

NOISE ESTIMATION FOR EFFICIENT SPEECH ENHANCEMENT AND ROBUST SPEECH RECOGNITION

Petr Motíček^{1,2}, Lukáš Burget^{1,2}

¹OGI School of Science and Engineering at OHSU
20000 NW Walker Road, Beaverton, OR 97006 USA

²Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66 CZ
E-mail: {petr,lukas}@asp.ogi.edu

ABSTRACT

Different approaches of minima tracking based noise estimation algorithms are compared and modifications increasing their efficiency are proposed. Estimated noise is used by noise suppression algorithm that is a part of speech recognition system. Moreover, the algorithms are developed to be applied in feature extraction of Distributed Speech Recognition (DSR). Therefore we propose such modifications to the 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 on five Speech-Dat Car (SDC) and Aurora 2 evaluation databases.

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 [6], [7] 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 [9], 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 system.

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.

2. EXPERIMENTAL SETUP

The noise estimation algorithms proposed for speech recognition system were tested on five SpeechDat - Car (SDC) databases used for Advanced DSR Front-End Evaluation: Italian SDC [1], Spanish SDC [2], Finish SDC, German SDC and Danish SDC [3]. 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. *Well-matched (wm)*: All the files (close-talk and hands-free microphones) were used for training and testing. *Medium mismatched (mm)*: Only recordings made with the hands-free microphone were used for training and testing. *Highly mismatched (hm)*: For the training only close-talk microphone recordings were used, whereas for testing the hands-free files were taken.

Furthermore Aurora 2 (noisy TI-digits) database, that is fully described in [5], was used for the evaluation in our experiments too. Here the conditions are divided into multi-condition training and clean training, covering eight noise environments and two types of convolutional distortion.

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 [4], which is used in our feature extraction, was 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 sub-band k , and $|N_k[n]|^2$ is the noise power spectrum. The noise

reduction algorithm can be viewed as filtering by filter with time varying frequency response. At particular time, the filter attenuates low SNR regions of the speech spectrum so that the noise energy portion is removed:

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

An oversubtraction factor *osub* is a filter parameter which varies with time and is estimated from energy of signal and noise. A spectral floor threshold β does not change with time and prevents the low energy regions unchanged.

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. Noise estimation methods without explicit VAD were tested in our feature extraction system.

4.1. Temporal minima tracking

Good estimation of noise in our experiments was obtained with temporal minima tracking algorithm [6]. This algorithm is applied consequently on smoothed power spectrum:

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

with forgetting factor α between 0.75...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 noise power spectrum $P_{nk}[n]$ for k^{th} subband is found as a minimum of $P_{xk}[n]$ within a temporal window of D previous and current power sample:

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

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 temporal 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.

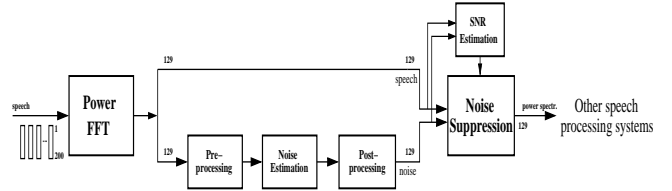


Fig. 1. Scheme of noise suppression system with noise estimation applied in power spectrum.

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 [6], 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.

The size of memory for minima tracking algorithm can be largely decreased when the spectral resolution of $P_{xk}[n]$ is reduced. In one experiment we tried to omit every second spectral subband of $P_{xk}[n]$, estimate the noise and apply some kind of interpolation technique with spectral smoothing to get the initial number of spectral subbands.

In another experiment we tried 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 ($1 \dots F_{sampling}/2$). 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:

$$Mfb_{MOD_k}[i] = \frac{Mfb_k[i]}{\sum_{j=1}^{129} Mfb_k[j]}, \quad (4)$$

where k denotes spectral subband in Mel-scaled spectrum, while i denotes the spectral subband in original power spectrum.

Tab. 1 contains the results for minima tracking based noise estimation algorithm, where $W = 10$ and the spectral resolution was reduced by 2. The results were slightly worse when applying the Mel filter bank instead of simple omitting every second spectral band.

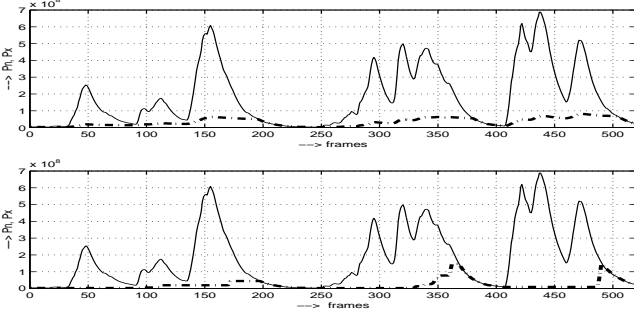


Fig. 2. Process of noise estimation in short-time power spectra (8th subband - related to 250 Hz). The solid lines represent the smoothed power spectra $P_{xk}[n]$, the dashed lines describe estimated $P_{nk}[n]$. *Upper panel:* Filtering of spectral subbands in temporal domain. *Lower panel:* Minima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

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

Often used noise estimation algorithm, proposed in [7], that does not need information about speech/non-speech segments, was 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:

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

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 [7] was not successful in our front-end system. 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 that would modify the trajectories of previously estimated noise. The used algorithm comes from [8] 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:

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

where

$$NX_k[n] = \frac{P_{nk}[n]}{P_{xk}[n]}. \quad (7)$$

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

$$P_{nk}[n] = \begin{cases} 0.4P_{nk}[n] & \text{if } NX_{relk}[n] < \text{thresh} \\ 1.2P_{nk}[n] & \text{else,} \end{cases}$$

where $k \in 1 \dots 129$. The threshold *thresh* 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.

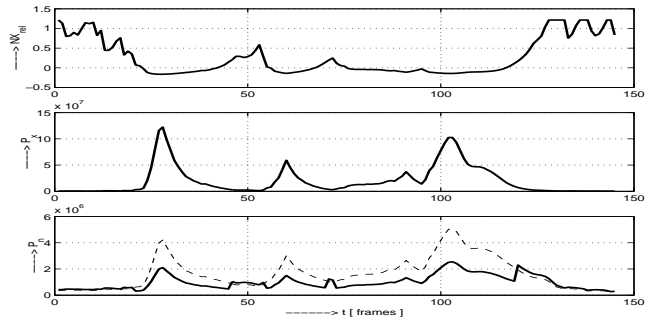


Fig. 3. Trajectories related to 15th spectral band (465Hz). *Upper panel:* $NX_{relk}[n]$ used for speech/pause detection. *Middle panel:* Trajectory of $P_{xk}[n]$. *Lower panel:* $P_{nk}[n]$ (dashed line) estimated using temporal filtering of $P_{xk}[n]$, and modified $P_{nk}[n]$ (solid line) by simple 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 output features for speech recognizer were based on MFCCs. We have used the set of 23 triangular band filters with projection of output log-energies into 15 cosine basis. 15 delta and 15 acceleration coefficients are computed.

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 were tuned while the rest was kept constant so that we did not have to retrain any data-dependent algorithms.

The results for all SDC and Aurora 2 databases are in Tab. 1. The robustness of our noise estimation followed by noise suppression system is compared to the *baseline* system where noise suppression was not applied. In the *system 1* the noise estimation algorithm was performed by temporal minima tracking (section 4.1) algorithm (the spectral resolution was reduced by 2, number of subwindows $W = 10$, the minimal size of processing memory buffer is $65 \times 10 \times 2$ words). In the *system 2* the estimation of noise was based on filtering of temporal trajectories and modified by simple speech/pause detector (section 4.2).

Accuracy	cond.	baseline	system 1	system 2
Aurora 2	set A	85.33%	87.67%	88.35%
Aurora 2	set B	85.41%	87.64%	87.99%
Aurora 2	set C	85.84%	87.11%	88.69%
Aurora 2	overall	85.46%	87.55%	88.07%
SDC	hm	77.44%	86.04%	87.41%
SDC	mm	86.23%	88.60%	88.79%
SDC	wm	94.66%	95.68%	95.66%
SDC	overall	87.40%	90.79%	91.19%
Aurora 2 + SDC	overall	86.62%	89.49%	89.94%

Table 1. Word recognition results for SDC and Aurora 2 databases. The weightings used to obtain the overall results are mentioned in [9].

6. CONCLUSIONS

Experimented noise estimation techniques for our Wiener filter based noise suppression algorithm of feature extraction DSR system were 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 several modifications in order to decrease this memory requirement. From our experiments we can see that 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. Hence 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 brought slightly worse results.

On the other side, standard temporal filtering based noise estimation method did not work as well as the previous noise estimation methods. However its advantage is that there is no need for any memory buffer for algorithm processing. It improved after incorporation of simple speech/pause detector.

The overall performance of described systems demonstrates that the proposals increase the robustness of feature extraction while their application in DSR front-end satisfy given limitations.

7. 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] B. Lindberg. Danish SpeechDat-Car Digits Database for ETSI STQ Aurora Advanced DSR, CPK, Aalborg University, January 2001.
- [4] QualComm-ICSI-OGI Aurora Advanced Front-End Proposal, Technical report, January 2002.
- [5] H. G. Hirsch & D. Pearce. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 2000.
- [6] 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.
- [7] G. Dobliger. 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, Madrid, Spain, September 1995.
- [8] H. G. Hirsch, C. Ehrlicher, Noise estimation techniques for robust speech recognition. *Proc. ICASSP'95* pp. 153-156, May 1995.
- [9] Advanced DSR Front-end: Definition of required performance characteristics, Technical report, version 3, Source: Motorola, October 2001.