NOISE COMPENSATION FOR SPEECH RECOGNITION IN CAR NOISE ENVIRONMENTS

Ruikang Yang and Petri Haavisto

Nokia Research Center
Tampere, SF-33720, Finland
email: ry@rpeltp.research.nokia.com

ABSTRACT

In this paper, a noise compensation algorithm for HMM based speech recognition systems, which utilizes the parallel model combination concept[6], is presented. The algorithm was tested using the TIDIGITS database with artificially added car noise. Very promising results were obtained. The results show that at -10 dB SNR the recognition accuracy could be improved from 34% to 89%. The noise compensation algorithm was also tested using a database which was recorded in a car. Improved performance was obtained, but the improvement was clearly smaller than with the artificially added noise.

1. INTRODUCTION

An important problem in speech recognition is the ability of the recognition algorithm to achieve reliable performance in noisy environments. One application is the hands-free mobile phone in a car where the user can access telephone functions through voice. It has been shown that the performance of speech recognition systems dramatically decreases when they are trained and used in different environments [1]. There have been many studies for achieving reliable performance under noise environments [2]-[7]. One of the major approaches is to modify the pattern matching stage to take the effects of noise into account. Parallel model combination (PMC) is one of such schemes which transforms a set of HMM word models trained with clean speech into a set of models which can be used under the noise conditions of interest [6].

Gales and Young applied PMC to the NOISEX-92 database where speech and noise were artificially added at different SNR levels, and they obtained quite good results. However, only a single Gaussian mixture was assumed in the noise model and the test was speaker-dependent. Since multiple Gaussian mixtures are widely used to achieve better performance, it is useful to study the noise compensation algorithm with multiple Gaussian mixtures. It would also be interesting to see how PMC performs in speaker-independent recognition.

In this paper, we study a noise compensation algorithm for speech and noise models having multiple Gaussian mixtures. It is shown that the compensated models will have $M_s \times M_n$ Gaussian mixtures, where M_s and M_n are the numbers of mixtures for clean speech and noise, respectively. The algorithm was applied to digit recognition in a car noise environments. The TIDIGITS database was used in the experiment. Car noise was added to the clean digits at different SNRs, and the compensated HMMs were used to recognize the noisy digits. Very good results were observed. For instance, the results show that at -10 dB SNR, the recognition accuracy was improved from 34% to 89%. The algorithm is also applied to a car speech database recorded in a parked car with the motor off, while driving downtown, and while driving on a highway. The database consists of over 100 male and female speakers.

2. PROPOSED ALGORITHM

In the noise compensation algorithm, it is assumed that the speech and noise are additive in the linear power domain, and the noise is stationary. Thus a single state noise model is sufficient. Mel-frequency cepstral coefficients are used in the recog-

0-7803-2431-5/95 $4.00 © 1995 IEEE

nition system. The front-end is shown in Fig.1.

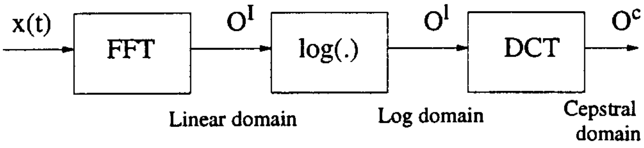

Figure 1: *Front-end of the recognition system*

Let O^c, O^l and O^I represent the observation vectors in the cepstral, log, and linear domain, respectively. Suppose in the cepstral domain the signal can be modeled by a set of Gaussian mixtures, i.e., the density function has the following form,

$$f^c(O^c) = \sum_{m=1}^{M} D_m \mathcal{N}(O^c, {\mu^c}_m, {U^c}_m). \qquad (1)$$

where $\mathcal{N}(O^c, {\mu^c}_m, {U^c}_m)$ denotes the Gaussian distribution, with ${\mu^c}_m$ and ${U^c}_m$ as its mean vector and covariance matrix for the m-th mixture. D_m is the scaling coefficient for the m-th mixture.

The mapping from the cepstral domain to the log domain is a linear transform. This inverse discrete cosine transform is represented by a matrix C^{-1} in

$$O^l = C^{-1} O^c.$$

Then the distribution in the log domain is still a set of Gaussian mixtures, i.e.,

$$f^l(O^l) = \sum_{m=1}^{M} D_m \mathcal{N}(O^l, {\mu^l}_m, {U^l}_m), \qquad (2)$$

with

$$\begin{aligned} {\mu^l}_m &= C^{-1} {\mu^c}_m \\ {U^l}_m &= C^{-1} {U^c}_m (C^{-1})^T. \end{aligned} \qquad (3)$$

When transforming O^l to O^I in the linear domain, it can be shown that the density function in the linear domain is a set of log normal mixtures, i.e.,

$$f^I(O^I) = \sum_{m=1}^{M} D_m \mathcal{L}(O^I, {\mu^I}_m, {U^I}_m). \qquad (4)$$

where $\mathcal{L}(O^I, {\mu^I}_m, {U^I}_m)$ denotes the log normal distribution with ${\mu^I}_m$ and ${U^I}_m$ as mean vector and covariance matrix of each mixture. They have the following relationship with the mean vectors and covariance matrices of Gaussian mixtures in the log domain:

$$\begin{aligned} {\mu^I}_m(i) &= \exp({\mu^l}_m(i) + \tfrac{{U^l}_m(i,i)}{2}) \\ {U^I}_m(i,j) &= {\mu^I}_m(i){\mu^I}_m(j)(\exp({U^l}_m(i,j)) - 1), \end{aligned} \qquad (5)$$

where the indices in the parentheses refer to the vector and matrix elements.

The above mapping process suggests that if the signal can be modeled by a set of Gaussian mixtures in the cepstral domain, then in the linear domain its distribution is a mixture of log normal distributions.

Let ${f_s}^I$ and ${f_n}^I$ denote the density functions for the clean speech and the noise in the linear domain,

$${f_s}^I(O^I) = \sum_{m=1}^{M_s} D_m \mathcal{L}(O^I, {\mu^I}_{sm}, {U^I}_{sm}). \qquad (6)$$

$${f_n}^I(O^I) = \sum_{m=1}^{M_n} E_m \mathcal{L}(O^I, {\mu^I}_{nm}, {U^I}_{nm}). \qquad (7)$$

According to the assumption that in the linear domain the speech and noise are additive and independent, the distribution for noisy speech will be the convolution of ${f_s}^I(O^I)$ and ${f_n}^I(O^I)$, i.e.,

$$\begin{aligned} {f_x}^I(O^I) = & \sum_{i=1}^{M_s}\sum_{j=1}^{M_n} D_i E_j \mathcal{L}(O^I, {\mu^I}_{sm}, {U^I}_{sm}) \\ & * \mathcal{L}(O^I, {\mu^I}_{sm}, {U^I}_{sm}), \end{aligned} \qquad (8)$$

where $*$ stands for the convolution operation. If the convolution of two log normal functions is assumed to be approximately log normal, as is assumed in the single mixture PMC [6], then the distribution for noisy speech in the linear domain is a set of log normal mixtures. The number of mixtures for noisy speech is $N = M_s \times M_n$, and the density is

$$f^I({O_x}^I) = \sum_{m=1}^{N} H_m \mathcal{L}(O^I, {\mu^I}_{xm}, {U^I}_{xm}), \qquad (9)$$

where

$$H_m = D_i E_j$$

$$\mu^I{}_{xm} = g\mu^I{}_{si} + \mu^I{}_{nj} \quad . \qquad (10)$$

$$U^I{}_{xm} = g^2 U^I{}_{si} + U^I{}_{nj}$$

where g is a gain matching term [6] and

$$i = 1, \cdots, M_s, \quad \text{and} \quad j = 1, \cdots, M_n$$

$$and \quad m = (i-1)M_s + j.$$

Therefore, the noise compensation process is straightforward. Given the HMMs for clean speech and noise in the cepstral domain, their model parameters in the linear domain can be calculated using Eq.(3) and Eq.(5). Then compensation of the clean speech model by the noise model in the linear domain according to Eq.(10) is performed to get the model for noisy speech. The model parameters in the cepstral domain can be calculated by inversing Eq.(3) and Eq.(5).

3. EXPERIMENT RESULTS

In order to evaluate the performance of the algorithm, we applied it to two databases. The first database is TIDIGITS with car noise artificially added. The second is a car speech database recorded in real car environments. In the experiments, only single digits were used, though the compensation algorithm is generally applicable for connected digit recognition.

As shown in Fig.2, the models of the clean speech are first obtained by training them with the single digits in the database. Each digit model had 10 states, and a single-state model was used for noise. The recognition system updates the composite models by modifying the clean speech models according to the noise model, and the composite models are used to recognize the noisy speech.

In the first experiment, car noise with different SNRs was added to the single digits in TIDIGITS. The clean digit models were trained with the training set of the database. For different SNRs the noise model was obtained by training with the non-speech segments of the noisy digits. Fig. 3 shows the recognition results with and without noise compensation. It also shows the HMMs with 5 Gaussian mixtures achieve better performance in both cases.

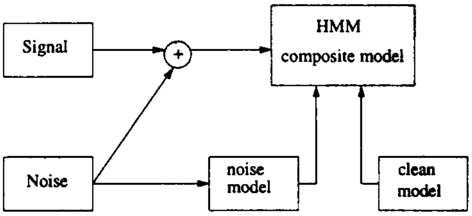

Figure 2: *Model updating according to the input noise*

In the earlier paper [6], only mean vectors were updated for each digit model. In order to see the difference it makes by updating state variances we have two options in our experiment, i.e. updating the mixture means only, or updating both the means and the variances. Table 1 shows the results for the two options. It is observed that updating both mean and variance improves the recognition performance.

It is interesting to note that the noise compensation algorithm is very effective in achieving noise robustness. For instance, it can increase the recognition accuracy from 34% to 89% under -10 dB SNR.

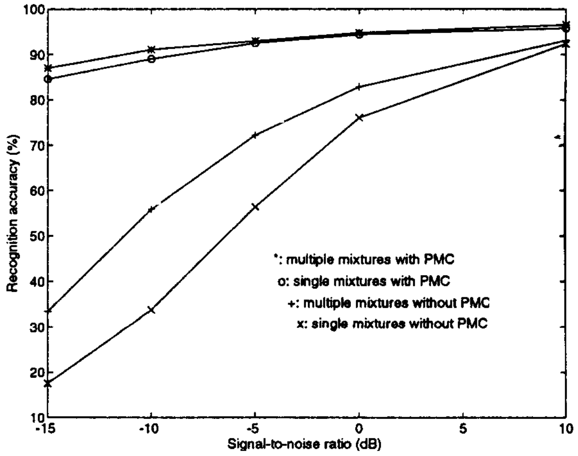

Figure 3: *Test results for TIDIGITS with additive car noise*

Table 1: *Test results by updating the mean only and updating both the mean and variance*

	without compensation	with compensation	
SNRs (dB)		mean	mean & var.
+10	92.3	95.1	95.8
0	76.1	92.5	94.4
-5	56.4	88.6	92.5
-10	33.8	77.0	89.0
-15	17.5	49.9	84.6

In the second experiment, we applied the algorithm to a car speech database. The database was recorded under different car environments: parking place, downtown, and highway. The clean speech models were trained with the digits recorded when the car was parked. The noise model was trained with the noise segments where no speech was present. The test results are shown in Table 2. It is observed that improved performance was obtained with PMC. However, the improvement is not as much as that for the first experiment. This can be explained by the fact that the speech and noise are not independent as are assumed in PMC. Therefore, other measures may have to be employed to improve the performance further.

In the noise compensation algorithm, only static cepstral coefficient compensation was used here. As is well known, delta and delta-delta cepstral coefficients can be used to improve the recognition performance. Therefore, it is necessary to extend PMC with delta and delta-delta coefficient compensation. Gales and Young proposed a method to compensate the delta cepstral coefficients[7]. However, their method applies only for a special case and is not valid for the general cases where delta cepstral coefficients are calculated by using longer linear regression. Thus there is still room to enhance the noise compensation scheme.

Table 2: *Test results for the car database*

No.of mixtures	Parking place	Highway	Highway with PMC
1	96.8	45.5	64.9
3	98.0	46.0	65.8
5	98.9	46.5	66.1

4. CONCLUSIONS

We have proposed a noise compensation algorithm to enhance the robustness of speech recognition systems under noisy environments. The algorithm applies to the cases where both speech and noise models use multiple Gaussian mixtures. Experiments have confirmed the effectiveness of the algorithm.

5. REFERENCES

[1] B.H.Juang, “Speech recognition in adverse environments”, *Computer Speech and Language*, vol.5. pp.275-294, 1991.

[2] D.Van Compernolle, “DSP Techniques for Speech Enhancement”, *Proc. of Speech Processing in Adverse Conditions*, pp.21-30, Nov. 1992.

[3] D.Van Compernolle, ”Spectral estimation using a log-distance error criterion applied to speech recognition”. *Proc. Int. Conf. Acoust. Speech Signal Process.*, 21.S6.2, pp.258-261, 1989, Glasgow. U.K..

[4] H.B.D. Sorensen, ”A cepstral noise reduction multilayer neural network”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp.933-936, 1991, Toronto, Canada.

[5] B.A. Mellor and A.P. Varga, ”Noise masking in the MFCC domain for the recognition of speech in background noise”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol.14, Part 6, pp.361-368, 1992, San Fransico, USA.

[6] M. Gales and S. Young, “Cepstral parameter compensation for HMM recognition in noise”, *Speech Communication*, vol.12, pp.231-239, 1993.

[7] M. Gales and S. Young, “HMM recognition in noise using parallel model combination,” *Proc. of Eurospeech-93*.