A ROBUST TRAINING ALGORITHM FOR NOISY MANDARIN SPEECH RECOGNITION

HONG Wei-Tyng and CHEN Sin-Horng
Department of Communication Engineering,
National Chiao Tung University, Hsinchu, Taiwan.
E-Mail: schen@cc.nctu.edu.tw

ABSTRACT

A new robust training algorithm to train a set of noise-suppressed speech HMM models directly from a noisy speech database to substitute the clean-speech HMM models for being used in the PMC method is proposed in this paper. The main idea is to incorporate the noise compensation operation, used in the PMC recognition test, into the training process so as to make the resulting speech HMM models better match with the PMC recognition test. The effect of imperfect PMC model combination operation on degrading the recognition performance can therefore be partially compensated. The robust training algorithm consists of an iterative procedure which alternatively performs the following three steps: optimally segment the training speech by using the PMC noise-compensated HMM models, enhance the speech by state-based Wiener filtering, and re-estimate the speech models. Experimental results confirm that it can efficiently generate better reference speech HMM models for the PMC-based Mandarin base-syllable recognition.

1. INTRODUCTION

Performance of speech recognition systems degrades rapidly when operating in adverse environment. In the past decade, many methods have been proposed [1] to make speech recognizers robust to a given noise environment. Among them, the parallel model combination (PMC) method [2] is a promising one. The basic idea of the PMC method is to use noise-compensated HMM models, generated by combining the clean-speech HMM models with the current noise model, to recognize the input testing utterance. The mismatch on acoustic characteristics between the testing utterance and the reference HMM models can hence be compensated. Although the PMC method is effective on noisy speech recognition, there still exists a flaw that the noise-compensated composite HMM models may not completely match to the input noisy testing speech due to the use of an imperfect noise-compensation operator. This will lead to a performance degradation.

To avoid this drawback, an alternate approach to train reference speech HMM models directly from a noisy speech training data set to substitute the clean-speech HMM models for being used in the PMC method is studied in this paper. A new robust training algorithm embedded with the PMC noise-compensation operation and the state-based Wiener filtering is proposed. The main idea is to incorporate the noise-compensation operation, used in the PMC recognition test, into the training process so as to make the resulting HMM models better match with the PMC recognition test. The effect of imperfect PMC model combination operation on degrading the recognition performance can therefore be partially compensated.

The organization of the paper is stated as follows. Section II discusses the proposed robust training algorithm in detail. Section III describes the PMC method used to test the reference speech HMM models generated by the training algorithm. Effectiveness of the proposed algorithm is evaluated by simulations discussed in Section IV. A summary is given in the last section.

2. THE ROBUST TRAINING ALGORITHM

The robust training algorithm consists of an iterative procedure which alternatively performs

the following three steps: optimally segment the training speech by using the PMC noise-compensated HMM models, enhance the speech by state-based Wiener filtering, and re-estimate the speech models. We discuss the training algorithm in more detail as follows.

The first step of the iterative procedure is to find the best state sequence of the noisy input speech by the PMC method using the previously estimated HMM models Λ_X^{k-1} and noise models N_r^{k-1}. The best state sequence $\zeta_r^k = \left(\zeta_{\Lambda_X}^k, \zeta_{N_r}^k\right)$ for the observation sequence $Y_r = \left(y_1^r, y_2^r, \cdots, y_T^r\right)$ of the *r*th utterance at the *k*th iteration can be expressed by

$$\left(\zeta_{\Lambda_X}^k, \zeta_{N_r}^k\right) = \arg\max_{\left(\zeta_{\Lambda_X}, \zeta_{N_r}\right)} \Pr\left(Y_r, \zeta_{\Lambda_X}, \zeta_{N_r} \middle| \Lambda_X^{k-1}, N_r^{k-1}\right)$$

where ζ_{Λ_X} and ζ_{N_r} represent, respectively, the speech and noise state sequences. The optimal segmentation is realized by a Viterbi search which finds the best state sequence to maximize the product of the state observation probability $b_j\left(y_t^r\right) = \Pr\left\{y_t^r \middle| j, \Lambda_X^{k-1} \otimes N_r^{k-1}\right\}$ evaluated using the composite HMM models formed by the previously estimated Λ_X^{k-1} and N_r^{k-1}. After finishing the segmentation, the noise model and the noise power spectral density function, $\{N_r^{k-1}(m), \Gamma_r^{k-1}(f)\}$, of the *r*th training utterance are updated to $\{N_r^k(m), \Gamma_r^k(f)\}$.

The second step is to enhance the noisy speech by state-based Wiener filtering in order to obtain the noise-suppressed enhanced speech for model re-estimation. The state-based Wiener filtering, working on linear-spectral domain, has been shown to be effective on suppressing noise in the HMM framework [4,5]. It uses the segmentation information of the noisy speech, obtained in the first step, to form the current Wiener filter

$$W_j^k(f) = \frac{S_j^{k-1}(f)}{S_j^{k-1}(f) + \Gamma_r^k(f)}$$

for the *j*th state and the *r*th training utterance. Here, $S_j^{k-1}(f)$ is the previously estimated power spectrum density functions of the *j*th state speech signal.

The third step is to re-estimate the HMM models and the state-based speech power spectral density functions based on the enhanced speech $\left\{X_r^k\right\}$ and the speech segmentation information $\left\{\zeta_{\Lambda_X}^k\right\}$.

The combination of the above three steps can be interpreted as a sequential maximum *a posterior* (MAP) estimation:

$$\left\{X_r^k, \zeta_r^k\right\} = \arg\max_{\left\{X_r, \zeta_r\right\}} \Pr\left(X_r, \zeta_r \middle| Y_r, \Lambda_X^{k-1}, S^{k-1}, \Phi^{k-1}\right)$$

$$\left\{\Lambda_X^k, S^k, \Phi^k\right\} = \arg\max_{\left\{\Lambda_X, S, \Phi\right\}} \Pr\left(\Lambda_X, S, \Phi \middle| X_r^k, \zeta_r^k\right)$$

where $S^k \equiv \left\{S_j^k(f)\right\}_{j=1,\cdots J}$ is the set of current speech power spectrum density functions, and $\Phi^k \equiv \left\{N_r^k(m), \Gamma_r^k(f)\right\}_{r=1,\cdots,R}$ is the set of noise models and the noise power spectrum density functions, and *J* and *R* are the total numbers of the HMM states and the training utterances, respectively. The sequential MAP estimation method was successfully used by Lim and Oppenheim [6] for iterative speech enhancement at frame level to sequentially estimate the linear prediction coefficients, gain, and the noise-free speech waveform.

Like other iterative algorithms, the robust training algorithm must be initialized to give the initial HMM models, the initial state-based speech power spectral density functions, and the initial noise models. The initial HMM models Λ_X^0 and the initial state-based speech power

spectral density functions $\left\{S_j^0(f)\right\}_{j=1,\cdots,J}$ can be obtained via a conventional maximal likelihood (ML) training using either an enhanced version of the given noisy training set or another training set with high SNR. In the study, we adopts the former method to use a spectral subtraction method to obtain the enhanced training set . The initial noise models and power density functions, $\left\{N_r^0(m),\Gamma_r^0(f)\right\}_{r=1,\cdots,R}$,can be obtained from non-speech segments determined manually or determined by the above-mentioned RNN-based speech segmentation.

In summary, the robust training algorithm is given below:

Initialization

Perform the RNN-based speech segmentation and estimate $\left\{N_r^0(m),\Gamma_r^0(f)\right\}_{r=1,\cdots,R}$. Use the spectral subtraction method to obtain the enhanced training set and generate $\left\{S_j^0(f)\right\}_{j=1,\cdots,J}$ and Λ_X^0. Set *k = 0.*

Iterative procedure

Set *k=k+1.*

For each utterance Y_r of the training set, do the following two steps:

1. Optimal segmentation

1.1 Form the composite HMM model by $\Lambda_{Y_r}=\Lambda_X^{k-1}\otimes N_r^{k-1}$.

1.2 Segment Y_r to obtain the optimal state sequence ζ_r^k and the likelihood scores using Λ_{Y_r}.

1.3 Estimate $\left\{N_r^k(m),\Gamma_r^k(f)\right\}$ using Y_r and ζ_r^k.

2. Wiener filtering

2.1 Construct the state-based Wiener filters by

$$\left\{W_j^k(f)=\frac{S_j^{k-1}(f)}{S_j^{k-1}(f)+\Gamma_r^k(f)}\right\}_{j=1,\cdots,J}$$

2.2 Generate the enhanced speech $X_r^k=(X_1^{k,r},\cdots,X_T^{k,r})$ by

$$X_t^{k,r}(m)=DCT\left(\log\left[W_{\zeta_r(t)}^k(f)\times e^{IDCT(y_t^r(m))}\right]\right)$$

for *t=1* to *T*.

3. Models updating

Re-estimate Λ_X^k and $\left\{S_j^k(f)\right\}_{j=1,\cdots,J}$ from $\left\{X_r^k,\zeta_{\Lambda_X}^k\right\}_{r=1,\cdots,R}$.

4. Termination test

If the increase rate of the average likelihood < 0.001, then stop; otherwise continue the iterative procedure.

3. THE PMC-BASED MANDARIN BASE-SYLLABLE RECOGNITION SYSTEM

A modified PMC method previously proposed in [7] is employed to test the reference speech HMM models generated by the proposed robust training algorithm. Fig. 1 displays its block diagram. It consists of six parts: RNN-based noisy speech segmentation, noise model estimation, PMC noise-compensation, clean-speech HMM models, likelihood compensation (LC), and one-stage DP search. The input utterance is first processed in the RNN-based noisy speech segmentation to detect non-speech frames. Noise models are then estimated recursively from non-speech frames and used in the PMC noise-compensation to adapt all reference speech HMM models to the current noise environment. These noise-compensated HMM models are then used in the one-stage DP search to generate the recognized base-syllable string. Meanwhile, the broad-class classification information provided by the RNN outputs are further used in the likelihood compensation to assist in the noisy speech recognition.

The RNN-based speech segmentation uses a three-layer simple RNN to discriminate each input frame among the three broad classes of

initial, *final*, and non-speech. Non-speech segments are then detected by comparing the RNN non-speech output with a pre-determined threshold. A recursive estimation method is then used to adaptively estimate noise models from these non-speech segments. The PMC method uses a noise-combination operator to generate noise-compensated composite HMM models. The log-normal approximation [2] is employed for the noise-combination operator. The likelihood compensation directly takes the three RNN outputs as weighting factors to add additional scores to the log-likelihood scores of HMM states associated with the three broad classes, i.e.,

$$\rho_j^c(y_t) = \begin{cases} \rho_j(y_t) + \alpha \log(W_I(t)), & j \in \text{Initial} \\ \rho_j(y_t) + \alpha \log(W_F(t)), & j \in \text{Final} \\ \rho_j(y_t) + \alpha \log(W_N(t)), & j \in \text{Non-speech} \end{cases}$$

where $W_I(t)$, $W_F(t)$ and $W_N(t)$ are the *initial*, *final*, and non-speech outputs of the RNN, $\rho_j(y_t)$ is the log-likelihood score of state j, and α is a scaling factor to control the degree of the likelihood compensation.

4. EXPERIMENTAL RESULTS

Efficiency of the proposed robust training algorithm on noisy speech recognition was examined by simulations on a Mandarin base-syllable recognition task using a multi-speaker (2 males and 2 females) database. The database contains in total 6197 syllables including 5124 training syllables and 1073 testing syllables. Each utterance comprises several syllables and is pronounced in such a way that every syllable is clearly pronounced. A set of recognition features including 12 MFCCs, 12 delta MFCCs, and a delta log-energy was computed. Two noisy speech databases were artificially generated by adding two types of noises into the above clean-speech database. The two noise types are the Lynx helicopter noise and a computer-generated white Gaussian noise. For simplicity, they are referred to as Lynx and White noises, respectively. The following two recognition schemes were tested: (S1) the basic PMC method with noise model being estimated based on the proposed RNN-based speech segmentation; (S2) an extended version of S1 with LC speech segmentation information assistance scheme. Both cases of matched and mismatched noise types used in the training and the testing phases were examined. Tables I and II show the base-syllable recognition accuracy of the open tests using the HMM models trained in the Lynx and White noises, respectively. The 'Baseline HMM' and 'RT HMM' denote the HMM models obtained by the traditional segmental k-means and the robust training algorithm, respectively. The 'Clean HMM' denote the clean HMM models trained according to the clean speech database. It can be found from these two tables that the proposed robust training algorithm performs much better than the baseline method for both cases of matched and mismatched noise types. For the robust training algorithm, the results of the matched noise-type case are better than those of the mismatched noise-type case. By comparing the results of RT HMM and Clean HMM shown in Tables I and II, we find that the HMM models obtained by the robust training algorithm performed better than the clean-speech HMM models in both of S1,S2 recognition scheme with matched noise condition.

5. SUMMARY

In this paper, a robust training algorithm is proposed to generate reference speech HMM models directly from a noisy speech data set for the PMC-based noisy Mandarin base-syllable recognition. It incorporates the noise compensation operation and the state-base Wiener filtering into the iterative training process with the goal to make the resulting speech HMM models better match with the PMC recognition test. Two advantages of the new training algorithm can be found. One is that the generated speech HMM models are better than the clean-speech HMM models for being used in the PMC-based noisy Mandarin base-syllable recognition. The other is that it can be used in the case when the clean-speech HMM models are not available.

6. ACKNOWLEDGEMENTS

This work was supported by the National Science Council, Taiwan, ROC under Contract No. NSC87-2213-E-009-056.

7. REFERENCES

[1] Gong, Y. (1995). "Speech recognition in noisy environments: a survey", Speech Communication, 16, 261-291.

[2] Gales, M. J. F., and Young, S. J. (1996). “Robust continuous speech recognition using parallel model combination,” IEEE Trans. Speech and Audio Processing. 5, 352-359.

[3] Junqua, J. S., Mak B., and Reaves, B. (1994). “A robust algorithm for word boundary detection in the presence of noise,” IEEE Trans. Speech and Audio Processing. 2, 406-412.

[4] Ephraim, Y. (1992). “Statistical-model-based speech enhancement systems,” Proc. IEEE 80, 1526-1555.

[5] Vaseghi, S. V., and Milner, B. P. (1997). “Noise compensation methods for hidden Markov model speech recognition in adverse environments,” IEEE Trans. Speech and Audio Processing. 5, 11-21.

[6] Lim, Jae S., and Oppenheim, A. V. (1978). “All-pole modeling of degraded speech,” IEEE Trans. Acoust. Speech Sig. Process. 26, 197-210.

[7] Hong, W. T. and Chen, S. H. (1997). “A robust RNN-based pre-classification for Noisy Mandarin speech recognition,” EuroSpeech 97, 3, 1083-1086.

TABLE I. The recognition results of the open tests using HMM models trained from noisy training set corrupted by Lynx noise.

		Baseline HMM	RT HMM		Clean HMM	
Testing Noise	SNR(dB)	S1	S1	S2	S1	S2
White	9	20.4	28.8	30.5	26.9	33.0
	18	46.1	51.4	54.9	48.3	52.0
	30	61.6	69.0	71.4	65.4	71.8
Lynx	9	35.0	43.6	48.7	39.1	42.3
	18	52.0	62.8	67.6	58.6	62.5
	30	65.1	73.6	78.3	71.2	75.1

TABLE II. The recognition results of the open tests using HMM models trained from noisy training set corrupted by White noise.

		Baseline HMM	RT HMM		Clean HMM	
Testing Noise	SNR(dB)	S1	S1	S2	S1	S2
White	9	23.1	35.0	38.1	26.9	33.0
	18	43.6	54.2	58.0	48.3	52.0
	30	55.4	68.2	73.8	65.4	71.8
Lynx	9	21.1	39.0	44.8	39.1	42.3
	18	38.1	54.0	59.4	58.6	62.5
	30	49.0	66.0	72.0	71.2	75.1

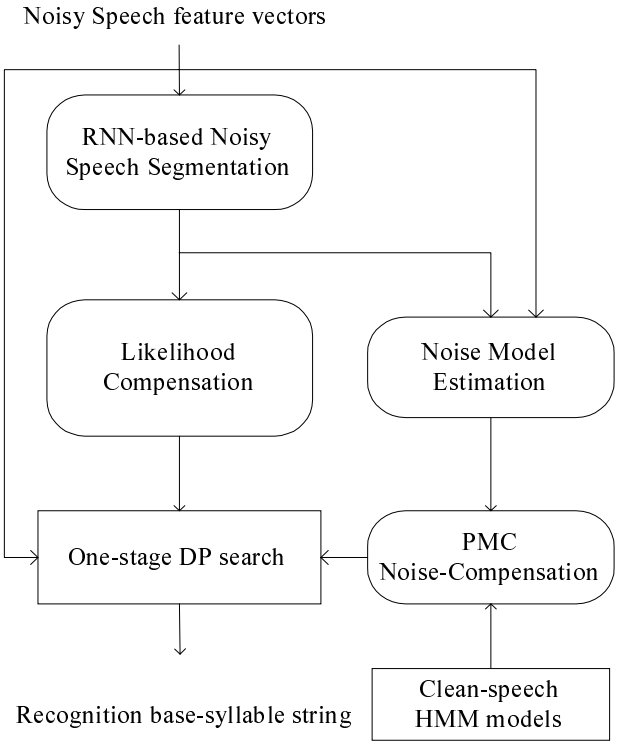

FIG. 1. A block diagram of the proposed PMC/RNN-LC noisy speech recognition method