Efficient Noise Estimation and its Application for Robust Speech Recognition*

Petr Motlíček[1,2], Lukáš Burget[1,2]

[1] OGI School of Science & Engineering at Oregon Health & Science University
20000 NW Walker Road, Beaverton, OR, 97006, USA
[2] Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66, Czech Republic
{petr,lukas}@asp.ogi.edu

Abstract. The investigation of some well known noise estimation techniques is presented. The estimated noise is applied in our noise suppression system that is generally used for speech recognition tasks. Moreover, the algorithms are developed to take part in front-end of Distributed Speech Recognition (DSR). Therefore we have proposed some modifications of noise estimation techniques that are quickly adaptable on varying noise and do not need so much information from past segments. We also minimized the algorithmic delay. The robustness of proposed algorithms were tested under several noisy conditions.

1 Introduction

The error rate of speech recognition systems increases dramatically in the presence of noise. It is therefore very convenient to use some noise reduction technique which can operate under adverse conditions. Often used speech enhancement systems based on spectral decomposition such as Wiener filtering or Spectral subtraction rely on an accurate estimation of the background noise energy as well as signal-to-noise ratio (SNR) in the various frequency bands.

A number of approaches were proposed to estimate the noise without the need for speech/pause detector. However the implementation of the front-end DSR system is limited by technical constraints, e.g. memory requirements, algorithmic delay, complexity. Since this limitation is given a-priori, we were supposed to come up with noise estimation algorithm that would satisfy the requirements and that would be the best for our noise suppression algorithm.

2 Experimental setup

The noise suppression algorithms proposed for speech recognition system were tested on three SpeechDat - Car (SDC) databases used for Advanced DSR Front-

* This research was supported by industrial grant from Qualcomm, DARPA N66001-00-2-8901/0006, and by the Grant Agency of Czech Republic under project no. 102/02/0124.

End Evaluation: Italian SDC [1], Spanish SDC [2], and Finish SDC. The recordings were taken from the close-talk microphone and from one of the hands-free microphones. Data were recorded at 16kHz, but downsampled to 8kHz. The databases contain various utterances of digits.

During experiments, the robustness was tested under three different training conditions. For each of these three conditions 70% of the files were used for training, 30% for testing.

- **Well-matched condition (wm):** All the files (close-talk and hands-free microphones) were used for training and testing.
- **Medium mis-matched condition (mm):** Only recordings made with the hands-free microphone were used for training and testing.
- **Highly mis-matched condition (hm):** For the training only close-talk microphone recordings were used, whereas for testing the hands-free files were taken.

3 Noise suppression system

Many of noise suppression schemes exist. Practically all of them share the common goal of attempting to increase the signal-to-noise ratio (SNR). They differ in complexity and suitability for real-time processing. The noise suppression algorithm [3], which is being used in our feature extraction, has been derived from standard Spectral subtraction and Wiener filtering. The algorithm supposes that the noise and the speech signal are uncorrelated. Moreover we assume that their power spectral contributions are additive: $|X_k[n]|^2 = |Y_k[n]|^2 + |N_k[n]|^2$, where $|Y_k[n]|^2$ denotes the clean speech power spectrum at the given time n in the frequency subband k, and $|N_k[n]|^2$ is the noise power spectrum. The noise reduction algorithm can be viewed as a filtering operation where high SNR regions of the measured spectrum are attenuated less than low SNR regions. The mathematic description of our noise suppression filter is:

$$|H_k[n]|^2 = max\left(\frac{|X_k[n]|^2 - osub|N_k[n]|^2}{|X_k[n]|^2}, \beta\right)^2. \tag{1}$$

An oversubtraction factor $osub$ is a filter parameter which varies with time and is estimated from energy of signal and noise. While a large $osub$ essentially eliminates residual spectral peaks, it also affects quality of speech so that some of the low energy phonemes are suppressed. This drawback is reduced by dependency of $osub$ on SNR. Yet a spectral floor threshold β does not change with time and prevents the filter components from small values.

In order to alleviate the influence of musical noise, the filter transfer function $|H_k[n]|^2$ is smoothed in temporal domain, whereas the following smoothness in spectral domain showed itself to be very useful for low SNR as well as clean speech recognition.

4 Noise estimation

As can be seen from Eq.1, the noise suppression algorithm requires the accurate estimation of the noise power spectrum $|N_k[n]|^2$. This is however difficult in practical situations especially if the background noise is not stationary or SNR is low.

A commonly used method for noise spectrum estimation is to average over sections which do not contain speech, i. e. voice activity detector (VAD) is required to determine speech and non-speech sequences. It relies on the fact that there actually exists a sufficient amount of non-speech in the signal. Standard noise estimation methods without explicit VAD were tested in our feature extraction system.

4.1 Temporal minima tracking

The best estimation of noise in our experiments has been obtained with standard temporal minima tracking algorithm [4]. This algorithm is applied consequently on smoothed power spectrum:

$$P_{xk}[n] = \alpha P_{xk}[n-1] + (1-\alpha)|X_k[n]|^2, \tag{2}$$

with forgetting factor α between $0.75 \ldots 0.8$. The algorithm is independently used on each spectral subband of $P_{xk}[n]$. The initial smoothing of power spectra slows down the rapid frame-to-frame movement. The estimated power spectrum of noise $P_{nk}[n]$ for k^{th} subband is found as a minimum of $P_{xk}[n]$ within a temporal window of D previous and current power sample:

$$P_{nk}[n] = min(P_{xk}[n-D] : P_{xk}[n]). \tag{3}$$

The processing window of D samples is at the beginning filled by first frame $P_{xk}[1]$. It reflects the assumption that the first frame of an utterance does not contain speech. The example of estimated noise $P_{nk}[n]$ is given in Fig 2 (lower panel). However the standard minima tracking algorithm causes problems of causality and large memory requirement. From many experiments we have observed that $P_{nk}[n]$ can be well estimated just from current and previous samples of $P_{xk}[n]$. But the necessity of large memory buffer makes this noise estimation technique not applicable for feature extraction part of DSR system.

The memory buffer size for minima tracking algorithm is given as D times r, where r denotes number of spectral subbands. In order to get sufficient estimation of $P_{nk}[n]$, D should not be smaller than 80. Usually r is 129. In [4], it is suggested to decompose one window of length D into W subwindows (for each spectral band independently), which brings some memory reduction but does not cause the system's degradation (*noise est. 2, 3, 4* in Tab.1).

In our previous experiments we have observed some good properties of noise estimation algorithms when processing in spectral domain as well. Spectral domain processing is another possibility to reduce memory buffer for noise estimation. We may decrease the spectral resolution of $P_{xk}[n]$, estimate the noise and

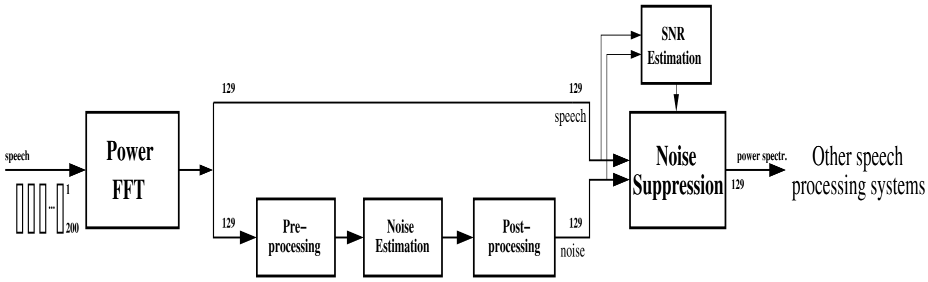

Fig. 1. Scheme of noise suppression system with noise estimation applied in Mel-scale filter bank domain.

apply some kind of interpolation technique with spectral smoothing to get the initial number of spectral subbands *noise est. 6* in Tab.1).

Another spectral processing we have tried in our experiments was to integrate the power spectra $|X_k[n]|^2$ into spectral bands applying the Mel filter bank. This operation can be viewed as a smoothing of power spectra in spectral domain. Then the noise estimation is done in this integrated spectrum. Number of spectral bands of initial power spectra $|X_k[n]|^2$ is 129 $(1 \ldots F_{sampling}/2)$. After application of Mel filter bank, number of bands was reduced to 23. The estimated noise in Eq. 1 is however expected in power spectral domain (again 129 subbands). Hence we applied inverse projection from 23 spectral bands into 129 subbands of power spectra, which caused the additional smoothing. In order to keep the same energies in bands, standard Mel filter bank for direct projection was modified, so that the areas under particular triangular weighting functions were normalized to unity:

$$Mfb_{MOD_k}[i] = \frac{Mfb_k[i]}{\sum_{j=1}^{129} Mfb_k[j]}, \quad k \in 1 \ldots 23, i \in 1 \ldots 129. \tag{4}$$

The results obtained with noise estimated using previously described approaches are in Tab. 1.

4.2 Noise estimation based on filtering of temporal trajectories of spectral coefficients

Often used noise estimation algorithm, proposed in [5], which does not need information about speech/non-speech segments, has been tested in our experiments. Here each spectral subband is filtered by nonlinear estimator that might be perceived as an efficient implementation of temporal minima tracking in power spectral domain. This temporal processing also requires a smoothed version of power spectrum $P_{xk}[n]$ pre-computed by Eq. 2. The algorithm can be described as follows:

$$P_{nk}[n] = \gamma P_{nk}[n-1] + \frac{1-\gamma}{1-\beta}(P_{xk}[n] - \beta P_{xk}[n-1]). \tag{5}$$

Accuracy [%]	**Italian**			**Finish**			**Spanish**			overall
conditions	hm	mm	wm	hm	mm	wm	hm	mm	wm	[%]
baseline	85.01	91.17	96.00	88.15	86.85	95.48	88.21	90.67	95.84	**91.81**
noise est. 1	86.77	92.77	96.90	91.17	88.92	96.67	92.51	92.85	96.43	**93.23**
noise est. 2	89.11	93.45	96.75	92.16	90.36	97.13	92.60	93.11	96.80	**93.89**
noise est. 3	89.21	93.41	96.74	92.19	90.22	97.13	92.72	93.22	96.60	**93.87**
noise est. 4	89.27	93.25	96.71	92.19	90.29	96.97	92.84	93.31	96.70	**93.83**
noise est. 5	88.24	92.53	96.59	92.12	88.17	96.97	91.94	93.55	96.86	**93.41**
noise est. 6	89.16	93.05	96.67	92.05	89.81	96.82	92.60	93.20	96.65	**93.71**

Table 1. Speech recognition results for Italian, Finish and Spanish databases with noise estimation based on temporal minima tracking algorithm. The experiments' conditions are explained in section 2. The detailed algorithm descriptions are in Tab. 3.

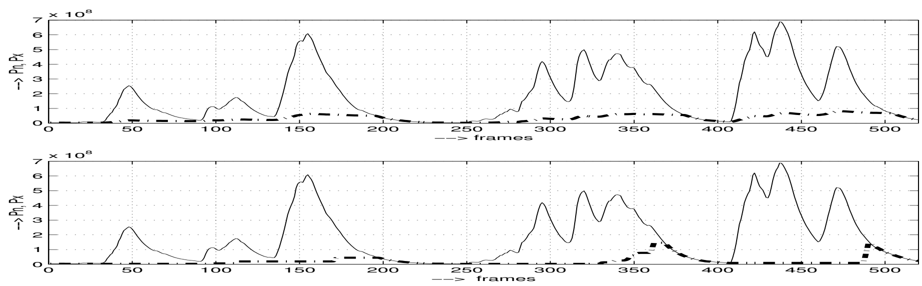

Fig. 2. Process of noise estimation in short-time power spectra (8^{th} subband - related to 250 Hz). The solid lines represent the smoothed power spectra $P_{xk}[n]$, the dashed lines describe estimated $P_{nk}[n]$:
Upper panel: Filtering of spectral subbands in temporal domain.
Lower panel: Minima tracking in temporal domain.

The minima tracking is ensured in this approach so that $P_{nk}[n] \leq P_{xk}[n], \forall k, n$ as can be seen in Fig. 2 (upper panel). Although this method does not bring any difficulties with memory size, the basic approach from [5] was not successful in our front-end system (*noise est. 7* in Tab. 2). That was mainly caused by high level of estimated noise in speech portions of processed sentences. Therefore we have experimented with implementation of some simple speech/pause detector. The used algorithm comes from [6] and is based on the evaluation of the SNRs in each spectral subband individually. We compute the relative ratio of noise energy to signal&noise energy NX for each subband:

$$NX_{relk}[n]) = \frac{NX_k[n] - NX_{mink}[n]}{NX_{maxk}[n] - NX_{mink}[n]}. \qquad (6)$$

NX_{min} and NX_{max} are originally determined from the past (at least 400ms) which can cause memory complexity. Therefore we have used NX_{min}, NX_{max} fixed. For calculation of NX ratio, $P_{xk}[n]$ from Eq. 2 and $P_{nk}[n]$ from Eq. 5

were taken. For each spectral subband independently the speech is indicated, and $P_{nk}[n]$ is modified so that

$$P_{nk}[n] = \begin{cases} 0.4P_{nk}[n] \; if \quad NX_{rel_k}[n] < thresh, \quad k \in 1 \ldots 129, \\ 1.1P_{nk}[n] \qquad\qquad else \end{cases} \tag{7}$$

The threshold is in our case equal to 0.15. The example of estimated and later modified trajectory of $P_{nk}[n]$ is given in Fig. 3.

Accuracy [%]	**Italian**			**Finish**			**Spanish**			overall
conditions	hm	mm	wm	hm	mm	wm	hm	mm	wm	[%]
noise est. 7	88.98	92.61	96.66	91.98	88.85	96.92	92.66	93.07	97.19	**93.60**
noise est. 8	88.87	93.57	96.65	92.69	89.67	96.97	91.49	94.74	97.00	**93.82**

Table 2. Speech recognition results for Italian, Finish and Spanish digit databases with application of temporal filtering based noise estimation system. The experiments' conditions are explained in Tab. 2. The algorithm descriptions are in Tab. 3.

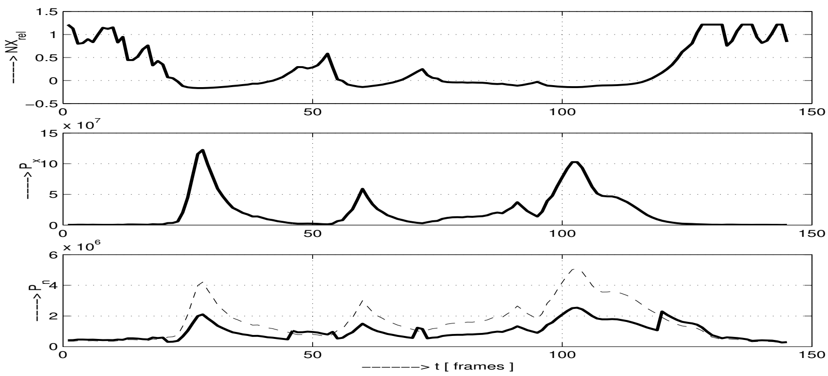

Fig. 3. Trajectories related to 15^{th} spectral band (465Hz).
Upper panel: $NX_{rel_k}[n]$ used for speech/pause detection.
Middle panel: Trajectory of $P_{xk}[n]$.
Lower panel: $P_{nk}[n]$ (dashed line) estimated using temporal filtering of $P_{xk}[n]$, and modified $P_{nk}[n]$ (solid line) by speech/pause detection.

5 Experimental results

The whole proposed noise reduction algorithm is shown in Fig. 1. At the beginning the power spectra $|X_k[n]|^2$ is computed using FFT algorithm. Then the input $|X_k[n]|^2$ is split into two branches. In the upper branch (Fig. 1), the signal goes directly into noise suppression system.

In the lower branch, the noise estimation algorithm is applied.

The noise suppression algorithm is in our experiments only part of the feature extraction. The whole feature extraction system consists of several processing blocks, such as voice activity detection, mean and variance normalization or application of temporal filter in auditory spectrum. The experimented noise estimation algorithms have been tuned while the rest was kept constant so that we did not have to retrain any data-dependent algorithms.

The output features for speech recognizer were based on MFCCs. We have used the standard set of 23 triangular band filters with projection of output log-energies into 15 cosine basis. Tab. 1 contains the results with temporal minima tracking noise estimation technique, while Tab. 2 describes the results of experiments with noise estimation based filtering in temporal domain. The overall results of our experiments are obtained so that the **wm** conditions are weighted by 0.4, **mm** by 0.35, and **hm** by 0.25 over average of all databases.

baseline	Not used noise estimation and noise suppression algorithm at all.
noise est. 1	The average of the first 15 frames of each sentence used, (1x129 f).
noise est. 2	The whole temporal minima tracking alg. [4] in smoothed power spectra (129 spectr. bands), temporal window $D = 80$, (80x129 f).
noise est. 3	Derived from *noise est. 2*, decomposition of temporal window D into 10 subwindows, (10x129 f).
noise est. 4	Derived from *noise est. 2*, decomposition of temporal window D into 5 subwindows, (5x129 f).
noise est. 5	Derived from *noise est. 2*, addition of spectral smoothing using modified Mel-filter bank projection (23 critical banks), decomposition of temporal window D into 10 subwindows, (10x23 f).
noise est. 6	Derived from *noise est. 2*, decreasing the spectral resolution of initial $P_{xk}[n]$ by 2, decomposition of temporal window D into 5 subwindows, linear interpolation into 129 bands, (5x65 f).
noise est. 7	Appl. of standard temporal filter [5] (129 spectr. bands), (1x129 f).
noise est. 8	Derived from *noise est. 7*, speech/pause detector applied, (1x129 f).

Table 3. Description of noise estimation experiments (results mentioned in Tab. 1 and Tab. 2). Each algorithm contains the approximate size of processing memory buffer in floats.

6 Conclusions

Experimented noise estimation techniques for modified Wiener filter based noise suppression algorithm of feature extraction DSR system have been described.

The standard temporal minima tracking noise estimation itself which is guaranteed to be very robust in our task does not satisfy the memory size limitation. Therefore we came up with modification in order to decrease this memory requirement. As can be seen from Tab. 1, the decomposition of one temporal window (applied for one spectral band) into several smaller ones does not bring almost any degradation. However such a memory reduction is not sufficient for our task. So we have experimented with algorithms estimating the noise from spectrum with reduced frequency resolution. Sufficient results were obtained with simple reduction of spectral resolution. The filtering of power spectra by modified Mel-filter bank seems to be applicable too.

On the other side, standard temporal filtering based noise estimation method did not work well. However its advantage is that there is no need for any memory buffer for algorithm processing. The results became interesting when we implemented simple speech/pause detector based on SNR estimation (Tab. 2). Its application for noise estimation based on minima tracking did not bring any improvement.

One of the goals of these experiments was to see if noise estimation techniques can be improved (better overall speech recognition) when doing additional spectral processing. Generally any other spectral processing algorithms, such as spectral smoothing or spectral resolution's reduction did not improve the noise estimation. The spectral processing seems to be good for clean speech (attenuate the noise suppression system's influence when clean speech is processed), but degrade robustness for noisy speech. However the complexity as well as memory size of such noise estimator is widely reduced. Very interesting fact is that spectral processing greatly increases the robustness of noise suppression algorithm when applied on its filter characteristics.

References

1. U. Knoblich. Description and Baseline Results for the Subset of the SpeechDat-Car Italian Database used for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation, Alcatel, April 2000.
2. D. Macho. Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation, Description and Baseline Results, UPC, November 2000.
3. QualComm-ICSI-OGI Aurora Advanced Front-End Proposal, Technical report, January 2002.
4. R. Martin. Spectral Subtraction Based on Minimum Statistics. In *Proc. of EUSIPCO-94*, Seventh European Signal Processing Conference, pp. 1182-1185, Edinburgh, Scotland, U. K., September 1994.
5. G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. In *EUROSPEECH'95* - Proceedings of the 4th European Conference on Speech Technology and Communication, pp. 1513-1516, Madrid, Spain, September 1995.
6. E. C. Hirsch H. G., Noise estimation techniques for robust speech recognition. *Proc. ICASSP'95* pp. 153-156, May 1995.