AN RNN-BASED COMPENSATION METHOD FOR MANDARIN TELEPHONE SPEECH RECOGNITION

Sen-Chia Chang, Shih-Chieh Chien, and Chih-Chung Kuo
E000/CCL, Building 51, Industrial Technology Research Institute,
Chutung, Hsinchu, Taiwan, ROC, 310.
saga1@atc.ccl.itri.org.tw

ABSTRACT

In this paper, a novel architecture, which integrates the recurrent neural network (RNN) based compensation process and the hidden Markov model (HMM) based speech recognition process into a unified framework, is proposed. The RNN is employed to estimate the additive bias, which represents the telephone channel effect, in the cepstral domain. Compensation of telephone channel effects is implemented by subtracting the additive bias from the cepstral coefficients of the input utterance. The integrated recognition system is trained based upon MCE/GPD (minimum classification error/generalized probabilistic descent) method with an objective function that is designed to minimize recognition error rates. Experimental results for speaker-independent Mandarin polysyllabic word recognition show an error rate reduction of 21.5% compared to the baseline system.

1. INTRODUCTION

It is generally agreed that the performance of automatic speech recognition systems often degrades due to a mismatch between the training and testing environments. In applications of speech recognition over the telephone network, the distortion introduced by the telephone channel (including both handset microphone and the transmission line) is one of the major sources of mismatching conditions. For robust telephone speech recognition, it is essentially important to remove the telephone channel effects.

In general, the telephone channel characteristics can be approximated with a linear time invariant system, $h(\tau)$, for a given time interval. Based on this assumption, the speech signal transmitted through the telephone channel is a convolution of $h(\tau)$ and the input speech signal. In the cepstral domain, this convolved component becomes an additive bias. It varies with each call and increases the variability in the observed feature space. In recent years, many compensation approaches have been developed to remove the additive biases from the observed feature vectors for robust telephone speech recognition [1][2][3][4][5]. Most of them were implemented by spectral parameter filtering in the front-end feature analysis and unrelated to the recognition process.

In this paper, a novel architecture, which integrates the RNN-based compensation process and the HMM-based speech recognition process into a unified framework, is proposed. The RNN is employed to estimate the additive bias. Compensation of telephone channel effects is implemented by subtracting the additive bias from the cepstral coefficients of the input utterance. The integrated recognition system is trained based upon MCE/GPD method with an objective function that is designed to minimize recognition error rate.

This paper is organized as follows. In section 2, the RNN-based compensation of telephone channel effects is discussed. In section 3, a speaker-independent Mandarin polysyllabic word recognition is conducted to evaluate the RNN-based compensation method. Finally, a conclusion is given in section 4.

2. RNN-BASED COMPENSATION OF TELEPHONE CHANNEL EFFECTS

Previous studies showed that the interaction between feature extraction and classification strongly affects speech recognition performance [6][7]. Aiming to improve the performance of speech recognition systems, MCE/GPD [10][11] has been used to the design of the feature extractor. All of the adjustable parameters of both feature extraction and speech recognition processes are

trained with a consistent objective function that is designed to minimize recognition error rate [7][8][9]. Since our proposed RNN-based compensation scheme is employed to remove the additive biases, it can be regarded as a part of the feature extraction. Figure 1 is a block diagram of showing the integration of RNN-based compensation scheme and HMM-based recognizer. Under a unified framework using the MCE/GPD training procedure, the goal is to adjust modifiable parameters of the RNN, Θ, and the HMMs, Λ, so as to minimize the recognition error rate.

Before simultaneously adjusting all parameters of the integrated recognition system by MCE/GPD, the RNN and the HMMs are trained individually to give good initialization. The error back-propagation algorithm [12] is used to train the RNN. The widely used Minimum Squared Error (MSE) criterion is adopted as the learning criterion. Because the RNN is used to estimate the additive bias, a linear activation function is used in the output nodes. The desired output target is set to be the average of the cepstral vectors of all training tokens for a given call. The assignment of the desired output target is according to the following two factors: (1) telephone channel effects are almost constant for a given call but vary with the calls; (2) the average of cepstral coefficients on a long period of speech signals is a reliable estimate of the additive bias. After the RNN has been trained, each training utterance is bias removed by applying RNN-based compensation process and then used in the maximum likelihood (ML) based HMM estimation.

Let $\mathbf{O} = \{o_1, \ldots, o_T\}$ and $\hat{\mathbf{O}} = \{\hat{o}_1, \ldots, \hat{o}_T\}$ be the observation sequences of T frames before and after bias removing by the RNN-based compensation process, respectively. $\hat{\mathbf{O}}$ can then be expressed as a transformation of $\mathbf{O}$ with parameters Θ

$$\begin{aligned} \hat{\mathbf{O}} &= F(\mathbf{O};\Theta) \\ &= \mathbf{O} - \overline{B} \end{aligned} \tag{1}$$

where $\overline{B}$ is the bias vector obtained by averaging output vectors of the RNN. Assume that $\mathbf{O}$ belongs to the *i*th class of M classes, then the objective in MCE/GPD training is to reduce the expected loss

$$L(\Phi) = E[l_i(\mathbf{O};\Phi)] \tag{2}$$

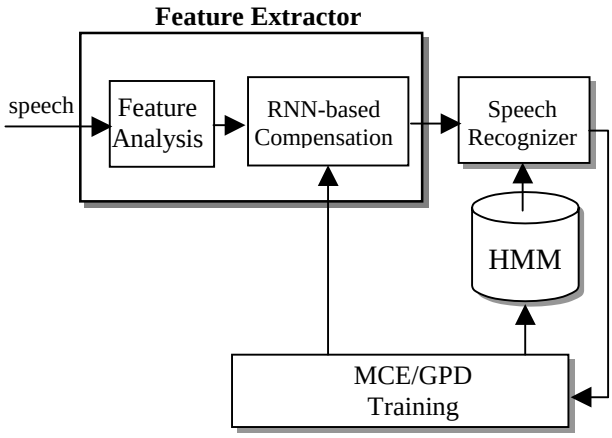

Figure 1: A block diagram of showing the integration of RNN-based compensation and HMM-based recognizer.

$\Phi = \{\Theta, \Lambda\}$ and $l_i\{\bullet\}$ is a loss function which has a form

$$\begin{aligned} l_i(\mathbf{O};\Phi) &= l(d_i(\mathbf{O})) \\ &= \frac{1}{1+\exp(-ad_i(\mathbf{O})+b)} \end{aligned} \tag{3}$$

where d_i is a class misclassification measure taking the following form:

$$\begin{aligned} d_i(\mathbf{O}) &= -g_i(\mathbf{O};\Phi) + \log\left[\frac{1}{M-1}\sum_{j,j\neq i}^{M}\exp\left[g_j(\mathbf{O};\Phi)\eta\right]\right]^{1/\eta} \\ &= -g_i(\mathbf{O};\Phi) + G_i(\mathbf{O};\Phi) \end{aligned} \tag{4}$$

with η is a positive number.

When segmental GPD [11] is applied to update Λ and Θ, $g_i(\mathbf{O};\Phi)$ can be defined as the log-likelihood of the optimal state sequence $\overline{\mathbf{q}}$

$$\begin{aligned} g_i(\mathbf{O};\Phi) &= \log\{g_i(\mathbf{O},\overline{\mathbf{q}};\Phi)\} \\ &= \log\{g_i(\hat{\mathbf{O}},\overline{\mathbf{q}};\Lambda)\} \end{aligned} \tag{5}$$

The loss function $L(\Phi)$ is minimized according to an iterative step

$$\Phi_{n+1} = \Phi_n - \varepsilon_n U_n \nabla l(d_i(\mathbf{O};\Phi))\Big|_{\Phi=\Phi_n} \tag{6}$$

where ε_n is a learning rate and U_n is a positive define matrix. Any parameter of Θ and Λ can be update according to Eqn. (6). In our work, the gradient can be written as the partial derivative

$$\frac{\partial l_i}{\partial \Phi} = \frac{\partial l_i}{\partial d_i}\left(\frac{\partial d_i}{\partial g_i}\frac{\partial g_i}{\partial \Phi} + \frac{\partial d_i}{\partial G_i}\frac{\partial G_i}{\partial \Phi}\right) \quad (7)$$

The derivations for updating parameters of Λ can be found in [11]. When each component of the observation vector is treated independently, adjustment of parameters of the RNN is as follows

$$\frac{\partial g_i}{\partial \Theta} = \sum_{t=1}^{T}\sum_{d=1}^{D} E_d \frac{\partial f_d(o_t)}{\partial \Theta} \quad (8)$$

where D is the dimension of the observation vector, $f_d(o_t)$ is the output of the *d*th node of the RNN with input o_t, and E_d has the following form

$$E_d = \frac{1}{T}\sum_{t=1}^{T} e_d(t) \quad (9)$$

with

$$e_d(t) = \sum_{k=1}^{K} \gamma_t(\bar{\mathbf{q}}_t, k)\frac{(\hat{o}_{t,d} - \mu_{\bar{\mathbf{q}}_t,k,d})}{\sigma^2_{\bar{\mathbf{q}}_t,k,d}} \quad (10)$$

is the error term propagating back to the RNN. K is the mixture number, μ and σ are the mean and variance of a Gaussian distribution, respectively.

3. EXPERIMENTAL RESULTS

Simulations on a speaker-independent Mandarin polysyllabic word recognition task are performed to examine efficiency of the proposed method. The vocabulary is 1038. The length of the vocabulary is ranging from two to four syllables. A database collected from 362 speakers calling from different regions of Taiwan is used for simulations. A subset of this database consisting of 7674 utterances spoken by 292 speakers is assigned for training and another subset of 1892 utterances spoken by 70 speakers is assigned for testing. All speech signals were sampled at a rate of 8 kHz and preemphasized with a digital filter, $1-0.95z^{-1}$. It was then analyzed for each Hamming-windowed frame of 20 ms with 10 ms frame shift. The recognition features consist of 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, the delta energy, and the delta-delta energy.

The HMM-based speech recognizer employed 138 sub-syllable models as basic recognition units, including 100 three-state right-context-dependent INITIAL models and 38 five-state context-independent FINAL models. The observation distribution for each state of the HMM was modeled by a multivariate Gaussian mixture distribution. The number of mixture components in each state varies from one to ten depending on the amount of training data, and each of the mixture components has a diagonal covariance matrix. For silence, a single-state model with ten mixtures is used. A three-layer RNN, which feeds back all outputs of its hidden layer to the input layer, is used to estimate the additive bias. In all experiments, the number of hidden nodes is set to be 100.

At first, a test using ML-trained recognition models without imposing any compensation scheme was performed and taken as a benchmark. A recognition rate of 87.0% was achieved. Next, three widely used compensation techniques, namely cepstrum mean normalization (CMN), relative spectral (RASTA) methodology, and signal bias removal (SBR), were tested for comparison. The recognition rates are summarized in Table 1. The results demonstrate that removing the telephone channel effects significantly improves the recognition performance. Thirdly, the ML-trained HMMs of the recognition system with CMN are further fine-tuned by MCE/GPD method. A recognition rate of 89.1% was obtained. It showes the powerfulness of MCE/GPD training method.

Finally, the proposed RNN-based compensation scheme was tested to demonstrate its efficiency. Prior to MCE/GPD training, the RNN is trained independently by using MSE criterion. A recognition rate of 88.9% was achieved when applying the RNN to remove the additive biases. As shown in Table 1, it is superior to the performances of CMN, RASTA and SBR. In the case of simultaneously training the RNN and the HMMs by MCE/GDP method, the recognition rate was up to 89.8%. It is correspond to an error rate reduction of 21.5%s compared to the baseline system.

4. CONCLUSION

The work presented in this paper is devoted to the problem of compensation of telephone channel effects for robust telephone speech recognition. The

speech *Method*	*Recognition Rate (%)*	*Error Reduction Rate (%)*
Baseline	87.0	-
CMN	87.8	6.2
RASTA	87.7	5.4
SBR	88.0	7.7
CMN+MCE	89.1	16.1
RNN	88.9	14.6
RNN+MCE	89.8	21.5

Table 1: Recognition results of various compensation approaches.

telephone channel introduces a convolved component in the observed speech signals. This convolved component becomes a additive bias in the cepstral domain. We use an RNN to estimate the additive bias from the cepstral coefficients of the input utterance. Aiming to improve the performance of the speech recognition systems, a novel architecture, which integrates the RNN-based compensation process and the HMM-based speech recognition process into a unified framework, is proposed. The integrated recognition system is trained based upon MCE/GPD method with an objective function that is designed to minimize recognition error rate. Experimental results for speaker-independent Mandarin polysyllabic word recognition show an error rate reduction of 21.5% compared to the baseline system.

5. ACKNOWLEDGMENT

This paper is a partial result of the project No. 3P11100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.
The authors thank ROC Computational Science Council Society in Taiwan for kindly supplying the database.

6. REFERENCES

[1] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," *Proc. EUROSPEECH'91*, pp. 1367-1370, 1991.

[2] M. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing,* Vol. 4, pp. 19-30, 1996.

[3] C. Mokbel, P. Paches-Leal, D. Jouvet and J. Monne, "Compensation of telephone line effects for robust speech recognition," *ICSLP*-94, pp. 987-990, 1994.

[4] C. Mokbel, D. Jouvet and J. Monne, "Deconvolution of telephone line effects for speech recognition," *Speech Communication* 19(1996) 185-196.

[5] M. Westphal, "The use of cepstral means in conversational speech recognition," *Proc. EUROSPEECH'97*, pp. 1143-1146, 1997.

[6] H. Leung, B. Chigier, and J. Glass, "A comparative study on signal representation and classification techniques for speech recognition," *Proc. ICASSP-93*, pp. 680-683, 1993.

[7] A. Biem, S. Katagiri, and B.-H. Juang, "Discriminative feature extraction for speech recognition," *Proc. 1993 IEEE Workshop on Neural Networks for Signal Processing*, pp. 392-401, Sept. 1993.

[8] M. G. Rahim and C.-H. Lee, "Simultaneous ANN feature and HMM recognizer design using string-based minimum classification error (MCE) training," *Proc. ICSLP-96*, pp. 1824-1827, 1996.

[9] M. Rahim, Y. Bengio and Y. LeCun, "Discriminative feature and model design for automatic speech recognition," *Proc. EUROSPEECH'97*, pp. 75-78, 1997.

[10] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, SP-40, no. 12, pp. 3043-3054, 1992.

[11] W. Chou, C.-H. Lee and B.-H. Juang, "Segmental GPD training of an hidden Markov model based speech recignizer," *Proc. ICASSP-92*, pp. 473-476, 1992.

[12] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proceeding, IEEE*, vol. 78, pp. 1550-1560.