A Robust RNN-based Pre-classification for Noisy Mandarin Speech Recognition

Wei-Tyng Hong and Sin-Horng Chen
Department of Communication Engineering,
National Chiao Tung University, Hsinchu, Taiwan.
TEL: 886-3-5731822, FAX: 886-3-5710116, E-Mail: jeff@cm.nctu.edu.tw

ABSTRACT

This paper addressed the problem of speech signal pre-classification for robust noisy speech recognition. A novel RNN-based pre-classification scheme for noisy Mandarin speech recognition is proposed. The RNN, which is trained to be insensitive to noise-level variation, is employed to classify each input frame into the three broad classes of initial, final and pure-noise. An on-line noise tracking and estimation for noise model compensation is then performed. Besides, a broad-class likelihood compensation based on the RNN outputs is also performed to help the recognition. Experimental results showed that a significant improvement on syllable recognition rate has been achieved under non-stationary noise environment.

1. INTRODUCTION

Background noise is a major source of corruption in speech signals collected in an adverse environment. It often dramatically degrades the performance of speech recognizers developed for clean speech. In recent years, many methods have been proposed [1] to make the recognition system robust to adverse conditions. They can be generally divided into three broad categories: (1) extraction of recognition features robust to noise, (2) speech enhancement prior to recognition and (3) adaptation of the clean speech reference models for compensating the noise effect. Most of these proposals need the noise statistics to be given in advance in order to help the estimation of the spectral characteristics of the clean speech from the corrupted input speech or to modify the reference models. For example, in the non-linear spectral subtraction (NSS) method [4], a frequency-dependent overestimation factor is determined according to the signal-to-noise ratio (SNR) in each band. In the MLP-based feature-transformation method [5], statistical information about the noise signal is included explicitly as additional input to the MLP filters; and in the parallel model combination (PMC) method [2] and the HMM state-integrated Wiener filters method [3], the HMM model parameters are adapted to new acoustic environment with the noise statistics being given.

But, in practical applications, the noise statistics may not be known in advance. A noise/speech detector is usually required to identify the noise segments for on-line acquiring the noise statistics from the noise contaminated speech. The quality of noise/speech decision will determine the accuracy of noise estimation, which in turn seriously affect the performance of the whole recognition system. Basically, it is a difficult task to precisely identify pure-noise or speech segments from a real noisy speech signal. This is especially true when the noise is time-varying or when the SNR is low.

In the past, many noise/speech detection algorithms adopted the rule-based approach [6] which makes decision based on comparing specific features, such as the short-term energy, zero-crossing rate and the subband spectral magnitude, with some pre-determined thresholds. The main drawback of those methods is that their thresholds must be empirically determined from the training data and can not be applied to the case when the environmental noise changes in level. Another disadvantage of those methods lies on the use of hard-decision segmentation. The information of how reliable the decision is has been lost. This information can be used to assist in speech recognition.

In addition to the difficulty of noise/speech segmentation, the noise may also degrade the discrimination power of HMM models, even when they are used in a matched conditions, i.e., the same environment for both training and testing. For instances, a pure-noise segment is easily mixed up with a fricative sound. A weak voiced sound tends to become a unvoiced one. This kind of confusion does not exist in clean speech recognition; however, it is a serious problem in noisy speech recognition.

This motivates our studies of using neural network-based broad-class pre-classification to estimate noise statistics for model adaptation as well as to compensate recognition scores for curing the noise-induced confusion across broad-classes.

2. THE PROPOSED SYSTEM

The block diagram of the proposed system is displayed in Fig. 1. It consists of six parts: RNN-based pre-classification, noise estimation, model compensation, clean speech models, broad-class likelihood compensation and one-stage DP search. Input speech feature vectors are firstly pre-classified into three classes of pure-noise, initial, and final. Noise model is then estimated from the pure-noise parts and used to adapt the clean speech HMM models for compensating the noise

effect. Meanwhile, the information of pre-classification is also used in likelihood compensation to help the one-stage speech recognition. In the following Subsections, we discuss the system in more detail.

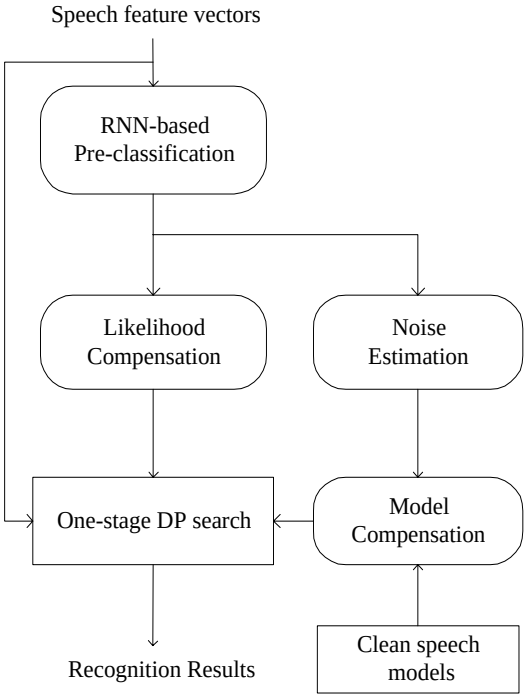

Fig. 1 The block diagram of the proposed system

2.1 The RNN-based Pre-Classification

A three-layer RNN which feeds back all outputs of its hidden layer to the input layer is employed to pre-classify each frame of the input speech into three broad classes of pure-noise, initial, and final. We note that all Mandarin syllables have the same initial-final structure. An RNN of this architecture is attractive in this problem because it has a distinct property that the outputs of the hidden layer at any time depend on a complex aggregate of all previous inputs. So it is suitable for discriminating dynamic speech patterns [8]. The RNN was trained by the back-propagation algorithm [7] with desired output targets being set according to the labelled positions of all training utterances determined by an initial-final based hidden Markov model (HMM) recognizer. An immunity learning procedure was applied for making it robust to the change of noise level. The learning process started from using the clean speech. Then, the speech signals were corrupted by background noise with the SNR being gradually decreased for obtaining the noise immunity. In this study, several SNR levels including 48 dB, 32 dB, 24 dB, 12 dB and 0 dB were applied sequentially to train the RNN.

After pre-classifying the input signal, noise estimations are then performed from the input signal feature vectors:

$$N(t) = \frac{1}{\lambda}\sum_{m=t-\lambda+1}^{t} X_m$$

where λ is the segment length for the noise estimation. A simple recursive smoothing scheme using the current and the previous noise estimates is then performed:

$$N(t) \leftarrow \begin{cases} \eta N(t) + (1-\eta) N(t-1), & if \quad C(t) > \gamma \\ N(t-1) & , \quad else \end{cases}$$

where *C(t)* is a confidence score to judge how reliable the current noise estimate *N(t)* is and γ is a pre-determined threshold. In this paper, the confidence score is defined as

$$C(t) = \frac{1}{\lambda}\sum_{n=t-\lambda+1}^{t} W_N(n)$$

where $W_N(t)$ is the noise discriminant function generated by the RNN. We note that to embed such a confidence score in the smoothing scheme can prevent us from obtaining a poor noise estimate when the current *N(t)* is estimated based on a pure-speech segment. This consideration is important for reliably tracking the noise statistics under nonstationary noise environment.

2.2 The Model Compensation

After on-line estimating noise statistics, all clean-speech HMM models are adapted to make them match the current environment for compensating the noise effect on speech recognition. Some previously proposed HMM model compensation methods [2,3] can be used. The observation likelihood of the composite model at state *j* can then be expressed by

$$\rho_j(x_t) = \log\left(prob\left\{x_t \middle| j, M \otimes N(t)\right\}\right)$$

where *M* represents a clean-speech HMM model and $\otimes$ denotes the model compensation operator. This model-compensated system is expected to perform better under a wide range of noise levels because of the on-line updating of *N(t).*

2.3 The Likelihood Compensation

As mentioned previously, the effect of noise perturbation on recognition features will result in the degradation of the discrimination capabilities of HMM models. This is true even when both the training and the testing are in a matched condition. Fig.2 shows the scattering plots of cepstral coefficients C_1-C_2 for some phones corrupted by additive white Gaussian noises with various SNRs. It can be seen from the figure that the C_1-C_2 distributions of the three classes become overlapped as the noise gain increases. A likelihood compensation scheme which uses the information of pre-classification to assist in speech recognition is therefore suggested. This is realised by incorporating the broad-class discriminant functions generated by the RNN pre-classifier into the original likelihood score $\rho_j(x_t)$ in the log-probability domain, i.e.,

$$\rho_j^c(x_t) = \begin{cases} \rho_j(x_t) + \alpha \log\left(W_I(t)\right), & j \in \text{Initial} \\ \rho_j(x_t) + \alpha \log\left(W_F(t)\right), & j \in \text{Final} \\ \rho_j(x_t) + \alpha \log\left(W_N(t)\right), & j \in \text{Noise} \end{cases}$$

where $W_I(t)$, $W_F(t)$ and $W_N(t)$ are respectively the initial, final, and noise discriminant functions and α is a scaling factor to control the degree of the likelihood compensation. Here, the value of α is empirically determined. We note that if a hard decision is made in the pre-classification to make $W_I(t)$, $W_F(t)$ and $W_N(t)$ become zero-one functions, the likelihood compensation is equivalent to restricting the recognition search only on the space set by the pre-classification.

3. EXPERIMENTAL RESULTS

Efficiency of the proposed method was examined by simulations on a Mandarin syllable recognition task using a multi-speaker (2 males and 2 females) database. The four speakers spoke a total of 1200 utterances. Each utterance comprises several syllables with a pause located in each pair of two contiguous syllables. The database contains in total 6197 syllables including 5124 training syllables and 1073 testing syllables. Speech signals were first pre-processed for each of 20-ms Hamming-windowed frame with 10-ms shift. A set of recognition features including 12 Mel-cepstral coefficients, 12 delta Mel-cepstral coefficients, energy, and delta energy was computed. Two types of noise were chosen and artificially added to these utterances. They are the car noise from NOISEX-92 database [9] and a computer-generated white Gaussian noise. The HMM recognizer employed 139 sub-syllable models as basic recognition units, including 100 3-state right-context-dependent initial models and 39 5-state context-independent final models. In each state, a partitioned mixture Gaussian distribution with diagonal covariance matrices is used. The number of mixture in each state is variable and is dependent on the number of training samples, but a maximum number of 4 mixtures is set. A single-state noise model with single mixture was used. A conventional one-stage DP search embedded with cumulative bounded-state-duration constraints was used in the recognition process. The syllable recognition rate defined below was used to evaluate the system performance

$$\textit{syllable accuracy} = 1 - \frac{\textit{Subs} + \textit{Dels} + \textit{Ins}}{\textit{num. of testing syllables}}$$

where *Subs*, *Dels*, and *Ins* denote the numbers of substitution, deletion, and insertion errors, respectively.

First, a matched-condition test using clean speech was performed and taken as a benchmark. A syllable recognition rate of 77.5% was achieved. The proposed method was then examined using two types of noises: white and car noises. A simple Log-Add PMC [2] model compensation, which adapts only the mean parameters of HMM, was used in this study. The following systems were tested:

(1) System A: the conventional Log-Add PMC method with noise model being trained according to the manual noise/speech segmentation.

(2) System B: the Log-Add PMC method with noise model being obtained by the proposed RNN-based on-line noise estimation.

(3) System C: an extended version of System B with likelihood scores being compensated by the initial/final/ noise discriminant functions generated by the RNN pre-classifier.

(4) System M0: the conventional HMM method trained and tested under matched condition.

(5) System M1: an extended version of System M0 with the recognition search space being constrained by a manual initial/final/noise pre-segmentation.

In this test, α is set to be 5.0 and λ (the segment length for the noise estimation) is set to be 100 frames. The noise estimate is updated every 150 frames. Table 1 shows the syllable accuracy rates for the five systems. It can be found from the table that the performance of System B is comparable to System A. This confirms that the proposed RNN-based noise estimation is very reliable under noisy environment with unknown SNR. It can also be found from the table that the recognition rate of System C is much better than System B and is slightly worse than System M0. This confirms the effectiveness of the proposed likelihood compensation. System M1 is slightly better than System M0. This shows that the information of broad-class segmentation is useful in recognition search to improve the performance.

Table 1. The experimental results for Car and White noises

Noise Type	SNR	A	B	C	M0	M1
Car	15	59.7	59.3	71.7	73.6	75.8
Car	27	64.6	64.4	73.0	74.4	76.8
Car	39	65.0	64.9	73.4	74.7	77.0
White	15	21.9	21.5	35.8	47.0	52.8
White	27	54.2	53.7	63.3	66.1	68.0
White	39	66.4	66.3	70.9	73.1	74.0

Next, an experiment designed to assess the noise tracking ability of the proposed system under the nonstationary noise environment with time-varying noise gain was performed. The noise gain $P_N(m)$ at sample m of the testing utterance was determined by

$$P_N(m) = P_S - \left[27 + 12\sin\left(\frac{2\pi m}{\beta \cdot f_s}\right)\right]$$

where P_S is the average power of the entire training clean speech, β is a parameter that controls the changing speed of noise level, and f_s is the sampling rate. It is noted that the noise level is sinusodially changed with values in the range of 15 dB and 39 dB. In this test, λ was set to the same value of 10 frames and a shorter updating period of 2 frames was used in order to track the rapidly changing noise level. Table 2 lists the recognition rates for Systems A, B, and C. It can be found from Table 2 that System C

outperforms both Systems A and B. The recognition rate of System B is comparable to System A. For the worst case of β=12, the recognition rates of System B drop 15.3% (from 77.5% to 62.2%) and 37.3% (from 77.5% to 40.2%) for car noise and white noise, respectively. But they only drop 4.4% and 21.8% for System C. This confirms that the proposed system performs very well for nonstationary noise environment with time-varying noise gain.

Table 2. The experimental results for nonstationary noises with time-varying gain

Noise Type	β	A	B	C
Car	12	62.4	62.2	73.1
Car	24	62.4	62.1	73.3
Car	24	62.5	62.7	73.3
Car	96	64.4	63.8	73.6
White	12	42.0	40.2	55.7
White	24	45.9	45.6	57.4
White	48	48.6	47.6	59.7
White	96	54.8	53.9	61.1

4. CONCLUSIONS

In this paper, a robust RNN-based pre-classification scheme for noisy Mandarin speech recognition has been discussed. It pre-segments each input frame into three broad classes in order to estimate noise model for noise compensation as well as to compensate likelihood scores for assisting in speech recognition. Its effectiveness has been demonstrated by incorporating it with an HMM-based 411 Mandarin base-syllables recognizer and tested under white Gaussian and car noises. Experimental results have confirmed that the proposed pre-classification scheme works very well. The whole system is very robust to both stationary and nonstationary noise environments.

5. REFERENCES

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

[2] P.C. Woodland, M. J. F. Gales and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," *ICASSP-96, Vol. 1, pp. 65-68. 1996.*

[3] 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.*

[4] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), hidden Markov models and the projection for robust speech recognition in cars," *Speech Communication, Vol. 11, pp. 215-228, 1992.*

[5] Fei Xie and Dirk Van Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," *ICASSP-94, Vol. 2, pp. 53-56, 1994.*

[6] M.H. SAVOJI, "A Robust algorithm for accurate endpointing of speech signal," *Speech Communication, Vol. 8, pp.45-60, 1989.*

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

[8] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. on Neural Networks, Vol.5, No.2, pp.298-305, 1994.*

[9] Andrew Varga, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication, Vol. 12., pp. 247-251, 1993.*

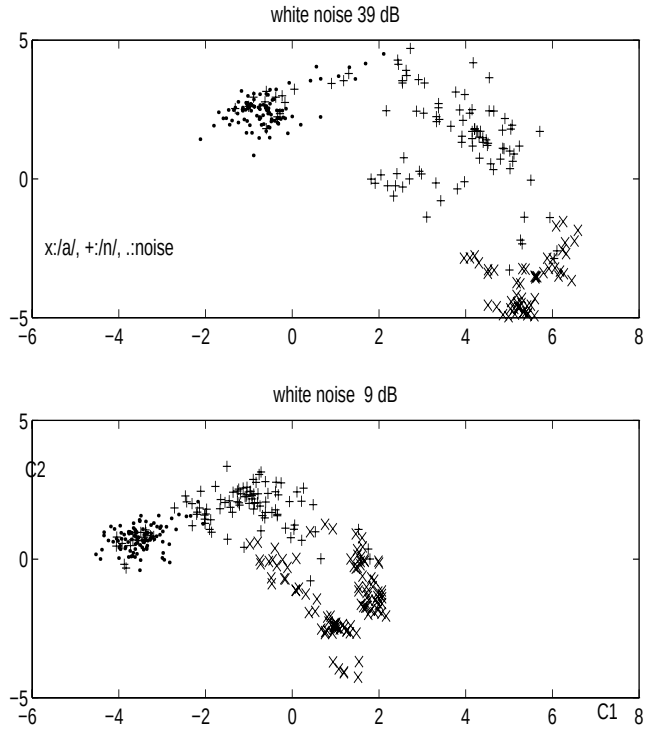

Fig. 2 The scattering plots of cepstral coefficients C_1-C_2 for a pure-noise segment and two phonemes, /a/ and /n/, corrupted by two white Gaussian noises.