APPLICATION OF MINIMUM STATISTICS AND MINIMA CONTROLLED RECURSIVE AVERAGING METHODS TO ESTIMATE A CEPSTRAL NOISE MODEL FOR ROBUST ASR

Veronique Stouten[‡]*, Hugo Van hamme, Patrick Wambacq*

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{vstouten,hvanhamm,wambacq}@esat.kuleuven.be

ABSTRACT

Many compensation techniques, both in the model and feature domain, require an estimate of the noise statistics to compensate for the clean speech degradation in adverse environments. We explore how two spectral noise estimation approaches can be applied in the context of model-based feature enhancement. The minimum statistics method and the improved minima controlled recursive averaging method are used to estimate the noise power spectrum based only on the noisy speech. The noise mean and variance estimates are non-linearly transformed to the cepstral domain and used in the Gaussian noise model of MBFE. We show that the resulting system achieves an accuracy on the Aurora2 task that is comparable to MBFE with prior knowledge on noise. Finally, this performance can be significantly improved when the MS or IMCRA noise mean is reestimated based on a clean speech model.

1. INTRODUCTION

Estimating the statistics of non-stationary noise based on only the noisy speech signal, is a very challenging task. Traditional voice activity detection (VAD) based methods do not update the noise estimate when speech is present. This implies that the noise is assumed stationary during these periods. In real life applications, the performance of these methods can be disappointing, especially at low input SNR levels. Recently, some noise estimation approaches have been introduced that are capable of tracking the noise during speech activity. In [1] an iterative stochastic approximation of the non-linear model of the acoustic environment is proposed to recursively estimate the noise cepstra with an EM-algorithm. In [2] a sequential Monte Carlo method is used in combination with a random walk noise model and an extended Kalman filter to estimate the time-varying log-spectral noise mean. In [3] the noise log-spectra are modelled by a first order auto-regressive (AR) process and tracked by a particle filtering algorithm. Rainer Martin developed a minimum statistics (MS) technique [4] to estimate the power spectral density of the noise. Based on an optimally smoothed version of the noisy speech power spectral density, the noise estimate is obtained by tracking the spectral minima in each frequency band and compensating for the bias. Israel Cohen introduced an improved minima controlled recursive averaging (IMCRA) approach [5] that consists of two iterations of smoothing and minimum tracking. The noise estimate is obtained by averaging past spectral power values, using a time-varying and frequency-dependent smoothing parameter that is adjusted by the speech presence probability in subbands. The latter is controlled by the minima values of a smoothed periodogram.

[‡] Veronique Stouten is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium) (F.W.O. - Vlaanderen).

In this paper, we use the MS or the IMCRA technique to estimate the noise spectral mean and propose an estimate for the variance of the noise spectrum w.r.t. this mean. These noise statistics are then non-linearly transformed to the cepstral domain, using the log-normal approximation [6]. Finally, the cepstral noise mean and variance are used as a time-varying Gaussian noise model in the context of model-based feature enhancement (MBFE) [7] for noise robust speech recognition. Also, a method is described to reestimate the initial noise model mean according to an MMSE-formula and a clean speech model. Because this allows to incorporate more prior knowledge about speech, a more accurate noise estimate is obtained.

The outline of this paper is as follows. First, the MS and the IMCRA technique will be briefly reviewed in section 2. Then, the transformation of the spectral estimate of the noise mean and variance to the cepstral domain will be explained in more detail in section 3. The use of the obtained noise statistics in the context of model-based feature enhancement can be found in section 4. The reestimation of the noise mean is explained in section 5. Finally, experimental results on the Aurora2 digit recognition task and conclusions are given in sections 6 and 7, respectively.

2. SPECTRAL NOISE ESTIMATION

Both the MS and the IMCRA noise estimation technique operate on the noisy speech power spectrum. The advantage of moving into the spectral domain is the good separability of speech and noise, compared to the cepstral domain where recognition takes place. For instance, narrow band noise will be spread over all cepstral components. Also, a lot of detail is lost after the mel-integration.

The calculation of the power spectrum slightly differs from the one in [4], but complies to the ETSI ES 201 108 standard for MFCC feature extraction. We will use superscripts to denote the domain, e.g. μ_n^{spec} is the noise mean in the power spectral domain, μ_n^{mel} is in the mel-power domain, μ_n^{lm} is in the log-mel domain and μ_n^{cep} is in the MFCC domain. A brief review of the MS and the IMCRA technique will be described in the next subsections.

2.1. Minimum Statistics

The MS algorithm [4] is based on the observation that even during speech activity the short term power spectrum of the noisy signal frequently decays to values which are representative of the noise power level. By tracking the minimum P_{min} of a smoothed version P of the noisy power spectrum $|X(t,k)|^2$ within a finite window, an estimate of the noise floor can be obtained. It is assumed that a sliding window of 96 consecutive values is large enough to bridge high power speech segments. This implies that a sudden increase of the

(smoothed) power spectrum that disappears within 96 frames, will also be neglected in the noise estimate. The tracking of an impulsive noise type, such as the firing of a machine gun, of which the statistics are not stationary within this sliding window, will consequently go wrong.

$$P(t,k) = \alpha(t,k)\, P(t-1,k) + (1-\alpha(t,k))\, |X(t,k)|^2 \quad (1)$$

in which t represents the frame index and k is the frequency bin. An optimal value for the time- and frequency dependent smoothing factor α is derived in [4]:

$$\alpha(t,k) = \frac{0.96\,\alpha_c(t)}{1+\left(\frac{P(t-1,k)}{\hat{\mu}_{n,1}^{spec}(t-1,k)}-1\right)^2} \quad (2)$$

in which α_c monitors the tracking error between the short term smoothed psd estimate and the actual averaged periodogram. The minimum P_{min} is multiplied by a bias correction factor B_{min} to obtain a mean estimate of the noise power spectrum $\hat{\mu}_{n,1}^{spec}$.

2.2. Improved Minima Controlled Recursive Averaging

The IMCRA [5] noise estimate is obtained by a recursive averaging procedure, using a time-varying and frequency-dependent smoothing parameter $\beta(t,k)$ that is adjusted by the speech presence probability $p(t,k)$ in subbands. The latter is controlled by the a priori and the a posteriori SNR-levels that are estimated based on the minimum values of the smoothed periodogram. By smoothing both in time and in frequency, the correlation of speech presence in neighbouring bins can be taken into account. To make the minimum tracking more robust during speech activity, two iterations of smoothing and minimum tracking are carried out. The first iteration provides a rough voice activity detection in each frequency band, such that strong (speech) components can be excluded in the second iteration.

$$\bar{\mu}_n^{spec}(t+1,k) = \beta(t,k)\bar{\mu}_n^{spec}(t,k) + (1-\beta(t,k))|X(t,k)|^2 \quad (3)$$

$$\beta(t,k) = \tilde{\beta} + (1-\tilde{\beta})\, p(t,k) \quad (4)$$

with $\tilde{\beta}$ a constant. Also in this case, the final mean estimate of the noise power spectrum $\hat{\mu}_{n,2}^{spec}$ is obtained after multiplying $\bar{\mu}_n^{spec}$ by a bias correction factor. An explicit expression for this bias is derived in [5]. The use of the speech presence probability for smoothing, implies that the noise estimator is based on a variable time segment in each frequency band.

2.3. Spectral Estimate of Noise Variance

Apart from an estimate of the mean noise power spectrum, also an estimate of the variance of the noise spectrum w.r.t. this mean is needed. Under the assumption that each periodogram bin $|N(t,k)|^2$ is an exponentially distributed random variable, this variance is equal to the squared mean. However, this assumption does not hold when the signal is non-stationary. We found empirically that the underestimation of the spectral variance can be corrected by introducing a scale factor. In our experiments, a good performance was obtained when the value of this factor was equal to 2 for all noise conditions of the Aurora2 and Aurora4 recognition tasks. Hence,

$$E\{(|N(t,k)|^2 - \hat{\mu}_n^{spec}(t,k))^2\} \approx 2\,(\hat{\mu}_n^{spec}(t,k))^2 \quad (5)$$

in which $\hat{\mu}_n$ without a subscript '$_1$' or '$_2$' indicates that this is valid for both the MS and IMCRA method.

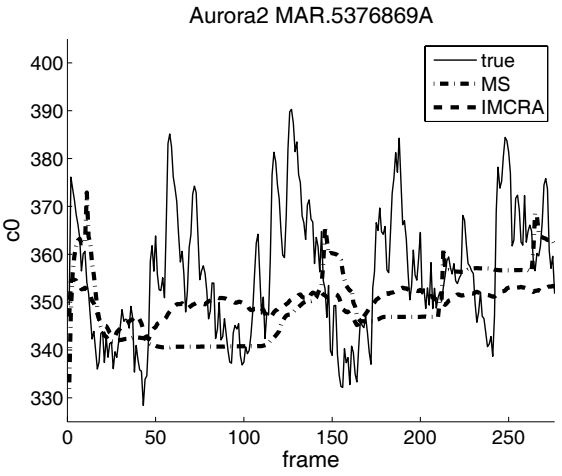

Fig. 1. Cepstral coefficient c_0 of the true noise, the noise mean estimate of MS and IMCRA. Aurora2 setA, subway noise, SNR5.

3. SPECTRAL TO CEPSTRAL TRANSFORMATION

The obtained estimate of the noise mean $\hat{\mu}_n^{spec}(t,k)$ and the diagonal covariance matrix $\hat{\Sigma}_n^{spec}(t)$ in the power spectral domain are first transformed to the log-mel spectral domain. Since the noise distribution is assumed to be Gaussian in the cepstral domain, it is log-normal in the linear spectral domain. Therefore, the log-normal approximation as in [6] can be used for the non-linear transformation to the logarithmic domain. Let K be the mel-matrix, then for each frame t we have $\hat{\mu}_n^{mel} = K\,\hat{\mu}_n^{spec}$ and $\hat{\Sigma}_n^{mel} = K\,\hat{\Sigma}_n^{spec}\,K'$.

$$\hat{\mu}_n^{lm}(i) = log\left(\hat{\mu}_n^{mel}(i)\right) - \frac{1}{2}\hat{\Sigma}_n^{lm}(i,i) \quad (6)$$

$$\hat{\Sigma}_n^{lm}(i,j) = log\left(1+\frac{\hat{\Sigma}_n^{mel}(i,j)}{\hat{\mu}_n^{mel}(i)\,\hat{\mu}_n^{mel}(j)}\right) \quad (7)$$

with $'$ indicating transpose. Finally, the DCT matrix C is used to obtain estimates of the cepstral noise statistics: $\hat{\mu}_n^{cep} = C\,\hat{\mu}_n^{lm}$ and $\hat{\Sigma}_n^{cep} = C\,\hat{\Sigma}_n^{lm}\,C'$. After diagonalising the covariance matrix, these statistics are used as a time-varying front-end noise model in the model-based feature enhancement algorithm, as described in the next section. An example of the estimated cepstral component c_0 of the noise mean for the MS and the IMCRA method is shown in figure 1. As can be seen, the IMCRA noise estimate responds more quickly to noise variations.

4. MODEL-BASED FEATURE ENHANCEMENT

MBFE is a front-end feature enhancement technique that generates estimates of the clean speech cepstra, based on the noisy speech features and two HMM models, namely λ_s for the clean speech and λ_n for the noise. In the remaining of this section, all features are in the cepstral domain and the superscript $(.)^{cep}$ will be dropped. The state-conditional pdfs of clean speech s_t and noise n_t are assumed to be Gaussian mixtures with means μ_s^i, μ_n^j and diagonal covariance matrices Σ_s^i, Σ_n^j, respectively. In this paper, the noise model λ_n is based on the time-varying estimates of the mean $\hat{\mu}_n^{cep}(t)$ and variance $\hat{\Sigma}_n^{cep}(t)$ from the MS or IMCRA method. In section 5, we explain how the noise model can be refined by an MMSE-reestimation of the noise mean.

The first step in the MBFE front-end [7], is to combine λ_s with λ_n to obtain an estimate of the noisy speech HMM λ_x. The non-linearity of the relationship $x_t = f(s_t, n_t, h)$ (between speech, noise, the channel h and the noisy speech x_t) is approximated by a first order Vector Taylor Series, with a state-dependent expansion point given by $(\mu_s^i, \mu_n^j, \overline{h})$:

$$x_t \approx f\left(\mu_s^i, \mu_n^j, \overline{h}\right) + F^{(i,j)}\left(s_t - \mu_s^i\right) + G^{(i,j)}\left(n_t - \mu_n^j\right) \quad (8)$$

in which the gradients of the combination function $f(s_t, n_t, h)$ have the closed form :

$$F^{(i,j)} = C\,\mathrm{diag}\left(\frac{1}{1 + exp\left(C^{-1}\left(\mu_n^j - \mu_s^i - \overline{h}\right)\right)}\right) C^{-1} \quad (9)$$

$$G^{(i,j)} = I - F^{(i,j)} \quad (10)$$

with I denoting the identity matrix. The Gaussian pdf of x_t then has a mean and a covariance matrix :

$$\mu_x^{(i,j)} \approx C\,\log\left(\exp\left(C^{-1}\left(\mu_s^i + \overline{h}\right)\right) + \exp\left(C^{-1}\,\mu_n^j\right)\right) \quad (11)$$

$$\Sigma_x^{(i,j)} \approx F^{(i,j)}\,\Sigma_s^i\,(F^{(i,j)})' + G^{(i,j)}\,\Sigma_n^j\,(G^{(i,j)})' \quad (12)$$

An estimate of the channel $\overline{h}$ is obtained by an iterative EM-algorithm to jointly remove additive and convolutional noise. Details on this channel estimation can be found in [7]. Finally, the global MMSE-estimate of clean speech is given by :

$$\hat{s}_t^{MMSE} = \sum_{(i,j)} P[i,j|x_1^T]\,E\left[s_t|x_1^T, i, j\right] = \sum_{(i,j)} \gamma_t^{(i,j)}\,\hat{s}_t^{(i,j)} \quad (13)$$

in which $_{(i,j)}$ denotes the combined (speech, noise) state. The state-conditional estimates are given by:

$$\hat{s}_t^{(i,j)} = \mu_s^i + \Sigma_s^i\,(F^{(i,j)})'\left(\Sigma_x^{(i,j)}\right)^{-1}\left(x_t - \mu_x^{(i,j)}\right) \quad (14)$$

5. REESTIMATION OF NOISE MEAN

The noise statistics from the MS or IMCRA method can be used to obtain an MBFE noise model. However, very little information about speech is used to derive this estimate. In this section, a method is described to reestimate the initial noise model mean according to an MMSE-formula and the clean speech front-end model λ_s. To this end, the framework of section 4 is extended. Using equations (10), (11) and (12), the new noise mean can be calculated :

$$\hat{n}_t^{MMSE} = \sum_{(i,j)} \gamma_t^{(i,j)}\,\hat{n}_t^{(i,j)} \quad (15)$$

$$\hat{n}_t^{(i,j)} = \mu_n^j + \Sigma_n^j\,(G^{(i,j)})'\left(\Sigma_x^{(i,j)}\right)^{-1}\left(x_t - \mu_x^{(i,j)}\right) \quad (16)$$

Figure 2 shows a spectrogram of the true noise, the MS estimate and the noise mean after reestimation. The time-varying estimate $\hat{n}_t^{MMSE}$ allows to calculate an improved version of the combined noisy speech model λ_x. To reduce the computational load, the update of λ_x is limited to blocks of 100 frames, over which the noise estimate is averaged. Preliminary experiments showed that performing more iterations to reestimate the noise model mean is beneficial only for large vocabulary recognition tasks (e.g. Aurora4). On the Aurora2 task, we did not observe any further improvement when more than one iteration is done. With these improved statistics, the clean speech estimate is obtained as in equation (13).

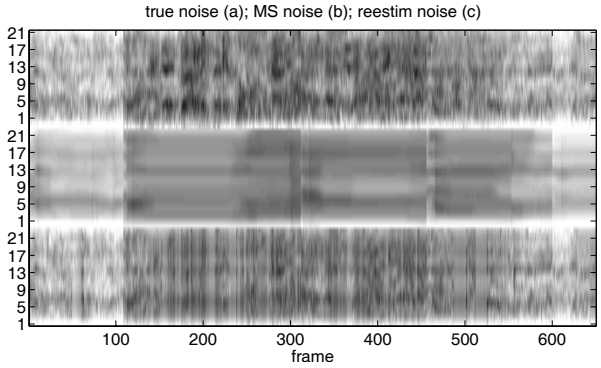

Fig. 2. From top to bottom: spectrogram of the true noise, MS estimate of noise and MMSE-reestimated noise. Aurora2 setA, babble noise, SNR10.

6. EXPERIMENTS

Experiments are conducted on the Aurora2 digit recognition task for the 4 noise types of setA and SNR-levels between 20 dB and 0 dB. Features are extracted by the mel-cepstrum front-end, complying to the ETSI ES 201 108 standard, except that a power spectrum instead of an amplitude spectrum is used. For completeness, the average accuracy on setA when no enhancement is performed (only cepstral mean subtraction), is given in table 5. All other results are obtained by enhancing the noisy speech by the MBFE-algorithm, using a front-end speech model with 128 fully connected single-Gaussian states, the parameters of which are obtained offline. A channel estimate is calculated online by the recursive EM-algorithm [7].

The noise model consists of a single-Gaussian state with a time-varying mean and diagonal covariance matrix. The latter are obtained by transforming the MS or the IMCRA noise estimates (section 2.1 and 2.2, respectively) to the cepstral domain. Optionally, a reestimation of the noise model mean is done as in section 5. The noise model statistics are updated in blocks of 100 frames with an exponential forgetting mechanism. It is possible to use a noise HMM with more Gaussians. Instead of averaging the noise statistics over all frames in a block, a clustering can be performed. If for instance three Gaussians are used, one of them can represent the high energy peaks of the noise, another the low energy frames and the third tracks an average energy level. However, the spectral noise mean is already a rather smooth version, such that the clustering in the cepstral domain does not result in significantly different noise Gaussians. Indeed, preliminary experiments revealed that no better performance is obtained when the MS or IMCRA noise estimate is clustered in more Gaussians.

Front-end estimates are evaluated by the complex back-end recognition system, with whole word digit models trained on the clean speech training database of Aurora2 using the HTK scripts with default settings. The digit models have 16 emitting states with 20 Gaussians per state, while the silence model has 3 states with 36 Gaussians per state. Also, a one-state short pause model, tied with the middle state of the silence model, is used.

Recognition results can be found in table 1 for the MBFE clean speech estimate obtained with a MS noise model and in table 2 for the MBFE clean speech estimate obtained with an IMCRA noise model. The reference in table 4 is obtained when the MBFE noise model consists of 1 fixed noise mean and variance that are trained for each noise condition using the true noise data. As can be seen,

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.17	99.12	99.34	98.83	99.11
15 dB	98.31	97.94	98.60	97.16	98.00
10 dB	96.10	94.01	96.96	93.30	95.09
5 dB	91.00	76.00	90.37	82.75	85.03
0 dB	74.76	47.64	66.54	63.25	63.05
Avg.	91.87	82.94	90.36	87.06	88.06

Table 1. *Recognition accuracy with MBFE and a MS noise model.*

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.20	99.06	99.28	98.86	99.10
15 dB	98.25	97.67	98.81	97.10	97.96
10 dB	95.61	92.87	96.96	93.21	94.66
5 dB	89.35	73.19	90.69	81.70	83.73
0 dB	72.77	45.07	70.09	62.73	62.66
Avg.	91.04	81.57	91.17	86.72	87.62

Table 2. *Recognition accuracy with MBFE and an IMCRA noise model.*

the performance with both the minimum statistics and the improved minima controlled recursive averaging method is comparable to the performance of the reference system, while the latter makes use of prior knowledge of the noise. Only the babble noise seems difficult to track accurately. Except for the car noise, the performance with IMCRA is slightly worse than with MS. Both methods achieve a better performance than the reference system on the car noise and on the subway noise, since the noise model can be adapted online. Table 3 shows that reestimating the noise mean improves the accuracy of babble noise and exhibition noise significantly.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have illustrated the use of two spectral domain noise estimation approaches to obtain the cepstral noise statistics that are needed in the context of model-based feature enhancement. The minimum statistics method of Martin and the improved minima controlled recursive averaging method of Cohen generate an estimate of the noise power spectrum based only on the noisy speech observation. We proposed an estimate of the variance of the noise spectrum w.r.t. this mean noise power and non-linearly transformed the corresponding statistics to the cepstral domain. The performance of MBFE with a front-end noise model based on the MS or IMCRA noise estimate was compared. Recognition results on the Aurora2 task indicated that the same level of accuracy could be obtained as MBFE with a fixed noise model that is trained on the true noise data for each noise condition, while the latter makes use of prior knowledge of the noise. Also, a reestimation of the initial noise mean is proposed that incorporates more knowledge about speech. A significant increase of the accuracy was observed.

A better discrimination between strong noise components and speech could increase the accuracy of the noise estimate. As already mentioned, a sudden increase of the noise power spectrum that disappears within approximately a second, will have a high chance to be neglected in the MS or IMCRA noise estimate. The tracking of very non-stationary noise types, such as machine gun noise, is still a challenging task. To this end, the incorporation of more knowledge of speech in the spectral domain should help to make a more accurate detection of non-speech events.

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.20	99.24	99.43	98.83	99.17
15 dB	98.25	97.97	98.87	97.38	98.12
10 dB	95.98	94.86	96.96	93.67	95.37
5 dB	91.28	82.95	90.07	84.85	87.29
0 dB	74.36	54.47	65.43	66.18	65.11
Avg.	91.81	85.90	90.15	88.18	89.01

Table 3. *Recognition accuracy with MBFE and a reestimated noise model (MS initialisation).*

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.17	99.06	99.28	98.86	99.09
15 dB	97.88	97.79	98.48	97.16	97.83
10 dB	95.21	94.20	96.09	92.90	94.60
5 dB	89.47	82.35	87.09	82.57	85.37
0 dB	72.06	52.00	63.32	63.59	62.74
Avg.	90.76	85.08	88.85	87.02	87.93

Table 4. *Recognition accuracy with MBFE and a fixed noise model trained on noise data.*

20 dB	15 dB	10 dB	5 dB	0 dB	Avg.
97.62	90.87	70.32	37.87	14.07	62.15

Table 5. *Average recognition accuracy on setA without feature enhancement (only CMS).*

8. ACKNOWLEDGEMENT

This work was partly supported by 'Research Fund (Onderzoeksfonds) K.U.Leuven', project no. OT/03/32/TBA.

9. REFERENCES

[1] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. on SAP*, vol. 11, no. 6, pp. 568–580, 2003.

[2] K. Yao and T.-W. Lee, "Time-varying noise estimation for speech enhancement and recognition using sequential monte carlo method," *EURASIP JASP*, vol. 15, pp. 2366–2384, 2004.

[3] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 965–968.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on SAP*, vol. 9, no. 5, pp. 504–512, 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. on SAP*, vol. 11, no. 5, pp. 466–475, 2003.

[6] M.F.J. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, Sept. 1995.

[7] V. Stouten, H. Van hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Comm.*, 2006, (accepted for publication).