Joint Model and Feature Based Compensation for Robust Speech Recognition under Non-stationary Noise Environments

Chuan Jia, Peng Ding and Bo Xu

High Technology Innovation Center; National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080
{cjia,pding,xubo@hitic.ia.ac.cn}

Abstract

This paper presents a novel compensation approach, which is implemented in both model and feature spaces, for non-stationary noise Due to the nature of non-stationary noise which can be decomposed into constant part and residual noise part, our proposed scheme is performed in two steps: before recognition, an extended Jacobian adaptation (JA) is applied to adapt the speech models for the constant part of noise; during recognition, the power spectra of noisy speech are compensated to eliminate the effect of residual noise part of noise. As verified by the experiments performed under different stationary and non-stationary noise environments, the proposed JA is superior to the basic JA and the joint approach is better than the compensation in single space.

1. Introduction

The performance of automatic speech recognizer degrades drastically under the environments mismatched to the quiet training environment. For practical use it is required for recognition systems to work robustly in interfering noise.

Most approaches to compensate for mismatch between training and deployment conditions due to additive noise assume that the noise is stationary. These approaches can be categorized into two types: model based or feature based. The former adapts the acoustic models to any kind of noise and the latter compensates features by noise reduction techniques. Under the non-stationary noise conditions, they can not be used indiscriminately.

Recently many efforts are made to extend the methods to cope with non-stationary noise. For instance, in [1], time-varying noise sources are modeled by Hidden Markov models or Gaussian mixture models that were trained by noise data before model compensation. Then parallel Viterbi decoding is needed to identify the optimum state sequences for both speech and noise. The use of mixture models for noise will result in a cost both in the evaluation of observation likelihoods and in the decoding. In [2], noise model parameters are assumed to be time varying and different parameter estimations are obtained for different frames. After the environmental parameters are estimated, they are used to estimate the clean features. In [3], the noise effects are decomposed into two parts in log-spectral domain: One represents stationary noise effects and the other represents effects from the residual time varying components of the noise. By the linearization of likelihood score, the residual noise can be compensated by computing the residual likelihood score during the recognition procedure.

Generally speaking, model-based compensation performs better than feature-based compensation [4]. However, its computational load is usually heavier than that of feature-based approach. Moreover, feature-based approach is more flexible than model-based approach for non-stationary noise.

To enjoy the merits of both methods, in this paper, we propose to apply a two-step scheme to solve the problems posed by the non-stationary noise. In the first step, model-based approach is applied to adapt the model parameters before recognition to compensate the “global” mismatch statistically. We adopted Jacobian adaptation (JA) [5] in this phase, and extended the algorithm to deal with the situation that the mismatch between reference environment and target environment is large. To retain the low computational cost merit of the original method, noise is still modeled by one component as in stationary noise environments. In the second step, to dynamically compensate the mismatch in the fine details of the temporal continuity, a feature space procedure, named as noise residue removal, is applied during recognition. The power spectra of noisy speech are modified so as to match the adapted models obtained in the first step. By this way, the compensation is implemented in both model and feature spaces.

The remainder of this paper is organized as follows: in order to facilitate the balance of this paper, the experiment settings are introduced firstly in Section 2. Section 3 describes the motivations and framework of the proposed joint compensation method. Section 4 and Section 5 present the extended JA technique and the noise residue removal method followed by corresponding experiment results respectively. Then the experiments and results of the proposed joint approach are shown in section 5. Finally, the conclusions are summarized in Section 6.

2. Experiment settings

The proposed methods described in this paper had been evaluated in large vocabulary continuous speech recognition. One triphone model set for recognition was trained using clean training sets from 85 speakers. A trigram language model was used in all the tests with a 40,000 words vocabulary. The HMM topology was the same as described in [6].

The feature vectors consist of 42 components including 13 cepstral coefficients and the pitch, their first order and second order time derivatives. The zeroth MFCC was retained to enable the inverse cosine transformation.

To investigate the performance of the proposed algorithms, two kinds of noise were employed in this work: stationary noise such as white and pink noise, non-stationary noise such as factory, babble and F16 noise etc. Noise was added into 240 clean testing utterances by varying the signal-to-noise ratio (SNR). The lengths of the testing utterances range from 5

This work was supported by the National Key Fundamental Research Program of China (the 973 Program) under the grant G19980300504.

seconds to 10 seconds. The recognition accuracy for the clean test set with the speech models trained by clean speech is 88.6%.

3. Motivations and framework

At the *t*th frame, power spectrum of noisy speech is represented as follows:

$$Y(t) = X(t) + N(t) \tag{1}$$

where *X*(*t*) and *N*(*t*) denote the power spectra of clean speech and noise, respectively. Inspired by [3], we decompose the power spectrum of noise into constant part (mean spectrum) N_0 and residual part $\Delta N(t)$:

$$N(t) = N_0 + \Delta N(t) \tag{2}$$

and in the log-spectral domain, the noisy speech is represented by

$$y(t) = x(t) + \log(1 + \exp(n(t) - x(t))) \tag{3}$$

where *y*(*t*), *x*(*t*) and *n*(*t*) represent the log-spectra of noisy speech, clean speech and noise respectively.

Then (3) is rewritten by introducing (2) as follows:

$$y(t) = \underbrace{x(t) + \log(1 + \exp(n_0 - x(t)))}_{part1} + \underbrace{\log(\frac{X(t) + N_0 + \Delta N}{X(t) + N_0})}_{part2} \tag{4}$$

where n_0 is log-spectrum of constant part N_0 of noise. Clearly, part one corresponds to the situation that the clean speech is corrupted by stationary noise N_0, which can be compensated by model adaptation techniques such as JA [5] or PMC [7]. Part two includes the fine details $\Delta N(t)$ of the non-stationary noise that requires the dynamical compensation. Ideally, this part will equal to zero if the residual part of noise $\Delta N(t)$ is eliminated for each frame *t*.

Therefore, a two-step framework, consisting of extended JA, named virtual Jacobian adaptation, followed by noise residue removal, is proposed to deal with the effect of constant and residual part of noise explained in (4), respectively. These two steps are detailed in the following sections.

4. Virtual Jacobian adaptation in model space

4.1. Virtual Jacobian adaptation

Jacobian adaptation [5] is proposed as an analytic approach to adapt an initial acoustic model under reference condition to a target condition. The underlying idea is to express the change in the noisy speech model given that the reference condition changes towards the target condition.

Let C_{S+Nr}, C_{S+Nt}, C_{Nr} and C_{Nt} denote cepstra of noisy speech under reference condition, noisy speech under target condition, reference noise and target noise, respectively. The relation between noisy speech cepstra is as follows:

$$C_{S+N_t} = C_{S+N_r} + J_r \cdot (C_{N_t} - C_{N_r}) \tag{5}$$

where $J_r = F \cdot diag\,(\frac{N_r}{S + N_r}) \cdot F^*$ denotes the Jacobian matrix for the reference noise, and F and F^* denote the discrete cosine transformation and its inverse, respectively.

However, if the difference between the reference and the target noise is not small enough to guarantee that the change in noisy speech model stays within the linear range of JA, the method does not works well. An extended JA named Virtual JA (VJA), which is an extension of [8], is proposed here to break this limit. Virtual JA assumes that a virtual intermediate noise condition N_m between N_r and N_t exists with which it is possible that the non-linearity of the changes in noisy speech models both from reference noise to intermediate noise and from intermediate noise to target noise is alleviated. Then, the cepstra of noisy speech under target condition may be presented as follows:

$$C_{S+N_t} = C_{S+N_r} + J_r \cdot (C_{N_m} - C_{N_r}) + J_m \cdot (C_{N_t} - C_{N_m}) \tag{6}$$

where $J_m = F \cdot diag\,(\frac{N_m}{S + N_m}) \cdot F^* \cdot$

4.2. Assumption of the virtual intermediate noise condition

It is optimal to make the noisy speech cepstra in (6) approach the exact noisy speech cepstra under target condition. Thus the optimal N_m verifies:

$$C_{S+N_r} + J_r(C_{N_m} - C_{N_r}) + J_m(C_{N_t} - C_{N_m}) = F \cdot \log(S + N_t) \tag{7}$$

Obviously it is not easy to obtain the optimal N_m. We can experiment with some solutions. For example, it is assumed that the J_m and C_{Nm} can be obtained alternately and iteratively or the N_m is weight sum of N_t and N_r. In order to keep the low cost of JA and simplicity, we suppose that N_m satisfy:

$$N_m = \alpha \cdot N_t \tag{8}$$

where coefficient α is within 0 and 1. Then the mean of cepstra of N_m is obtained by:

$$\mu_{N_m} = \alpha \cdot \mu_{N_t}\,, \quad \Sigma_{N_m} = \alpha^2 \cdot \Sigma_{N_t} \tag{9}$$

$$\mu_{C_{N_m}} = \log \mu_{N_m} - 0.5 \cdot \log(\Sigma_{N_m} / \mu^2_{N_m} + 1) \tag{10}$$

For recognition, means of speech distributions are adapted by the proposed VJA and variances are still adapted by JA.

Once the coefficient α is determined, the computational cost of VJA for adapting means of speech distributions doubles than that of JA, but is still less than that of PMC.

4.3. Implementation issues

The HMMs of the reference noise environment was trained using the training sets to which the white noise was added such that the resulting SNR was 40db. From these HMMs, the vector S+N_r were obtained for each mixture component of the HMMs, which were computed by transforming cepstra from the static part of the Gaussian mean vectors into mel-power spectra. The noise vector N_r and C_{Nr} were estimated by averaging the mel-power spectra and cepstra from the beginning segments of 100 utterances in the training sets corrupted with white noise. The Jacobian matrix J_r was computed before recognition.

For each testing utterance, the means of C_{Nt} and N_t were obtained from the beginning of the utterance for the target noise. The means of N_m and C_{Nm} were estimated by equations (8)-(10) for the virtual intermediate noise environment and another Jacobian matrix J_m was obtained. The means and variances of static cepstra in speech models were adapted in all of the following experiments.

Care must be taken to select α to assure good performance of VJA. In this paper, the coefficient α in equation (8) was determined on experiment basis. Fig. 1 presents the recognition results for the different target noise environments with different α. For example, the lowest solid line represents the recognition accuracy for testing set corrupted by white noise and target SNR 5db. The abscissa denotes the virtual SNR of the testing set corrupted by virtual intermediate noise generated by different α . Fig.1 shows that the recognition accuracy was maximized when the coefficient α is set to make the SNR of the intermediate noise condition approximatively 5-10db higher than the SNR of the target noise environments.

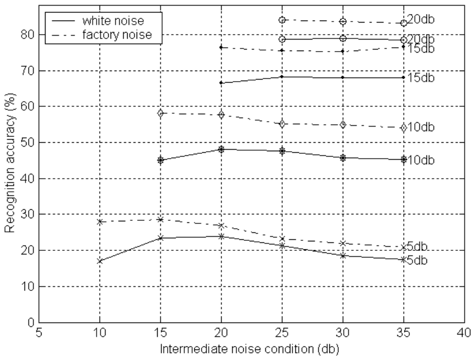

Figure 1: the relationship between recognition accuracy with different α

4.4. Experiments for stationary noise

The task of this experiment is to compare the VJA with JA and PMC under stationary noise. We applied the testing set with white noise. In this table, results are shown for no adaptation for HMM (NA), adaptation with PMC, JA and the proposed VJA. All the adaptation was only applied on the static part of the cepstral means and variances. In VJA, the coefficient α was selected to make the SNR of the intermediate noise condition 10db higher than the target SNR.

Table 1: recognition accuracy (%)

SNR(db)	NA	JA	VJA	PMC
20	68.28	77.86	78.98	78.96
15	37.13	61.58	68.23	71.42
10	18.88	36.34	48.10	58.37
5	6.80	9.69	23.91	34.27

It can be seen in Table 1 that along with the enlargement of the mismatch between target noise environment and reference noise environment the performance of JA deteriorated. However, VJA improved the performance of JA especially under lower SNR environment. The results showed the validity of virtual intermediate noise environment in VJA for stationary noise. PMC outperformed JA and VJA with higher computational cost.

5. Noise residual removal in feature space

5.1. Noise residual removal

The part two of (4) indicates the reason why the performance of model compensation degraded in non-stationary noise environments. In order to match the non-stationary noise, model compensation can be applied dynamically. However, the computational cost is heavy. On the other hand, processing in feature space is rather plastic for non-stationary noise with lower computational cost. According to (4), during recognition the removal of noise residual from power spectra of noisy speech results in matching better with the compensated models obtained by model adaptation. The power spectrum of noisy speech is modified as follows:

$$\hat{Y}_t = \begin{cases} Y_t - \Delta \ddot{N}_t = Y_t - \beta \cdot (\ddot{N}_t - N_0) & \text{if} \quad \hat{N}_t > N_0 \\ Y_t & \text{else} \end{cases} \tag{11}$$

where $\hat{y}_t$ is the modified power spectrum of noisy speech and $\hat{N}_t$ is the estimated noise power spectrum. Theoretically speaking, the additive relationship among the power spectra of noisy speech, clean speech and noise is not strictly held. In our practice, β is introduced to consider the error.

The validity of this step combined with model adaptation techniques for non-stationary noise is verified by the following experiment. In Table 2 it is assumed that the noise power spectra are known in advance and SNR is about 8 db.

Table 2: recognition accuracy (%)

Noise type	F16	Babble	White+ Babble	Factory
SS	39.93	40.41	37.61	37.78
JA	45.11	44.26	40.41	39.68
JA+Ideal	46.63	45.91	43.72	42.33

In Table 2, SS denotes spectral subtraction with known noise power spectra. JA+Ideal refers to adopt the noise residual removal technique during recognition after JA. Obviously, JA+Ideal is superior to JA and SS. As we expected, the recognition results indicated that the joint processing in both model and feature spaces does better for non-stationary noise.

Generalized from the above ideal situation, the accuracy of the estimated noise power spectra is the key factor for this step. In this paper, we use the sequential estimation algorithm that is based on the Kullback-Leibler information measure [9] in power spectral domain to track the non-stationary noise environments.

5.2. Estimation of noise power spectra

In the work described in this paper, the noise spectrum is assumed to be a deterministic and time-varying vector. The estimation of noise spectrum is based on Maximum likelihood (ML) criterion as follows:

$$\hat{N}_{t+1} = \underset{\hat{N}}{\arg\max}\, p(Y_1^{t+1} \mid \hat{N}, \Lambda_X, \hat{N}_1^t) \tag{12}$$

where Y_1^{t+1} represents the sequence of power spectra of noisy speech $\{Y_1, Y_2, \ldots, Y_{t+1}\}$, $\hat{N}_1^t$ is the sequence of estimated noise power spectra $\{\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_t\}$ and Λ_X is the set of clean speech models. We assume that the power spectral space of clean speech is represented by N Gaussian mixture models, e.g. Λ_X, and each model has M components with mixture coefficients, means and diagonal covariance matrices $\{w_{n,m}^x, \mu_{n,m}^x, \Sigma_{n,m}^x, 1 \le n \le N, 1 \le m \le M\}$.

The objective function above is optimized indirectly using the EM algorithm. The ML auxiliary function can be expressed out as

$$Q_{ML,t+1}(\hat{N} \mid \hat{N}_t) = E\{\log p(Y_1^{t+1}, S_1^{t+1}, C_1^{t+1} \mid \hat{N}) \mid \hat{N}_1^t, \Lambda_X, Y_1^{t+1}\} \tag{13}$$

where $S_1^{t+1} = \{s_1, s_2, \ldots s_{t+1}\}$ be the sequence of state indices, $C_1^{t+1} = \{c_1, c_2, \ldots c_{t+1}\}$ be the sequence of indices of mixture components in the clean speech models.

In the E-step, the auxiliary function is simplified as

$$Q_{ML,t+1}(\hat{N} \mid \hat{N}_1^t) = \sum_{\tau=1}^{t+1} \xi^{t+1-\tau} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{\tau|t+1,\hat{N}_1^{\tau-1}}(n,m) \{ - \frac{(Y_\tau - \hat{N} - \mu_{n,m}^x)^2}{2\Sigma_{n,m}^x} \} \quad (14)$$

In (14), ξ is the forgetting factor, which is to reduce the effect of past data to the new input data, and $\gamma_{\tau|t+1,\hat{N}_1^{\tau-1}}(n,m) = p(s_\tau = n, c_\tau = m \mid Y_1^{t+1}, \hat{N}_1^{\tau-1}, \Lambda_X) \cdot$

To carry out the M-step, we can use second-order Taylor series expansion and the Newton-Raphson technique [9] to sequentially estimate the noise power spectra via the following recursive form

$$\hat{N}_{t+1} = \hat{N}_t + F_{t+1}^{-1} \cdot S_{t+1} \quad (15)$$

where the fisher information item F_{t+1} and score item S_{t+1} are as follows

$$F_{t+1} = \xi \cdot F_t + \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \gamma(n,m) \cdot \frac{1}{\Sigma_{n,m}^x} \quad (16)$$

$$S_{t+1} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \gamma(n,m) \cdot \frac{(Y_{t+1} - \hat{N}_t - \mu_{n,m}^x)}{\Sigma_{n,m}^x} \quad (17)$$

The estimation process described here was adopted for each testing utterance during recognition.

Concretely speaking, the output energies of Mel-scaled filterbank were used to replace the power spectra due to its lower dimension and the still kept linear relationship among mel-power spectra of noisy speech, noise and original speech. The details of the estimation process can refer to [10].

6. Experiments and results

The objective of the experiments in this section is to investigate the performance of the proposed joint algorithm under non-stationary noise environments.

The recognition results for testing sets with additive babble noise and F16 noise are shown in Table 3 and 4. In these tables, JA+Nrr denotes combining JA and noise residual removal; JA+Ideal denotes the same process as in JA+Nrr, except that the noise power spectra were assumed to be known in advance. VJA+Nrr represents the combination of VJA and noise residual removal. The coefficient β defined in (11) was set as 0.8. The forgetting factor ξ in (16) was set at 0.8 and 0.9 under babble and F16 noise environment respectively.

It can be concluded from table 3, 4 that the VJA outperformed JA, as had been shown in table 1, especially under low SNR. Moreover, comparison between JA+Nrr and JA+Ideal shows the efficacy of the estimation process of noise power described in 5.2. Compared with JA and VJA, the corresponding joint compensation approaches such as JA+Nrr, VJA+Nrr can obtain better performance. The results of JA+Nrr and VJA+Nrr illustrated that the processing in feature space can improve the recognition performance obtained by model-based compensation for non-stationary noise and VJA+Nrr achieved the best results, which verify the validity of the proposed joint compensation approach for non-stationary noise environments.

7. Conclusions

In this paper, two new techniques are proposed to improve the recognition performance under noise environments, especially non-stationary noise. Virtual Jacobian adaptation outperforms the Jacobian adaptation if the mismatch between reference noise environment and target noise environment is large. Noise residual removal technique makes the speech signal contaminated by non-stationary noise match the adapted HMMs better than before. The joint compensation by Virtual Jacobian adaptation and noise residual removal demonstrates that the advantages of joint compensation both in model space and feature space for non-stationary noise environments.

Table 3: recognition accuracy with babble noise (%)

SNR(db)	NA	JA	VJA	JA+Nrr	JA+Ideal	VJA+Nrr
20	83.13	83.56	83.81	83.66	84.54	84.22
15	70.14	73.44	76.02	74.02	75.26	76.66
10	42.28	46.16	55.29	48.52	48.74	56.59
5	17.17	21.56	27.28	22.70	22.89	28.58

Table 4: recognition accuracy with F16 noise (%)

SNR(db)	NA	JA	VJA	JA+Nrr	JA+Ideal	VJA+Nrr
20	81.90	83.40	82.92	83.49	83.62	83.38
15	68.48	72.68	73.29	73.10	73.57	76.12
10	42.38	51.09	53.00	52.20	52.30	54.78
5	12.94	18.21	24.13	19.93	20.60	25.46

8. References

[1] T.Kristjansson, B.Frey, L.Deng and A.Acero, "Towards non-stationary model-based noise adaptation for large vocabulary speech recognition", In *Proceeding of ICASSP*, 2001.

[2] Nam Soo Kim, "Nonstationary environment compensation based on sequential estimation", *IEEE signal processing letters*, Vol.5, No.3, March 1998.

[3] Kaisheng Yao, Bertram E.SHI, Satoshi NAKAMURA and Zhigang Cao, "Residual noise compensation by a sequential EM algorithm for robust speech recognition in nonstationary noise", In *Proceeding of ICSLP* 2000.

[4] A.Sankar, L.Neumeyer and M.Weintraub, "An experimental study of acoustic adaptation algorithms", In *Proceeding of ICASSP*, 1996.

[5] Shigeki Sagayama, Yoshikazu Yamaguchi, Satoshi Takahashi and Jun-ichi Takahashi, "Jacobian approach to fast acoustic model adaptation", In *Proceeding of ICASSP,* 1997.

[6] Gao sheng, Tan Lee, Y.W. Wong, Bo Xu, P.C. Ching and Taiyi Huang, "Acoustic modeling for Chinese speech recognition: a comparative study of Mandarin and Cantonese", In *Proceeding of ICASSP*, 2000.

[7] M.Gales, "Predictive model-based compensation schemes for robust speech recognition ", *Speech communication*, vol.25, pp.49-74, 1998.

[8] Kimmo Pärssinen, Petri Salmela, Mikko Harju and Imre Kiss, "Comparing Jacobian adaptation with cepstral mean normalization and parrel model combination for noise robust speech recognition ", In *Proceeding of ICASSP*, 2002.

[9] Vikram Krishnamurthy, John B. Morre, "On-line estimation of Hidden Markov Model parameters based on the Kullback-Leibler information measure", *IEEE Trans. on Signal Processing*, pp.2557-2573, vol. 41, no. 8, 1993.

[10] Chuan Jia, Peng Ding and Bo Xu, "Sequential MAP estimation based speech feature enhancement for noise robust speech recognition", In *Proceeding of ICASSP*, 2003.