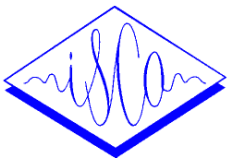

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

AN INVESTIGATION ON WIRELESS SPEECH RECOGNITION BY DATA CONTAMINATION AND ROBUST TRAINING TECHNIQUES

Wei-Tyng HONG[1] *Ke-Shiu CHEN*[2]

[1]PenPower Technology, Hsinchu

[2]Advanced Technology Center/CCL, Industrial Technology Research Institute, Hsinchu

jfhong@seed.net.tw koche@itri.org.tw

ABSTRACT

This paper is concerned with the robust endpoint detection and noisy speech recognition over wireless network. Firstly, the MLP-based and GMM-based endpoint detection incorporated with data contamination and continuous spectral subtraction techniques were investigated. Then, for noisy wireless speech recognition, a combined technique of data contamination and robust training was proposed to separately model the environmental characteristics and phonetic information. According to the results from an abbreviated stock name recognition task, we observe that the proposed techniques has the potential to improve robustness not only on diverse data contaminated training data, but also on the unmatched noise-type condition between training and testing environments.

1. INTRODUCTION

Many studies have been devoted to the field of robust speech recognition in noisy environment. In this paper, the robust endpoint detection and model-compensation based noisy speech recognition over wireless environment are investigated. In the robust endpoint detection, the MLP-based and GMM-based approaches incorporated with data contamination [1] and noise subtraction techniques were proposed. For the model-compensation based noisy speech recognition, a combined technique of data contamination and REST (Robust Environment-effects Suppression Training) training [2] was investigated. The main concern of the combined technique is to train a set of compact speech models for model-compensation based noisy speech recognition directly from noisy training data, which are artificially contaminated by some specific noise environment. The data contamination techniques have been widely used in attempting to improve the robustness on a specific environment. They make recognizer be trained in noisy conditions similar to that of testing for reducing the mismatch between testing data and speech models. However, noisy training data will make speech patterns distribute more widely in the feature space so as to overlap to each other more seriously and cause the speech models degrade on their discriminative capabilities.

The REST algorithm is applied for the generation of a set of environment-effects suppressed and compact speech HMMs directly from a contaminated database suffering with both channel bias and additive noise. Its principle is through separate modeling of different effects (e.g., the phonetic variability and environment-effects) of speech signal to make the training process emphasize phonetic variation modeling instead of taking into account disturbed variability from environment-effects. The design goal of the combined algorithm is threefold. The first is to countervail the large variability of the contaminated training samples for obtaining a set of compact speech HMMs with both signal bias and noise being suppressed. The second is to make the generated compact speech HMMs better for a given robust speech recognition method. The third is for improving the discrimination capability. This paper is organized as follows. The proposed techniques are presented in section 2. In section 3, we analyze the experimental results. In section 4, some conclusions are drawn.

2. THE PROPOSED TECHNIQUES

2.1 The Robust MLP-based Endpoint Detection

The MLP-based classification method is extended to attack the difficult task of endpoint detection on noisy wireless speech. The MLP adopts a noise-immunity training algorithm [3] on a multi-SNR database which is artificially contaminated for some specific noise environment. It is employed for classifying each input frame into *speech*, *non-speech* and *transition* classes, which was implemented by comparing the classification scores with two thresholds, T_H and T_L. While one output was higher than T_H and the other was lower than T_L, the input frame was classified into one of the *speech* and *non-speech* classes. Otherwise, it went into the *transition* class. To improve the robustness of the endpoint detection, the continuous spectral subtraction (CSS) [4] was applied on each input frame as following:

$$Y_t(f) = \begin{cases} Z_t(f) - \alpha \cdot N_t(f), & \text{if } Z_t(f) > \frac{\alpha}{1-\beta} N_t(f), \\ \beta \cdot Z_t(f), & \text{otherwise} \end{cases} \quad (1)$$

where $Y_t(f)$ and $Z_t(f)$ are the power spectrum of the enhanced speech and noisy speech at frame t, respectively; α and β are the over-estimation and spectral flooring factors, respectively. $N_t(f)$ is the noise power spectrum at frame t, which is obtained from moving average of w consecutive frames on noisy speech:

$$N_t(f) = \frac{1}{w} \sum_{j=t-w+1}^{t} Z_j(f). \quad (2)$$

CSS was shown to have advantages on noisy speech recognition [4]. In this paper, we apply it for improving the performance of the endpoint detection.

A finite state machine (FSM) based decision logic [5], shown in Fig. 1, is applied for endpoint detection. It contains four states: *non-speech* (N), *speech* (S), *non-speech* to *speech* (NS) and *speech* to *non-speech* (SN) states, respectively. We

assume the *non-speech* state is the starting state. The rules of the state transition in the FSM are presented below:

Path 1	Go to NS if the input frame is *transition* or *speech* class. Accumulate the number of frames in NS.
Path 2	Go back to N if the number of the accumulative frames in NS is smaller than L1. Set all the frames in NS to be *non-speech* frames.
Path 3	Go to S if the input frame is *speech* class and the number of the accumulative frames in NS is larger than L1. Set all the frames in NS to be *speech* frames.
Path 4	Go to SN if the input frame is *transition* or *non-speech* class. Accumulate the number of frames in SN.
Path 5	Go back to S if the number of the accumulative frames in SN is smaller than L2. Set the all frames in SN to be *speech* frames.
Path 6	Go to N if the input frame is *non-speech* class and the number of the accumulative frames in SN is larger than L2. Set the first half frames in SN to be *speech* frames and the others be *non-speech* frames.

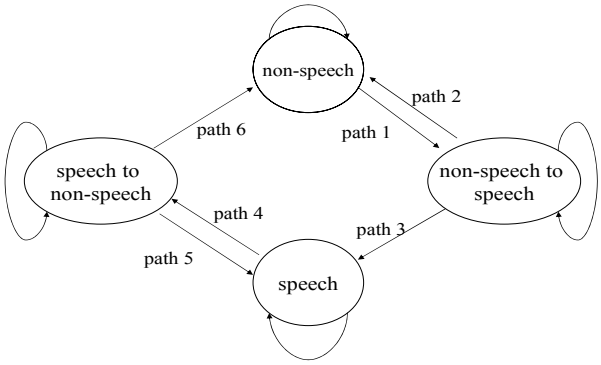

Figure 1: The state transition diagram of the FSM.

2.2 The REST Algorithm

The REST algorithm is derived based on a presumed noisy speech realization model, which assumes that, the observed speech Z is generated from the clean X corrupted by ambient noise and channel bias. Our target is to estimate the compact speech models of speech X given observed speech Z, and to make this model as a robust seed model for noisy speech recognition using model-compensation based approach. Assume that the environment-effects comprise a convolutional channel b and an additive noise n on speech. Here b is assumed to be time-invariant and n is stationary throughout the utterance. Assume that the training data set contains R utterances. Let $\Lambda_e \equiv \left\{\Lambda_n^{(r)}, b^{(r)}\right\}_{r=1,\cdots,R}$ 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; Let $Z^{(r)}$ be the observed speech feature vector sequence of the r-th utterance, and Λ_x denote the set of environment-effect normalized speech HMMs that we want to generate. Based on the ML (maximum likelihood) criterion, the goal of an ideal robust training algorithm is to jointly estimate Λ_x and Λ_e with given $\{Z^{(r)}\}_{r=1,\cdots,R}$ by

$$(\Lambda_x, \Lambda_e) = \arg\max_{(\overline{\Lambda}_x, \overline{\Lambda}_e)} \Pr\left(\left\{Z^{(r)}\right\}_{r=1,\cdots,R} \middle| \overline{\Lambda}_x, \overline{\Lambda}_e\right). \quad (3)$$

An iterative training procedure was proposed in [2] to sequentially optimize Eq. (3). It includes the following three operations: parameter estimation for environment characterization, environment-effects compensation of speech models for speech segmentation, and environment-effect suppression for HMMs re-estimation. 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 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.

2.3 The Recognition Stage

An integrated PMC (parallel model combination) [6] and SBC (signal bias compensation)-based recognition method, referred to as the PMC-SBC method is employed in this work to test the compact speech HMM models generated by the proposed REST training algorithm. The testing speech is firstly processed by the proposed MLP-based endpoint detection. The detected non-speech frames used to estimate the on-line noise model. The input utterance Z is then enhanced by state-based Wiener filtering method [7] to obtain an enhanced speech based on the state sequence obtained in the previous iteration of the PMC-SBC on Z; and transforming the enhanced speech to cepstral domain to estimate the bias by the SBR [8] 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 HMM models. The bias estimate is used to convert all speech HMMs into bias-compensated speech HMM models. These models are further converted, in the PMC, into noise- and bias-compensated speech HMMs using the above noise model estimate. These noise- and bias-compensated speech HMMs are then used in recognition for the input testing utterance Z.

3. EVALUATION

3.1 Database

The training database were collected through 800 toll-free calls, which were comprised different telephony networks in Taiwan and made by various GSM (Global System for Mobile Communication)-based mobile phones in open environments such as the office, department and streets. The speech signals were received and digitally recorded using a PC-based speech server with a Dialogic D/41ESC card. A sampling rate of 8 kHz was used. The speakers were demanded to read the sentences from designate scripts in several categories. The training database comprises 23,534 utterances made by 492 speakers. The recording scripts of the database were drawn from a mixed type comprising 2% digits, 2.6% person names, 3.2% Taiwan city names, 3.2% phrases, 7% continuous speech corpus and 82% abbreviated Taiwan stock names. Most of the mobile phone calls of the database were made with hand-held devices. For open tests, a set of the abbreviated Taiwan stock names was recorded through GSM network in quite office to

form the clean GSM testing database. These calls were made by a hands-free mobile phone, and 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 car noise were applied to generate the in-car noisy speech databases. They were VOLVO noise from SPIB (Signal Processing Information Base) [9], CIVIC_1 and ROVER_4 noises from NTT-AT ambient noise database [10]. Both the recording conditions in cars of VOLVO and CIVIC_1 were running in highways with close windows. The main noise sources were engines and tires. Although the two noises were recorded in different cars, they were similar in spectral characteristics. On the other hand, the recording of ROVER_4 was performed in district road under right rain and its main noise sources were come from wiper, engine and tires. Hence, the spectral characteristics of ROVER_4 were quite different from ROVER_2 and VOLVO noises. All the noise data were converted to a sampling of 8 kHz for artificially adding to speech in some particular SNRs.

3.2 Evaluation I

We started with examining the performance of the endpoint detection using the robust MLP-based classifier and ML (Maximum Likelihood)-based classifier with Gaussian Mixture models (GMM) respectively. Both the two classifiers used the same recognition features and all were trained on a multi-SNR training database. The multi-SNR database consists of the original training data and the noisy training data that were generated from artificially corrupting the original training data by the CIVIC_1 noise with 3, 9 and 15 dB in SNR. The number of hidden nodes of the MLP was empirically set to be 50. The GMM-based classifier used two mixture Gaussian distributions to model the likelihood probabilities for the speech and non-speech broad classes to be classified. The number of mixture components in each distribution was set to be 16 for making the MLP-based and GMM-based classifiers be roughly equal in number of parameters. The same FSM-based logic, shown in Fig. 1, was applied for performing endpoint detection.

The average deviations of the detected endpoints by MLP-based and GMM-based approaches with respective to the targeted results under different types and levels of noise are presented in Fig. 2. The 'MLP/CSS' and 'GMM/CSS' denote that the MLP-based and GMM-based detection schemes combined with the CSS method, respectively. It can be found that the MLP-based approach significantly outperforms the ML-based approach for all testing conditions. The best performance of each testing environment was achieved by the 'MLP/CSS' scheme. Furthermore, these results show that the CSS technique and MLP-based classifier can be combined properly for dealing with endpoint detection in unmatched noise-type condition between training and testing environments.

3.3 Evaluation II

A speaker-independent recognition task for abbreviated Taiwan stock names was applied in the evaluation. 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 [11]. 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 performance of the proposed recognition system was examined on a multi-SNR training database, which consists of the original training data and the noisy training data that were generated from artificially corrupting the original training data by the VOLVO noise with 3, 9 and 15 dB in SNR. The ROVER_4 noise with levels of 3, 6, and 12 dB in SNR were added to the clean GSM testing database for testing data. The following recognition schemes were used for comparing the performances of the proposed algorithms: (1) The 'Base' scheme: The conventional HMM recognition method without any noise model compensation. The HMMs were trained from the GSM training database by ML-based segmental k-means algorithm. (2) The 'Multi' scheme: The conventional HMM recognition method without any noise model compensation. The HMMs were trained from the multi-SNR training database by ML-based segmental k-means algorithm. (3) The 'REST' scheme: The PMC-SBC recognition method with the HMMs trained from the multi-SNR training database by the REST algorithm.

Table 1 presents the word error rates (%) of the recognition tests on noisy GSM testing speech, which are corrupted by ROVER_4 noise with 3, 6 and 12 dB in SNR, respectively. '+ED' represents the recognition scheme incorporated with the 'MLP/CSS'-based endpoint detection and 'AVE' denotes the average error rate over the three SNRs. The spectral characteristics of ROVER_4 were quite different from the multi-SNR training data, which were contaminated by VOLVO noise. The speech models in the 'Base' scheme were trained from the original training database without the car noise. Hence, it performed badly due to the remarkable environmental mismatch between training and testing speech. In comparison, the speech models in the 'Multi' scheme were trained in the similar environment to noisy testing speech; and thus it amounts to a 44.8% decrease in average error rate compared with that of the 'Base' scheme. In noisy speech recognition, the noise will degrade the discrimination power of HMM models, even when training and testing are performed in a similar conditions. A pure-noise segment is easily mixed up with a fricative sound. A weak voiced sound tends to become an unvoiced one. This kind of confusion does not exist in clean speech recognition; however, it is a serious problem in noisy speech recognition. The problem can be partially solved by a robust and precise endpoint detector. It can restrict recognition search and thus the consonant segment will not be easily mixed up with pure-noise segment. This inference can be proved by comparing the results between the 'Multi' and 'Multi+ED' schemes: the scheme with a robust MLP-based endpoint detector could obtain a significant drop in average error rate by 43.9%.

To countervail the diverse characteristics of the multi-SNR training speech and to increase the discriminative capability of trained HMMs, the REST algorithm was applied for generating a set of compact HMMs. It can be seen by the 'REST' and 'REST+ED' schemes, we could obtain significant drops in average error rate by 56% and 27.8%, respectively, compared with the 'Multi' and 'Multi+ED' schemes. This shows that the REST algorithm is a very efficient training

algorithm to generate environment-effects suppressed HMMs directly from a noisy training speech with diverse noise levels. This is because that the environment-specific characteristics and phonetic variation are modeled separately by the REST algorithm. This makes the reference models be trained emphatically on the speech whose environment-effects are suppressed. Hence, the resulting HMMs can perform well in the PMC-SBC model compensation scheme for testing noisy speech with untrained noise characteristics.

SNR (dB)	Base	Base +ED	Multi	Multi +ED	REST	REST +ED
3	79.6	61.7	36.1	19.3	16.3	15.5
6	69.1	53.0	34.9	19.6	15.6	14.4
12	46.1	40.1	36.5	21.3	15.2	13.7
AVE	64.9	51.6	35.8	20.1	15.7	14.5

Table 1: Word error rates (%) of the tests for noisy GSM testing speech corrupted by ROVER_4 car noise. '+ED' represents the recognition scheme with the robust MLP-based endpoint detection and 'AVE' denotes the average error rate over the three SNRs.

4. SUMMARY

A MLP-based endpoint detection and a combined technique of the REST and data contamination are proposed to enhance the robustness for noisy speech recognition. In the evaluation I, the data contamination and noise subtraction techniques were applied into the MLP-based and GMM-based endpoint detectors, respectively. The results show that the proposed MLP-based approach can significantly outperform the GMM-based approach in different SNRs with different car noise types. In the evaluation II, the experimental results showed that the usage of the REST combined with data contamination technique amounted to a 27.8% reduction in average error rate over the performance by the conventional data contamination technique. Furthermore, we observe that the REST algorithm has the potential to improve robustness not only on diverse data contaminated training data, but also on the unmatched noise-type condition between training and testing environments.

5. REFERENCES

[1] Morii, S. *et al*, "Noise robustness in speaker independent speech recognition", *ICSLP-90*, 1145-1148, 1990.

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

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

[4] Flores, J. A. Nolazco *et al*, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation", *ICASSP-94*, 1, 409-412, 1994.

[5] You, S.R., "RNN-based processing for speech recognition", master thesis, National Chiao-Tung University, Taiwan, 2000.

[6] Gales, M.J.F. and Young, S.J., "Robust continuous speech recognition using parallel model combination", *IEEE Trans. Speech and Audio Process.*, 5, 352-359, 1996.

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

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

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

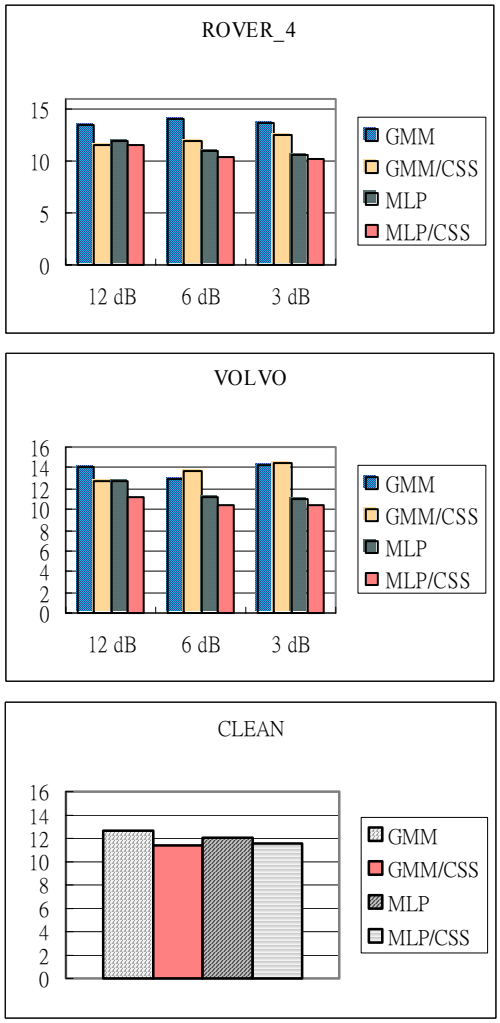

Figure 2: The average deviations (unit: frame) of the detected endpoints by MLP-based and GMM-based approaches for testing speech under CIVIC_1, ROVER_4, VOLVO and CLEAN environments. The '/CSS' denotes that the detection schemes combined with the CSS method.