Including uncertainty of speech observations in robust speech recognition

M.C. Benítez, J.C. Segura, A. dela Torre J. Ramírez, A. Rubio.

Department of Electronics and Computer Technology
University of Granada, Spain

carmen@ugr.es, segura@ugr.es, atv@ugr.es, javierrp.@ugr.es, rubio@ugr.es

Abstract

Noise compensation methods for speech recognition provide a cleaned version of the speech representation. Usually this cleaned version is the expected value of the speech parameters given the observed noisy speech and the noise statistic. A more realistic representation should include the probability distribution of the cleaned speech instead of its expected value in order to represent the uncertainty associated to the compensation process due to the variability of the noise process. Recently, the inclusion of the uncertainty in the recognition process has been studied. Some approaches represent the uncertainty in the HMM parameters values. Other approaches represent it in the feature space. This second approach offers a much simpler system implementation and lower computational cost.

In this paper we have developed a noise compensation technique that incorporates the variance of the cleaned speech into the speech representation. The variance is estimated using a Wiener filter during the speech feature enhancement process. This way of including the uncertainty implies the modification of the decoding rule. Experimental results using AURORA 2 database demonstrate a sustained improvement of the performance in the recognition system (about 21% word error rate reduction) when uncertainty is considered in the decoding rule.

1. Introduction

Noise significantly degrades the quality of speech and modifies the statistics of its feature vectors. For this reason automatic speech recognizers suffer a loss of accuracy. The goal of robust speech recognition methods is to reduce the effect of the noise. Some methods work in the feature space with the aim of finding feature vectors that are less sensitive to the noise. Others attempt to find transformations which adapt the noisy feature vectors to the reference models. A third type of methods try to adapt the models to the noise conditions.

The effect of the noise over the log-energy representation consists on a non-linear transformation. The noise causes a distortion in the representation space (the low energy region is compressed to the noise level while the high level energies are not altered) and moves the optimal decision border for classification [1, 2]. If the noise level is constant (with zero variance) it can be shown that the error probability (the overlap between the clean speech and the noise distributions) is not affected by the noise [1, 2]. In this case the problem consists on the estimation of the noise energy level, with the highest possible precision, in order to obtain the best estimation of the clean speech from the noisy signal. Moreover, the performance of the recognizer would not be affected by the noise if the compensation of the features was performed with the exact estimation of the noise level but, of course, this is an ideal situation.

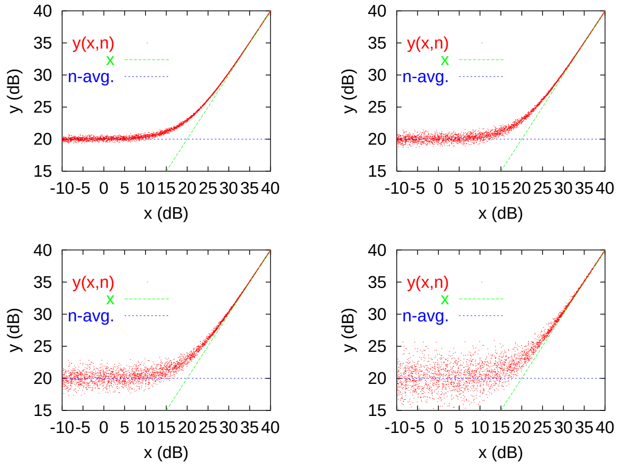

Figure 1: *Mapping of the clean speech signal and contaminated speech signal in the log energy domain. The speech has been contaminated using additive noise with a Gaussian distribution; $\mu_n = 20dB$; $\sigma_n = 0.25, 0.50, 1.0$ and 2 dB.*

A more realistic situation is to consider that the variance of the noise is not zero. In this case, a probability distribution (in the log-filter-bank-energy domain) $p_n(n_b)$ describes the noise energy at the band b, and the probability distribution of the contaminated signal y_b is a function of the clean speech energy x_b, and the noise distribution:

$$p_y(y_b|x_b) = p_n(n_b(y_b))\frac{\partial n_b}{\partial y_b} \tag{1}$$

where

$$n_b(y_b) = y_b + \log(1 - \exp(x_b - y_b)) \tag{2}$$

and

$$\frac{\partial n_b}{\partial y_b} = \frac{1}{1 - \exp(x_b - y_b)} \tag{3}$$

Figure 1 shows a plot of the clean and contaminated signals in the log energy domain. The contaminated one has been obtained by adding Gaussian noise with mean $\mu_n = 20dB$ and standard deviations $0.25, 0.50, 1.0$ and $2.0dB$, respectively. Note that: (i) each value of the clean signal is transformed into a probability distribution; (ii) when the energy of the clean signal is higher than the mean noise level, the noisy signal shows a narrow distribution; as the clean signal energy approximates the mean noise level, the noisy signal distribution gets broader and, (iii) a loss of information is caused by the random nature of the noise; this loss of information depends on the standard deviation of the noise and the energy of the clean signal.

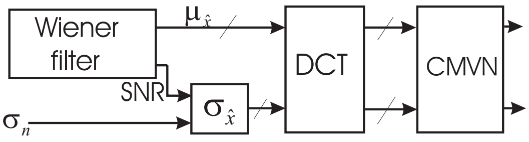

Figure 2: *Block Diagram of the feature extraction algorithm*

The inclusion of the uncertainty has been recently studied by different researchers [3, 4, 5, 6]. Some approaches represent the uncertainty in the HMM parameters, and use the Bayesian Predictive Classification (BPC) rule for decoding [3, 4]. Other approaches represent the uncertainty in the feature space [5, 6]. The main advantage of working in the feature space is the simplicity and its low computational cost. In both cases, results have shown the utility of their inclusion in the recognition process.

Following the second approach, there are two questions related with the idea of exploiting the uncertainty in the feature space; the first one is how to estimate it, and the second one is how to incorporate it into the recognizer. In this work we carry out a preliminary study of the inclusion of the uncertainty in the decoding process. We present a procedure to estimate the mean and the variance of the cleaned signal using Wiener filtering. With this approximation we have obtained an important improvement of the performance in the recognition system.

2. Feature extraction algorithm

Classical feature enhancement algorithms as spectral subtraction, Wiener filtering or VTS (Vectorial Taylor Series) [8, 9, 10, 11] discard the uncertainty information. In this section we propose a new technique, based on Wiener filtering, to update mean and variance of the clean speech estimations. The goal is to obtain a probability density function (which we assume to be Gaussian) to describe the statistics of the enhanced speech and use both, its mean value and variance, in the decoding process.

Figure 2 presents a block diagram of the feature extraction algorithm. After the Wiener filter we obtain an estimation of the mean value of the Gaussian and an estimation of the local SNR which will be used to estimate the variance of the Gaussian.

2.1. Computing the expected values for the clean speech

Assuming that the speech and noise are uncorrelated signals, the relation in the log-Filter-bank-energy domain between clean speech, noisy speech and noise for each filter is:

$$y_b = \log(\exp(x_b) + \exp(n_b)) \tag{4}$$

Wiener filtering [9, 10] provides an estimation of the clean speech given the noisy speech and the average SNR:

$$\hat{x}_b = y_b - \log(1 + \exp(\overline{n_b} - \overline{x_b})) \tag{5}$$

We have estimated the expected value, $\mu_{\hat{x}}$, as the average of the clean speech estimations over three consecutive frames.

2.2. Computing variance of the enhanced speech

In our previous work [12], we developed an expression for the expected value and variance of the estimator of the clean speech using VTS noise compensation algorithm. As it was also pointed out in the paper, the expression of the estimator is similar to the one of the Wiener filter. From these results we assume that the influence of the noise in the variance of the estimator is:

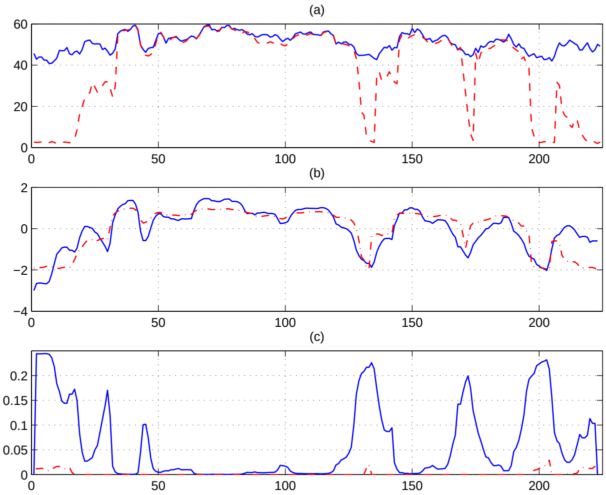

Figure 3: *(a) logarithmic energies for a clean signal (dashed) and a contaminated signal (5dB) (solid). (b) logarithmic energies of the signals after the Wiener filtering and the MVN (same color code).(c) estimation of the variances of both signal.*

$$\sigma_{\hat{x}}{}^2 = \left(\frac{1}{1 + \exp(\mathrm{SNR})}\right)^2 \sigma_n{}^2 \tag{6}$$

This equation depends on the variance of the noise ($\sigma_n{}^2$) and the local signal to noise ratio (SNR). This expression is also consistent with the noise effects discussed in Figure 1. That is, for high SNR values, the variance of the estimator goes to zero; for SNR$=0$, $\sigma_{\hat{x}}{}^2 = \sigma_n{}^2/4$ and finally, when the SNR goes to $-\infty$, $\sigma_{\hat{x}}{}^2 = \sigma_n{}^2$.

2.3. Cepstral mean and variance normalization of the expected features

After the computation of the expected values and variances of the log energies at the output of a 23 channel Mel scaled filter bank, a DCT transformation is applied to both, means and variances, so the number of parameters is reduced to 26 (13 means and 13 variances); derivatives and accelerations are also calculated. The cepstral coefficients are then normalized to have zero mean and unity variance using cepstral Mean and Variance Normalization algorithm (CMVN) as in [13]. Figure 3 shows a graphic example for two signals; a clean speech signal (selected from the clean training set of AURORA 2 database) and the same speech signal contaminated with a noise at 5dB. In Fig 3 (a) logarithmic energies of clean and noisy sentences are represented. In Fig 3 (b) both signals have been enhanced with the Wiener filtering and then normalized with the CMVN criterion. In Fig 3 (c), we plot the estimated variances for the log energies. We can observe that for the clean signal case, the variance is very low. These results are consistent because the signal has not been contaminated and so there is no uncertainty. For the contaminated signal, the variance changes with time: the non speech regions correspond to higher variance values, and the variance of the speech region depends on the local SNR.

3. Including uncertainty in the decoding process

In the previous sections we have studied the uncertainty in the enhanced speech signal due to the noise, and the convenience to represent the clean speech estimation by a probability density function instead of just the mean value. In this paper, we assume a Gaussian distribution to represent the clean signal given the noisy signal, and we have described a method to estimate its parameters (mean and variance). Now the problem is how to incorporate it into the decoding algorithm of the recognizer.

In the classical noise reduction methods for robust speech recognition, the input of the recognizer is just the output of the noise reduction algorithm. The recognition process relies on the evaluation of a set of Gaussian mixture components evaluated at the cleaned observations. The probability density of the cleaned observation $\hat{x}$ given the Gaussian k is:

$$p(\hat{\mathbf{x}}|k) = \mathcal{N}(\mu_{\hat{\mathbf{x}}}; \mu_k, {\sigma_k}^2) \tag{7}$$

If a Gaussian $p(x) = \mathcal{N}(x; \mu_x, {\sigma_x}^2)$ is used to represented the probability of x, instead of the expected value, Eq. (7) can be modified:

$$p(\mathbf{x}|k) = \int \mathcal{N}(x; \mu_x, {\sigma_x}^2)\mathcal{N}(x; \mu_k, {\sigma_k}^2)dx =$$

$$= \mathcal{N}(\mu_{\mathbf{x}}; \mu_k, {\sigma_k}^2 + {\sigma_x}^2) \tag{8}$$

and therefore, including uncertainty on the recognition process implies that the variance of the Gaussian is modified according to Eq. (8). Note that according to Eq. (6), ${\sigma_x}^2$ depends on the local SNR and the variance of the noise. Therefore, Eq. (7) can be considered a particular case of Eq. (8) valid for high values of local SNR or when the variance of the noise is very low.

4. Speech recognition experiments on the AURORA 2 taks

Recognition experiments were conducted on the AURORA2 database. The AURORA task consists of connected digit strings, artificially contaminated with various noise types at multiple SNR values. In this database two training sets (clean training and multi condition training) and three test sets (A, B and C) are available. Noises used in test set A are the same that are used in multi condition training. Noises used in B do not appear in the multi condition training set, and test set C includes channel effects. In both types of recognition experiments (clean and multi condition) the models are trained without including uncertainty.

The baseline recognition system is based on HTK. The system uses continuous density HMM models with 6 Gaussians per state. There are 11 digits models with 16 states, one silence model with 3 states and one inter-digit pause model with one state. In order to enhance the noisy signal, a Wiener filter is used. The 12 basic MFCC parameters plus the log-energy are computed by means of a standard feature extraction scheme. This basic set of 13 parameters is increased with its corresponding deltas and accelerations coefficients; finally, the whole set is normalized using the mean and variance normalization criterion (CMVN).

We have used Eq. 6 to estimate the variance of the enhanced speech. In order to evaluate the SNR, the mean of the noise is calculated from the first 10 frames of each sentence (100ms) which are assumed to be silence. In this equation, σ_n is assumed to be a constant value for all the recognition experiments and it was obtained from all the set of noises of the AURORA 2 multi condition training set.

Table 1: *Recognition results for AURORA 2 task, clean condition training, using the baseline (Wiener filtering, not including uncertainty).*

Baseline				
	setA	setB	setC	Average
Clean	98.96	98.96	98.95	98.95
20dB	97.84	98.17	97.59	97.92
15dB	96.10	96.59	96.04	96.28
10dB	91.90	92.32	90.20	91.73
5dB	80.18	80.18	77.42	79.63
0dB	55.22	53.38	50.68	53.57
Average	84.25	84.13	82.38	83.83

Table 2: *Recognition results for AURORA 2 task, clean condition training, including Wiener filtering and uncertainty.*

With variances				
	setA	setB	setC	Average
Clean	98.94	98.94	98.94	98.94
20dB	97.25	97.84	97.58	95.55
15dB	96.49	96.47	96.25	96.43
10dB	93.28	93.27	91.69	92.96
5dB	85.93	85.25	83.67	84.80
0dB	66.38	63.68	64.43	64.91
Average	87.87	87.10	86.72	87.33

We have fully characterized the statistical distribution of the enhanced speech features (means and variances) and this distribution could be used to perform speech recognition. For every frame t, Eq. 8 is implemented in the conventional HMM decoder. The estimated variance ${\sigma_{\hat{x}}}^2$ is added to all the Gaussian of the pool, and the expected value $\mu_{\hat{x}}$ is used as the observation vector. Some routines of the HTK software were adequately modified to calculate Eq 8.

Table 1 shows recognition results for clean training experiments using AURORA II with the previously described baseline (i.e. including Wiener filtering but not the uncertainty of the observations). The average word accuracy over all SNR from 0 to $20dB$ obtained is 83.83%. Table 2 presents recognition results when uncertainty is included in the decoding process. In this case the average word accuracy is 87.33%. The overall improvement in the recognition rate corresponds to a reduction of 21.64% in the word error rate (WER). Note that the effect of including uncertainty is similar for the three test sets, A, B and C, being the average WER reduced in all the cases. For clean speech experiments the variance estimation is very low and consistently, results are the same in both, when the uncertainty is included and when it is not. For SNR of $20dB$ and $15dB$ recognition results are also very similar in both cases; however when the SNR decreases the inclusion of the variance estimations is more important. Word error rate reductions for 10dB, 5dB and 0dB are 14.87%, 25.38% and 24.42% respectively.

In Table 3 the results with clean training condition and multi condition training are compared. We observed that, in the first case, the word of accuracy is increased when uncertainty is included and in the second case it is almost the same for both types of experiments. Obviously, in the multi condition case

Table 3: *Average recognition results for clean condition training and multi condition training .*

Clean Condition training				
	setA	setB	setC	Average
Wiener filt.	84.25	84.13	82.38	83.83
Wiener filt. + uncert.	87.87	87.10	86.72	87.33
Multi Condition training				
Wiener filt.	91.09	89.56	89.05	90.07
Wiener filt. + uncert.	91.13	89.66	89.31	90.18

Table 4: *Recognition results for AURORA2 SDC Spanish.*

	WM	MM	HM
Wiener filt.	95.59	92.47	87.28
Wiener filt. + uncertainty	95.46	92.27	89.92

some information about the noise is included during training and, for this reason, the inclusion of the uncertainty is less important.

Additionally, recognition experiments have been performed using Spanish SDC-Aurora Database [15] which is a subset of SpeechDat car database and contains only digit utterances. This database was recorded in car environments under several driving condition using two microphones (close talking and hands free). The experiments are defined with increasing level of mismatch between train and test condition: well matched (WM), medium mismatch (MM) and high mismatch (HM). Table 4 presents recognition results using this database. We observe that recognition results are almost the same for WM and MM experiments, but for HM experiments, the inclusion of the uncertainty yields a reduction of the word error rate.

5. Summary and conclusions

In this work we have carried out a study of the inclusion of the uncertainty in the decoding process for robust speech recognition. If the variance of the noise is not zero the enhanced speech is better described by a probability density function. Assuming a Gaussian distribution, we have developed a statistical feature extraction algorithm to estimate mean and variance of this Gaussian. The algorithm is based on the Wiener filtering. To calculate the expected value we have considered the average of the clean speech estimations over three consecutive frames, while to estimate the variance we have used the results obtained in [11]. The recognition decision rule has been modified to include the statistical characterization of the cleaned speech.

The experimental evaluation using the AURORA 2 database demonstrates a 21.64% average word error rate reduction for all the noises and SNR conditions, when clean train condition is performed. We have also reported results for multi condition training, being the inclusion of the uncertainty less important in this case.

6. Acknowledgements

This work has been supported by the Spanish Government under TIC201-3323 research project.

7. References

[1] A. de la Torre, “Técnicas de mejora de la representación en los sistemas de reconocimiento automático de voz ”, Phd Thesis Universidad Granada, España. Abril 1999.

[2] A. de la Torre, J.C. Segura,C. Bentez, A.M. Peinado, A.J. Rubio,“Non-linear transformations of the feature space for robust speech recognition”, Proc of the ICASSP-2002, Orlando, Florida.

[3] Q. Huo, C. Lee,“A Bayesian predictive Approach to robust speech recognition”, IEEE Trans. on Speech and Audio Processing, vol. 7, pag. 426-440, July, 1999.

[4] C.H. Lee et al. “Upper and Lower bounds on the mean of the noisy speech: Application to minimax clasification”, IEEE Trans. on Speech and Audio Processing, vol. 10, n.2, pag. 79-86, February, 2002.

[5] L. Deng, J. Droppo, A. Acero.,“Exploting variances in robust feature extraction based on a parametric model for speech distortion”, Proc of the ICSLP, Denver, Colorado, September 2002, pag 2449-2452.

[6] J. Droppo, A. Acero, L. Deng,“Uncertainty decoding with splice for noise robust speech recognition”, Proc of the ICASSP-2002, Orlando, Florida.

[7] J.A. Arrow, M. A. Clements, “Using observation uncertainty in the HMM decoding”, Proc of the ICSLP, Denver, Colorado, September 2002, pag 1561-1563.

[8] S. Vaseghi, “Advanced digital signal processing for noise reduction”,Wiley, 2000.

[9] ETSI ES 202 108,“Speech Processing, Transmission and Quality aspects (QST); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithm”, 2002.

[10] J.C. Segura, J. Ramírez, C. Benítez, A. de la Torre, A. Rubio, “Improved feature extraction based on spectral noise reduction and non-linear feature normalization”, Proc. of EUROSPEECH, GENEVA, 2003.

[11] J.C. Segura, A. de la Torre, M.C. Benítez, A. M. Peinado, “Model based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora II database and tasks”,Proc of the EUROSPEECH, Alborg, Denmark, September 2002.

[12] J.C. Segura, M.C. Benítez, A. de la Torre, S. Dupont, A.J. Rubio, “VTS Residual Noise Compensation”, Proc of the ICASSP-2002, Orlando, Florida.

[13] J.C. Segura, M.C. Benítez, A. de la Torre, A.J. Rubio, J. Ramírez,“Cestral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition”, to appear in IEEE Signal Processing Letters, vol. 11, No. 5, May 2004.

[14] H.G. Hirsch, D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition system under noise condition”, ISCA ITRW ASR 2000. Paris, France, September 2000.

[15] D. Macho, “Spanish SDC-AuroraDatabase for ETSI STQ WI008 front-end standarization”, November 1999.