NON-STATIONARY NOISE MODEL COMPENSATION IN VOICE ACTIVITY DETECTION

Mikko Myllymäki and Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, 33720, Tampere, Finland
email: tuomas.virtanen@tut.fi

ABSTRACT

This paper proposes methods for acoustic pattern recognition in dynamically changing noise. Parallel model combination and vector Taylor series model compensation techniques are used to adapt acoustic models to noisy conditions are applied together with a time-varying noise estimation algorithm. The noise estimation produces biased noise estimates and therefore we propose methods to accommodate the compensation to the bias. We apply the methods in robust voice activity detection, where frame-wise speech/non-speech classifier is first trained in clean conditions and then tested in and adapted to non-stationary noise conditions. The simulations show that a model compensation with the time-varying noise estimator improves clearly the accuracy of voice activity detection.

1. INTRODUCTION

Mobile communication devices can be used in environments with highly varying background noise conditions. Many devices apply voice activity detection or automatic speech recognition algorithms, which performance is significantly affected by the noise. Dynamic noise conditions are especially difficult for these algorithms, because it is not possible to train the algorithms beforehand to match the noisy conditions. Therefore, the algorithms must be compensated so that they match the new noise conditions. Algorithm adaptation to noisy conditions can be split into two separate stages: noise estimation and model compensation.

Previous approaches estimate the noise spectrum during noise-only segments in speech, such as pauses in speech, and therefore need a voice activity detector [1]. However, when the level of the noise is high, the activity detection is difficult to perform robustly. Recently, algorithms (see for example [2], [3] and the review in [4]) have been proposed to update the noise spectrum continuously, even during speech segments. This can be achieved by tracking the minimum of the spectrum, which can then be used as an estimate of the noise, because of the sparsity of the speech spectrum. Section 2 presents briefly the noise estimation algorithm used in our study.

Adaptation to noisy conditions can be done by either subtracting the noise estimate from the noisy features, or compensating the noise in the model that describes the features. Model compensation techniques have proved to be a superior alternatives to feature subtraction in many cases. Parallel model combination (PMC) [5] and vector Taylor series (VTS) approaches [6, 7] use a model of clean speech as the starting point and then adapt this model to fit a new noise environment, as explained in Section 3. PMC and VTS have been widely used in robust speech recognition.

The original versions of PMC and VTS do not compensate the models continuously to fit the dynamic noise conditions. Furthermore, their performance relies on the noise estimate obtained during speech pauses indicated by a voice activity detector.

The proposed method applies a noise estimation algorithm that produces an estimate of the noise spectrum in every frame. This allows the speech models compensated with PMC or VTS methods to be time-varying. The noise estimate is biased in speech segments, and therefore in Section 4 we propose a method to compensate the bias. In Section 5 the model compensation scheme is applied to robust voice activity detection. Simulation experiments in Section 6 show that the compensation method outperforms the basic PMC. When the bias is taken into account, both compensation methods produce results which are significantly better than those obtained without compensation or with a stationary noise estimate.

2. NON-STATIONARY NOISE ESTIMATION

We use the noise estimation algorithm by Rangachari and Loizou [4]. The basic idea behind the algorithm is that the spectrogram of speech is sparse, and local minima of the spectrum in a window of multiple frames can be used as an estimate of the noise spectrum.

An overview of the algorithm is provided below. The algorithm operates on power spectrum calculated in 64 linearly spaced frequency bands, and the calculations are done for every frequency bin $k = 1, \ldots, 64$ in every frame t.

1. Calculate temporally smoothed power spectrum $\hat{x}(t,k)$ by filtering the power spectrum of the observed noisy signal $x(t,k)$ with a first-order recursive filter.
2. If $\hat{x}(t,k)$ is smaller than the current estimate of the noise power spectrum minimum $x_{\min}(t-1,k)$, replace $x_{\min}(t,k)$ with $\hat{x}(t,k)$, else use a first-order recursive filter to calculate a new estimate for $x_{\min}(t,k)$
3. Calculate the ratio between the smoothed power spectrum $\hat{x}(t,k)$ and the current estimate of the minimum $x_{\min}(t,k)$ and threshold it to make a decision between speech present and speech absent.

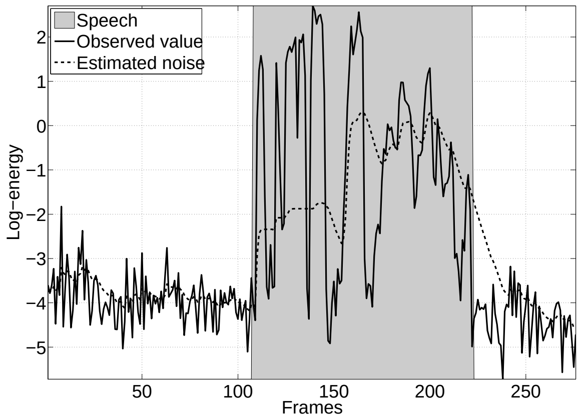

Figure 1: Example of a signal log-energy and the estimated noise log-energy within a frequency band. The signal-to-noise ratio of the signal is 5 dB.

4. Calculate the speech presence probability by smoothing the speech present/absent decision in time.
5. Calculate a frequency dependent smoothing factor using the speech presence probability.
6. Update the noise spectrum estimate $x_{\mathrm{n}}(t,k)$ by first-order recursive filtering the observed power spectrum $x(t,k)$ using the estimated smoothing factor.

Details of the algorithm can be found in [4]. The adaptation time of the algorithm to new noise conditions is about 0.5 s.

The rest of the paper operates with log-energies $n(t,i)$ calculated on 10 mel-frequency bands. The estimated noise spectrum is decimated to mel scale $i = 1,\ldots,10$ by windowing the bands with triangular windows and calculating the log-energy from the windowed bands. An example of observed log-energies and the corresponding noise estimates is illustrated in Figure 1.

Time-varying mean $\mu_{\mathrm{n}}(t,i)$ and variance $\sigma^2_{\mathrm{n}}(t,i)$ of the noise log-frequency features are calculated as

$$\begin{aligned}\mu_{\mathrm{n}}(t,i) &= \delta\mu_{\mathrm{n}}(t-1,i) + (1-\delta)n(t,i)\\ \sigma^2_{\mathrm{n}}(t,i) &= \delta\sigma^2_{\mathrm{n}}(t-1,i) + (1-\delta)[n(t,i)-\mu_{\mathrm{n}}(t,i)]^2,\end{aligned}$$

where $\delta = 0.9$ is a smoothing parameter.

3. MODEL COMPENSATION

In model compensation the models estimated for clean speech are adapted to match the noisy conditions using an estimate of the noise statistics. The distributions of features are modeled with Gaussian mixture models (GMMs) and in both methods used here, PMC and VTS approaches, the basic idea is to modify the means and variances of the GMMs so that they model the distributions of the noisy features.

3.1 Parallel model combination

In PMC, the log-normal approximation [8, pp. 47-48] assumes that the sum of normally distributed speech and noise is also normally distributed. The means and variances of the noise corrupted GMMs are calculated by assuming that speech and noise are additive in the power spectral domain and matching the first two moments of the noisy distribution with the sum of the moments of the speech and noise distributions.

We perform the compensation separately for each Gaussian in the speech model GMMs. In the following, the compensation is presented for an individual Gaussian. Let us denote the original clean speech model mean and variance by subindex s, the estimated noise distribution parameters by subindex n, and the resulting noisy speech model parameters by subindex y.

First, the clean speech model means and variances are transformed from the log to the linear power spectrum spectrum domain as

$$\hat{\mu}_{\mathrm{s}}(t,i) = \exp(\mu_{\mathrm{s}}(t,i) + \sigma^2_{\mathrm{s}}(t,i)/2) \quad (1)$$

$$\hat{\sigma}^2_{\mathrm{s}}(t,i) = \hat{\mu}^2_{\mathrm{s}}(t,i)[\exp(\sigma^2_{\mathrm{s}}(t,i)) - 1]. \quad (2)$$

The same transformation is applied also to the noise means and variances. The noisy speech model parameters in the linear power spectrum domain are obtained as

$$\begin{aligned}\hat{\mu}_{\mathrm{y}}(t,i) &= \hat{\mu}_{\mathrm{s}}(t,i) + \hat{\mu}_{\mathrm{n}}(t,i)\\ \hat{\sigma}^2_{\mathrm{y}}(t,i) &= \hat{\sigma}^2_{\mathrm{s}}(t,i) + \hat{\sigma}^2_{\mathrm{n}}(t,i).\end{aligned}$$

The log-normal approximation assumes that the sum of two log-normally distributed variables is also log-normally distributed, therefore the means and variances of the noisy speech model in the log-spectral domain are obtained as

$$\mu_{\mathrm{y}}(t,i) = \log(\hat{\mu}_{\mathrm{y}}(t,i)) - \frac{1}{2}\log\left(\frac{\hat{\sigma}^2_{\mathrm{y}}(t,i)}{\hat{\mu}^2_{\mathrm{y}}(t,i)} + 1\right) \quad (3)$$

$$\sigma^2_{\mathrm{y}}(t,i) = \log\left(\frac{\hat{\sigma}^2_{\mathrm{y}}(t,i)}{\hat{\mu}^2_{\mathrm{y}}(t,i)} + 1\right). \quad (4)$$

3.2 Vector Taylor series

Vector Taylor series (VTS) approach [9] models the noisy speech features $y(t,i)$ as

$$y(t,i) = s(t,i) + g(s(t,i), n(t,i)), \quad (5)$$

where $s(t,i)$ is the clean speech feature and $g(s(t,i), n(t,i))$ is an environmental function depending on the clean speech and noise. Contrary to the original VTS formulation [9], the effect of the transmission channel is omitted here, because in our case the training and testing channels for speech are identical. The environmental function is approximated with the VTS and the approximation is then used to calculate the corrupted speech model. Similarly to the PMC, the compensation is done individually for each Gaussian in the speech models.

The zeroth-order VTS expressions for the mean and variance vectors are [9, p.83]

$$\begin{aligned}\mu_{\mathrm{y}}(t,i) &= \mu_{\mathrm{s}}(t,i) + g(s_0(t,i), n(t,i))\\ \sigma^2_{\mathrm{y}}(t,i) &= \sigma^2_{\mathrm{s}}(t,i),\end{aligned}$$

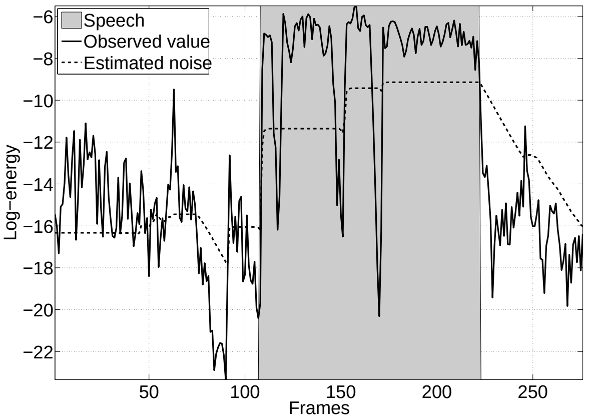

Figure 2: Clean speech log-energy and the estimated noise log-energy within a frequency band. During speech activity the noise estimate is biased.

where $s_0(t,i) = \mu_\mathrm{s}(t,i)$ is the VTS expansion point and

$$g(s_0(t,i), n(t,i)) = \ln(1 + e^{n(t,i) - s_0(t,i)}).$$

Similarly, the first-order VTS expression [9, p.84] for the means and variances results to

$$\begin{aligned}\mu_\mathrm{y}(t,i) =& [1 + g'(s_0(t,i), n(t,i))]\mu_\mathrm{s}(t,i) \\ &+ g(s_0(t,i), n(t,i)) - g'(s_0(t,i), n(t,i))s_0 \\ \sigma^2_\mathrm{y}(t,i) =& [1 + g'(s_0(t,i), n(t,i))]^2 \sigma^2_\mathrm{s}(t,i),\end{aligned}$$

where

$$g'(s_0(t,i), n(t,i)) = -\frac{1}{1 + \exp(s_0(t,i) - n(t,i))}.$$

We also tested the method [7] that uses noise means $\mu_\mathrm{n}(t,i)$ and variances $\sigma^2_\mathrm{n}(t,i)$, instead of point-estimates $n(t,i)$. This approach produced similar results as the method used here, and therefore we use the above compensations.

4. NOISE BIAS SUBTRACTION

A problem in the time-varying noise estimator is that it produces non-zero values even when applied to clean speech signals. In other words, the noise estimate is biased. The noise estimate $\hat{n}(t,k)$ in the linear power spectrum domain is considered to be composed of two parts as

$$\hat{n}(t,k) = \hat{n}_\mathrm{b}(t,k) + \hat{n}_\mathrm{e}(t,k), \tag{6}$$

where $\hat{n}_\mathrm{b}$ is the noise bias and $\hat{n}_\mathrm{e}$ is the environmental noise. The bias in the noise estimation algorithm is illustrated using a clean speech signal is in Figure 2. Only the environmental noise, but not the bias should be used in the compensation. We tested three alternative techniques to compensate the bias.

The first approach models the bias with a single Gaussian, which is then subtracted from the speech model using PMC. First, we train mean and variance for the noise estimated clean speech training data of each model class (to be explained later). The bias mean and variance are transformed to linear-frequency domain according to equations (1)-(2) and then the subtraction is done in the linear-frequency domain for every GMM component as

$$\hat{\mu}_\mathrm{z}(t,i) = \hat{\mu}_\mathrm{s}(t,i) - \hat{\mu}_\mathrm{b}(t,i) \tag{7}$$

$$\hat{\sigma}^2_\mathrm{z}(t,i) = \hat{\sigma}^2_\mathrm{s}(t,i) + \hat{\sigma}^2_\mathrm{b}(t,i), \tag{8}$$

where $\hat{\mu}_\mathrm{b}(t,i)$ and $\hat{\sigma}^2_\mathrm{b}(t,i)$ are the linear mean and variance of the bias model.

We call this result the noise-bias-subtracted GMM and denote the corresponding parameters with a subindex z. The noise-bias-subtracted GMM is transformed back to log-frequency domain according to equations (3)-(4), and the model compensation is done using noise bias-subtracted GMMs.

Second, we tested using more than one GMM components to model the bias. In this case the estimation of the noise bias subtracted GMM becomes ambiguous. We tested a method were the noise bias subtracted GMM had MN components, where M is the number of clean speech and N the number of noise bias GMM components, respectively. The noise bias subtracted GMM is calculate separately for every Gaussian in the noise bias model.

The third option, which produced the best results at least in the case of PMC, was to subtract all the noise bias GMM components from each clean speech GMM component. Thus, the linear domain parameters are obtained according to Eq. (7), but $\hat{\mu}_\mathrm{b}(t,i)$ and $\hat{\sigma}^2_\mathrm{b}(t,i)$ are now the sums of all the noise bias GMM means, and variances, respectively. This approach retains the number of GMM components in the speech models. In our simulations we obtained good results by using 5 noise bias GMM components.

In the case of all the bias compensation methods, the obtained noise bias subtracted GMMs are used as a starting point for the environmental noise compensation instead of the original clean speech models. In practice, this means replacing the mean $\mu_\mathrm{s}(t,i)$ and variance $\sigma^2_\mathrm{s}(t,i)$, $i = 1, \ldots, I$, vectors in the PMC and VTS algorithms with the corresponding noise bias subtracted versions $\mu_\mathrm{z}(t,i)$ and $\sigma^2_\mathrm{z}(t,i)$, $i = 1, \ldots, I$.

5. APPLICATION TO ROBUST VOICE ACTIVITY DETECTION

We apply the proposed method in noise-robust voice activity detection targeted to a communication device and applications where there can be a significant amount of user-produced noise, for example breathing [10]. The user-produced noise has specific characteristics for which we have to train a model in order to perform robust voice activity detection (VAD).

The proposed VAD algorithm is a hidden Markov model (HMM) consisting of speech and non-speech states, whose state emission distributions are modeled with GMMs, which parameters are trained beforehand using material of both classes. In the training phase we also train two bias GMMs using noise estimated from clean material of both classes. The bias GMMs of each class are subtracted from the corresponding original GMMs to obtain noise bias subtracted GMMs for

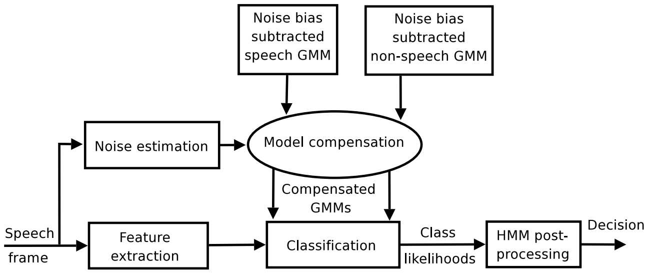

Figure 3: Block diagram of the used VAD algorithm.

both classes. The acoustic material used to train the VAD is explained in Section 6.

The frame-wise processing is illustrated in Figure 3. The input signal is processed in 16 ms frames that do not overlap. Noise estimation is performed using the algorithm explained in Section 2. The observed noisy speech and estimated noise features are log-energies within 10 mel-frequency bands which overlap by 50%.

The noise features or the noise means and variances work as an input to the model adaptation block, were it is used to adapt the original clean speech and non-speech GMMs with PMC or VTS approach to match the noisy speech and non-speech distributions.

Given an observed feature vector, the noisy speech and non-speech GMMs are then used to calculate the likelihoods for the two classes. Finally, the class likelihoods work as an input to the two-state hidden Markov model, where state transition probabilities are used to obtained the probabilities of speech and non-speech state for the current frame, given the probabilities of the previous frame.

6. SIMULATIONS

Simulations using acoustic material corresponding to the final usage situations of the communication device were conducted. The device is used in physically demanding situations and the microphone is located directly in front of the speaker's mouth, which results in high-level breathing noise (see [10] for an illustration of a signal).

Signals from five different speakers were recorded, the total amount of data being 43 minutes. The percentage of speech in the signals is 2-20% depending on the speaker. The recorded signals were manually labeled into speech and noise segments with a temporal resolution of 10 ms. A 5-component clean speech GMM was trained using the speech frames to model the emission probability density function (pdf) of the speech state in the VAD HMM, and similarly a 5-component non-speech GMM was trained using non-speech frames to model the non-speech state emission pdf. The expectation-maximization algorithm was used to train the GMMs.

The recorded speech signals did not have environmental noise, but in the testing we used four different types of noise signals which were mixed with the speech signals. The noise signals are from the study [11], and they include construction site and bus environments noise. The signals were mixed to obtain signal-to-noise ratio of 5 dB.

6.1 Methods

The following methods were tested:

- "No compensation" means that the models are not compensated but the clean speech and non-speech models are used to classify the noisy signals.
- PMC is the proposed VAD algorithm that uses the PMC as the model adaptation method. The method was tested with and without noise bias subtraction (NBS).
- VTS is the proposed VAD algorithm that uses the zeroth-order VTS approach as the model adaptation method. The method was also tested with and without noise bias subtraction (NBS).
- STATIONARY is the original PMC algorithm that estimates a stationary noise model from the beginning of the noise signal before mixing it with the speech signal and uses this model to adapt the clean speech model to a noisy speech model.

The noise bias model was a 5-component GMM, trained separately for speech and non-speech frames. The subtraction was done by subtracting all the Gaussians in the bias GMM from the corresponding speech/non-speech model, as explained in Section 4.

In VTS, we used point-estimates of the noise $n(t, i)$ instead of mean and variance, since it resulted in slightly better results. We used the zeroth-order VTS, because it produced better results than the first-order VTS.

6.2 Evaluation

The performance evaluation of the VAD algorithm was done using a leave-one-out cross-validation method where the signal of one speaker was regarded as a test set and the rest as the training set. The GMMs of speech and non-speech states were trained using the clean signals and the annotations in the training set. The noise-corrupted test signal was processed using each tested VAD algorithm, which produce speech/non-speech decision for each frame.

The classification accuracy was measured by comparing the classifications to the annotated speech activity. The following four measures were used to judge the classification accuracy:

- *Sensitivity* gives the percentage of the frames correctly classified as speech from all the speech frames in the signal
- *Specificity* gives the percentage of the frames correctly classified as noise from all the noise frames in the signal
- *Positive predictive value* gives the percentage of the frames that actually are speech from all the frames classified as speech
- *Negative predictive value* gives the percentage of the frames that actually are noise from all the frames classified as noise

The speech/non-speech decision was tuned so that the average sensitivity was always 97% or higher and the specificity as high as possible. Having an average sensitivity of 97% retains the intelligibility of the speech and

Algorithm	Sens.	Spec.	PPV	NPV
No compensation	97.0	37.5	20.7	98.7
PMC without NBS	97.1	27.8	18.1	97.9
PMC with NBS	97.0	45.4	22.2	98.8
VTS without NBS	97.2	48.1	24.6	99.1
VTS with NBS	97.3	46.2	24.1	99.1
STATIONARY	97.0	24.0	18.2	98.3

Table 1: VAD algorithm results (%), construction site noise

Algorithm	Sens.	Spec.	PPV	NPV
No compensation	97.1	37.4	20.8	98.7
PMC without NBS	97.3	21.7	17.1	97.6
PMC with NBS	97.0	45.1	22.3	99.0
VTS without NBS	97.0	56.2	28.3	99.1
VTS with NBS	97.1	56.7	28.6	99.1
STATIONARY	97.0	33.9	19.9	98.6

Table 2: VAD algorithm results (%), bus noise

also facilitates direct comparison between the different methods.

6.3 Results

The results are illustrated in Tables in 1 and 2. All proposed dynamic model compensation methods except PMC without NBS improve the performance in comparison with the case where no compensation is done. Taking into account the noise bias in PMC improves clearly its performance. Clearly the best results are obtained with VTS. The noise bias does not have a big effect in its performance. This might be because the proposed noise bias subtraction methods are motivated by the processing principles of PMC. The stationary noise model method performs clearly worse than the non-stationary noise compensation methods.

7. CONCLUSIONS

We have proposed a method to compensate acoustic models to non-stationary environmental noise. We apply a noise estimation algorithm, and then compensate the clean acoustic models with the time-varying noise estimate. Parallel model combination and vector Taylor series methods were tested in the compensation. A method to compensate the bias of the noise estimator was found to be necessary at least in the case of parallel model combination. The developed methods were tested in robust voice activity detection, where acoustic models trained on clean speech and non-speech were adapted to noisy signals. The proposed non-stationary model compensation methods were found to be succesful in comparison with the stationary compensation. The best results were obtained with the vector Taylor series compensation.

REFERENCES

[1] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, 2004.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, 2001.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, 2003.

[4] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, 2006.

[5] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, 1996.

[6] M. Gales, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[7] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Sixth International Conference on Spoken Language Processing*, 2000.

[8] M. J. F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, Cambridge University, 1995.

[9] P. J. Moreno, *Speech Recognition in Noisy Environments.* PhD thesis, Carnegie Mellon University, 1996.

[10] M. Myllymäki and T. Virtanen, "Voice activity detection in the presence of breathing noise using neural network and hidden Markov model," in *European Signal Processing Conference*, 2008.

[11] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, and T. Sorsa, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.