# Noise Robust Speech Recognition based on Spectral Reduction Measure

Mayumi Beppu*, Koichi Shinoda*, and Sadaoki Furui*
* Tokyo Institute of Technology, Tokyo, Japan
E-mail: shinoda@cs.titech.ac.jp

***Abstract*—It has been reported that spectral features of spontaneous speech shrink from those of read speech and that this spectral reduction causes degradation in speech recognition performance. This spectral reduction is expected to be observed with other factors than spontaneity, and cause degradation. In this study, we analyze spectral reduction in noisy speech and its correlation to speech recognition accuracies. Based on the analysis result, we propose a model adaptation method to noisy environment. In our evaluation using noisy speech data, our method increased the recognition accuracies by 2.6 points.**

## I. Introduction

Although state-of-the-art automatic speech recognition technology can achieve high accuracy in quiet conditions, its performance is drastically degraded in noisy environment. This is a serious problem in many applications such as car navigation systems and mobile device interfaces. To solve it, many studies have been done. For example, model adaptation methods such as MLLR and MAP adapt an acoustic model to noisy environment [1][2]. Gales *et al.* proposed PMC to model noisy speech, speech in noisy environment [3].

The influence of noise on speech spectral features may differ in different classes of noise and different classes of speech. While most model adaptation methods are oriented to be universal, that is, to be applied to any kinds of speech and noise, we are more interested in analyses of the differences among them and in utilizing their results to improve recognition performance.

In the area of speech spectral analysis, van Son *et al.* [4] found that the distances between formant frequencies in spontaneous speech became smaller than those in read speech and called such phenomena *spectral reduction*. They also concluded that this spectral reduction was caused by stronger co-articulation in spontaneous speech. Nakamura *et al.* [6] found that such reduction was also observed in mel-frequency cepstral coefficients (MFCCs), and that the rate of reduction was positively correlated with the degradation rate of speech recognition accuracies.

Since the spectral reduction in spontaneous speech is one reason for the degradation in speech recognition accuracy, it may also be observed in noisy speech and may deteriorate speech recognition performance. To the best of our knowledge, however, no studies have been done in this respect.

In this paper, we compare the spectrum of noisy speech with that of clean speech to examine the spectral reduction. We further propose an adaptation method for acoustic models based on the results obtained by the analysis.

## II. Speech data

In this paper, we use speech data from Corpus of Spontaneous Japanese (CSJ)[7]. We use 600-hour speech data of academic presentations (AP) and extemporaneous presentations (EP) for acoustic model training. We use four hours of them as speech data for our analysis (analysis data). To evaluate our adaptation method in Section V, we use 5.5 hours of AP and EP data not involved in the 600-hour training data as test data. All utterances are digitized by 16kHz sampling.

We synthesize the CSJ speech data with noise data to make noisy speech data. As noise data, we use 29 classes of noise data, such as "computer room", "train station", and "elevator hall", in Denshikyo noise database [5]. The length of each noise data is 80 seconds. We add the noise to the speech data with three different signal-to-noise ratios (SNRs), 10, 20, and 30 dB. The total number of noise classes is 29 x 3 = 84.

From the speech data, we extract 12-dim (1st to 12-th) MFCC features, their first derivatives, and the power derivative to make 25-dim feature vectors for each speech frame, where the frame period is 10 ms, and the window width is 25ms.

In our analysis, we select those phones which occur more than 0.5% of the occurrences of all the phones. They are listed in Table I, where the number of vowels is 10 and that of consonants is 17.

## III. Cepstral reduction rate

In this section we explain a cepstral reduction rate (CRR), which we use in our analysis of spectral reduction in noisy speech data.

First, we calculate a *phone-average vector* for each phone as follows:

1) Using the training data without noise, construct continuous density HMMs with Gaussian-mixture observation probability, where the number of Gaussian per state is 64.
2) Carry out Viterbi alignment between models and the analysis data.
3) For each segment of the analysis data aligned with one phone, identify its middle frame between its beginning and its end and extract from that frame a 12-dim feature vector which consists of 1-st to 12-th MFCC static coefficients.
4) Average the 12-dim feature vectors which belong to each phone in the analysis data to make a phone-average vector.

| | |
|---|---|
| vowel | /a, i, u, e, o, a:, i:, u:, e:, o:/ |
| consonant | /y, w, j, r, t, k, b, d, g, ts, ch, s, ʃ, h, N, m, n/ |

TABLE I
PHONE CLASSES USED IN OUR ANALYSIS

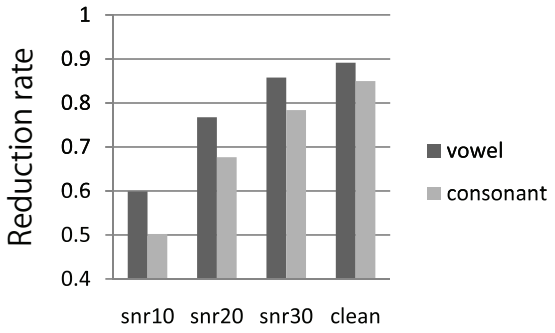


Fig. 2. CRRs for different SNRs, where CRRs for all phone classes are averaged.

Then, we calculate the reduction rate of the phone-average vector from clean speech and call it a *cepstral reduction rate* (CRR). The CRR of phone $p$ in noisy speech $X$ from clean speech $R$ is defined as:

$$\mathrm{CRR}_p(X) = \frac{||\mu_p(X) - \mathrm{Av}(\mu_p(X))||}{||\mu_p(R) - \mathrm{Av}(\mu_p(R))||}, \quad (1)$$

where $\mu_p(X)$ and $\mu_p(R)$ are the phone-average vector of $p$ of the noisy speech $X$, and that of the clean speech $R$, respectively. "Av" indicates average operation over all phones.

## IV. Analysis of noisy speech

### A. Cepstral reduction rate for each phone

We expect that the cepstral feature vectors of speech in noisy environment shrinks from those in quiet conditions. We examine this phenomenon by calculating the CRR defined by Eq. (1).

We show the CRR of each phone in Fig. 1 and their average over all phones in Fig. 2. For both vowels and consonants, the CRR becomes smaller as SNR becomes smaller. Fig. 3 shows CRRs for 10 noise class. The CRRs are largely different among them. They seem to be not large in such noise classes whose dominant frequency band is not so wide. We would like to examine the relation between noise spectrum and its related CRR in the future work.

### B. CRRs and speech recognition accuracies

Next, we analyze the relationship between CRRs and speech recognition accuracies.

For this analysis, we employ concatenated-phone recognition to measure phone recognition accuracies, where a phone network grammar designed for Japanese is employed. As acoustic models, we use monophone HMMs in which the output probability distribution of each state is a single Gaussian.

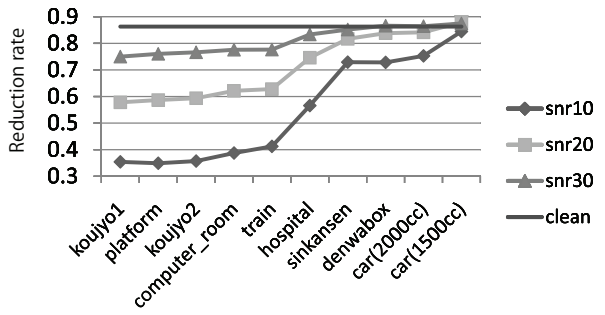


Fig. 3. CRRs averaged over all phone classes. We show them the results for 10 of the 29 noise classes; five classes with larger CRRs than others and five classes with smaller CRRs than others.

Fig. 4 shows the correlation coefficients between CRRs and phone recognition accuracies. High correlation larger than 0.8 is observed in short vowels and twelve consonants. In some long vowels, /a:, i:, e:/, CRRs and recognition accuracies are negatively correlated. The reason for this phenomena should be investigated in future.

## V. Acoustic model adaptation

In Subsection IV-B, we found that there exists a strong correlation between CRR and phone recognition accuracies in 17 of the 27 phones. Based on this observation, we propose an acoustic model adaptation method which shrinks the HMM parameters in noisy environments.

### A. Model reduction rate

In our analysis, we found that the cepstral features for each phone are reduced in noisy environments and its reduction rate is represented by CRRs. According to this observation, we first tried to improve the recognition accuracies by multiplying the CRR to its corresponding mean vectors in HMMs. In our preliminary experiments, however, this method did not bring any improvements in speech recognition accuracies. Here, we introduce *model reduction rate* (MRR), which is a coefficient for each mean vector of HMMs. We further define MRRmax, which is the MRR maximizing the recognition accuracy, and examine whether it correlates with the CRR.

For each phone, we carry out the following process to obtain the relationship between its CRR and MRRmax. As we explained in Section II, we have 87 different noise conditions.

1) For each noise condition, we change MRR from 0.2 to 1.2 with a step size 0.05 to find the MRRmax.
2) Using 87 pairs of CRRs and MRRmax's, we calculate the correlation coefficient between them.

We find that most phones have strong correlations between CRRs and MRRmax's. Here we use those phones whose correlation coefficients are larger than 0.85, namely the four vowels /a,i,e,o/ and the eleven consonants /y,j,t,ch,d,g,m,n,s,ʃ,r/. In these 15 phones, 14 phones are selected from 17 phones with high correlations between CRRs and recognition accuracies in Fig. 4, and one phone /t/ is newly added.

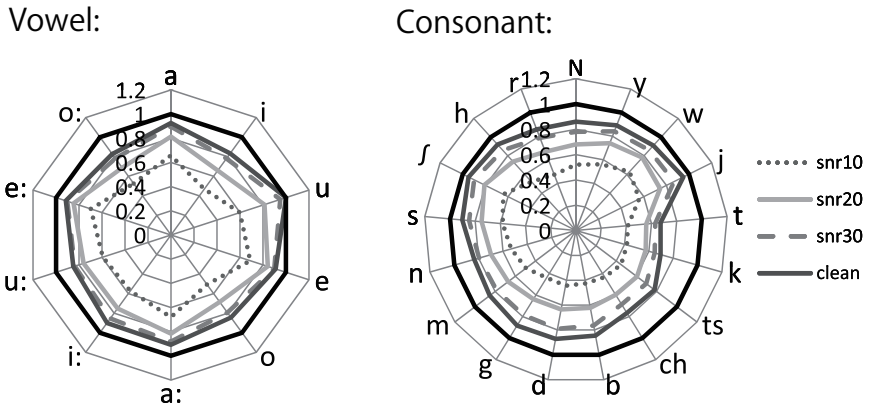


Fig. 1. Cepstral Reduction Rate (CRR) for each phone. For each SNR, the CRRs for all the noise classes are averaged. The thick lines indicate CRRs of clean speech (CRR$=1$).

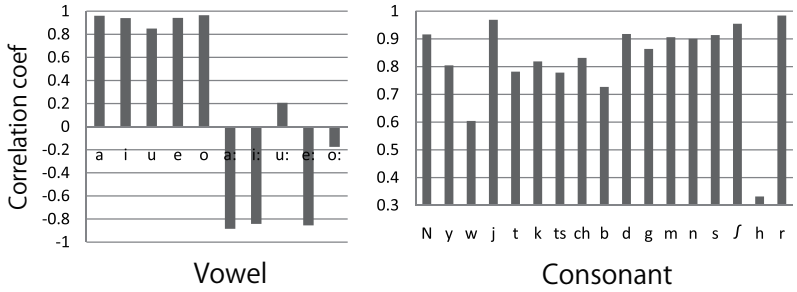


Fig. 4. Correlation coefficients between CRRs and phone recognition accuracies

Next, for each phone, we obtain a regression expression estimating a MRRmax given a CRR. We use a leaving-one-out process as follows.

1) For each noise class, we obtain the CRR averaged over the rest of 84 noise conditions (28 noise classes with three SNR, 10, 20, 30).
2) For each of 84 conditions, we estimate the MRRmax.
3) We obtain a first-order regression expression using 84 pairs of CRRs and MRRmax's.

### *B. Adaptation algorithm*

Here, we explain our method to adapt the acoustic model to speech data in a given noisy environment. The model adaptation process for each phone is as follows:

1) Estimate the CRR for the phone using a small amount of the speech data in the target noise environment.
2) Estimate the MRRmax using the CRR and the regression expression described in Subsection V-A.
3) Multiply mean vectors of Gaussian-mixture distribution of all the states of the corresponding phone HMMs by the MRRmax.

This process is applied only to static cepstral coefficients in the mean vector and not to their derivatives.

### *C. Phone recognition experiments*

We evaluated our model adaptation method by phone recognition experiments. We used the same acoustic model and language grammar as those in Subsection IV-B. We used a part of the training data with additional noise for estimating the MRRmax. We used the CSJ test set for recognition tests. The amount of adaptation data was 0.5 hour.

Fig. 5 shows the results of the experiments, where "Estimated" is the results of our adaptation methods using the estimated MRRmax, and "Optimized" is the results with the optimal MRRmax's maximizing the recognition accuracies (oracle). In "Estimated", the smaller CRRs become, the more significant the improvements become. In "Optimized", the same tendencies were observed in lower CRRs, but the recognition accuracies degraded for noise conditions with higher CRRs. We should investigate the reason for this in the future work.

### *D. Amount of adaptation data and recognition performance*

Next, we investigated the amount of adaptation data needed to obtain significant improvements in recognition accuracies. We used the same configuration as in Subsection V-C for the recognition experiments. We chose $n$-minutes speech data for estimating the MRRmax, where we tested three cases: $n = 5, 15, 30$.

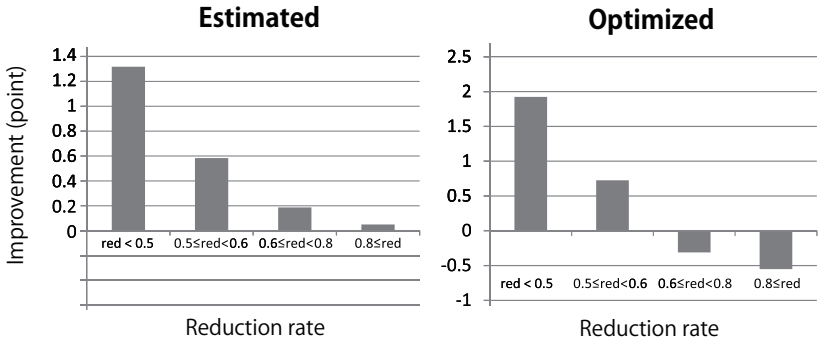


Fig. 5. Improvement in phone recognition accuracies obtained by our adaptation method. We categorize 87 groups of noise conditions into four groups according to the their CRR values (red) averaged over all the phones. "Estimated" is the results of our adaptation methods. "Optimized" is the results with the optimal MRRmax's maximizing the recognition accuracies (oracle).

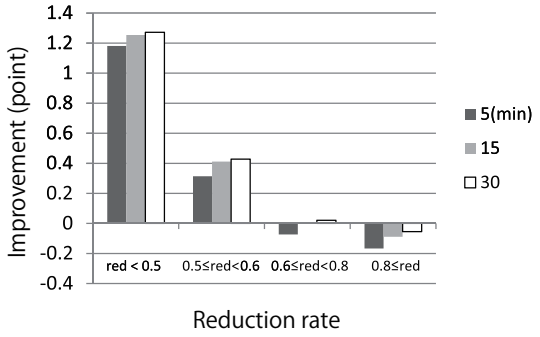


Fig. 6. Amounts of adaptation data and improvements obtained by adaptation. We categorize 84 groups of noise conditions into four groups according to the their CRR values averaged over all the phones.

Fig 6 shows the improvement obtained by our adaptation method. As the amount of adaptation increase, the improvements become more significant. When CRRs are less than 0.5, however, five minutes of adaptation data seemed to be enough; more data did not bring any significant improvements. It was observed that the amount of adaptation data needed was rather small when CRR was small.

The recognition accuracies are slightly degraded when CRR is more than 0.6, It might be effective to apply our proposed adaptation method only to those noise class whose CRRs are small.

Correlation coefficients between the improvements in phone recognition accuracy and averaged CRRs over the 16 phones for $n = 5, 15, 30$ were $-0.852$, $-0.862$, $-0.859$, respectively. The improvements became larger as the CRRs were smaller. This result indicates that from CRR values we can predict the degree of improvements given the amount of adaptation data.

## VI. Conclusion

We analyzed the difference in speech features between clean speech and noisy speech. We observed that the cepstral features of noisy speech shrink from those of clean speech (cepstral reduction). This phenomena was observed in both vowels and consonants. We also found that this cepstral reduction rate was largely different when noise classes and SNRs were different.

Next we analyzed the relationship between the cepstral reduction and the recognition accuracy of each phone. We observed that their correlation was high for most phones.

Based on these analysis results, we have proposed a model adaptation method utilizing correlations between cepstral reduction rates and recognition accuracies. It was proved to be effective especially when the cepstral reduction was significant. The largest improvement was 2.6 points. We further confirmed that the amount of adaptation data needed became smaller as the cepstral reduction became larger.

We observed that the dominant frequency in noise spectrum is related to the degree of cepstral reduction. We will examine this relationship in our future work. Interestingly, we found strong negative correlation between cepstral reduction and recognition accuracy in long vowels. We have to find the reason for this. We used single-Gaussian monophone HMMs in our experiments of model adaptation. We would like to confirm the effectiveness of our adaptation method when the number of Gaussians is more than one and when triphone HMMs are used. Based on these developments, we plan to apply our approach to large vocabulary continuous speech recognition and verify its effectiveness in real applications.

## References

[1] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Proc. of ICSLP1994, pp 451-454, 1994.

[2] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. SAP, vol. 2, no. 2, pp. 291-298, 1994.

[3] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, pp. 352-359, 1996.

[4] van Son, R. J. J. H. and Pols, L. C. W., "An acoustic prole of consonant reduction," In Proc. ICSLP 1996, pp. 1529-1532, Philadelphia, U.S.A., 1996.

[5] Denshikyo noise database, http://research.nii.ac.jp/src/list/detail.html#JEIDA-NOISE.

[6] M. Nakamura et al. "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," Computer Speech and Language, pp.171-184, 2008.

[7] Corpus of spontaneous Japanese (CSJ), http://www2.kokken.go.jp/csj/public/.