音声言語情報処理 39－8
(2001. 12. 20)

Noise Speech Recognition based on Robust Features and A Model-Based Noise Compensation evaluated on Aurora-2 Task

Kaisheng Yao†, Jingdong Chen †, Kuldip K. Paliwal†‡, and Satoshi Nakamura†

† ATR Spoken Language Translation Research Laboratories, Kyoto, Japan

‡ School of Microelectronic Engineering, Griffith University, Australia

E-mail: †{kyao, jchen, nakamura}@slt.atr.co.jp, ‡k.paliwal@me.gu.edu.au

Abstract We have evaluated several feature-based and a model-based method for robust speech recognition in noise. The evaluation was performed on Aurora 2 task. We show that after a sub-band based spectral subtraction, features can be more robust to additive noise. We also report a robust feature set derived from differential power spectrum (DPS), which is not only robust to additive noise, but also robust to spectrum colorization due to channel effects. When the clean training set is available, we show that a model-based noise compensation method can be effective to improve system robustness to noise. Given the testing sets, as a whole, the feature-based methods can yield about 22% relative improvement in accuracy for multi-condition training task, and the model-based method can have about 63% relative performance improvement when systems were trained on clean training set.

Key words Speech recognition, Noise, Robust speech recognition

AURORA-2 タスク雑音環境下音声認識における雑音にロバストな特徴抽出法とモデル補償

ヤオ カイシェング†, チェン ジンドン†, パリワル クルディップ†‡, 中村 哲†

† ATR 音声言語コミュニケーション研究所

‡ グリフィス大学（オーストラリア）

E-mail: †{kyao, jchen, nakamura}@slt.atr.co.jp, ‡k.paliwal@me.gu.edu.au

あらまし: 雑音環境下の音声認識において，筆者らはこれまで雑音にロバストな特徴量抽出法や音響モデルの補償方法について AURORA2 タスクを用いた評価を行ってきた. Multi condition 学習データを利用できる場合においては，加法性雑音に対して DPS (Differential Power Spectrum) に基づいた特徴量が最も良い性能を示し，さらにこの方法が伝送経路による歪みに対しても有効であることがわかった. これに対し，学習データとして clean speech しか使えない場合には，音響モデルの補償を行う方法が良い性能を示した. 評価実験の結果，提案する特徴量抽出法を用いることで Multi condition 学習データを用いる場合，22%の改善を得た. また，提案するモデル補償法を用いることで clean な学習データを用いる場合，63%の改善が見られた.

キーワード: 音声認識，雑音，耐雑音性

1. Introduction

Speech recognition systems have to be operated in situations where it is not possible to control the acoustic environments. This may cause mismatch between the training and testing conditions. Varieties of methods have been proposed to cope with this problem. They can be categorized into three approaches. One approach can be denoted as front-end preprocessing of noisy speech signals. The second approach is robust feature extraction approach, and the third one is denoted as the model-based noise compensation approach. We report our evaluation results of robust feature extraction methods and a model-based noise compensation method on Aurora 2 task [1], which is a noise-contaminated version of the TI-DIGITS database. The speech in the task have been down-sampled from 20kHz to 8kHz and contaminated by eight kinds of noise in seven different Signal-to-Noise (SNR) conditions. Speech signals in the task were also contaminated by channel distortion, resulting in spectrum colorization.

Generally speaking, if speech signals in testing environments can be collected, it would be straightforward to train a speech recognition system based on the collected data. This is denoted as multi-condition training. Methods based on robust feature extraction can go beyond in performances obtained by standard speech features, such as MFCC and LPCC, by introducing robustness to the extracted features. For example, spectral subtraction is carried out in the linear spectral domain, and as a result, the features extracted are more robust to additive noise than the standard features. In the spectral subtraction method, noise spectrum is required to be estimated. In this paper, we report two methods for noise spectrum estimation. One method makes use of long-term average of noisy spectrum, and the other method makes use of long-term Fourier analysis. Once the noise spectrum has been estimated, it is applied to a sub-band based spectral subtraction method, where spectral subtraction is carried out in each sub-band in mel-scale. Besides the above method, a differential procedure is proposed before the binning stage of MFCC feature extraction process. The obtained linear spectrum is denoted as Differential Power Spectrum (DPS). We will show that MFCCs derived from DPS can effectively improve system robustness to noise.

In many situations, a system only has training data collected in clean conditions. Given the speech features are simple, e.g., MFCCs, it is possible to explicitly make a mathematical model approximating the noise effects on speech features. This is one of the motivations for the model-based approach. Varieties of methods have been proposed in this category. In this paper, we employ a non-linear mean transformation method, Log-Add noise compensation method [2], where mean of clean speech models is transformed before speech recognition. We will show through experiments on the Aurora 2 task that it can effectively improve system robustness to additive noise.

2. Feature based robust speech recognition

2.1 Spectral subtraction based approach

Spectral subtraction has been proven to be an effective technique to enhance speech. This method involves two problems, i.e., the estimation of noise power spectrum and its subtraction from the power spectrum of noisy speech signal. We have investigated in [3] several noise estimation methods and a sub-band based subtraction strategy. Experiments showed that this approach outperformed the traditional spectral subtraction [4] for various tasks. Here we assess two ideas for noise estimation and the sub-band based spectral subtraction.

2.1.1 Noise estimation

a) Long-term average

An intuitive way to achieve noise spectrum is to estimate it from the non-speech portion of the signal. This, however, will need a speech signal detector (SSD), which can distinguish speech segments from non-speech segments. To avoid a SSD, Flores *et al* [5] proposed to use the average of M consecutive frames of short-term power spectra of noisy speech signal as the noise estimate, *i.e.*,

$$\hat{N}(i,f)=\frac{1}{M}\sum_{j=i-M+1}^{i}Y(j,f) \tag{1}$$

where $\hat{N}(i,f)$ is the noise estimate at the time instant i, $Y(j,f)$ is the power spectrum of the jth frame of noisy speech. Inspired from this idea, we assume that all frames of speech for a test utterance are available simultaneously and use the average of the short-term power spectra over all frames as the noise estimate, namely,

$$\hat{N}(i,f)=\hat{N}(f)=\frac{1}{N}\sum_{j=1}^{N}Y(j,f) \tag{2}$$

where N is the number of frames for the given utterance. Note that in this approach, all frames in one utterance share a single noise estimate. We call this method long-term average (LTA).

b) Long-term Fourier analysis method

It was found that phonetic information of speech is encoded in the changes of the speech spectrum over time. Relatively less phonetic information is encapsulated in the long-term speech spectrum. Noise, however, can be treated as a stationary process. Long-term spectrum will provide a good estimate of noise. Based on this fact, we propose to estimate noise using long-term Fourier analysis, i.e.,

$$\hat{N}(i,f)=\hat{N}(f)=\frac{1}{\zeta}\left|\mathcal{F}[y(l)w(l)]\right|^2 \tag{3}$$

where $\mathcal{F}[\cdot]$ denotes Fourier Transform, $y(l)$ is the discrete speech signal for a whole utterance, $w(l)$ is a window function, ζ a normalization factor which is defined as $\zeta=\mathrm{L}\sum_l w(l)$, and L is the length of the Fourier Transform.

Noting that the noise effect estimated from Eq. (3) has a much longer length than that of short-term power spectrum, we therefore need to warp it to have a same length as the power spectrum of each frame. We should also point out that the short-term power spectra should be normalized in a similar way as in Eq. (3) before one subtracts this noise estimate from them.

2.1.2 Sub-band based spectral subtraction

In [3], we proposed a sub-band based spectral subtraction strategy. We show that if we split the full-band signal into sub-band on the mel-scale and the cutoff frequencies of each sub-band are set to be same as what adopted in the mel-scale triangle filter banks, the sub-band based subtraction can be explicitly expressed as

$$E^i_{ss}(k)=\begin{cases}E^i_Y(k)-\alpha E^i_{\hat{N}}(k), \text{if } E^i_Y(k)>\frac{\alpha^k}{1-\beta^k}E^i_{\hat{N}}(k)\\ \beta^i E^i_Y(k), \quad otherwise\end{cases} \tag{4}$$

where $E^i_Y(k)$ is the output of the kth triangle filter when the noisy speech power spectrum $Y(i,f)$ is passed through the triangle filter bank, and $E^i_{\hat{N}}(k)$ the output of the same filter with its input being $\hat{N}(i,f)$, and α^k and β^k are sub-band dependent over-estimation factor and spectral flooring respectively.

2.2 MFCCs derived from Differential Power Spectrum

If we denote the power spectrum of the ith frame of speech signal as $Y(i,k)$, the differential power spectrum (DPS) can be defined by following difference equation

$$D(i,k)=\sum_{l=P_1}^{P_2}b_l Y(i,k+l) \tag{5}$$

We have investigated to use the DPS to represent speech signal and found that the MFCC derived from some special forms of DPS resulted features more robust to noise than the conventional MFCCs [6].

DPS1: $D(i,k)=Y(i,k)-Y(i,k+1)$ (6a)

DPS2: $D(i,k)=Y(i,k)-Y(i,k+2)$ (6b)

In this paper, we investigate to use DPS1 for robust speech recognition. We pass $|D(i,k)|$ defined in Eq. (6a) through a mel-scaled triangular filter band and convert it into some MFCC-like feature coefficients. We assess this new MFCC feature vector and compare them with the conventional MFCCs on Aurora 2 task.

3. Model based approach

3.1 Log-Add noise compensation

The Log-Add noise compensation [2] is a non-linear transformation of mean vector in clean speech models. It is carried out in the log-spectral domain. After estimation of noise parameters μ^l_{nj} in the log-spectral filter bank j, the mean μ^l_{imj} of static MFCCs in each Gaussian mixture m at state i is transformed by the following formula,

$$\hat{\mu}^l_{imj}=\mu^l_{imj}+\log\left(1+\exp\left(\mu^l_{nj}-\mu^l_{imj}\right)\right) \tag{7}$$

where $1\le j\le J$, and J is the total number of log-spectral filter banks. Superscript l indicates that parameters are in the log-spectral domain. The transformed mean in the log-spectral domain is further transformed back into the cepstral domain to substitute the original mean vector in clean speech models.

This transformation assumes that the noise variance is very small, and accordingly, only the mean of the acoustic models are transformed. Other model based methods employ similar functional formula with different complexities.

The noise parameter estimation is the key point to the success of the method. If the noise is stationary in the testing set, we can do mean transformation of clean speech models once and then carry out speech recognition on the testing set.

4. Experiments

4.1 Performance of the feature extraction methods

4.1.1 Experimental setup

The Aurora 2 task has provided baseline system performances to calculate relative performance improvements automatically, so that methods proposed from different organizations can be compared. In this subsection, the baseline system was trained on multi-condition training set with 8840 utterances containing subway, babble, car and exhibition hall noise in five different SNR conditions from 5dB to clean condition in 5dB step. Speech features are 13 MFCC coefficients with energy extracted from amplitude of power spectrum, along with the delta and acceleration coefficients. The MFCC coefficients are derived from 23 mel-spaced triangular filters. The baseline system thus uses a 39-dimentional feature vector per frame. Speech models are eleven whole word HMMs fixed to 16 states 3 diagonal Gaussian mixtures per state. Two silence model, one with 3 states and 3 mixtures each state modeling the utterances beginning and end, and the other with one state and six mixtures modeling the interword silence have been used. Recognition system for robust feature extraction was also trained on the multi-condition training set. Testing was carried out on three testing sets. Each testing set contains noisy utterances with seven SNR conditions from –5 dB to clean conditions. Testing set A contains noises in the multi-condition training, whereas testing set B contains four kinds of noises unseen in the testing set A. Testing set C is more difficult since signals were distorted by MIRS channel instead of G712 channel in the training set, testing set A and B. Thus, it has channel effects besides the additive noise effects on the speech features.

4.1.2 Performance of the sub-band based spectral subtraction method

Figure 1 and Figure 2 each plot the word accuracies versus SNR in three testing sets (denoted as Set A, Set B, and Set C) of the spectral subtraction method with noise spectrum estimated by long-term average and long-term Fourier analysis, together with their overall word accuracies (denoted as Overall) and that of the baseline (denoted as Reference).

As a whole, the total relative performance improvement from –5dB to clean condition are 19.35% and 19.32% for the method with noise spectrum estimated by long-term

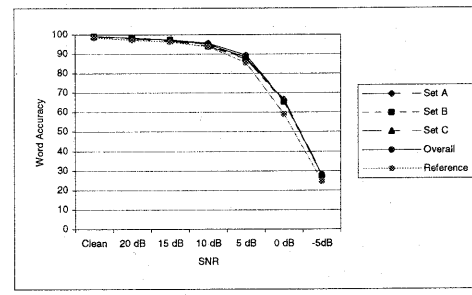

Figure 1. Word Accuracy by sub-band based spectral subtraction, where $\alpha^k = 0.5$ *and* $\beta^k = 0.1$, *and noise*

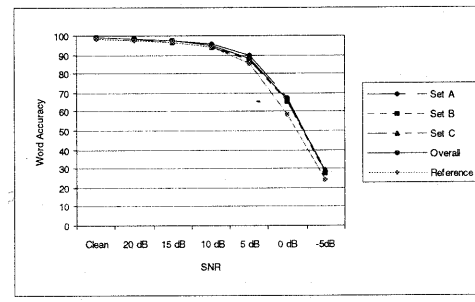

Figure 2. Word accuracy by sub-band based spectral

average and long-term Fourier analysis, respectively.

4.1.3 Performance of the MFCCs derived from DPS

Table 1 shows the absolute values of word accuracy for systems trained with MFCCs derived from Differential Power Spectrum, together with the relative improvements in each testing set for the average scores from 20dB to 0dB along the bottom (in bold font), and similarly the relative improvements in the overall scores for each SNR on the right. The total relative improvement is shown in the bottom right corner. It is observed that the proposed feature extraction is not only robust to additive noise but also robust to spectrum colorization, since the relative performance improvement in testing set C is 36.82%, which is higher than that in testing set A, 15.78%, and testing set B, 17.92%. As a whole, a total relative performance improvement of 21.66% can be obtained by the proposed method.

Figure 1~3 show the averaged recognition accuracies of the described three feature extraction methods in the evaluation. It can be seen that all of them can improve system robustness over that trained with standard MFCCs.

	A					B					C				Percentage Improvement
	Subway	Babble	Car	Exhib.	Average	Rest.	Street	Airport	Station	Average	Sub.	Street	Average	Overall	
Clean	99.08	98.76	99.11	99.26	99.05	99.08	98.76	99.11	99.26	99.05	99.05	98.85	98.95	99.03	34.01%
20 dB	97.94	98.31	98.54	98.03	98.21	98.04	98.00	98.42	98.67	98.28	98.13	97.64	97.89	98.17	30.01%
15 dB	97.39	97.70	97.88	96.98	97.49	97.36	97.22	97.55	97.13	97.32	97.64	96.74	97.19	97.36	27.25%
10 dB	95.70	96.34	95.68	94.72	95.61	94.57	95.19	95.38	93.98	94.78	95.33	94.32	94.83	95.12	20.23%
5 dB	91.46	88.91	90.46	87.44	89.57	85.42	88.60	89.11	86.02	87.29	90.42	88.00	89.21	88.58	20.31%
0 dB	75.13	64.93	62.48	68.74	67.82	62.67	68.20	71.25	61.83	65.99	72.09	67.20	69.65	67.45	19.61%
-5dB	34.05	27.45	20.46	29.10	27.77	27.48	28.93	31.35	21.85	27.40	32.42	27.93	30.18	28.10	4.70%
Aver.	91.52	89.24	89.01	89.18	89.74	87.61	89.44	90.34	87.53	88.73	90.72	88.78	89.75	89.34	
	24.62%	10.69%	18.43%	9.62%	15.78%	15.20%	18.55%	21.85%	16.78%	17.92%	44.63%	28.49%	36.82%		21.66%

Table 1. Word accuracy and relative performance improvement of the MFCCs derived from Differential Power Spectrum. Baseline system and the recognition system were trained on multi-condition training set. Aver. denotes the averaged score in each testing set. Overall denotes the overall score in each SNR condition for all testing sets.

	A					B					C				Percentage Improvement
	Subway	Babble	Car	Exhib.	Average	Rest.	Street	Airport	Station	Average	Sub.	Street	Average	Overall	
Clean	99.26	99.27	99.27	99.35	99.29	99.26	99.27	99.27	99.35	99.29	99.37	99.34	99.36	99.30	27.55%
20 dB	98.71	98.72	98.92	98.36	98.68	98.94	98.46	98.77	98.68	98.71	98.46	97.80	98.13	98.58	70.25%
15 dB	94.59	98.15	97.95	97.12	96.95	98.08	97.40	98.13	98.00	97.90	97.20	95.05	96.13	97.17	73.63%
10 dB	94.33	94.69	91.71	93.57	93.58	95.38	92.35	95.69	94.37	94.45	91.73	91.09	91.41	93.49	78.78%
5 dB	85.19	82.35	70.44	82.99	80.24	85.51	78.04	80.82	78.35	80.68	75.55	77.34	76.45	79.66	66.02%
0 dB	60.32	58.15	51.17	58.25	56.97	62.83	60.73	58.90	55.94	59.60	54.78	57.27	56.03	57.83	48.88%
-5dB	48.30	47.55	44.39	44.90	46.29	49.76	46.08	47.04	46.31	47.30	48.15	46.50	47.33	46.90	41.85%
Aver.	86.63	86.41	82.04	86.06	85.28	88.15	85.40	86.46	85.07	86.27	83.54	83.71	83.63	85.35	
	56.18%	72.89%	54.41%	59.71%	61.93%	75.00%	62.05%	71.04%	66.35%	68.97%	51.37%	51.92%	51.64%		63.31%

Table 2. Word accuracy and relative performance improvement of the system compensated by the model-based method. Baseline system and the recognition system were trained on clean training set. Aver. denotes the averaged score in each testing set. Overall denotes the overall score in each SNR condition for all testing sets.

4.2 Performance of the model-based noise compensation method

4.2.1 Experimental setup

In many situations, a system has only clean data for model training. We thus carried out experiments here to show that model-based methods can be effective to improve system robustness to noisy environments in this kind of situations. In this subsection, the baseline system provided by the Aurora 2 task was trained on clean training set. Our recognition system had HMMs with the same structure as the baseline system and the models were trained on the same clean training set, expect that the speech feature had some differences from the baseline system. First, the MFCCs were generated from the power of the Fourier transform in the binning stage in the MFCC extraction procedure. Second, the static coefficient was MFCC plus C0. Last, the number of filter-banks was 26 instead of 23 in the baseline system.

4.2.2 Noise parameter estimation for the model-based noise compensation method

The Log-Add noise compensation by Eq. (7) requires noise parameter estimation. In the evaluation, it was estimated by the method shown below.

We assume that, in each SNR, the noise is stationary. Thus, a single Gaussian distribution can possibly represent its statistics. Accordingly, the frames at the beginning and end of utterances as a whole were used to estimate the statistics. We thus made use of the transcript of clean speech utterance output from the recognizer, and estimated the noise statistics from noisy utterance given the transcript. In the experiments, we only picked up one utterance in each SNR condition for each noise situation in each testing set. The estimated mean vectors of the MFCC plus C0 in the segmented noise frames were transformed to the log-spectral domain, and plugged into Eq. (7) to transform the mean vectors of clean speech

models for each condition.

4.2.3 Performances of the Log-Add noise compensation method

Experiments were carried out on the same three testing sets as the previous subsection. Performances of the system compensated by the model-based method are shown in Table 2, which has the same structure as that in Table 1. It is observed that the method is very effective in improving system robustness to additive noise, since the relative performance improvement over the provided baseline system are 61.93% in testing set A and 68.97% in testing set B. As a whole, the method yields a total relative performance improvement of 63.31% on three testing sets. [1] Fig. 4 apparently show that the method can effectively improve system robustness to noise, compared to systems without noise compensation.

5. Conclusions

Evaluations of a sub-band based spectral subtraction, MFCCs derived from Differential Power Spectrum, and a model-based method have been carried out on Aurora 2 task. When multi-condition training set is available, we suggest using the first two methods to improve system robustness to noise. Specifically, MFCCs derived from Differential Power Spectrum can be robust to additive noise and spectrum colorization. When only clean training set is available, we suggest a model-based method, where mean vectors of clean speech models are transformed. Experiments carried out have validated their effectiveness in improving system robustness to noise.

References

[1] H. G. Hirsch and D. Pearce, ``The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", in *ISCA ITRW ASR2000.*, Sep., 2000.

[2] K. Yao, B. E. Shi, S. Nakamura, and Z. Cao, ``Residual noise compensation by a sequential EM algorithm for robust speech recognition in non-stationary noise", in *ICSLP.*, 2000, vol. 1, pp. 770-773.

[3] J. Chen, K. K. Paliwal and S. Nakamura, "Sub-Band Based Additive Noise Removal for Robust Speech Recognition", *in Eurospeech'2001.*

[4] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, 27(2): 112-120, April 1979.

[1] The baseline was trained on clean training set, and therefore, its performance in testing sets were lower than the baseline system in the previous subsection, which was trained on multi-condition training set.

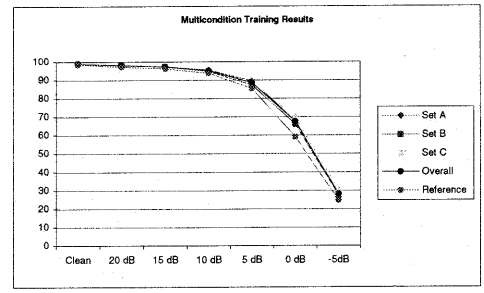

Figure 3. Word Accuracy by MFCC derived from Differential Power Spectrum

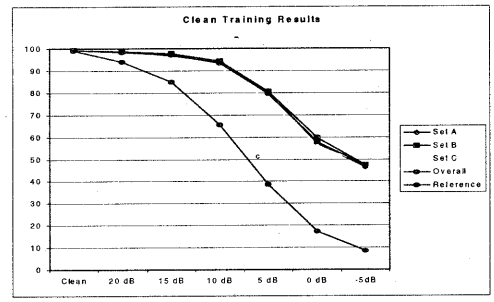

Figure 4. Word Accuracy by the model-based method

[5] J. A. Nolazco Flores, *et al*, "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation", in *ICASSP.*, 1994, vol. 1, pp. 409-412.

[6] J. Chen, K. K. Paliwal, M. Mizumachi, and S. Nakamura, "Robust MFCCs Derived from Differentiated Power Spectrum", *in ICSP'2001.*