping of a noisy speech environment into a clean speech environment

(3) Adapting model from environment 1 to environment 2

Fig. 1 Classification of techniques for improving robustness.

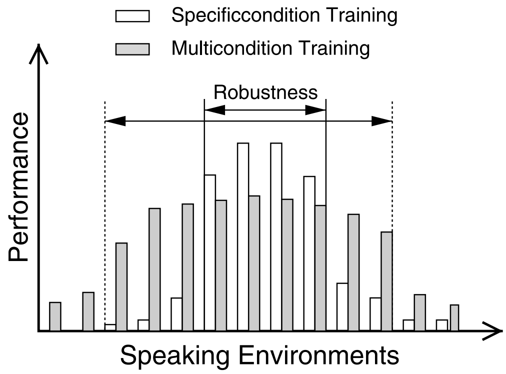


Fig. 2 Performance of models trained with speech data from single and multiple environments.

tem works with multiple acoustic models. A wide variety of speaking environments are handled by using many acoustic models, which are trained and constructed for each specific environment (noise type, SNR, speaker gender and speaking style) using different acoustic features. For each test utterance, noise-adapted acoustic models are generated, then multiple hypotheses obtained from these acoustic models are selected and combined by hypothesis selection and hypothesis combination [11] based on lattice reconstruction. To adapt acoustic models to noisy environments, we proposed a new fast noise adaptation technique using noise GMMs that can be applied to any acoustic features, not only MFCC without Cepstrum Mean Subtraction (CMS). Moreover, we proposed a new hypothesis combination technique using Generalized Word Posterior Probability (GWPP). Even though the robustness of each single acoustic model is limited, an ASR system with many specific models can handle various conditions.

We demonstrated that a parallel decoding-based ASR system can achieve higher performance, than in a single decoding-based ASR system, even though speech is contaminated and distorted by both noise and speaking styles.

In Sect. 2, we describe techniques used in our ASR system. In Sect. 3, we evaluate our system's noise robustness. In Sect. 4, in addition to noise, the system's robustness to speaking style is evaluated. Conclusions are drawn in Sect. 5.

2. System Description

2.1 Fast Noise Adaptation

For fast noise adaptation, we propose a Gaussian Mixture Model (GMM)-based technique. The technique consists of two steps. First, noise GMMs and noise-dependent speech HMMs are prepared using various types of noise in advance. During recognition, given the first 500 ms of speech data, weights to individual noise-dependent GMMs are estimated using the Expectation Maximization (EM) algorithm. Then, one HMM is composed from the noise-dependent HMMs using those estimated weights. This procedure is illustrated in Fig. 3, where, $P(\boldsymbol{x}|s_{n,i})$ is the state output probability of the i-th state in the n-th noisy speech HMM and w_n is the weight for the n-th noise GMM. $P(\boldsymbol{x}|\hat{s}_i)$ is calculated as sum of these state output probabilities as follows:

$$P(\boldsymbol{x}|\hat{s}_i) = \sum_{n=1}^{N} w_n P(\boldsymbol{x}|s_{n,i}), \tag{1}$$

where N is number of noisy speech HMMs.

State transition probabilities in the noisy speech HMM are calculated using state durations obtained by using the following equation:

$$l_{n,i} = \frac{1}{1 - a_{n,i}}, \tag{2}$$

where $a_{n,i}$ and $l_{n,i}$ are respectively the self-state transition probability and the state duration of the i-th state in the n-th noisy speech HMM. Then, the state transition probability $\hat{a}_i$ in a noise-adapted HMM is calculated by using these state duration times as follows:

$$\hat{l}_i = \frac{1}{N} \sum_{n=1}^{N} l_{n,i}, \tag{3}$$

$$\hat{a}_i = \frac{\hat{l}_i - 1}{\hat{l}_i}, \tag{4}$$

Therefore, state output distributions are adapted using weights for noise GMMs; however, these transition probabilities are not adapted because the average duration of all noisy speech HMMs is used.

This technique belongs to the third category in Fig. 1. This approach is similar to the HMM composition-based

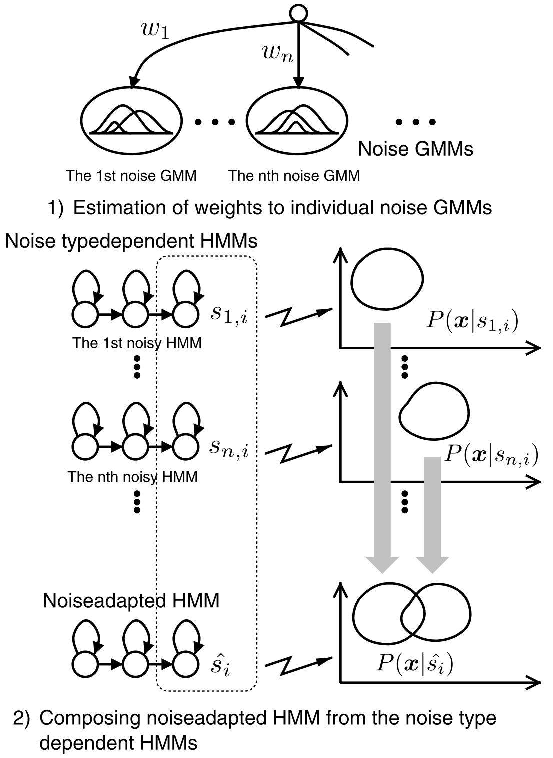

w_1
w_n
The 1st noise GMM
The nth noise GMM
Noise GMMs
1) Estimation of weights to individual noise GMMs
Noise typedependent HMMs
$s_{1,i}$
The 1st noisy HMM
$P(\boldsymbol{x}|s_{1,i})$
$s_{n,i}$
The nth noisy HMM
$P(\boldsymbol{x}|s_{n,i})$
Noiseadapted HMM
$\hat{s}_i$
$P(\boldsymbol{x}|\hat{s}_i)$
2) Composing noiseadapted HMM from the noise type dependent HMMs

Fig. 3 Generation of noise-adapted HMM.

technique [13]. This technique cannot be applied to an acoustic feature other than an MFCC feature, since it is based on the PMC method. Our technique, however, can be applied to any feature such as an MFCC feature normalized by CMS and DMFCC.

In the previous ATR system presented in [11], MLLR online adaptation is used for adaptation of noise type, speaker, and channel characteristic. A model is adapted using an utterance, which consists of noise and noisy speech periods. Then the utterance is decoded again using the adapted model. On the other hand, fast noise adaptation can generate an adapted model using only the noise period. The fast noise adaptation technique can generate a noise-adapted model before the decoding process. Therefore, it is more suitable for realtime applications.

2.2 Hypothesis Selection

Here we implement the hypothesis selection technique based on posterior probability. The hypothesis with the highest score is selected from multiple hypotheses as follows:

$$\hat{k} = \operatorname*{argmax}_{k=1}^{K} H_k, \tag{5}$$

where H_k is the score of a hypothesis obtained from the k-th decoder. K denotes the number of decoders. A hypothesis obtained from the $\hat{k}$-th decoder has the highest score, which is defined as the sum of the log acoustic model likelihood and the log language model probability of a hypothesis as follows:

$$H = \log P(\boldsymbol{X}|\boldsymbol{W}) + \lambda \log P(\boldsymbol{W}), \tag{6}$$

where $\boldsymbol{X}$, $\boldsymbol{W}$, and H are respectively an observed feature vector sequence, a hypothesis represented by a word sequence, and the score for the hypothesis. $\log P(\boldsymbol{X}|\boldsymbol{W})$ and $\log P(\boldsymbol{W})$ denote respectively a log acoustic model likelihood and a log language model probability. λ denotes a language model weight used during the decoding process.

This technique is used for expanding an area that can be recognized robustly and accurately. Therefore, this technique belongs to the first category in Fig. 1. Even though the robustness of each acoustic model is limited, an area that can be recognized robustly is expanded effectively when the technique can select an appropriate hypothesis.

Note that a period of silence basically does not depend on the SNR level, speaker gender or speaking style. In hypothesis selection, we experimentally confirmed that an incorrect hypothesis is often selected using the likelihoods of silence models estimated from different speech databases at a very significant rate. In the experiments described later, we estimated a common silence model to avoid this performance degradation.

2.3 Differential MFCC

Our previous research showed that some modifications to the MFCC algorithm can yield better performance in noisy speech conditions [12]. The so-called differential spectrum MFCC is calculated from the differential power spectrum of speech, which is defined as:

$$D(i,k) = |Y(i,k) - Y(i,k+1)|, \tag{7}$$

where $D(i,k)$ is the differential spectrum, $Y(i,k)$ is the power spectrum for the i-th frame and k is the spectrum bin index. This technique belongs to the second category in Fig. 1. We denote this type of differential spectrum MFCC feature as DMFCC.

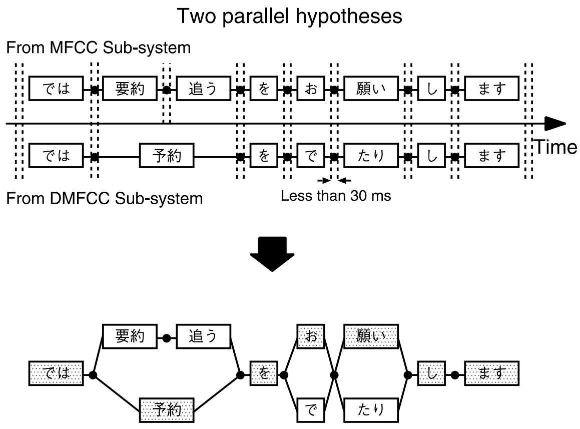
2.4 Hypothesis Combination

In our system, we propose a hypothesis combination technique based on GWPP (Generalized Word Posterior Probability) [14], which is used as a confidence measure. This technique combines multiple hypotheses obtained from different decoders, and if these hypotheses are complementary to each other, it is possible to obtain a more correct result. Figure 4 depicts an example of hypothesis combination. Our system uses this technique to combine hypotheses from the two different sub-systems.

Initially, each word in each of the hypotheses is represented by an arc in the graph, where the weighted confidence score of that word is associated with the arc. The score A_n for n-th word is defined as:

$$A_n = T_n \log C_n, \tag{8}$$

where C_n and T_n are respectively the GWPP and duration of the n-th word in a hypothesis. In the next step, all arcs representing identical words hypothesized between the same time instants are collapsed into a single arc. Finally, nodes



Creating a word lattice from hypotheses. The best path is shown with gray boxes.

Fig. 4 Examples of a hypothesis combination.

are formed between all arc pairs where the word-end time of one arc and the word start time of the next arc are within 30 ms of each other.

After the word graph is constructed in this manner, weighted confidence score and language model probability for each word are combined to form the word score. Finally, a DP search is performed to find the best scoring path through the graph.

This technique can also be used for expanding an area that can be recognized robustly and accurately. Therefore, this technique belongs to the first category in Fig. 1.

The previous ATR system presented [11] includes the likelihood-based hypothesis combination technique that uses a normalized acoustic likelihood obtained from the log acoustic model likelihood $P(\boldsymbol{X}|\boldsymbol{W})$. On the other hand, GWPP is the approximation of posterior probability $P(\boldsymbol{W}|\boldsymbol{X})$. In this paper, we evaluate the improvement in both the GWPP-based and likelihood-based hypothesis combination techniques.

2.5 Acoustic Model for Hyper-Articulated Speech

When using real ASR system, if a recognition error occurs, the user has to repeat the last utterance. Okuda et al. [8] reported that a short pause is usually inserted after vowels in the repeated utterances, and consequently the ASR performance degrades. To recognize such utterances robustly, they proposed a new acoustic model which allows short-pause insertion after vowels for hyper-articulated speech. The structure of this acoustic model is illustrated in Fig. 5. Our system employs such an acoustic model for recognition of hyper-articulated utterances such as repeated speech. This technique belongs to the first category.

3. Evaluation of Robustness to Noise

3.1 Experimental Conditions

Our system was tested on the AURORA-2J task [15]. This database consists of Japanese connected digit corpus for training and testing. The ATRASR version 3.3 developed by ATR Spoken Language Communication Laboratories was used as a decoder. The clean-training set in the AURORA-2J was employed for estimating acoustic models. There is a total of 8,440 utterances by 110 speakers (55 male and 55 female). The training set was contaminated by four types of noise as shown in Table 1 at four types of SNR (20, 15, 10 and 5 dB). Digits and silence were modeled by different HMMs as shown in Table 2, with the total number of acoustic models being 2 features × 4 SNRs × 4 noises = 32. For each noise type, a noise GMM with eight Gaussian distributions was estimated.

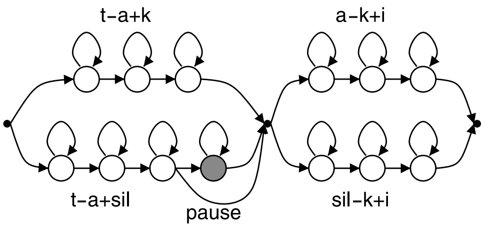


Fig. 5 Structure of an acoustic model for hyper-articulated speech.

Table 1 Noise types used for the AURORA-2J task.

Training	Restaurant, Street, Airport, Station
Testing	Subway, Babble, Car, Exhibition

Table 2 HMMs for the AURORA-2J task.

	Number of states	Number of mixtures
digits	16	20
silence	3	36
short pause	1	36

For a baseline system, we prepared an acoustic model trained using a multi-condition training set, in which four types of noise are added to the clean speech at five types of SNR (Clean, 20, 15, 10 and 5 dB). Each noise and SNR condition included 422 utterances. The test set-B in the AURORA-2J was used for testing. In this set, speech utterances were contaminated by four types of noise, which were different from the noises used for the training, at different SNRs ranging from 0 dB to 20 dB as shown in Table 1. The G.712 filter was applied to all training and testing speech data. The feature vector consists of 12 MFCCs, Δ pow, 12 ΔMFCCs, ΔΔ pow and 12 ΔΔMFCCs calculated with a 10-ms frame period and 25-ms frame length. The DMFCC features have the same structure as the MFCC features, and the Cepstrum Mean Subtraction (CMS) was applied to both types of feature. We denote these features as MFCC_CMS and DMFCC_CMS, respectively. In addition, before feature extraction, we applied a two-stage Wiener-filtering: the AFE (ES 202 050 front-end) [3] distributed by ETSI. We denote MFCC and DMFCC features extracted from noise-suppressed speech by the AFE as MFCC_AFE and DMFCC_AFE, respectively.

In [15], AURORA-2J results were reported where multi-conditional trained models with MFCC_AFE features are obtained using the HTK toolkit [16]. In order to be able to compare our system with that from [15], we also trained a simple model with the multi-condition data and MFCC_AFE features. The ATRASR achieved an 89.09% average word accuracy, which is comparable with the 88.98% reported in [15].

3.2 System Structure

To investigate the effect of the parallel-decoding based speech recognition, we evaluated the performance of several ASR systems as follows.

System A) Multi-conditional trained AM

In this system, a single acoustic model estimated with the multi-condition training set is used only, thus this system is based on single-decoding. MFCC_CMS is used as the acoustic feature.

System B) SNR-independent AM

Also, this system is based on single-decoding, having only a single acoustic model composed from 16 noise type and SNR-dependent models described in Sect. 3.1. Each state of the model consists of mixture components of individual environment-dependent models, and has 20 mixture × 4 SNR levels × 4 noise types = 320 Gaussian distributions.

System C) SNR-dependent AMs

This system contains four decoders with acoustic models for four SNR levels, thus there are four recognizers. Each SNR-dependent model is composed from models which depend on four noise types for each SNR level. Each state has 20 mixture × 4 noise types = 80 distributions. The hypothesis with the highest score is selected as the final result from the hypotheses of these models.

System D) Noise-adapted SNR-dependent AMs

This system contains the four SNR-dependent acoustic models of the system C which are adapted using the fast 500 ms of speech data by the fast noise adaptation.

System E) Overall system

The overall system consists of two sub-systems, one for MFCC_CMS and one for DMFCC_CMS as shown in Fig. 6. Each sub-system has four SNR-dependent models, which are adapted with the first 500 ms of speech data. Each sub-system outputs a hypothesis obtained by noise-adapted SNR-dependent models. The final result is obtained by the hypothesis combination technique.

As an another combination method, system E without noise adaptation, the system consists of system C using MFCC and system C using DMFCC. However, a situation where an ASR system without noise adaptation is used is not realistic, because speech is inevitably contaminated by background noise in real environment. Therefore, the fast noise adaptation technique is essential for ASR systems in real environments.

3.3 Evaluation

Figure 7 shows the average word accuracies of several systems. The performance of system A, which has a multi-conditional trained AM, is very similar to system B, which has an SNR-independent AM. System C, which is based on the parallel-decoding using SNR-dependent AMs, reduces the errors by 14.7% compared with system A. It is clear that the recognition accuracy is improved by parallel-decoding using multiple models. Moreover, system D, which has noise-adapted models, reduces the error by 22.3% compared to system A. Clearly, then the accuracy of each models is improved by fast noise adaptation.

Figure 8 shows the performances of individual SNR-dependent models in system D and hypothesis selection technique. We can see that a model depending on a SNR level which is close to that of input speech achieves the best performance. The hypothesis selection performance is almost equal to the best performance, meaning that it can se-

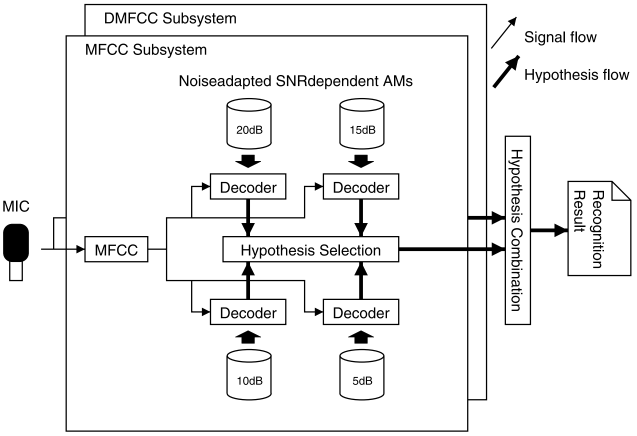


Fig. 6 Structure of the ASR system used for noise robustness experiments.

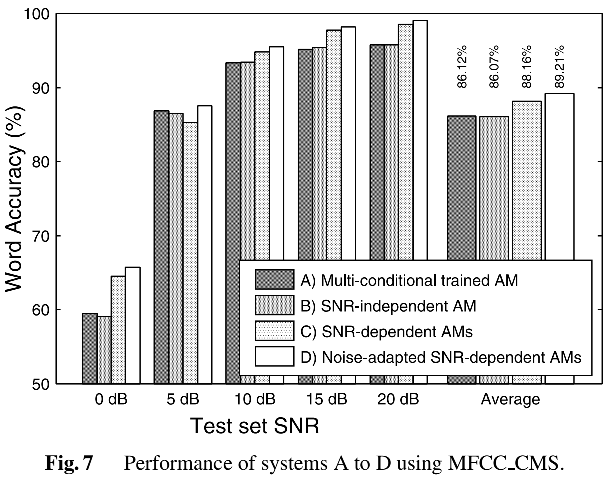


Fig. 7 Performance of systems A to D using MFCC_CMS.

lect a best hypothesis effectively.

We evaluated the performance of the overall system (System E), with Fig. 9 showing performances. By combining hypotheses from two sub-systems, our system performed better than systems A to D. Our system reduces the errors by 27.2% in comparison to system A. Furthermore, our system achieved higher performance than a system containing an MFCC_AFE acoustic model trained with a multi-condition training set. Clearly, performance is improved by applying the hypothesis combination to hypotheses from multiple features.

We compared the performances of both the GWPP-based and the likelihood-based hypothesis combination as shown in Table 3. In the evaluation of ASR systems that use acoustic features normalized by CMS, the performance of the likelihood-based technique was the same as the DMFCC sub-system. On the other hand, the GWPP-based technique could achieve higher performance than the DMFCC sub-system.

Moreover, we applied the AFE as the noise suppression technique to our ASR systems. Figure 10 shows the average word accuracies of several systems using MFCC_AFE. Just like with ASR systems using MFCC_CMS, it is clear that the recognition accuracy is improved by the fast noise adaptation technique, the hypothesis combination and the parallel decoding-based system.

Finally, we evaluated the performance of system E consisting of MFCC_AFE and DMFCC_AFE sub-systems, and the results are shown Fig. 11. As can be seen there, using AFE is effective and reduces the errors by respectively 41.6% and 21.4% compared with system A and a system containing an MFCC_AFE acoustic model trained using the multi-condition training set.

In the evaluation for hypothesis combination techniques, Table 3 shows that the GWPP-based hypothesis combination achieves higher performance than the

Table 3 Word accuracy (%) of both the GWPP-based and the likelihood-based hypothesis combination techniques in the evaluations for robustness to noise.

	+CMS	+AFE
MFCC sub-system	89.21	91.24
DMFCC sub-system	89.38	91.17
GWPP-based H.C.	89.80	91.90
Likelihood-based H.C.	89.38	91.81

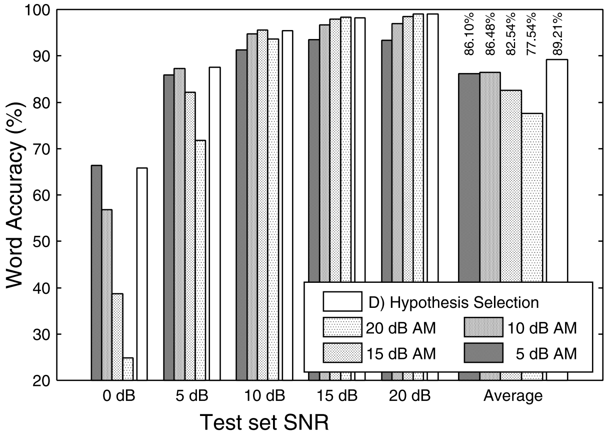


Fig. 8 Performance of individual noise-adapted SNR-dependent AMs and the hypothesis selection in system D using MFCC_CMS.

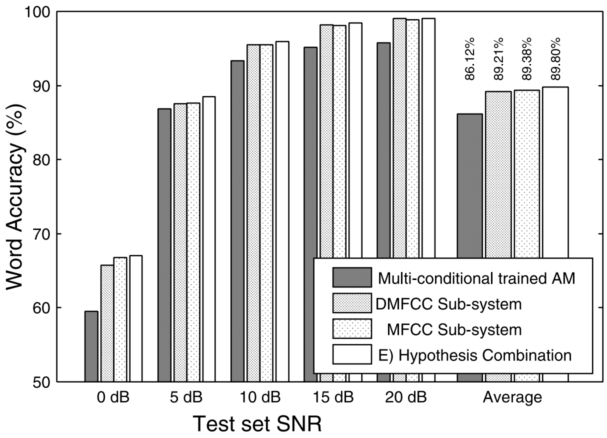


Fig. 9 Performance of the system E, which consists of the MFCC_CMS and the DMFCC_CMS sub-systems.

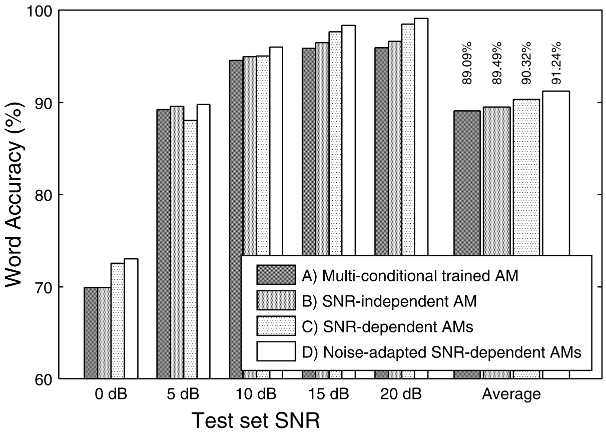


Fig. 10 Performance of systems A to D using MFCC_AFE.

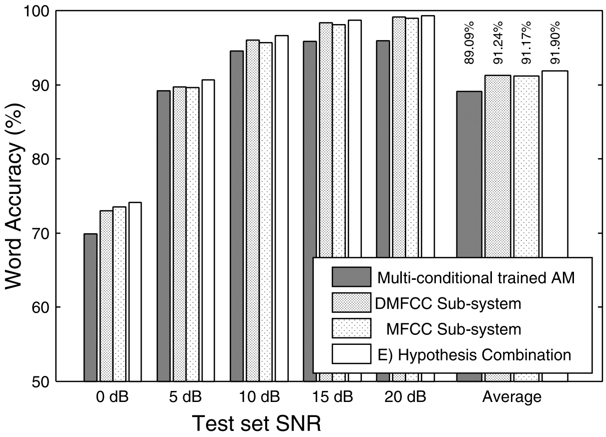


Fig. 11 Performance of the system E, which consists of the MFCC_AFE and the DMFCC_AFE sub-systems.

likelihood-based technique.

4. Evaluation of Robustness to Noise and Speaking Style

4.1 Experimental Conditions

To evaluate of robustness to both noise and speaking style, we tested our system using normal and hyper-articulated speech. Figure 12 shows the structure of the system. It consisted of two sub-systems as before, but there were now six decoders in each sub-system. Noise-dependent acoustic models for fast noise adaptation were trained using dialog speech from the ATR travel arrangement task database (5 hours), read speech of phonetically balanced sentences (25 hours) and 12 types of noise listed in Table 4. A state-tying structure with 2,089 states was generated by using the MDL-SSS technique [17] where each state had five Gaussian components. All acoustic models were trained from data contaminated by different noises and different SNR levels (10, 20 and 30 dB), and hyper-articulated acoustic models were generated from acoustic models of normal speech. Parameters of each distribution were kept the same but the HMM topology was different. Each acoustic model was gender dependent, and the generated acoustic models depended on 3 SNR levels, 12 types of noise, MFCC and DMFCC features, speaker gender and speaking style. Therefore, the total number of acoustic models was $3 \times 12 \times 2 \times 2 \times 2 = 288$. For each noise type, a noise GMM with eight Gaussian distributions was estimated. Noise-adapted acoustic models were generated using the first 500 ms of each test utterance. The MFCC feature consisted of 12 MFCCs, 12 ΔMFCCs and a Δ*pow* extracted with a 10-ms frame period and a 20-ms frame length. The DMFCC feature also had 12 DMFCCs, 12 ΔDMFCCs and a Δ*pow*. CMS was applied to both features.

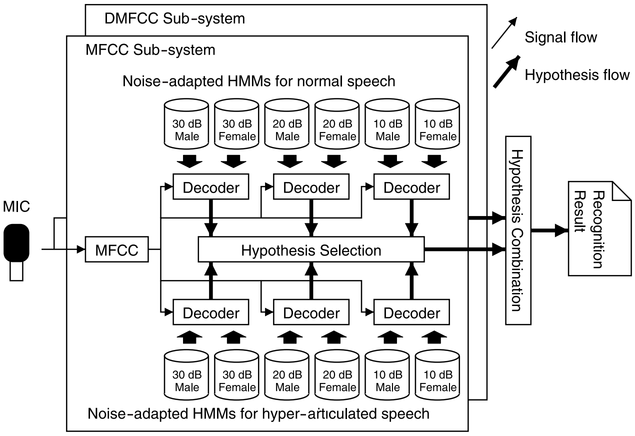


Fig. 12 Structure of the ASR system used for experiments on robustness to noise and speaking style.

Table 4 Noise types used for experiments on robustness to noise and speaking styles.

Training	Airport lobby, Airbus, Underground city, Car driving, Food counter, Square, Station yard Platform at station, High-speed railway, Boiler room, Rice Paddies, Forest
Testing	Public bus, In front of a station, Construction site

Our system uses a multi-class composite word bi-gram [18] and a word tri-gram language model. Each language model is trained from the spontaneous speech database (SDB), language database (LDB) and spoken language database (SLDB) [19], with a total of 6.1 M words. Lexicon size is 34 k words.

For normal speech testing, we used the basic travel expression corpus (BTEC) testset-01 (510 sentences, four males and six females) [20], and for hyper-articulated speech testing, we collected 40 syllable-stressed sentences (two males and two females). The normal speech for testing was contaminated by three types of noise at four different SNR levels; also, the hyper-articulated speech was contaminated by three types of noise at three different SNR levels, as shown in Table 4.

4.2 Evaluation for Normal Speech

We evaluated the recognition performance of a system which contains acoustic models for normal speech and a system which contains acoustic models for both normal and hyper-articulated speech. All acoustic models in both systems were adapted to the noise environment using the first 500 ms of the input noisy speech. Figure 13 shows the average word accuracies for each of the individual sub-systems and the overall systems. Both of our systems, which have the fast noise adaptation, achieved higher performance than the that which consists of a clean MFCC acoustic model

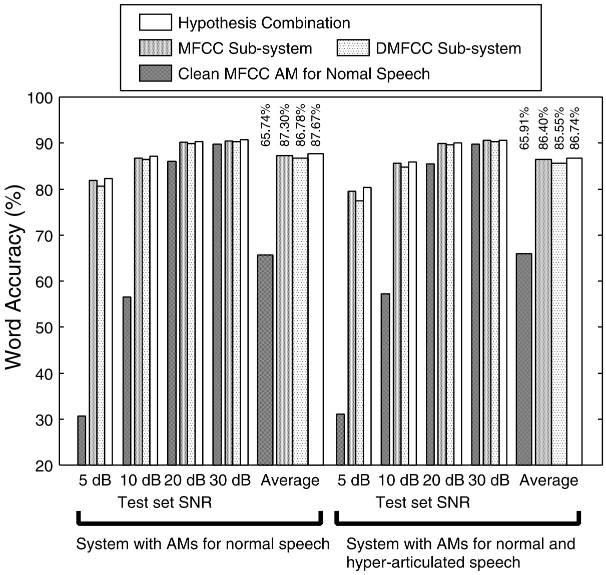


Fig. 13 Performances of the system with acoustic models for normal speech and the system with acoustic models for normal and hyper-articulated speech, for normal speech data contaminated by noise.

for normal speech. Table 5 shows the performances of both the GWPP-based and likelihood-based hypothesis combination techniques. In this table, “normal” means the ASR system including acoustic models for normal speech only, and “both” means the ASR system including acoustic models for normal and hyper-articulated speech. As the table shows, the GWPP-based technique performed better than individual sub-systems, even though the performance of the likelihood-based technique was lower than that of the MFCC sub-system. It is clear that the performance of both systems is similar, suggesting that the integration of models for hyper-articulated speech did not affect system performance.

4.3 Evaluation for Hyper-Articulated Speech

We evaluated the recognition performance of our system using hyper-articulated speech data contaminated by three types of noise. Figure 14 shows the average word accuracy for the evaluation data. Even though the word accuracies of the system with an acoustic model for normal speech only were less than 10%, our system could achieve a word accuracy of about 40%. The performance was improved further by applying the GWPP-based hypothesis combination. Clearly, then our ASR system can handle both normal and hyper-articulated speech contaminated by noise.

Table 5 Word accuracy (%) of both the GWPP-based and the likelihood-based hypothesis combination techniques in the evaluations using normal speech contaminated by noise.

Acoustic models	normal	both
MFCC sub-system	87.30	86.40
DMFCC sub-system	86.78	85.55
GWPP-based H.C.	87.69	86.74
Likelihood-based H.C.	86.84	85.89

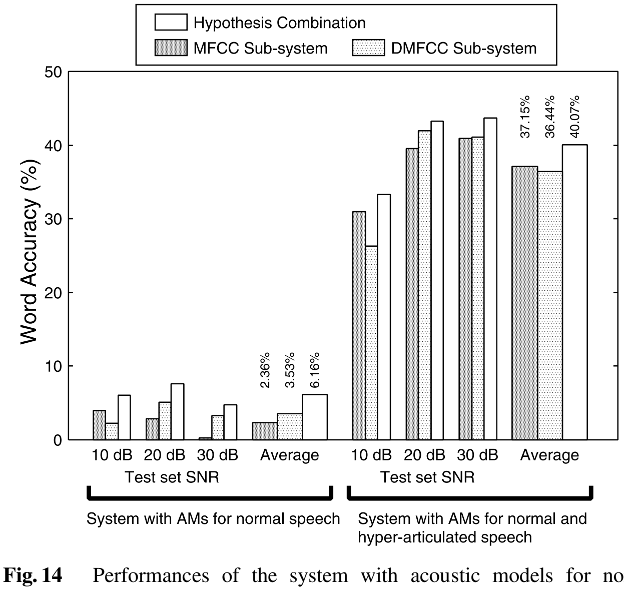


Fig. 14 Performances of the system with acoustic models for normal speech and the system with acoustic models for normal and hyper-articulated speech, for hyper-articulated speech data contaminated by noise.

5. Conclusion

In this paper, we described an ASR system robust to noise, SNR and speaking styles. Our system has multiple acoustic models, each of which depends on the noise, SNR, speaker gender and speaking style. The GMM-based fast noise adaptation technique was used to improve robustness to noise. Also, to improve robustness to hyper-articulated speech, we employed the acoustic model for hyper-articulated speech. In addition, we used two acoustic features as different “views” of the speech signal.

Experimental results of noise-robustness show that both recognition accuracy and noise-robustness can be improved significantly by parallel-decoding using multiple SNR-dependent models which are adapted by fast noise adaptation. Also, we found that the ML-based selection and word-graph hypothesis combination techniques are effective tools for obtaining recognition output from multiple hypotheses. Moreover, on the experiments of speaking style-robustness, our ASR system was able to recognize accurately both normal and hyper-articulated speech contaminated by noise.

Future work will include applying a microphone array that significantly reduces back-ground noise, and developing a method to generate a set of acoustic models that efficiently covers a wide variety of noises and speaking styles.

Acknowledgements

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled ‘A study of speech dialogue translation technology based on a large corpus’.

References

[1] Y. Gong, “Speech recognition in noisy environments: A survey,” Speech Commun., vol.16, no.3, pp.261–291, 1995.

[2] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, pp.113–120, 1979.

[3] ETSI ES 202 050 v1.1.1 Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI, April 2002.

[4] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech Audio Process., vol.2, no.4, pp.587–589, 1994.

[5] M. Gales and S. Young, “Robust continuous speech recognition using parallel model combination,” IEEE Trans. Speech Audio Process., vol.4, no.5, pp.352–359, 1996.

[6] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Comput. Speech Lang., vol.9, pp.171–185, 1995.

[7] J.C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizer,” J. Acoust. Soc. Am., vol.93, pp.510–524, 1993.

[8] K. Okuda, T. Matsui, and S. Nakamura, “Towards the creation of acoustic models for stressed Japanese speech,” Eurospeech2001, vol.3, pp.1653–1656, 2001.

[9] K. Okuda, T. Kawahara, and S. Nakamura, "Speaking rate compensation based on likelihood criterion in acoustic model training and decoding," ICSLP2002, vol.4, pp.2589–2592, 2002.

[10] H. Nanjo, K. Kato, and T. Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," Eurospeech 2001, pp.2531–2534, 2001.

[11] K. Markov, T. Matsui, R. Gruhn, J. Zhang, and S. Nakamura, "Noise and channel distortion robust ASR system for DARPA SPINE2 task," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, March 2003.

[12] J. Chen, K.K. Paliwal, and S. Nakamura, "Cepstrum derived from differentiated power spectrum for robust speech recognition," Speech Commun., vol.41, no.2-3, pp.469–484, 2003.

[13] M. Ida and S. Nakamura, "HMM composition-based rapid model adaptation using a priori noise GMM adaptation evaluation on Aurora2 corpus," ICSLP2002, vol.1, pp.437–440, 2002.

[14] F.K. Soong, W.K. Lo, and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," CD-ROM Proc. SWIM2004, 2004.

[15] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Syst, vol.E88-D, no.3, pp.535–544, March 2005.

[16] S. Young, D. Kershow, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, 2000.

[17] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," IEICE Trans. Inf. & Syst., vol.E87-D, no.8, pp.2121–2129, Aug. 2004.

[18] H. Yamamoto, S. Isogai, and Y. Sagisaka, "Multi-class composite N-gram language model," Speech Commun., vol.41-2003, pp.369–379, Oct. 2003.

[19] T. Takezawa, T. Morimoto, and Y. Sagisaka, "Speech and language databases for speech translation research in ATR," Proc. Oriental COCOSDA Workshop, pp.148–155, 1998.

[20] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," Proc. LREC, vol.I, pp.147–152, 2002.

Shigeki Matsuda received his B.S. degree from the Department of Information Science, Teikyo University, in 1997, completed his doctoral program at the Japan Advanced Institute of Science and Technology in 2003, and joined ATR Spoken Language Communication Research Laboratories as a researcher. He holds a doctoral degree in information science. He is engaged in research on speech recognition, and is a member of the Acoustic Society of Japan and Information Processing Society of Japan (IPSJ).

Takatoshi Jitsuhiro received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Japan, in 1991 and 1993. In 1993, he joined the Human Interface Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Kanagawa, Japan, and began work on speech recognition. From 2000, he has been a researcher at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He received the Ph.D. degree in engineering from Nara Institute of Science and Technology in 2005. His research interests include speech recognition and speech signal processing. He is a member of the Acoustical Society of Japan and IEEE.

Konstantin Markov was born in Sofia, Bulgaria. After graduating with honors from the St. Petersburg Technical University, he worked for several years as a research engineer at the Communication Industry Research Institute in Sofia. He received his M.Sc. and Ph.D. degrees in electrical engineering from Toyohashi University of Technology, Japan, in 1996 and 1999, respectively. In 1998, he received the Best Student Paper Award from the IEICE Society. In 1999, he joined the research development department of ATR, Japan, and in 2000 became an invited researcher at the ATR Spoken Language Communication (SLC) Research Laboratories. Currently, he is a senior research scientist at the Acoustics and Speech Processing Department of ATR SLC. He is a member of ASJ, IEEE and ISCA. His research interests include signal processing, automatic speech recognition, Bayesian networks and statistical pattern recognition.

Satoshi Nakamura received his B.S. degree in electronic engineering from Kyoto Institute of Technology in 1981 and a Ph.D. degree in information science from Kyoto University in 1992. Between 1981–1993, he worked with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986–1989, he worked with ATR Interpreting Telephony Research Laboratories and from 1994–2000, he was an associate professor of the Graduate School of Information science, Nara Institute of Science and Technology, Japan. In 1996, he was a visiting research professor of the CAIP center of Rutgers University of New Jersey USA. He is currently the director at ATR Spoken Language Communication Laboratories, Japan. He also serves as an honorary professor at the University of Karlsruhe, Germany since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya award from the Acoustical Society of Japan in 1992, and the Interaction2001 best paper award from the Information Processing Society of Japan in 2001. He served as an associate editor for the Journal of the IEICE Information in 2000–2002. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society in 2001–2004. He is a member of the Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and IEEE.