Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks

J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado

Dpto. de Electrnica y Tecnologa de Computadores
Universidad de Granada, 18071 GRANADA (Spain)
segura@ugr.es

Abstract

In this paper we apply a model-based compensation method to cancel the effect of the additive noise in Automatic Speech Recognition systems. The method is formulated in a statistical framework in order to perform the optimal compensation of the noise effect given the observed noisy speech, a model describing the statistics of the speech recorded in a clean reference environment and the estimation of the noise in the noisy recognition environment. The noise is estimated using the first frames of the sentence to be recognized and a frame-by-frame noise compensation algorithm is performed, so that the compensation procedure does not constrain real-time speech recognition systems and is compatible with emerging technologies based on distributed speech recognition.

We have performed recognition experiments under noise conditions using the AURORA II database for the recognition tasks developed for this database as a standard reference. Experiments have been carried out including both, clean and multicondition training approaches. The experimental results show the improvements in the recognition performance when the proposed model-based compensation method is applied.

1. Introduction

Noise degrades significantly the performance of Automatic Speech Recognition (ASR) systems running in real conditions [1]. Usually methods to compensate the effect of noise must be applied in order to perform an accurate enough recognition process. Otherwise, the degradation of the performance due to the noise would cause an improper operation of the ASR based application. The degradation of the recognizers is mainly due to the mismatch between training and recognition conditions and most of the methods for robust speech recognition are focused in the minimization of the mismatch and can be categorized in one of these groups [2]:

- Robust parameterizations: the speech signal is represented using parameters that are minimally affected by the noise.
- Compensation of the noise effect over the representation of the speech: these methods try to remove the noise from the parameters representing the speech.
- Adaptation of the models to noise conditions: the models in the recognizer are contaminated in order to properly model the noisy speech.

This way, for example, front-ends based on Mel Frequency Cepstral Coefficients (MFCC) are widely extended since these parameters are shown to be robust against different kinds of noise [3]. In the second category, methods like Cepstral Mean Normalization or Spectral Subtraction [3] [4] have been considered to compensate the effect of the noise, and methods like Parallel Model Combination [5] are useful to adapt the speech pattern to noise conditions.

In this paper we present a model-based noise compensation method to remove the effect of the noise over the representation of the speech signal. The method is based on a statistical formulation and provides the expected value of the clean speech representation given (a) the noisy speech, (b) an estimation of the noise and (c) a model describing the clean speech. The clean speech model is based on a set of multivariate Gaussian pdf's and have a precedent in the Vector Taylor Series approach [6] [7] [8]. The compensation procedure is performed in the logarithmic filter-bank energies (log-FBE) domain, and is compatible with filter-bank based representations, like the MFCC front-end. We have combined the model-based noise compensation method with a band pass filtering procedure in the log-FBE domain in order to reduce the residual noise.

The compensation method has been tested under several noise conditions using the AURORA II database and task set. The speech recognition experiments have been carried out for clean and multicondition training conditions as proposed in the AURORA test. The results presented in this work show that important improvements in the recognition performance can be obtained when the proposed compensation method is applied.

2. Effect of the additive noise in ASR

In real applications of speech recognition, the speech signal is usually affected by additive background noise, due to other audio sources in the environment where the speaker is. Assuming that the speech and noise signals are uncorrelated, the output energy of the filter b in the filter bank at frame t, corresponding to the noisy speech $Y_b(t)$ can be written as a function of the energy of the clean speech $X_b(t)$ and the noise $N_b(t)$,

$$Y_b(t) = X_b(t) + N_b(t) \tag{1}$$

and the relation in the log-FBE domain ($x_b = \log(X_b)$) is described by the equation,

$$y_b(t) = \log[\exp(x_b(t)) + \exp(n_b(t))] \tag{2}$$

Therefore, the effect of the additive noise consists of a non-linear transformation of the representation space in the log-FBE domain which is propagated to the MFCC-based representation, producing a mismatch between the clean and the noisy conditions. This degrades the performance of the recognizer when it is trained with clean speech and the the recognition is performed using speech acquired in noisy environments.

Additionally, as the noise is a random process, the additive noise affecting a band, $n_b(t)$ does not take a constant value, and therefore the relation between $x_b(t)$ and $y_b(t)$ cannot be described as a deterministic transformation, but in a probabilistic framework. From the noisy signal, an estimation of some parameters describing the noise statistics $p(n_b)$ is possible, but not the estimation of the exact value $n_b(t)$ of the noise in the band b at frame t. So, given a value of the observed noisy speech $y_b(t)$, and an estimation of the noise statistics $p(n_b)$, a proper compensation method should estimate the expected value of the clean speech parameters constrained to the observed noisy speech and the noise statistics. In addition, the use of a model describing the statistics of the clean speech $p(x_b)$ would provide a more accurate estimation of the clean speech, that should be calculated as the expected value $\hat{x}_b(t) = E[x_b|y_b(t), p(n_b), p(x_b)]$.

3. Model-based compensation of the additive noise

The compensation method we propose is based on a statistical model describing the distribution of the clean speech in the log-FBE domain as a K-Gaussians mixture,

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(v_k)\mathcal{N}(\mathbf{x}, \mu_{x,k}, \Sigma_{x,k}) \tag{3}$$

where v_k is the k-th Gaussian pdf, with mean $\mu_{x,k}$ and covariance matrix $\Sigma_{x,k}$ (assumed to be diagonal). From the equation (2), the relation between x_b and y_b can be written in vectorial notation,

$$\mathbf{y}(\mathbf{x}, \mathbf{n}) = \mathbf{x} + \mathbf{g}(\mathbf{x}, \mathbf{n}) \tag{4}$$

where the mismatch function is,

$$g_b(\mathbf{x}, \mathbf{n}) = \log[1 + \exp(n_b - x_b)] \tag{5}$$

and the definition of an auxiliary function $\mathbf{f}(\mathbf{x}, \mathbf{n})$ with the components,

$$f_b(\mathbf{x}, \mathbf{n}) = \frac{1}{1 + \exp(x_b - n_b)} \tag{6}$$

is useful since $\partial g_b/\partial n_b = -\partial g_b/\partial x_b = f_b$. By applying a 1st order Taylor series approach to the equation (4), it is possible to estimate how the mean and covariance matrix $\mu_{x,k}$ and $\Sigma_{x,k}$ of the Gaussian v_k are affected by an additive noise with mean μ_n and covariance matrix Σ_n,

$$\mu_{y,k}(b) \approx \mu_{x,k}(b) + g_b(\mu_{x,k}, \mu_n) \tag{7a}$$

$$\begin{aligned}\Sigma_{y,k}(b,b) \approx\ & [1 - f_b(\mu_{x,k}, \mu_n)]^2 \Sigma_{x,k}(b,b) + \\ & [f_b(\mu_{x,k}, \mu_n)]^2 \Sigma_n(b,b)\end{aligned} \tag{7b}$$

where b denotes the component associated to the band b in the filter bank and the covariance matrices are assumed to be diagonal. The estimation of the contaminated Gaussians allows an estimation of the clean speech given the observed contaminated speech, the clean speech model and an estimation of the noise,

$$\hat{\mathbf{x}}(t) = E[\mathbf{x}|\mathbf{y}(t), p(\mathbf{x}), p(\mathbf{n})] = E[\mathbf{y} - \mathbf{g}|\mathbf{y}(t), p(\mathbf{x}), p(\mathbf{n})]$$

$$\approx \mathbf{y}(t) - \sum_{k=1}^{K} P(v_k|\mathbf{y}(t))\mathbf{g}(\mu_{x,k}, \mu_n) \tag{8}$$

where the probabilities $P(v_k|\mathbf{y}(t))$ are estimated using the Gaussian pdfs contaminated according to the equations (7a) and (7b),

$$P(v_k|\mathbf{y}) = \frac{P(v_k)\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})}{\sum_{k'=1}^{K} P(v_{k'})\mathcal{N}(\mathbf{y}, \mu_{y,k'}, \Sigma_{y,k'})} \tag{9}$$

In contrast to the iterative algorithm proposed in [6] [7] [8] we estimate the noise parameters using the first 10 frames of signal (that are assumed to be silence) which allow the parameterization and the compensation procedures to run in real time. This provides an accurate enough estimation of the mean vector of the noise in the log-FBE domain. However the accuracy in the estimation of the covariance matrix is very poor.

4. Band pass filtering

According to a previous analysis of the distribution of the mutual information between the phonetic classes and the parameters in the logarithmic filter-bank energies domain, most of the acoustic information is contained in the frequency range between 2 and 15 Hz. Variations in the log-FBE domain slower than 2 Hz are associated to the average energy of the background noise and those variations faster than 15 Hz are mainly due to the random behavior of the noise. For this reason, we have combined the model-based noise compensation method with a band pass filtering procedure in the log-FBE domain in order to reduce the residual noise that the compensation method cannot remove. We have applied a 41 coefficients FIR band pass filter with a band between 2 and 15 Hz to each component of the compensated vectors in the log-FBE domain.

5. Experimental results

5.1. The recognition task

The model-based compensation method has been tested with the AURORA II database and the standard recognition tasks prepared for this database. This database contains English connected digits recorded in clean environments. Three sets of sentences (set A, set B and set C) have been prepared for the recognition experiments. The sentences have been artificially contaminated by adding noise recorded under several conditions (subway, babble, car, exhibition, restaurant, etc.). The sentences are contaminated for different noise levels, with SNRs ranging from -5 dB to 20 dB. Recognition experiments using clean sentences (no noise added) have also been carried out.

The recognition systems used as reference are based on continuous hidden Markov models (CHMM). One CHMM has been trained for each digit. Both, training and recognition processes are performed using the HMM Tool Kit (HTK) software, as proposed in the AURORA II documentation. The speech parameterization used for the reference experiments is based on standard Mel Frequency Cepstral Coefficients (MFCC). Two groups of experiments have been performed: (a) using a recognizer trained with clean speech (Clean training) and (b) training the recognizer with sentences contaminated with different kinds and levels of noise (Multicondition training).

The methods explored to improve the recognition performance under noise conditions only affects the front-end. Therefore, all the procedures for training and recognition are identical to the reference experiments with the exception of the front-end, which includes the model-based compensation procedure.

5.2. Band Pass Filtering of log-FBE parameters

We have performed a preliminary recognition experiment (only for set A) in order to understand the effect of the band pass filtering over the log-FBE parameters. In this experiment, the log-FBE parameters are filtered with a 2-15 Hz band pass filter before computing the Mel-cepstrum parameters, and no other compensation method is considered. Figure 1 compares the performance of the recognizer when Band Pass Filtering is applied. This results are the average over the different noise types considered in set A. As can be observed an improvement is provided by the Band Pass Filtering under Clean Training conditions and the performance is similar to the reference for Multicondition Training conditions. Therefore, we can conclude that the selected band preserves most of the discriminative information relevant for speech recognition and reduces part of the noise effects. For this reason, we have combined the Band Pass Filtering and the Model-Based Compensation method in the following experiments.

5.3. Model-Based Compensation of the noise

We have applied the model-based noise compensation method described in section 3 in order to compensate the effect of the noise over the log-FBE representation of the speech. The noise parameters are estimated using the first 10 frames of the signal (100 ms) which are assumed to be silence. This allow the compensation procedure to run in real time. Even though this provides an accurate enough estimation of the mean vector of the noise in the log-FBE domain, the accuracy in the estimation of the covariance matrix is very poor. For this reason, we only apply a 0th order Taylor series approach and we compensate the mean of the Gaussians according to equation (7a) but the covariance matrices of the Gaussians are not modified. We have estimated a set of 128 diagonal Gaussian pdfs in the log-FBE domain as a clean speech model to be used for the model-based compensation procedure. The Gaussians has been estimated using the Clean conditions training database of AURORA II.

Figure 2 shows the recognition results (averaged over the sets A, B and C and different types of noises) as a function of the SNR. The results applying the model-based compensation method and the band pass filtering are compared to the reference. An important improvement with respect to the reference can be observed for Clean training conditions. However, a degradation with respect to the reference is appreciated for Multicondition training.

We have investigated which factors are involved in this degradation in the Multicondition training case. We have observed that in the Multicondition case, the rates of correct words are similar for both, with and without compensation, and the degradation observed in the Word Accuracy is mainly due to an increment in the number of insertions when the compensation method is applied. We have also observed that the increment in the number of insertions is associated to improper estimation of the noise level affecting each band in the filter bank. This is due to the fact that most of the considered noises are not stationary. We have observed important differences between the noise levels at the beginning (first 10 frames) and at the end (last 10 frames) of the sentence: the typical difference between the noise estimation at the beginning and at the end is 2 dB, but there are sentences for which the difference exceeds 10 dB. This means that using a constant value for the noise level along all the sentence leads to an improper estimation of the noise in some parts of the sentence and therefore to an improper behavior of the compensation procedure.

5.4. Linear interpolation of the estimation of the noise

In order to deal with a non-stationary noise contaminating the speech signal, we have estimated the noise at the beginning and at the end of the sentence (using the first and last 10 frames, which are assumed to be silence) and we have estimated the noise affecting each frame as a linear interpolation between these values. This method for the noise estimation provides a more accurate model of the noise, specially interesting for non-stationary noises, but present the drawback that the compensation cannot be performed until all the sentence has been recorded, which should be considered for real-time applications. In addition, since the estimated noise level is different for each frame, the set of noisy Gaussians must be re-computed for each frame, and this increases significantly the computing load of the procedure. The experimental results when the linear interpolation is used for the noise estimation are shown in Figure 3. As can be observed, this approach improves the performance of the model-based compensation procedure.

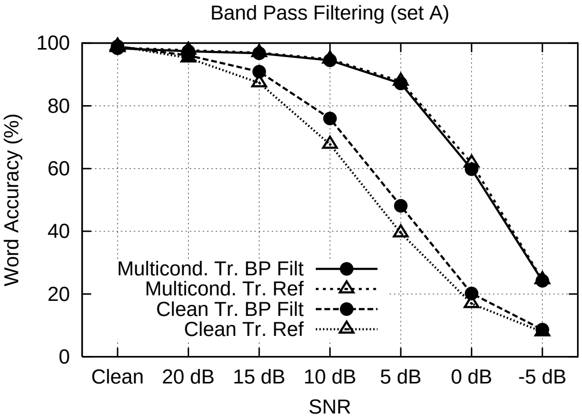

Figure 1: *Recognition results when the Band Pass Filtering is applied (only for set A).*

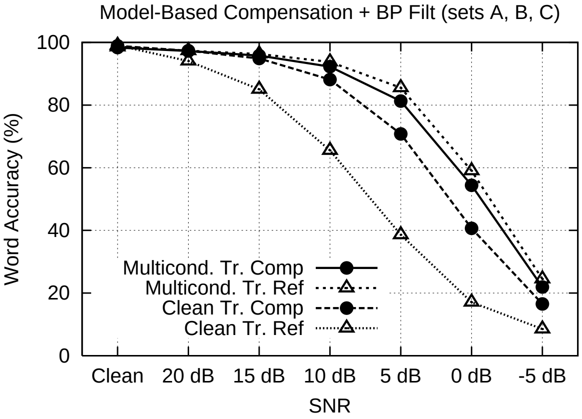

Figure 2: *Recognition results when the Model-Based Compensation method is applied (average over sets A, B, C). The compensation method is combined with Band Pass Filtering.*

5.5. Summary of recognition results

The results are summarized in Tables 1 and 2. Table 1 presents the recognition results averaged over the SNR range 0 - 20 dB and Table 2 shows the performance of the studied methods relative to the reference system (Mel-cepstrum front-end).

6. Conclusions

In this paper we have presented a model-based noise compensation method for automatic speech recognition systems. This method estimates the expected value of the clean speech given the observed noisy speech, an estimation of the noise and an

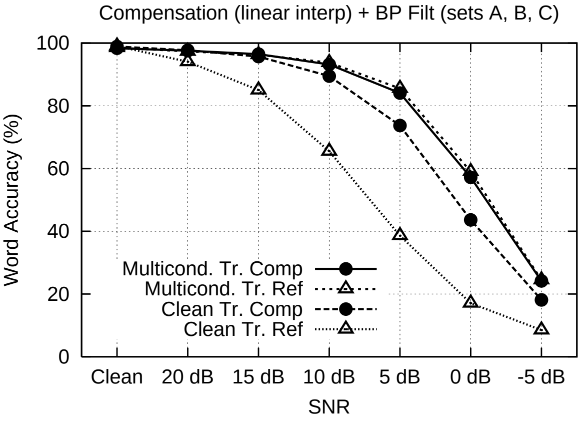

Figure 3: *Recognition results for Model-Based Compensation method where the noise level is estimated using a linear interpolation between the beginning and at the end of the sentence (average over sets A, B, C). The compensation method is combined with Band Pass Filtering.*

statistical model of the clean speech. The method is combined with a band pass filtering procedure in the log-FBE domain.

The proposed noise compensation method has been tested using the AURORA II data base and speech recognition set up. The method provides important improvements when the recognizer is trained in clean conditions. However, under multicondition training, a degradation is observed with respect to the reference recognizer. We have found that this degradation is associated to the evolution of the noise level along the sentence, due to the non-stationary properties of the noises considered in the AURORA II recognition test.

We have proposed a method to deal with non-stationary noises, which estimates the noise for each frame as the linear interpolation between the noise levels estimated at the beginning and at the end of the sentence. The use of linear interpolation for the estimation of the noise improves significantly the performance of the recognizer.

The experiments described in this paper shows the importance of the proper modeling of the noise and the adequate noise estimation procedures in order to improve the performance of speech recognizers working under noise conditions.

7. Acknowledgments

This work has been supported by the Spanish Government under the CICYT project TIC99-0583.

8. References

[1] Y. Gong, “Speech recognition in noisy environments: A survey”, Speech Communication, vol 16, no 3, pp. 261-291, 1995.

[2] Bellegarda, J. R., “Statistical techniques for robust asr: review and perspectives”, Proc. of EuroSpeech-97, pp. KN 33-36, 1997.

[3] C. Jankowski, J. Hoang-Doan and R. Lippmann, “A comparison of signal processing front ends for automatic word recognition”, IEEE Trans. on Speech and Audio Processing, vol. 3, pp. 286-293, July 1995.

[4] S. Vaseghi and B. Milner, “Noise compensation methods for hidden markov model speech recognition in adverse environments”, IEEE Trans. on Speech and Audio Processing, vol. 5, pp. 11-21, Jan. 1997.

[5] M. F. J. Gales and S. J. Young, “HMM recognition in noise using parallel model combination”, Proc. of EuroSpeech-93, vol. 2, pp. 837-840, 1993.

[6] P. J. Moreno, “Speech Recognition in Noisy Environments”, PhD thesis, Carnegie Mellon University, Pittsburgh, Pensilvania, April 1996.

[7] P. J. Moreno and B. Eberman, “A new algorithm for robust speech recognition: the delta vector Taylor series approach”, Proc. of EuroSpeech-97, vol 5, pp. 2599-2602, 1997.

[8] R. M. Stern, B. Raj, and P. J. Moreno, “Compensation for environmental degradation in automatic speech recognition”, ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Chanels, pp. 33-42, April 1997.

Reference (Mel-cepstrum)				
Training mode	set A	set B	set C	overall
Multicondition	87.82	86.27	83.78	86.39
Clean only	61.34	55.75	66.14	60.06
Average	74.58	71.01	74.96	73.23

Model-Based Compensation + BP Filt				
Training mode	set A	set B	set C	overall
Multicondition	85.60	82.85	84.06	84.19
Clean only	79.07	78.42	76.87	78.37
Average	82.33	80.63	80.46	81.28

Model-Based Comp. (linear interp) + BP Filt				
Training mode	set A	set B	set C	overall
Multicondition	86.81	84.48	85.89	85.69
Clean only	80.08	80.63	78.87	80.06
Average	83.44	82.55	82.38	82.87

Table 1: *Performance of the recognition systems averaged over the SNR range between 0 and 20 dB.*

Model-Based Compensation + BP Filt				
Training mode	set A	set B	set C	overall
Multicondition	-18.2%	-24.9%	1.7%	-16.1%
Clean only	45.9%	55.2%	31.7%	45.8%
Average	13.8%	13.2%	16.7%	14.8%

Model-Based Comp. (linear interp) + BP Filt				
Training mode	set A	set B	set C	overall
Multicondition	-8,3%	-13.0%	13.0%	-5.1%
Clean only	48.5%	56.2%	37.6%	50.1%
Average	20.1%	21.6%	25.3%	22.5%

Table 2: *Performance of the recognition systems relative to the reference system (Mel-cepstrum).*