Speech Recognition System Robust to Noise and Speaking Styles

Shigeki Matsuda, Takatoshi Jitsuhiro, Konstantin Markov and Satoshi Nakamura

ATR Spoken Language Translation Research Labs
Kyoto, 619-0288, Japan
{shigeki.matsuda, takatoshi.jitsuhiro, konstantin.markov, satoshi.nakamura}@atr.jp

Abstract

It is difficult to recognize speech distorted by various factors, especially when an ASR system contains only a single acoustic model. One solution is to use multiple acoustic models, one model for each different condition. In this paper, we discuss a parallel decoding-based ASR system that is robust to the noise type, SNR, speaker gender and speaking style. Our system consists of two recognition channels based on MFCC and Differential MFCC (DMFCC) features. Each channel has several acoustic models depending on SNR, speaker gender and speaking style, and each acoustic model is adapted by fast noise adaptation. From each channel, one hypothesis is selected based on its likelihood. The final recognition result is obtained by combining hypotheses from the two channels. We evaluate the performance of our system by normal and hyper-articulated test speech data contaminated by various types of noise at different SNR levels. Experiments demonstrate that the system could achieve recognition accuracy in excess of 80% for the normal speaking style data at a SNR of 0 dB. For hyper-articulated speech data, the recognition accuracy improved from about 10% to over 45% compared to a system without acoustic models for hyper-articulated speech.

1. Introduction

In recent years, ASR systems have been used in various applications like location setting for car navigation systems, speech input to word processors, etc. However, in order to achieve high recognition performance, ASR systems are subjected to various constraints on noise environments and speaking styles. 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 utterance, 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 and speech analysis, SS (Spectrum Subtraction) [2] and RASTA processing [3] have been proposed as acoustic feature extraction techniques robust to noise. PMC (Parallel Model Combination) [4] and MLLR (Maximum Likelihood Linear Regression) [5] have been proposed for model adaptation to a particular noise environment.

To deal with speaking style variations, there are some techniques for robust recognition of speech distorted by the Lombard effect [6], for hyper-articulated speech [7], and fast spontaneous speech [8, 9].

Most of these techniques improve robustness to a specific noise environment or a specific speaking style only, if a user speaks in a different environment or style, it is difficult to maintain the recognition performance as in the matched condition. Furthermore, noise and speaking styles change with time in the real environment. To recognize distorted speech, noise and speaking styles have to be predicted for each utterance in advance. In practice, this is difficult to do. Recently, however, a parallel decoding using multiple acoustic and language models has become popular. A final recognition result is obtained by combining multiple hypotheses from multiple decoders using models, trained for different environments. Even though robustness of each acoustic model is limited, an ASR system based on parallel decoding can handle various conditions.

We describe an ASR system based on parallel decoding with improved robustness to both noise and speaking styles. Our system works with multiple acoustic models. Acoustic models are trained for specific environments (noise type and speaking style) using different acoustic features, and multiple hypothesis obtained from these acoustic models are selected and combined by maximum likelihood criteria and hypothesis combination [11].

In Section 2, we describe the structure of our ASR system and each technique used. In Section 3, we evaluate our system's recognition performance evaluation, and conclude in Section 4.

2. System Description

2.1. Differential MFCC

Our previous research showed that some modifications to the MFCC algorithm can yield better performance in noisy speech conditions [10]. The so-called differential spectrum MFCC is calculated from the differential power spectrum of speech, which is defined as:

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

where $D()$ is the differential spectrum, $Y()$ is the power spectrum for the ith frame and k is the spectrum bin index. This simple modification was efficient for the AURORA2 task [12]. We denote this type of differential spectrum MFCC feature as DMFCC.

2.2. Fast Noise Adaptation

For fast noise adaptation, we use HMM a composition-based technique [13]. This technique comprises two steps. In the first step, an initial noise GMM is trained with prior data containing various types of noises. Using this GMM and a clean-speech HMM, noise-dependent HMMs are composed by the PMC technique. In the second step, we do MAP adaptation of the mixture weights of the noise GMM with a small amount of noise test data. Then, one HMM is composed from the noise-dependent HMMs using estimated GMM mixture weights. Figure 1 illustrates this procedure. In this figure, P_{λ_i} is the output distribution of the composed HMM and w_{N_i} is the estimated mixture weight for the ith GMM component.

In our system, since PMC technique cannot be applied to the DMFCC feature, noise-dependent HMMs are trained from data contaminated by different noises and different SNR levels.

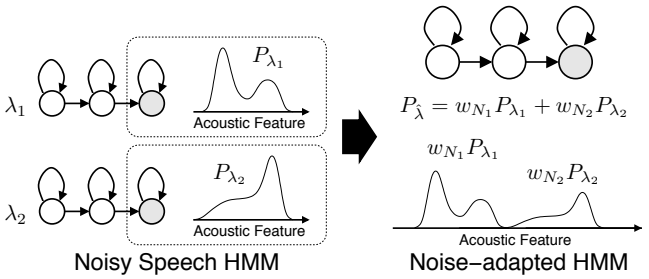

Figure 1: *Generation of noisy speech HMM.*

2.3. Acoustic Model for Hyper-articulated Speech

When using an ASR system, if a recognition error occurs, the user must repeat the last utterance. Okuda et. al. [7] reported that a short pause is usually inserted after vowels in the repeated utterance, and to recognize such an utterance robustly, they proposed a new acoustic model for hyper-articulated speech. The structure of the acoustic model is illustrated in Figure 2. By using this acoustic model, it is possible for our system to recognize hyper-articulated speech such as repeated speech.

2.4. Hypothesis Combination

In our system, we implement the hypothesis combination technique based on word graph construction [11]. This technique combines multiple hypotheses obtained from different decoders. If these hypotheses are complementary to each other, it is possible that a more correct result can be obtained.

Figure 3 depicts an example of hypothesis combination. This technique consists of two steps: in the first step, a word graph is created from two hypothesis, then in the second step, the best path through the graph is selected using word's acoustic and language model scores.

2.5. ASR based on Parallel Decoding

Figure 4 illustrates the structure of our system. It contains two parallel channels, one for each MFCC and DMFCC feature. Parallel decoding is applied using models for each SNR, gender and speaking style, and there are 24 decoders in each channel, which is the product of two speaker genders, six SNR levels (0, 5, 10, 20, 30 dB and clean), under normal and hyper-articulated speech conditions. Therefore, the total number of decoders used in our system is 48. All acoustic models are adapted to the test noise environment by using the HMM composition-based fast noise adaptation technique. In each channel, one hypothesis is selected based on its score. Then, the final result is obtained by combining hypotheses from the two channels [11].

3. Experiments

3.1. Experimental Conditions

Noisy acoustic models 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 1. A state-tying structure is generated by using the ML-SSS technique[14], with 2100 states. Each state has five distributions. The MFCC feature models are composed for different SNR levels (0, 5, 10, 20, 30 dB). DMFCC feature models are trained from data contaminated by different noises and different SNR levels, and hyper-articulated acoustic models are generated from normal speech acoustic models. Parameters of each distribution are

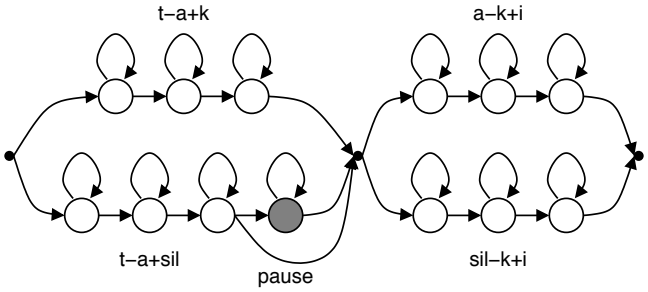

Figure 2: *The structure of an acoustic model for hyper-articulated speech.*

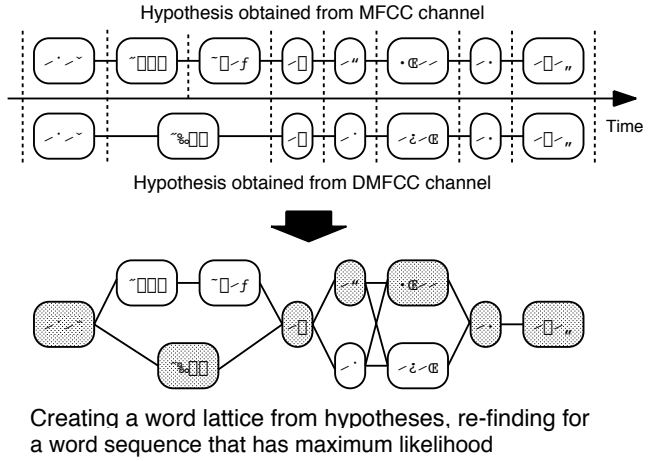

Figure 3: *Examples of a hypothesis combination.*

kept the same except for the HMM topology. Each acoustic model is gender dependent. Generated acoustic models depend on 5 SNR levels and clean conditions, 12 types of noises, MFCC and DMFCC features, speaker gender and speaking style. Therefore, the total number of noisy acoustic models is $5 \times 12 \times 2 \times 2 \times 2 = 480$ and the number of clean acoustic models is 8. During recognition, however, 40 noise-adapted acoustic models are generated from the 480 models by the fast adaptation techniques. MFCC features consists of 12 MFCCs, 12 ΔMFCCs and a $\Delta C0$ extracted with a 10 ms frame period and 20 ms frame length. DMFCC features also have 12 DMFCCs, 12 ΔDMFCCs and a $\Delta C0$.

For testing of normal speech, we used the basic travel expression corpus (BTEC) testset-01 (510 sentences, four males and six females, each speaker uttered 51 sentences). For testing on hyper-articulated speech, we collected 40 syllable-stressed sentences spoken consciously (two males and two females, each speaker uttered 10 sentences). These testing data were contaminated by three types of noises at five different SNR levels, as shown in Table 1.

Our system uses a word bi-gram and a composite word tri-gram language models [15]. Each language model is trained from the spontaneous speech database (SDB), language database (LDB) and spoken language database (SLDB), with the total number of words standing at 6.1M words. Lexicon size is 34k words.

3.2. Evaluation for Noisy Speech

We evaluated the recognition performance of our system using noisy speech data. All acoustic models in our system were adapted using one-second data from the test environment. First, Figure 5 shows the average word accuracies for each of normal speaking

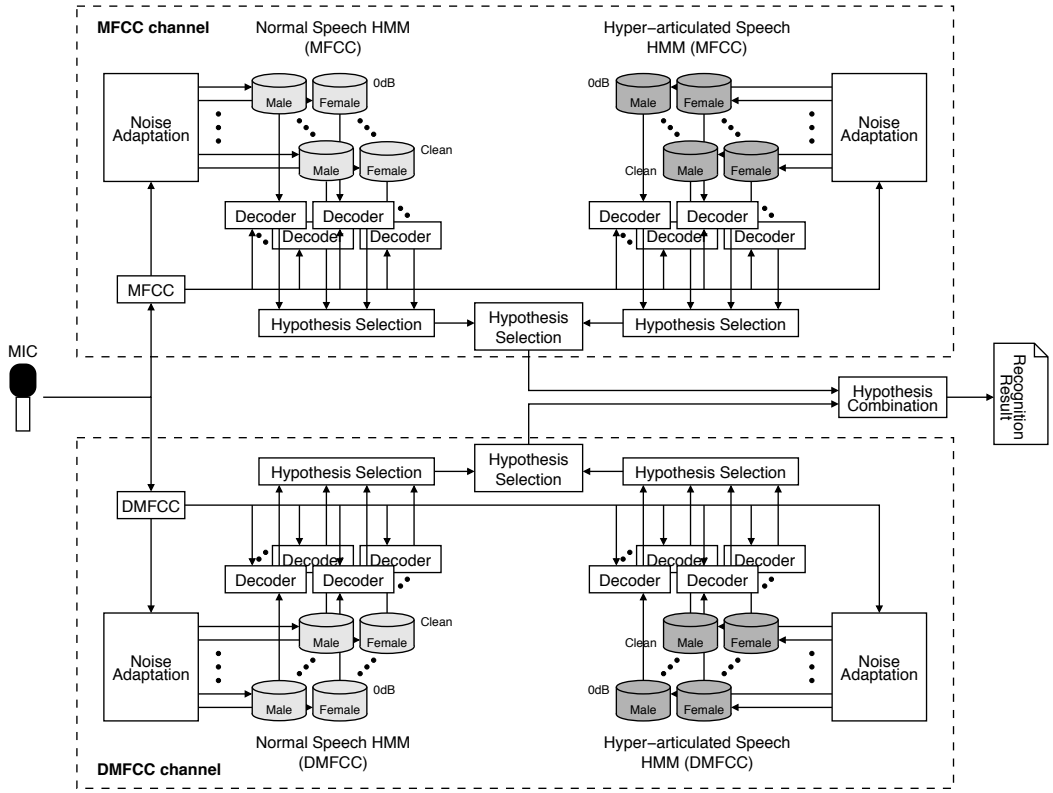

Figure 4: *Structure of our ASR system implemented by the parallel-type' method.*

Table 1: *Noise types used in experiments.*

For training	
Airport lobby	Airbus
Underground city	Car driving
Food counter	Square
Station yard	Platform at station
High-speed railway	Boiler room
Rice Paddies	Forest

For testing		
In front of a station	Public bus	Construction site

style recognizers and hypothesis selection. It is clear that the performance of the MFCC recognizer is similar to that of the DMFCC recognizer. Each channel achieved a word accuracy higher than 80%, even though the performances of the clean MFCC acoustic model and the clean DMFCC acoustic model were about 40% and about 60%, respectively. Figure 6 shows the word accuracies of the overall system. The performance was improved further by using the hypothesis combination.

3.3. Evaluation for Hyper-articulated Speech

We evaluated the recognition performance of our system using hyper-articulated speech data contaminated by three types of noises. Figure 7 shows the average word accuracy for the evaluation data. Even though the word accuracies of the system for normal speaking style only were about 10%, our system could achieve a word accuracy of about 45%.

4. Conclusion

In this paper, we described an ASR system robust to both noise and speaking styles. Our system has multiple acoustic models, each of which depends on the noise, SNR and speaking style. The total number of acoustic models is 488. The HMM composition-based noise adaptation technique was used in our system to improve robustness to noise. However, 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.

Experiments demonstrated that the system could achieve a word accuracy exceeding 90% for the normal speaking style data and a SNR of over 10 dB, and in excess of 80% for a SNR of 0 dB. Our system achieved a word accuracy of over 45% for hyper-articulated speech.

Future work includes study on the generation method of a set of acoustic models that efficiently covers a wide variety of noise and various speaking styles.

ACKNOWLEDGMENTS

This research was supported in part by the National Institute of Information and Communications Technology.

5. References

[1] Y. Gong, "Speech recognition in noisy environments: a survey," Speech Communication, 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 Processing, vol. ASSP-27, pp. 113–120, 1979.

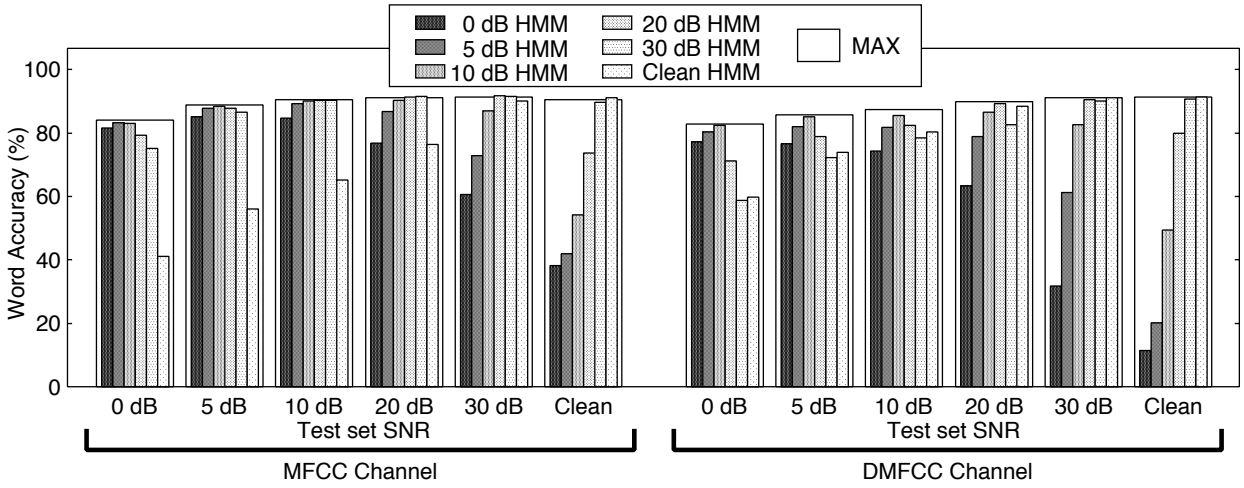

Figure 5: *Recognition performance of each channel for noisy speech data.*

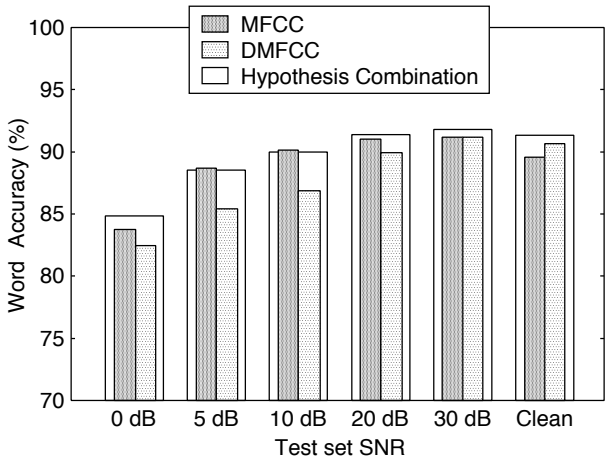

Figure 6: *Recognition performance for noisy speech data.*

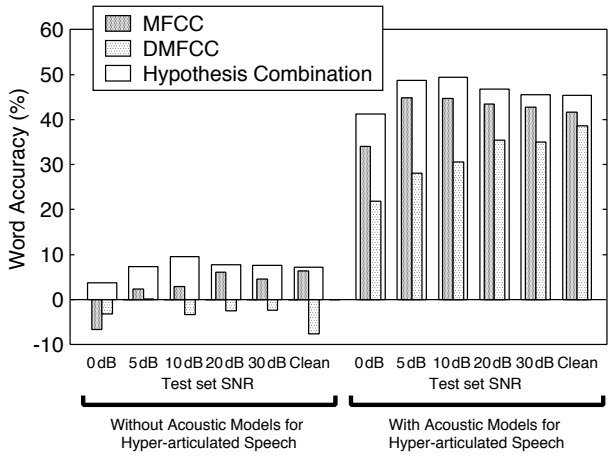

Figure 7: *Recognition performance for hyper-articulated speech data contaminated by noise.*

[3] H. Hermansky and N. Morgan, "RASTA Processing of Speech," IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 587–589, 1994.

[4] M. Gales and S. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEEE Trans. on Speech and Audio Processing, vol. 4, no. 5, pp. 352–359, 1996.

[5] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, vol. 9, pp. 171–185, 1995.

[6] J.C. Junqua, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizer," J. Acoustic. Soc. Amer., vol. 93, pp. 510–524, 1993.

[7] K. Okuda, T. Matsui, S. Nakamura, "Towards the Creation of Acoustic Models for Stressed Japanese Speech," Eurospeech2001, vol. 3, pp. 1653–1656, 2001.

[8] K. Okuda, T. Kawahara, S. Nakamura, "Speaking Rate Compensation Based on Likelihood Criterion in Acoustic Model Training and Decoding," ICSLP2002, vol. 4, pp. 2589–2592, 2002

[9] H. Nanjo, K. Kato, T. Kawahara, "Speaking Rate Dependent Acoustic Modeling for Spontaneous Lecture Speech Recognition," Eurospeech 2001, pp. 2531–2534, 2001.

[10] J. Chen, K.K. Paliwal, S. Nakamura, "Cepstrum Derived from Differentiated Power Spectrum for Robust Speech Recognition," Speech Communication, vol. 41, no. 2-3, pp. 469–484, 2003.

[11] K. Markov, T. Matsui, R. Gruhn, J. Zhang, S. Nakamura, "Noise and Channel Distortion Robust ASR System for DARPA SPINE2 Task," IEICE Trans. Inf. & Syst., vol. E86-D, no. 3, 2003.

[12] K. Yao, J. Chen, K.K. Paliwal, S. Nakamura, "Feature extraction and model-based noise compensation for noisy speech recognition evaluated on AURORA 2 task," Eurospeech, vol. I, pp. 233–236, 2001.

[13] M. Ida, 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] M. Ostendorf and H. Singer, "HMM Topology Design Using Maximum Likelihood Successive State Splitting," Computer Speech and Language, vol. 11, no. 1, pp. 17–41, 1997.

[15] H. Yamamoto, Y. Sagisaka, "Multi-class Composite N-gram Language Model Based on Connection Direction," Proc. ICASSP, pp. 533–536, 1999.