Noise Estimation for Spectral Subtraction in Speech Processing

Ing. Petr Motlíček, PhD student,
Faculty of Information Technology,
Supervisor: Dr. Ing. Jan Černocký

This contribution presents two often used algorithms for enhancement of noisy speech signal in spectral domain. These algorithms do not differ in removing noise from the input speech. Different parts are processes of noise estimation which are used before noise suppression. Both algorithms have ability to track non stationary noise signal and do not need voice activity detector. Both methods have been tested in front-end distributed speech recognition (DSR) system. Algorithm parameters, as CPU complexity or memory requirements, as well as the need of delay in processing, will be discussed.

1 Introduction

Spectral subtraction [4], which is well known method for speech enhancement, provides the computationally efficient tool for suppression of an additive noise in a speech signal. The necessary section before spectral subtraction algorithm is a noise power estimator and the section which translates the subband SNRs into spectral weighting factors. These factors control the "efficiency" of the subtraction algorithm for different SNR levels of each spectral subband.
In this paper we compare and evaluate two methods of noise estimation that are supposed to be applied in spectral subtraction algorithm of front-end DSR. The first approach, well described in [1], tracks the minima independently in each subband of the short time power spectrum of the input speech signal. The supposition is that while the peaks of the power spectrum represent the speech activity, the valleys correspond to noise power of each subband. Minima search is realized by shifting a window with a given length in time.
The second approach [2] of noise estimation, which can also be reasonably used in spectral subtraction algorithm, represent in basic principle minima tracking in subbands of power spectrum, but different way. The short time power spectrum of the estimated noise of the speech signal is computed by nonlinear method, which is simply realized by the first residual filter with previously estimated filter coefficients.

2 Methods

2.1 Minima tracking in temporal domain (Method 1)

Let us assume that we have a speech signal $s[n]$ corrupted by an additive noise $n[n]$, where n denotes the time index. Moreover we assume that $s[n]$ and $n[n]$ are uncorrelated. Therefore $E\{x^2[n]\} = E\{s^2[n]\}+E\{n^2[n]\}$. The input speech is projected into spectral domain so that short-time magnitude spectrum $X_k[n]$ and short-time power spectrum $X_k^2[n]$ are obtained, where k denotes frequency index (spectral subband). For the purpose of noise suppression algorithm, the absolute values of $X_k[n]$ and $X_k^2[n]$ are taken into account. However for subjective evaluation of resulting enhanced speech, the phase is kept for restoration of speech.

The estimation of noise is done in power spectrum. First each subband is smoothed by following recursive filter:

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

Then for each k of $P_{xk}[n]$ the minima tracking algorithm is applied. At the beginning a memory buffer of given length D is filled up by previously computed value which is supposed to be first minima value. Then the buffer is shifted frame by frame and the "oldest" value is replaced by actual $P_{xk}[n]$. The output value of each subband $P_{nk}[n]$ is found:

$$P_{nk}[n] = min(P_{xk}[n-D] : P_{xk}[n]). \qquad (2)$$

The noise estimated using this process in short-time power spectra is in Fig. 1 (lower panel).

2.2 Filtering of spectral subbands in temporal domain (Method 2)

Another alternative noise estimation based on minima tracking in short-power spectrum subbands can be done by residual filter with long time forgetting constant. This processing requires smoothed version of power spectrum $P_{xk}[n]$ pre-computed by Eq. 1. The algorithm can be described as follows:

$$P_{nk}[n] = \gamma P_{nk}[n-1] + \frac{1-\gamma}{1-\beta}(P_{xk}[n] - \beta P_{xk}[n-1]). \qquad (3)$$

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. 1 (upper panel).

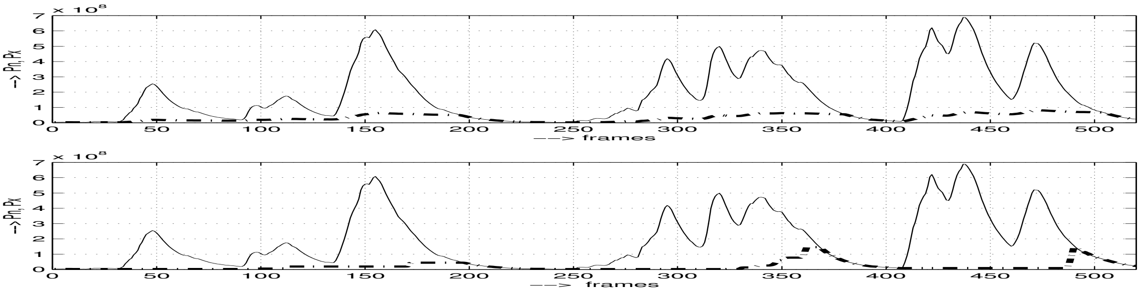

Fig. 1: Process of noise estimation for speech enhancement in short-time power spectra (8th subband - related to 250 Hz). The solid lines represent the smoothed power spectra P_x, the dashed lines describe estimated P_n:
Upper panel: Filtering of subbands in temporal domain (method 2).
Lower panel: Minima tracking in temporal domain (method1).

2.3 Spectral subtraction

The power spectrum of estimated noise $P_{nk}[n]$ has been used in standard spectral subtraction algorithm [4]. The spectral magnitudes are subtracted with an over subtraction factor *osub*. The size of this factor affects a lot the resulting quality of enhanced speech. If *osub* is too small, the noise subtraction algorithm is "soft", which means it doesn't remove the noise. But if *osub* is large, not only noise but partly a pure speech are being subtracted so that a "musical" noise will appear in resulting speech. Therefore *osub* has to be reasonably pre-estimated.

Although the noise estimation has been done in power spectra, the noise spectral subtraction is applied in magnitude spectral domain:

$$|Y_k[n]| = \begin{cases} \sqrt{P_{nk}[n]}, & if subf.\sqrt{P_{nk}[n]} \geq |X_k[n]|Q_k[n] \\ |X_k[n]|Q_k[n], & else, \end{cases} \tag{4}$$

where

$$Q_k[n] = \Big(1 - \sqrt{osub\frac{P_{nk}[n]}{|X_k[n]|^2}}\Big). \tag{5}$$

subf is a spectral floor constant (limitation of the maximum subtraction). In spectral subtraction algorithm, the smoothed version of $|X_k[n]|$ by first residual filter is applied, such like in Eq. 1 (with different time constant).

2.4 SNR estimation

The overestimation factor *osub* in Eq. 5 is in our experiments estimated from SNR. The approximated SNR is computed from energy of input speech and estimated noise:

$$SNR[n] \approx 10.log\left(\frac{\sum_k |X_k[n]|^2}{\sum_k P_{nk}[n]}\right). \tag{6}$$

SNR values are kept in range $0dB...20dB$. *osub* is then for each frame derived from SNR values by scaling, in order to keep the values in range $osub \in (1,5 \div 3,5)$. As we can see from Eq. 6, *osub* does not change over the subbands.

2.5 Estimation of parameters

Parameters of noise estimation as well as spectral subtraction algorithms have been tuned in order to get the most similar enhanced speech to the clean speech. Therefore the parameters were tuned using minimum distance as a criterion. Two experiments of distances' definitions were examined. In the first, the distances were computed between Mel-frequency cepstral coefficients (Eq. 7). In the second (Eq. 8), the distances were computed on the level of short-time power spectra. As a "tuning" algorithm, we have used the Matlab optimization toolbox. The optimized system parameters used in our experiments will be mentioned later.

$$D = \sum_n \Big(\sqrt{\sum_k (C_{cl}(k,n) - C_{out}(k,n))^2}\Big) \tag{7}$$

$$D = \sum_n \Big(\sqrt{\sum_k (|X_{cl}(k,n)|^2 - |X_{out}(k,n)|^2)^2}\Big) \tag{8}$$

Because of construction our noise suppression for front-end DSR system, we have not been allowed to bring delay into the algorithm. The above described techniques satisfy this requirement. At the beginning the used filters (Eq. 1, Eq. 3) need to be initialized. We have tried several experiments with different initializations in oder to preserve the beginning frames of sentences. Good results were obtained with a periodogram (Welch's method of estimating power spectrum). The periodogram was computed from the first two frames of speech. However initialization using just first two frames (averaging) was satisfied too.

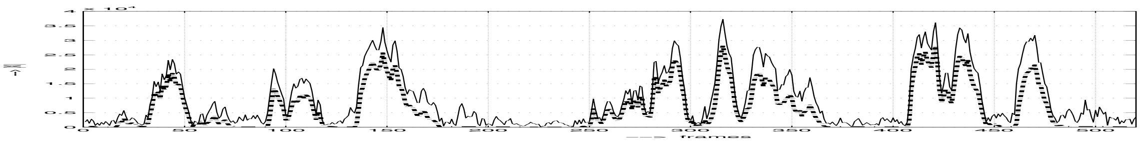

Fig. 2: The comparison of input P_x (solid line) and P_x obtained from the power spectra computed after noise suppression (dashed thick line) for the 8th subband.

3 Testing data

The noise suppression algorithm was tested on part of Ti-digit database. The experiment task was the speaker independent recognition of digit sequences with sampling frequency 8 kHz. The database consist of clean speech as well as speech with noise artificially added at several SNRs (20dB, 15dB, 10 dB, 5dB, 0dB). Four noises were used in training part: recording inside a subway, babble, car noise, recording in an exhibition hall.

The speech recognizer was trained on all noisy data. Furthermore three different sets of speech data were taken for the recognition:

- Set "a": The noises are the same as for training.
- Set "b": Different types of noises were used for this test (restaurant, street, airport, train station), which should represent realistic scenarios for application in a mobile terminal.
- Set "c": The same noise as in test "a" and "b" were used. But the different frequency characteristics were applied (speech transmitted through a different channel).

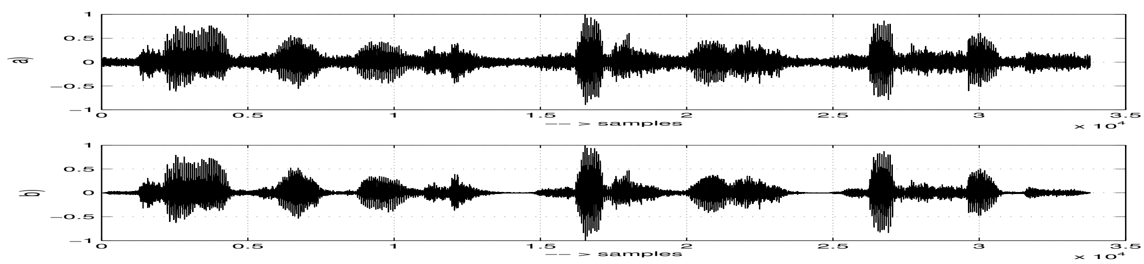

Fig. 3: Input-noisy and output-enhanced speech.

4 Experimental results

The experimental results for Ti-digits database are mentioned in (Tab. 1). Two previously described noise estimators combined with spectral subtraction algorithm have been examined. Mel-frequency cepstral coefficients (MFCC) computed from enhanced short-time magnitude spectra $|Y_k[n]|$ (Eq. 4) were applied as the input features for speech recognizer.

Tab. 1 contains the best scores for algorithm coefficients we have obtained from the optimization toolbox.

The system parameters used in experiments were: $D = 80, \alpha_{method1} = 0.8, \alpha_{method2} = 0.75, \beta = 0.99, \gamma = 0.998, subf = 0.02$.

method	TEST	CLEAN	SNR20	SNR5	SNR0
baseline		98.95	98.32	86.84	58.06
method 1	"a"	98.70	98.11	88.70	69.62
method 2		98.84	98.25	89.94	73.44
baseline		98.95	98.19	86.16	57.49
method 1	"b"	98.70	97.95	87.70	68.69
method 2		98.84	98.17	89.26	71.85
baseline		98.78	98.06	86.02	56.31
method 1	"c"	98.20	97.70	85.86	62.73
method 2		98.39	97.88	87.40	67.18

Table 1: WER [%] for MFCC with different noise suppression system: Baseline-no noise suppression, Method 1-minima tracking noise estimation and spectral subtraction, Method 2-filtering in temporal domain.

5 Conclusions

In this paper we compared the noise estimation systems followed by spectral subtraction algorithm. As can be seen from Tab. 1, better results are obtained when the noise is estimated by filtration of subbands in temporal domain (method 2). The results were examined against baseline system (no noise suppression). As we expected, we did not get better WER when clean speech has been recognized, but the degradation was very small. The improvement is significant for speech with low SNRs. An example of subband of noisy and enhanced short-time power spectra is in Fig. 2.

According to a memory consumption, which is an important specification for front-end systems, the second method does not need any buffer on contrary to the first one, where the previous D values of each subband have to be stored in order to get actual value.

The best results have been achieved with system parameters optimized according to Eq. 8, where short-time power spectra were compared. The enhanced synthesized speech (original phase) is shown in Fig. 3.

References

[1] R. Martin. Spectral Subtraction Based on Minimum Statistics. In *Proc. of EUSIPCO-94*, pp. 1182-1185, Edinburgh, Scotland, U. K., September 1994.

[2] G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. In *EUROSPEECH'95* , pp. 1513, Madrid, Spain, September 1995.

[3] E. C. Hirsch H. G., Noise estimation techniques for robust speech recognition. *Proc. ICASSP'95*, pp. 153-156, May 1995.

[4] Boll S. F., Suppression of acoustic noise in speech using spectral subtraction. *IEEE ASSP*, vol. 2, no. 27, 1979.