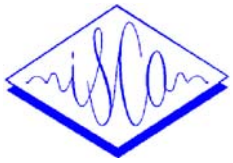

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A Robust Environment-effects Suppression Training Algorithm for Adverse Mandarin Speech Recognition

Wei-Tyng Hong and Sin-Horng Chen
Department of Communication Engineering,
National Chiao Tung University, Hsinchu, Taiwan.
email: schen@cc.nctu.edu.tw

10.21437/Eurospeech.1999-545

ABSTRACT

In this paper, a new robust training algorithm for the generation of a set of bias-removed, noise-suppressed reference speech HMM models directly from a training database collected in adverse environment suffering with both convolutional channel bias and additive noise is proposed. Its main idea is to incorporate a signal bias-compensation operation and a PMC noise-compensation operation into its iterative training process in order to make the resulting speech HMM models more suitable to the given robust speech recognition method using the same signal bias-compensation and PMC noise-compensation operations in the recognition process. Experimental results showed that the speech HMM models it generated outperformed both the clean-speech HMM models and those generated by the conventional k-means algorithm for two adverse Mandarin speech recognition tasks. So it is a promising robust training algorithm.

1. INTRODUCTION

Recently, many studies have been devoted to the field of robust speech recognition in adverse environment [1]. In this paper, we focus on the robust training issue for adverse speech recognition when the clean-speech reference models are not available. Its main concern is to train a set of robust reference speech models directly from a database collected in adverse environment for adverse speech recognition. For example, a training data set collected from telephone calls through the public switching network will suffer diverse recording conditions caused by different background noises, different types of transducers, different telephone 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.

A new robust training algorithm, referred to as the robust environment-effects suppression training (REST) algorithm, is proposed in this paper. The REST algorithm is applied for the generation of a set of bias-removed, noise-suppressed reference speech HMM models directly from a training database suffering with both convolutional channel bias and additive noise. The design goal of the REST algorithm is twofold. One is to countervail the large variability of the corrupted training samples for obtaining a set of compact reference speech HMM models with both signal bias and noise being suppressed. The other is to make the generated compact reference speech HMM models better for a given robust speech recognition method. The REST algorithm is an iterative training procedure that sequentially optimizes the following three operations: parameter estimation for environment characterization, environment-effect compensation for speech segmentation, and environment-effect suppression for HMM model re-estimation. The parameter estimation for environment characterization is to detect the signal bias and to estimate the noise statistics for each training utterance. It assumes that each utterance has its own environmental characteristics. Based on an assumed environment contamination model, the environment-effect compensation uses the estimated environment characterization parameters to adapt the HMM models to match with the current training utterance for optimally segmenting it. Using the segmentation results and the same environment contamination model, the environment-effect suppression is to remove the signal bias and the noise out of the corrupted speech for updating the HMM models. Owing to the involvement of the environment-effect compensation operation in the training process of the REST algorithm, we expect that it will generate better reference speech HMM models for the robust recognition method which 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.

2. THE REST ALGORITHM

The REST training algorithm are derived based on a presumed environment contamination model, which assumes that, for each utterance, the observed speech z is generated from the clean speech x by corrupting first with a convolutional channel b and then with an additive noise n. Here b is assumed to be time-invariant and n is stationary throughout the utterance. In the following discussions, we specify a signal in linear spectrum domain and in cepstrum domain by attaching

it with parameters *f* and *m*, respectively.

Assume that the training data set contains *R* utterances. Let $\Lambda_e \equiv \{\Lambda_n^{(r)}, b^{(r)}\}_{r=1,\Lambda,R}$ denote the set of environmental interference models of the whole training data set, where $b^{(r)}$ and $\Lambda_n^{(r)} = \{\mu_n^{(r)}, \Sigma_n^{(r)}\}$ are, respectively, the signal bias and the noise model of the *r*-th training utterance; $\mu_n^{(r)}$ and $\Sigma_n^{(r)}$ are the mean vector and covariance matrix of $\Lambda_n^{(r)}$. Let $Z^{(r)} = (z_1^{(r)}, \Lambda, z_{T_r}^{(r)})$ and $X^{(r)} = (x_1^{(r)}, \Lambda, x_{T_r}^{(r)})$ be, respectively, the observed and clean-speech feature vector sequence of the *r*-th utterance, and Λ_x denote the set of environment-effect normalized speech HMM models that we want to generate. Based on the maximum likelihood criterion, the goal of an ideal robust training algorithm is to jointly estimate Λ_x and Λ_e with given $\{Z^{(r)}\}_{r=1,\Lambda,R}$ by

$$(\Lambda_x^*, \Lambda_e^*) = \arg\max_{(\Lambda_x, \Lambda_e)} L(\{Z^{(r)}\}_{r=1,\Lambda,R} | \Lambda_x, \Lambda_e), \quad (2)$$

where $L(.)$ is the likelihood function of the observation sequence $Z^{(r)}$ given the parameter set of (Λ_x, Λ_e). But, due to the fact that it is generally difficult to derive a close form solution for the above joint maximization problem, we therefore use a three-step iterative training procedure in the REST algorithm to obtain a sub-optimal solution. The three steps are: (1) Form the environment-compensated speech HMM models $\Lambda_z^{(r)}$ by using the current (Λ_x, Λ_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 [2]-[3]; and then, estimate $b^{(r)}$ and further enhance the speech $Y^{(r)}$ to obtain $X^{(r)}$ by the SBR method [5]; (3) Update the current speech HMM models Λ_x using the enhanced speech $\{X^{(r)}\}_{r=1,\Lambda,R}$. We discuss these three steps in more detail as follows.

The first step of the REST algorithm is to optimally segment each training utterance using the current speech HMM models $\Lambda_{x,k-1}$ and the environmental interference model $\Lambda_{e,k-1}$ given by the previous iteration, where the subscript "*k*" denotes the index of iteration. The task can be accomplished, using Viterbi search, to find the best state sequence $U_k^{(r)} = (u_{1,k}^{(r)}, \Lambda, u_{T_r,k}^{(r)})$ and the best mixture component sequence $V_k^{(r)} = (v_{1,k}^{(r)}, \Lambda, v_{T_r,k}^{(r)})$ of the optimal segmentation based on the environment-compensated speech HMM models $\Lambda_{z,k-1}^{(r)}$, which is obtained from $\Lambda_{x,k-1}$ and $\Lambda_{e,k-1}$. The formation of $\Lambda_{z,k-1}^{(r)}$ from $\Lambda_{x,k-1}$ and $\Lambda_{e,k-1}$ is based on the assumed environment contamination model defined previously and realized by the following two sub-steps:

(a) Calculate $\Lambda_{y,k-1}^{(r)}$ in cepstrum domain by

$$\mu_{y,j,q,k-1}^{(r)}(m) = \mu_{x,j,q,k-1}(m) + b_{k-1}^{(r)}(m)$$

where $\mu_{y,j,q,k-1}^{(r)}(m)$ denotes the mean vector of the *q*-th Gaussian mixture in the *j*-th state of $\Lambda_{y,k-1}^{(r)}$; and $b_{k-1}^{(r)}(m)$ is the bias vector given in $\Lambda_{e,k-1}$.

(b) Use the PMC method [4] to form $\Lambda_{z,k-1}^{(r)}$ by first transforming $\Lambda_{y,k-1}^{(r)}$ from cepstrum domain to linear spectrum domain, then combining it with $\Lambda_{n,k-1}^{(r)}$ in linear spectrum domain, and lastly transforming the result back to cepstrum domain.

The second step of the REST algorithm is to enhance the adverse speech by first suppressing the noise using the state-based Wiener filtering method [2] and by then removing the signal bias by the SBR method [5]. It consists of the following two sub-steps:

(a) Noise Suppression: Given the segmentation information, estimate the noise model $\Lambda_{n,k}^{(r)}$ and eliminate it from the input adverse speech $z_t^{(r)}(f)$, in linear-spectrum domain, by the state-based Wiener filtering method to obtain the intermediate signal $y_{t,k}^{(r)}(f)$. The noise model $\Lambda_{n,k}^{(r)}$ and its average power spectrum density $P_{n,k}^{(r)}(f)$ of the *r*-th utterance are re-estimated from the non-speech frames decided by $U_k^{(r)}$. Basing on the assumed environment contamination model, the Wiener filter for the *j*-th state of speech model and the *r*-th training utterance is constructed and expressed by

$$W_{j,k}^{(r)}(f) = \frac{P_{y,j,k-1}(f)}{P_{y,j,k-1}(f) + P_{n,k}^{(r)}(f)},$$

where $P_{y,j,k-1}(f)$ is the average power density spectrum corresponding to the *j*-th state of the bias-compensated speech HMM models. After forming all state-based Wiener filters, we calculate the enhanced signal by

$$y_{t,k}^{(r)}(f) = W_{u_t,k}^{(r)}(f) \times z_t^{(r)}(f).$$

(b) Signal Bias Removal: Given with the segmentation information $(U_k^{(r)}, V_k^{(r)})$, estimate the signal bias and remove it from the intermediate signal $y_{t,k}^{(r)}(f)$ to obtain the environment-normalized speech estimate. The SBR method is realized by first transforming $y_{t,k}^{(r)}(f)$ to $y_{t,k}^{(r)}(m)$, then making a simplified assumption of $\Sigma_{z,j,q}^{(r)} =$ identity matrix to obtain

$$b_k^{(r)}(m) = \frac{\sum_{t=1}^{T_r} \left(y_{t,k}^{(r)}(m) - \mu_{x,u_{t,k}^{(r)},v_{t,k}^{(r)},k-1}(m)\right) \times I\left(u_{t,k}^{(r)} \notin \text{non-speech}\right)}{\sum_{t=1}^{T_r} I\left(u_{t,k}^{(r)} \notin \text{non-speech}\right)}$$

and lastly removing the signal bias by

$$x_{t,k}^{(r)}(m) = y_{t,k}^{(r)}(m) - b_k^{(r)}(m).$$

The third step of the REST algorithm is to re-estimate the speech HMM models $\Lambda_{x,k}$ and the average power density spectrum $\{P_{y,j,k-1}(f)\}_{j=1,\Lambda,N_j}$ using, respectively, the enhanced speech signals $\{X_k^{(r)}(m)\}_{r=1,\Lambda,R}$ and $\{Y_k^{(r)}(m)\}_{r=1,\Lambda,R}$ based on the current segmentation information $\{(U_k^{(r)}, V_k^{(r)})\}_{r=1,\Lambda,R}$, where N_j denotes the total number of states in the HMM models.

An integrated PMC-based Mandarin base-syllable recognition method, which is a modified version of the PMC method for additive and convolutional noise [6] by additionally considering board-class based likelihood compensation [7], is employed in this work to test the reference speech HMM models generated by the proposed REST training algorithm. It can be regarded as the combination of the PMC method and a signal bias compensation (SBC) method and is referred to as the PMC-SBC method. Each input testing utterance is first processed in the robust RNN-based speech segmentation [7] to detect non-speech frames. The RNN-based speech segmentation uses a three-layer simple RNN to discriminate each input frame among three broad-classes of *initial*, *final*, and non-speech. Non-speech frames are then detected by comparing the RNN non-speech output with a pre-determined threshold and used to estimate the on-line noise model. The input utterance is then processed by first subtracting the noise model estimate to obtain an enhanced speech and then transforming to cepstrum domain to estimate the signal bias by the SBR method. The SBR method estimates the signal 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 of all reference speech HMM models. The bias estimate is then used to convert all reference speech HMM models into bias-compensated speech HMM models. These models are then further converted, in the PMC, into noise- and bias-compensated speech HMM models using the above noise model estimate. These noise- and bias-compensated speech HMM models are then used in an one-stage DP search to generate the recognized base-syllable sequence for the input adverse testing utterance. The one-stage DP Search uses a Viterbi search algorithm invoking with cumulative bounded-state-duration constraints [8] to accomplish its task with the help of the likelihood compensation (LC). The LC scheme used is the one proposed previously for improving the PMC-based recognition method for noisy Mandarin speech [7]. The LC scheme uses the broad-class classification information, provided by the RNN outputs, to help reducing the recognition errors caused by the misalignments of syllable boundaries.

3. EVALUATION

Performance of the proposed REST algorithm was evaluated on a multi-speaker (8 males and 2 females) Mandarin base-syllable recognition task. It contained, in total, 3050 utterances including 2572 training utterances and 478 testing utterances. A set of sub-syllable HMM models containing 100 3-state right-*final*-dependent *initial* models and 39 5-state context-independent *final* models is used as basic recognition units [8]. In each state, a mixture Gaussian distribution with diagonal covariance matrices is used. The number of mixture in each state is variable and depends on the number of training samples, but a fixed maximum value is set for it. Besides, a single-state, single-mixture, utterance-dependent model is used for noise.

A simulated telephone-speech database generated by corrupting the multi-speaker clean-speech database with both convolutional channel bias and additive white noise was used in this study. To generate the adverse speech database, each clean-speech utterance was first corrupted by a computer-generated white Gaussian noise and then passed through a filter which simulated a telephone channel. In the training database generation, noises with levels of 12dB, 24dB, and 36dB in SNR were separately added to three subsets of the clean-speech training database. In the testing database generation, noises with levels of 9dB, 18dB, and 30dB in SNR were added to the whole clean-speech testing database. All speech signals were first pre-processed for each of 20-ms Hamming-windowed frame with 10-ms shift. Then, a set of 25 recognition features including 12 MFCC, 12 delta MFCC, and a delta log-energy was computed for each frame. To simulate the channel variations on the telephone speech through the public switching network, a set of 227 simulated filters was generated from a large telephone-speech database provided by Chunghwa Telecommunication Laboratories. Among these 227 channel filters, 195 were used to generate the training database while all others were used in the testing database generation.

In this test, the following recognition schemes were compared: (1) The 'BASELINE' scheme: The conventional HMM method using the reference speech models trained directly from the adverse training database by the segmental k-means algorithm. (2) The 'CLEAN' scheme: The bias- and PMC-compensated recognition method using the clean-speech reference HMM models. (3) The 'REST' scheme: The bias- and PMC-compensated recognition method using the REST-trained reference HMM models. (4) The 'REST-bias' scheme: The bias-compensated recognition method using the reference HMM models trained by the REST algorithm without considering noise suppression. (5) The 'REST-noise' scheme: The PMC-compensated recognition method using the reference HMM models

trained by the REST algorithm without considering signal bias removal. (6) The 'REST/LC' scheme: The PMC-SBC recognition method using the REST-trained reference HMM models.

Table 1 shows the base-syllable recognition results of these six schemes for adverse speech corrupted with channel bias and White noise. It can be found from Table 1 that, according to the recognition rate, these six schemes can be ordered as: REST/LC > REST > REST-noise or REST-bias > BASELINE > CLEAN. Based on the experimental results, several conclusions can be made. First, the conventional HMM method using the reference models trained by the K-means algorithm performed fair in adverse speech recognition. Second, the result that the CLEAN scheme performed much worse than the BASELINE scheme is mainly due to the imperfection of the channel bias compensation. Third, although the channel bias-compensation operation is imperfect, the REST training algorithm can still take the advantage of embedding it into the iterative training process to make the resulting HMM models more suitable to the recognition test using the same bias compensation operation. Fourth, the HMM models generated by the REST algorithm which considers both noise suppression and signal bias removal are better than those obtained by the REST algorithm considering only noise suppression or signal bias removal. Lastly, the likelihood compensation scheme is still effective on assisting in the adverse speech recognition.

4. CONCLUSIONS

A robust training algorithm for generating a set of speech HMM models directly from a training database collected in adverse environment for adverse speech recognition has been discussed in this paper. Its main advantage lies on the incorporation of the signal bias-compensation and PMC noise-compensation operations of a given robust adverse speech recognition method into its iterative training process so as to make the resulting speech HMM models more suitable to be used in the given robust adverse speech recognition method. Experimental results showed that the HMM models it generated were even better than the clean-speech HMM models for use in the given robust adverse speech recognition method when the PMC noise-compensation and/or channel bias-compensation operations are imperfect. So it is a promising robust training algorithm.

ACKNOWLEDGEMENT

This work was supported by National Science Council, Taiwan, ROC under project with contract #87WFA06C0180023.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey", *Speech Communication*, Vol. 16, pp. 261-291, 1995.

[2] Saeed V. Vaseghi and Ben P. Milner, "Noise compensation methods for hidden Markov model speech recognition in adverse environment," *IEEE Trans. Speech and Audio processing*, Vol. 5, pp.11-21,1997.

[3] J.H.L. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.* Vol. 39, pp. 795-805, 1991.

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

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

[6] M.J.F. Gales and S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*. Vol. 9, pp. 289-307, 1995.

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

[8] Y.R. Wang and S.H. Chen, "Mandarin telephone speech recognition for automatic telephone number directory service," *ICASSP-98*, Vol. 2, pp. 841-844, 1998.

Table 1. The recognition results of the open tests for adverse speech corrupted with channel bias and White noise

unit: %

SNR (dB)	BASELINE	CLEAN	REST-bias	REST-noise	REST	REST/LC

9	23.4	14.8	24.5	29.3	33.0	35.2
18	46.7	27.3	50.2	48.4	53.7	56.5
30	60.2	45.6	62.7	61.8	65.5	66.7