A DISCRIMINATIVE AND ROBUST TRAINING ALGORITHM FOR NOISY SPEECH RECOGNITION

Wei-Tyng Hong [1]

Industrial Technology Research Institute, Taiwan

ABSTRACT

A combined technique of discriminative and robust training algorithms, referred to as the D-REST (Discriminative and Robust Environment-effects Suppression Training), is proposed for noisy speech recognition. The D-REST technique can separately model the environmental characteristics and phonetic information and thus it can train speech models discriminatively on phonetic variability by eliminating the disturbance of environment-specific effects. According to the experimental results of Taiwan stock name recognition task over wireless network, the proposed D-REST algorithm has the potential to improve performance not only on diverse training data but also on noise-type unmatched environments between training and testing. Furthermore, the usage of the D-REST algorithm amounted to a 60% reduction in average word error rate over the performance by the conventional MCE/GPD-based training approach without environment-effects suppression training technique.

1. INTRODUCTION

This paper is concerned with the robustness issues of applying discriminative training algorithm to adverse speech recognition. Discriminative training techniques [1, 2, 3] have been widely used recently in attempting to improve the discrimination capabilities for increasing the recognition accuracy on a given vocabulary. These techniques have been shown to give significant improvement for clean speech or speech in homogeneous environment. However, there are few reports [4] discussing with the issues of applying discriminative training techniques on noisy data. The improvement of directly applying discriminative training techniques on diverse and noisy data could be insignificant because that the noisy speech are deeply influenced by environment-specific characteristics. Another concern of the discriminative training on noisy data is the generalization or over-fitting issue. The discriminative training process might adapt the models to the specific environmental characteristics of noisy training data; and it results in degrading performance on testing speech in other environment.

An extended version of the REST (Robust Environment-effects Suppression Training) algorithm [5, 6], referred to as the D-REST (Discriminative REST), is proposed in this paper to enhance both the discrimination and robust capabilities for adverse speech recognition. Its main function is to train a set of compact speech models directly from diverse training data by a separate modeling technique focused on environmental characteristics and phonetic information. The compact speech models are used as the seed models for noisy speech recognition with model compensation-based approach. The diverse training data are collected in open environments such as the office environment, department and streets. For example, a training data set collected through the wireless network will suffer diverse recording conditions caused by different background noises, different types of hand-held and hands-free transducers, various wireless channels, etc. This will make speech patterns distribute more widely in the feature space so as to overlap to each other more seriously and cause the trained speech models degrade on their discrimination capabilities.

The D-REST algorithm is applied to generate of a set of environment-effects suppressed HMMs discriminatively and directly from training data suffering with both channel bias and additive noise. Its principle is to make the training process discriminatively emphasize the modeling on phonetic variation instead of taking into account disturbed noise from environment-effects. This is achieved from separate modeling of different effects (e.g., the phonetic variability and environment-effects) embedded in training speech signals. The design goal of the D-REST algorithm is threefold. The first is to countervail the large variability of the corrupted training samples for obtaining a set of compact speech HMMs with both signal bias and noise being suppressed. The second is to make the compact speech HMMs better for a given robust speech recognition method. The other is to enhance the generalization capability of discriminative training process on noisy data. This paper is organized as follows. The proposed algorithm is presented in section 2. In section 3, we analyze the experimental results. In section 4, some conclusions are drawn.

2. THE D-REST ALGORITHM

The D-REST algorithm is derived based on a presumed noisy speech realization model in which we assume that the observed speech feature vector sequence of the r-th utterance $Z^{(r)}$ is generated from the corruption of the homogeneous and clean speech $X^{(r)}$. Consider the set of discriminant functions $\{g_i, i=1,2,..,M\}$ with the environment-compensated speech HMM models $\Lambda_z^{(r)}$ of $Z^{(r)}$ defined by

$$g_i(Z^{(r)};\Lambda_z^{(r)}) \equiv \log\left[\Pr\left(Z^{(r)},U_i^{(r)}\middle|\Lambda_z^{(r)}\right)\right] = \log\left[\Pr\left(Z^{(r)},U_i^{(r)}\middle|\Lambda_x\otimes\Lambda_e\right)\right], \quad (1)$$

where $U_i^{(r)}$ is the maximum likelihood state sequence of $Z^{(r)}$ with the ith HMM of $\Lambda_z^{(r)}$ by Viterbi algorithm; Λ_x denote the set of environment-effect suppressed HMMs (i.e., the compact model)

[1] The author is now with PenPower Technology Ltd. (email: jfhong@seed.net.tw)

and Λ_e is the set of environmental interference models. The symbol $\otimes$ denotes the operator of model compensation which is also employed in its recognition process. The target of the D-REST is to estimate Λ_x and Λ_e with the set of discriminant functions g_i, and to make Λ_x as a robust and discriminative seed model for model compensation-based noisy speech recognition.

The first stage of the D-REST algorithm is to jointly estimate the compact speech models and environmental interference models. Assume that the environment-effects comprise a convolutional channel b and an additive noise n on each utterance. Let $\Lambda_e \equiv \{\Lambda_n^{(r)}, b^{(r)}\}_{r=1,\cdots,R}$ denote the set of environmental interference models of the whole training data set, where $b^{(r)}$ and $\Lambda_n^{(r)}$ are, respectively, the signal bias and the noise model of the r-th training utterance. Based on the ML (maximum likelihood) criterion, the goal is to jointly estimate Λ_x and Λ_e with given $\{Z^{(r)}\}_{r=1,\cdots,R}$ by

$$(\Lambda_x, \Lambda_e) = \arg\max_{(\overline{\Lambda}_x, \overline{\Lambda}_e)} \Pr\left(\{Z^{(r)}\}_{r=1,\cdots,R} \middle| \overline{\Lambda}_x, \overline{\Lambda}_e\right). \quad (2)$$

An iterative training procedure, referred to as the REST technique was proposed in [6] to sequentially optimize Eq. (1). It includes the following three operations: (1) Form the compensated HMMs $\Lambda_z^{(r)}$ by using the current estimate $\{\Lambda_x, \Lambda_e\}$ and use it to optimally segment the training utterance $Z^{(r)}$; (2) Based on the segmentation result, estimate $\Lambda_n^{(r)}$ and enhance the adverse speech $Z^{(r)}$ to obtain $Y^{(r)}$ by the state-based Wiener filtering method [7]; and then, estimate $b^{(r)}$ and further enhance the speech $Y^{(r)}$ to obtain $X^{(r)}$ by the SBR method [8]; (3) Update the current speech HMM models Λ_x using the enhanced speech $\{X^{(r)}\}_{r=1,\cdots,R}$. Owing to the involvement of the environment-effect compensation operation in the training process, we expect that it will generate better reference speech HMM models for the robust recognition method that employs the same environment-effect compensation operation in its recognition process. This is especially true for the case when the environment-effect compensation operation is not perfect due either to the nonexistence of a perfect one or to the use of an inaccurate environment contamination model in its derivation. Furthermore, the separate modeling of Λ_x and Λ_e allows the training process to focus on the modeling of phonetic variation without the irrelevant influence come from environment-effects.

The second stage of D-REST is to perform the minimum classification error (MCE)–based discriminative training on the observed speech Z with its environment-compensated speech HMM models $\Lambda_z^{(r)}$. We adopted the segmental GPD (generalized probabilistic decent)-based training procedure [2] with the following misclassification measure of $Z^{(r)}$:

$$d_i(Z^{(r)} | \Lambda_z^{(r)}) = -g_i\left(Z^{(r)}; \Lambda_z^{(r)}\right) + g_k\left(Z^{(r)}; \Lambda_z^{(r)}\right), \quad (3)$$

where $k = \arg\max_{j, j\neq i} \Pr\left(Z^{(r)}, U_j^{(r)} \middle| \Lambda_z^{(r)}\right)$; By assuming that the state-based Wiener filtering is the inverse operation of the PMC [9], we can express the compensated cepstral mean vector $\mu_{z,j,q}^{(r)}$ by $\mu_{z,j,q}^{(r)} = \mu_{x,j,q} + b^{(r)} - h_j$, where $\mu_{x,j,q}$ be the mean vector of the q-th mixture component in the j-th state of Λ_x; h_j is the cepstral coefficients of the state-based Wiener filter of the j-th state. Based on the above expression and by further assuming that $\Sigma_{z,j,q}^{(r)} = \Sigma_{x,j,q}$, the term $\Pr\left(Z^{(r)}, U_i^{(r)} \middle| \Lambda_z^{(r)}\right)$ in Eq. (1) can be rewritten as:

$$\begin{aligned} \Pr\left(Z^{(r)}, U_i^{(r)} \middle| \Lambda_z^{(r)}\right) &= \Pr\left(Z^{(r)}, U_i^{(r)} \middle| \{\mu_{x,j,q}^{(r)} + b^{(r)} - h_j, \Sigma_{z,j,q}^{(r)}\}\right) \quad (4) \\ &= \Pr\left(X^{(r)}, U_i^{(r)} \middle| \{\mu_{x,j,q}, \Sigma_{x,j,q}^{(r)}\}\right) = \Pr\left(X^{(r)}, U_i^{(r)} \middle| \Lambda_x\right) \end{aligned}$$

Accordingly, Eq. (3) can be expressed by:

$$d_i(Z^{(r)} | \Lambda_z^{(r)}) = d_i(X^{(r)} | \Lambda_x), \quad (5)$$

This shows that, to perform the MCE-based training on Z is equivalent to performing the MCE–based training on the environment-effects suppressed speech X with given compact model Λ_x.

Fig. 1 presents the general block diagram of the implementation of the D-REST algorithm. Firstly, $\{\Lambda_x, \Lambda_e\}$ is jointly estimated by the REST technique. Then, X is obtained by suppressing the environmental effects of Z with given Λ_e. Finally, the GPD-based training technique is accomplished on Λ_x with the environment-effects suppressed speech X.

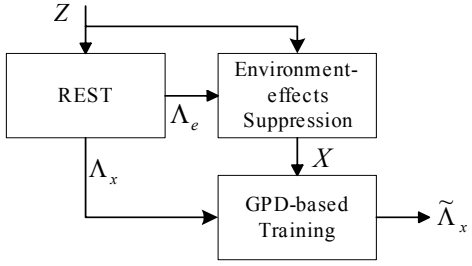

Figure 1: The general block diagram of the implementation of the D-REST algorithm

An integrated PMC (parallel model combination) [9] and SBC (signal bias compensation)-based recognition method, referred to as the PMC-SBC method [6] is employed in this work to test the compact speech HMMs generated by the proposed D-REST algorithm. Non-speech frames are detected by comparing an RNN (recurrent neural network) [10] non-speech output with a pre-determined threshold and used to estimate the on-line noise model. The input utterance Z is then processed by state-based Wiener filtering method to obtain an enhanced speech based on the state sequence which is obtained in the previous iteration of the PMC-SBC on Z; and then transforming the enhanced speech to cepstrum domain to estimate the bias by the SBR method. The SBR estimates the bias by first encoding the feature vectors of the enhanced speech using a codebook and then calculating the average encoding residuals. The codebook is formed by collecting the mean vectors of mixture components in compact speech HMMs. The bias estimate is then used to convert all speech HMMs into bias-compensated speech HMM models. These models are then further converted, by the PMC method, into noise- and bias-compensated speech HMMs using the on-line noise model. These noise- and bias-compensated speech HMMs are then used in recognition for the input testing utterance Z.

3. EVALUATION

3.1. Databases

The training data were collected through different telephony networks in Taiwan. They were made by various GSM (Global System for Mobile Communication)-based mobile phones in open environments such as the office, department and streets; and most of these phone calls were made with hand-held devices. The speech signals were received and digitally recorded at a sampling rate of 8 kHz using a PC-based speech server with a Dialogic D/41ESC card. The speakers were demanded to read the sentences from designate scripts in several categories. The recording scripts consist of 2% digits, 2.6% person names, 3.2% Taiwan city names, 3.2% phrases, 7% continuous speech corpus and 82% abbreviated Taiwan stock names. The training database comprises 23,534 utterances made by 492 speakers. For open tests, a set of clean testing speech made by a hands-free mobile phone over GSM network with the scripts of the abbreviated Taiwan stock names was recorded in quite office environment. The testing database consists of 724 utterances made by 7 speakers.

All speech signals were first pre-processed for each of 20-ms Hamming-windowed frame with 10-ms shift. A set of 26 recognition features including 12 MFCC, 12 delta MFCC, and a delta log-energy and a delta-delta log-energy was computed for each frame. Three types of noise were applied to simulate the in-car noisy environments. They were VOLVO noise from SPIB (Signal Processing Information Base) [11], ROVER_2 and ROVER_4 noises from NTT-AT ambient noise database [12]. These noise data were converted to a sampling of 8 kHz for artificially adding to speech at some particular SNRs. VOLVO and ROVER_2 were recorded in running cars on highways with close windows. Their main noise sources were engines and tires. Although the two noises came from different cars, their spectral characteristics were quite similar. The recording of ROVER_4 was performed in district road under right rain. Its main noise sources were from wiper, engine and tires. Hence, the spectral characteristics of ROVER_4 were quite different from ROVER_2 and VOLVO noises.

3.2. Experimental Results

A speaker-independent recognition for a query task of abbreviated Taiwan stock names was applied for the evaluations. We use the sub-syllable-based HMMs with 100 3-state right-final-dependent initial models and 38 5-state context-independent final models as the basic recognition units [13]. In each state, a mixture Gaussian distribution with diagonal covariance matrices is used. The number of mixture in each state was variable and depended on the number of training samples, but a maximum number of 32 mixtures was set for initial and final models and 96 mixtures for non-speech (or silence) models. The vocabulary contained 963 words, and each word consisted of 2 to 4 syllables. Although, the word set is only in medium size, its recognition is actually difficult because it comprises many highly confusable words.

The proposed D-REST algorithm was examined on a VOLVO multi-SNR training database, which was generated from artificially corrupting the original GSM training data by the VOLVO noise with 3, 9 and 15 dB in SNR. The ROVER_2 and ROVER_4, respectively, were added to the clean GSM testing speech with levels of 3, 6, and 12 dB in SNR to form the ROVER_2 and ROVER_4 noisy testing speech. The following five schemes were used for the evaluations: (1) The 'Base' scheme: The conventional HMM recognition method without any noise model compensation technique. The HMMs were trained from the original GSM training database by ML-based segmental k-means procedure. (2) The 'Multi' scheme: The same training and recognition methods as the 'Base' scheme were used. However, the HMMs were trained from the VOLVO multi-SNR training database by ML-based segmental k-means procedure. (3) The 'D-Multi' scheme: The conventional HMM recognition method without any noise model compensation technique. The HMMs were directly trained from the VOLVO multi-SNR training database by segmental GPD procedure. (4) The 'REST' scheme: The PMC-SBC recognition method with the HMMs trained from the VOLVO multi-SNR training database by the REST algorithm. (5) The 'D-REST' scheme: The PMC-SBC recognition method with the HMMs trained from the VOLVO multi-SNR training database by the proposed D-REST algorithm.

Table 1 shows the recognition results of the five schemes tested on the clean GSM testing speech. The performance of the 'Base' scheme was not good as expected in clean condition because of the mismatch between the hands-free device of testing speech and the devices used in training speech. It can be also found from the results of the 'D-Multi' and 'D-REST' schemes, both adopted the MCE/GPD training approach, gave about 12% and 19% reduction respectively in error rates compared with the counterparts by ML-based schemes. This presents that the MCE/GPD-based models outperform the ML-based models for clean testing speech.

Base	Multi	D-Multi	REST	D-REST
18.7	18.0	15.9	16.9	13.7

Table 1: Word error rates (%) of the clean GSM testing speech

SNR (dB)	Base	Multi	D-Multi	REST	D-REST
3	89.8	43.2	43.1	19.5	16.3
6	78.0	38.8	38.1	17.0	15.2
12	52.7	35.2	33.4	15.2	13.7
AVE	73.7	39.1	38.2	17.2	15.1

Table 2: Word error rates (%) of the noisy testing speech under ROVER_2 car noise. AVE denotes the average error rate over the three SNR tests.

Table 2 presents the word error rates (%) of the ROVER_2 noisy testing speech with 3, 6 and 12 dB in SNR. AVE denotes the average error rate over the three SNRs. The 'Base' scheme performed badly due to the remarkable environmental mismatch between the training data and noisy testing speech. The 'Multi' scheme were trained in the environment that was similar to noisy testing speech, and so it amounts to a 46.9% decrease in average error rate compared with that of the 'Base' scheme. Comparing the results of the 'D-Multi' scheme with those of the 'Multi' scheme, a drop in error rate by 5% in SNR = 12 dB was obtained by the 'D-Multi' scheme; however, the improvement in SNR = 3dB and 6 dB by the 'D-Multi' scheme were slight. This is mainly owing to the disturbance of environment-effects on training data and it counteracts the effects the MCE training scheme. To countervail the issue, the REST algorithm was applied for generating a set of compact

HMMs for noisy speech recognition. Can be found from the results, a significant drop was obtained by the 'REST' scheme. It gave 56% reduction in average error rate compared with that of the 'Multi' scheme; and this shows that the REST algorithm is an efficient training algorithm to generate environment-effects suppressed HMMs directly from a noisy training speech with diverse noise levels. The best performance was achieved by the 'D-REST' scheme, which led to 12% and 60% decreases in average error rate, respectively, compared with the results by 'REST' and 'D-Multi' schemes.

Then, we investigated the generalization capability of the D-REST algorithm in dealing with the noise-type unmatched conditions between training and testing environments. Fig. 2 presents the average error rate (%) of the ROVER_4 noisy testing speech over 3, 6 and 12 dB with the VOLVO multi-SNR training speech. Note that the spectral characteristics of ROVER_4 noise were quite different from VOLVO noise. Let us firstly compare the results of the 'Multi' scheme with the 'D-Multi' scheme. We observed that an increase in error rate was obtained by the 'D-Multi' scheme. This is due to the generalization problem of directly applying discriminative training techniques on diverse and noisy data. In such situation, the MCE training scheme might adjust the speech models for fitting the specific noise characteristics of the training data (the VOLVO noise in the evaluation). Consequently, this leads to degrading robustness for independent testing speech under other noisy environments that are unmatched to that of training speech. In the evaluation, the 'REST' scheme still could amount to a significant drop in average error rate by 58% compared with the result of the 'Multi' scheme. This shows that the resulting compact HMMs of the REST can perform well for testing noisy speech with untrained noise characteristics. Furthermore, we observed that the 'D-REST' scheme outperformed the other schemes. It could give 60% in average error rate reduction over the 'D-Multi' scheme. Note that the both schemes adopt the same MCE/GPD-based training approach. However, the 'D-REST' scheme can model separately the environment-specific characteristics and phonetic variation; and thus it make the speech models be trained discriminatively on phonetic variability instead of unrelated environmental characteristics. According to these results, the over-fitting issue of the MCE-based training techniques induced from noise characteristics can be suppressed by the proposed 'D-REST' scheme.

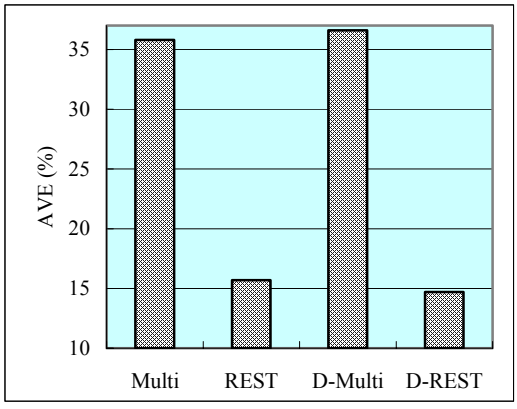

Figure 2: Average word error rates (%) of the ROVER_4 noisy testing speech for the 'Multi', 'REST', 'D-Multi' and 'D-REST' schemes.

4. CONCLUSIONS

The D-REST algorithm, which is an extended version of the REST technique, is proposed in this paper. It is applied to generate a set of environment-effects suppressed HMMs discriminatively and directly from training data suffering with both channel bias and additive noise. The principle of the D-REST algorithm is to make the training process discriminatively emphasize the modeling on phonetic variation instead of taking into account disturbed noise from environment-effects. This is achieved from separate modeling of different effects embedded in speech signals. The experimental results showed that the discriminative and robust techniques can be combined properly for dealing with noisy speech recognition by the D-REST technique. We also show that the D-REST algorithm has the potential to improve robustness not only on noisy training data, but also on noise-type unmatched conditions between training and testing environments. Furthermore, the D-REST algorithm amounted to a 60% reduction in average error rate over the performance by the conventional MCE/GPD-based training approach.

5. REFERENCES

[1] Katagiri, S., Lee, C.-H., and Juang, B.-H., "New discriminative training algorithms based on the generalized descent method", *IEEE Workshop on Neural Networks for Signal Processing*, 229-308, 1991.

[2] Chou, W., Juang, B.H. and Lee, C.H., "Segmental GPD training of an HMM-based speech recognition", *ICASSP-92*, 473-476, 1992.

[3] Woodland, P. C. and Povey, D., "Large scale discriminative training for speech recognition", *Proc. ASR-2000*, 7-16, Paris, 2000.

[4] Meyer, C. and Rose, G., "Improved noise robustness by corrective and rival training", *ICASSP-2001*, 1, 293-296, 2001.

[5] Hong, W.-T. and Chen, S.-H., "A Robust Environment-effects Suppression Training Algorithm", *Eurospeech-1999*, 6, 2495-2498, 1999.

[6] Hong, W.-T. and Chen, S.-H., "A robust training algorithm for adverse speech recognition", *Speech Communication*, 30(4), 273-293, 2000.

[7] Vaseghi, S.V. and Milner, B.P., "Noise compensation methods for hidden Markov model speech recognition in adverse environments". *IEEE Trans. Speech and Audio Process.*, 5, 11-21, 1997

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

[9] Gales, M.J.F. and Young, S.J., "Cepstral parameter compensation for HMM recognition in noise", *Speech Communication*, 12, 231-240, 1993.

[10] Hong, W.-T. and Chen, S.-H., "A robust RNN-based pre-classification for Noisy Mandarin speech recognition", *EuroSpeech-97*, 3, 1083-1086, 1997.

[11] SPIB: http://spib.rice.edu

[12] NTT-AT, "Ambient noise database for telephonometry", 1996.

[13] Lee, L.S., "Voice dictation of Mandarin Chinese", *IEEE Sig. Process. Magazine*, 17-34, 1994.