Methods for Speech SNR estimation: Evaluation Tool and Analysis of VAD Dependency

Martin VONDRÁŠEK, Petr POLLÁK

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Prague, Dejvice, Czech Republic

vondram3@fel.cvut.cz, pollak@fel.cvut.cz

Abstract. *This paper describes the algorithms for speech SNR estimation and the tool **snr** where these methods are implemented. The definitions of SNR optimized for speech application are summarized and implemented in above mentioned tool. The described tool can estimate the SNR of noisy speech signal with or without reference signal. The tool also can be used to create a speech and noise mixture with required SNR. Analysis of the influence on voice activity detector for the described criteria estimation was provided with artificially mixed data. Finally, the experiments with the measurements on real noisy speech were performed and presented. The described tool is implemented in the C and it can be used for different platforms. It is available for free on the Internet.*

Keywords

Signal-to-noise ratio, SNR, software tools, speech enhancement, voice activity detection, VAD.

1. Introduction

Looking out our ordinary life at present, we can meet different speech technology products, typically speech coding in communication (GSM), voice driven devices, synthetic voice output, etc. Working with a speech input in a real environment, i.e. with some noisy background typically, the robustness of speech technology algorithms with respect to background noise is one of the most studied fields. Noise robustness means mainly reliability of the system working with a noisy background similarly to high quality speech input, but it may also mean high quality speech transmission through a communication channel even with a noise background in input.

The design and development of noise robust systems brings about need for the algorithm behavior analysis in the noise background with different level. It means that the noise level in speech signal should be well measured. This evaluation is, of course, based on standard *signal-to-noise ratio* (SNR), well known criterion, however we encounter two basic difficulties in practical usage with speech signals. Firstly, we must take into account high non-stationarity speech and secondly, we must solve the problem of SNR estimation when only noisy signal is available.

The main task of this paper is to summarize suitable criteria based on SNR for speech applications commonly with the algorithms for their evaluation. Finally, the evaluation tool solving this problem is presented.

2. Algorithm Description

2.1 Criteria Definition

The standard well known SNR definition, for speech signals denoted as the *global SNR* (*GSNR*) is defined as

$$GSNR = 10\ \log \frac{\sigma_s^2}{\sigma_n^2} \tag{1}$$

where σ_s^2 is the power of speech signal and σ_n^2 the power of noise. It corresponds to the properties of the entire signal. The standard SNR definition optimized for speech signals denoted as *SNR* is based on evaluation of *GSNR* only from speech activity parts of the analyzed signal. Formula (1) can be rewritten in this case as

$$SNR = 10\ \log \frac{\sum_{n=0}^{l-1} s^2[n] \cdot vad[n]}{\sum_{n=0}^{l-1} n^2[n] \cdot vad[n]}, \tag{2}$$

where $s[n]$ is the n-th speech sample, $n[n]$ the n-th noise sample, and $vad[n]$ is the information about speech presence for the n-th sample of the signal.

Speech is quasi-stationary signal which is mainly processed in short frames, typically with approximately 30 ms length. The computation of SNR in these segments called *Local SNR* (*LSNR*) is another frequently used criterion. For the i-th segment, it is defined as

$$SNR_i = 10\ \log \frac{\sum_{n=0}^{M-1} s_i^2[n]}{\sum_{n=0}^{M-1} n_i^2[n]} = 10\ \log \frac{\sigma_{s,i}^2}{\sigma_{n,i}^2}, \tag{3}$$

where $s_i[n]$ and $n_i[n]$ are speech and noise samples in the i-th segment of analyzed signal or $\sigma_{s,i}^2$ and $\sigma_{n,i}^2$ are powers in the i-th frame respectively.

Finally, averaged *local SNR* is widely used criterion for speech SNR. The well known *Segmental SNR* (*SSNR*) is defined as the average of SNR_i values over segments with speech activity. It can quantify real level of non-stationary noise in speech more precisely and it was found that it correlates with the perception of the noisy speech by humans [3]. Unfortunately, standard *SSNR* has worse numerical properties because it is based on principle of geometrical averaging of linear signal-to-noise ratio, see (4) where sum of logarithms can be replaced by logarithm of multiplication of linear SNR_i. That is the reason for using of *Arithmetic SSNR* (*SSNRA*), see (5), which is defined as the arithmetic average of linear SNR_i values followed by logarithm, i. e,where K is

$$SSNR = \frac{1}{K} \sum_{i=0}^{L-1} \left(10 \log \frac{\sum_{n=0}^{M-1} s_i^2[n]}{\sum_{n=0}^{M-1} n_i^2[n]} \cdot VAD_i \right) \quad (4)$$

$$SSNRA = 10 \log \left(\frac{1}{K} \sum_{i=0}^{L-1} \frac{\sum_{n=0}^{M-1} s_i^2[n]}{\sum_{n=0}^{M-1} n_i^2[n]} \cdot VAD_i \right) \quad (5)$$

the number of segments with speech activity and VAD_i gives the information about voice activity for the i-th segment.

2.2 Estimation Algorithms

The SNR according to definitions mentioned above can be computed when both parts of the analyzed signal are available, i.e. noise-free speech and additive noise. If the mixture of these two signals is only available in practical applications, the speech SNR must be estimated [7].

The estimation of particular signals in the time-domain is not easy, so the computation is running on the basis of the estimation in the power-domain. Generally, the SNR estimation is working on the basis of following formula

$$\widehat{SNR} = 10 \log \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_n^2} = 10 \log \frac{\sigma_x^2 - \widehat{\sigma}_n^2}{\widehat{\sigma}_n^2}. \quad (6)$$

It means that the estimation of basic SNR is based on the estimation of noise power $\widehat{\sigma}_n^2$. We are assuming uncorrelated additive noise in speech, so signals are additive also in the power-domain. Speech power $\widehat{\sigma}_s^2$ can be then easily estimated by the subtraction presented in equation (6). Basicly, it is a very simple task but it has some limitations to possible non-stationarity of both signal parts. Moreover, when the level of the noise is high, it is very difficult to estimate correctly both the speech power and noise power.

Evaluating the real data with finite frame length, it may appear that $\widehat{\sigma}_n^2 > \widehat{\sigma}_x^2$. It is caused by the estimation error in the situation when the real power of speech is lower than noise one, i.e. $\sigma_s^2 < \sigma_n^2$. We can overcome this problem during the estimation procedure by thresholding, i.e. we set the lowest SNR which can be estimated. A finite value for each short-time SNR is required for further averaging in *SSNR* and *SSNRA* evaluation or also for statistical analysis of a block of experiments.

This paper is focused on the methods when noise power estimation is computed from non-speech parts of the analyzed signal. These methods are very popular and frequently used, however, different implementations may give the results with different precision and numerical stability.

For the estimation of global criterion, i.e. *SNR* (6), the powers are estimated at once, $\widehat{\sigma}_s^2$ from the parts with speech activity - see (7), $\widehat{\sigma}_n^2$ from the speech pause - see (8).

$$\widehat{\sigma}_s^2 = \frac{1}{l_s} \sum_{n=0}^{l-1} x^2[n] \cdot vad[n] - \widehat{\sigma}_n^2, \quad (7)$$

$$\widehat{\sigma}_n^2 = \frac{1}{l_n} \sum_{n=0}^{l-1} x^2[n] \cdot |1 - vad[n]|, \quad (8)$$

where l is the length of a signal in samples, l_s and l_n are the numbers of samples with or without speech activity and $vad[n]$ gives the information of speech presence in the n-th sample. The second approach, which is used much more frequently, is based on power estimation on the frame basis usually followed by the evaluation of segmental SNR. The analyzed segments are divided into two groups, with and without speech activity. Noise power is then estimated as the average of powers in non-speech frames. From the implementation point of view, it is suitable to realize it as exponential averaging, i.e.

$$\widehat{\sigma}_{n,j}^2 = p\, \widehat{\sigma}_{n,j-1}^2 + (1-p)\, \sigma_{x,j}^2. \quad (9)$$

The setting of the parameter p is dependent on segmentation parameters such as frame size, frame overlap, and also on sampling frequency. Generally, this value should be equivalent to time constant of exponential averaging of around 0.5 s. Variable j represents the index of speech-free frame. The estimation of noise power can also be done non-recursively by moving average, by minima tracking [5], or by signal energy statistics analysis [4] and [8]. Details are described in [6] and [7].

Having estimation of the power in particular frames, the speech power estimation is defined as the difference between the power of noisy speech and estimated noise power, so any local or global SNR can be evaluated using formula (6). Target criteria such as *SSNR* or *SSNRA* are then evaluated by definition formulae (4) and (5).

2.3 Voice Activity Detection

The decision about voice activity presence is the sensitive part of the whole algorithm as the noise power estimation can be degraded by the errors in Voice Activity Detection (VAD). There are many approaches to evaluate VAD information and they work with different reliability. We will summarize the most important algorithms for speech activity detection, mainly the approaches which are the most suitable to use during SNR evaluation.

- *Energy based algorithms* - These algorithms are based on simple idea that speech activity means an increase of the signal energy. They are frequently used because of their low computation costs. Their usage with SNR evaluation is often motivated by the fact, that whole algorithm can be based on energy evaluation. The disadvantage is the limitation of their usage with high level noise.

- *Cepstral based algorithms* - They work on the basis of spectral difference between background environment and speech, which can be detected also with higher SNR in comparison of energy VADs. On the other hand, the higher reliability brings also higher computational costs.

Other algorithms of VAD are not suitable for usage in SNR evaluation due to higher complexity or some other specific requirements, e.g. *coherence detectors* which need multichannel signal or *HMM or neural-net based detectors* requiring training and further usage in the same environment etc. We will analyze the influence of two basic VADs on the precision of SNR estimation in the experimental part of the paper.

3. Evaluating Tool '*snr*'

The above described algorithms for SNR estimation are incorporated in the tool ***snr*** [9]. It is implemented in the C and it has three basic modes of operation: evaluation of SNR criteria with reference signal, estimation of SNR criteria from one noisy signal, and finally also creation of noisy speech with given SNR from separate noise-free speech and background signal, as illustrated in Fig. 1.

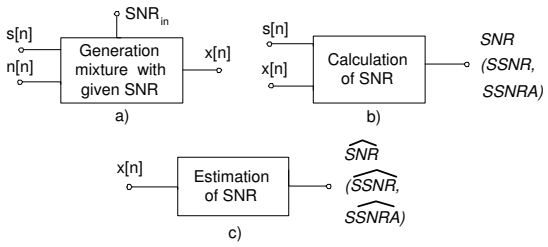

Fig. 1. Functions of program ***snr***.

The noisy speech $x[n]$ with the specified SNR can be created according to all criteria mentioned in the previous section. It is defined as $x[n] = s[n] + k{\cdot}n_0[n]$, i.e. by addition of scaled noise $n_0[n]$. Keeping clean speech $s[n]$ without any modification allows further usage of SNR evaluation with the clean speech as the reference signal.

3.1 Tool Distribution

The described tool is publicly available and interested reader can download the package from the section "*Download*" at the WEB page - **http://noel.feld.cvut.cz/speechlab** . The package consists of:

- *source code* separated into several files,
- *makefile* for the compilation on the most frequent platforms,
- *documentation* in the form of Unix manual page which is also distributed in PDF and PostScript.

The program is typically called from command line with modifiers by standard optional parameters. The required format of signal files is standard 16 bit linear PCM without any header. Working with another sound formats, any tool for sound file conversion may be used, e.g. frequently used ***sox*** [10] which is also available for all platforms. Estimated SNR is typically sent to standard output.

The optional parameters determine strongly the behavior of the program ***snr*** and they are dependent on sampling frequency. Default values are set for $f_s = 8000$ Hz and for adaptation to relatively slow changes in noise characteristics. VAD required for SNR evaluation are implemented by following three algorithms [9]:

- *internal energy detector* based on observations of maxima and minima of the short time signal energy,
- *internal cepstral detector* based on evaluation of cepstral distance from differential cepstral coefficients [1], [2] - default detector,
- *external VAD* which may be produced by independent tool is read from file.

3.2 Typical Application in the Evaluation

Figure 2 gives the example of noise suppression algorithm analysis. Creating artificial noisy speech, we have available clean speech and it allows to use it as the reference signal during analysis of SNR at the output of noise suppression system. In this case, it must be guaranteed that clean speech passes through the noise suppression system without delay, without amplification, or attenuation. Otherwise the usage of SNR evaluation with reference signal is not correct and the estimation of SNR without reference should be preferred, even if affected by an error of the estimation.

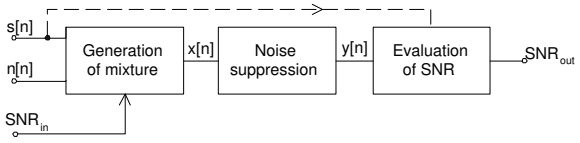

Fig. 2. Possible setup of noise suppression evaluation.

Above described block scheme could be simply realized by following commands with default settings, i.e. frame length 256 samples with step by 128 samples, internal cepstral detector, exponential averaging. *SSNRA* of input speech enhancement system is 5 dB.

Example of tool usage:

snr **-SSNRA -m** *5 noise.sig clean.sig in.sig enhance in.sig out.sig* (test tool for noise suppression)
snr **-SSNRA** [*clean.sig*] *out.sig*

4. Experiments

The functionality of SNR estimation evaluated by the tool ***snr*** was verified by two different experiments [9]. The first group of tests was done with simulated data. In this case we knew the *SNR*, *SSNR* or *SSNRA* respectively of analyzed signal so we could compare estimated and reference values of SNR. Especially these tests were focused on the impact of voice activity detector. The second group of experiments was done with real data recorded in noise environment to confirm the applicability of these methods in the analysis of real noisy speech SNR. The block scheme of realized experiments is depicted as follows in Figure 3.

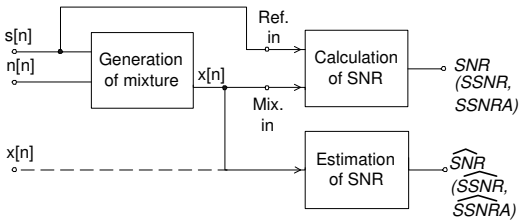

Fig. 3. Block scheme of realized experiments.

4.1 General Numerical Properties

Evaluating particular criteria, we can meet the following general properties of the above described criteria discussed in more details in [6] and [7]:

- *SNR* gives the information robust to speech pauses, but it often fails when $\widehat{\sigma}_n^2 > \widehat{\sigma}_x^2$, which is the case especially for low *SNR* and highly non-stationary noise.
- *SSNR* is a more robust criterion. In relation to *SNR*, it gives approximately 5 dB lower value [6]. The estimation of *SSNR* is very sensitive to the error in short-time SNR_i estimation.
- *SSNRA* gives value close to *SNR* and it may be also assumed as estimation of *SNR*. It is less sensitive to the error of short-time SNR estimation because the influence of the error in the i-th frame, especially when $\widehat{\sigma}_{n,i}^2 > \widehat{\sigma}_{x,i}^2$, is smoothed due to averaging before the computation of logarithm.

4.2 Used VAD

The performance of described SNR estimation methods depends strongly on VAD. It is clear that due to failure of VAD, the results of $\widehat{\sigma}_n^2$ estimation must be influenced by an error. Also the estimation of *SSNR* provided by averaging in speech activity is dependent on VAD results. We analyzed the influence of three basic VADs in our experiments.

1. *ideal voice activity detection*

   This detector is used for analyzing pure properties of SNR estimation algorithm. Finally, we worked with VAD over clean speech signal (which is available during artificial simulation) performed by cepstral detector. The error of cepstral VAD was very low in this case and we could assume this detection as ideal.

2. *energy voice activity detector*

   This is the efficient and most frequently used algorithm. It's main disadvantage is small accuracy for low SNR which influences the estimation of SNR.

3. *cepstral voice activity*

   This algorithm is more reliable in wider range of SNR [2]. On the other hand it's implementation is more complicated and the usage in real time processing more difficult. It is the question of required accuracy of SNR estimation to be reasonable to use this approach.

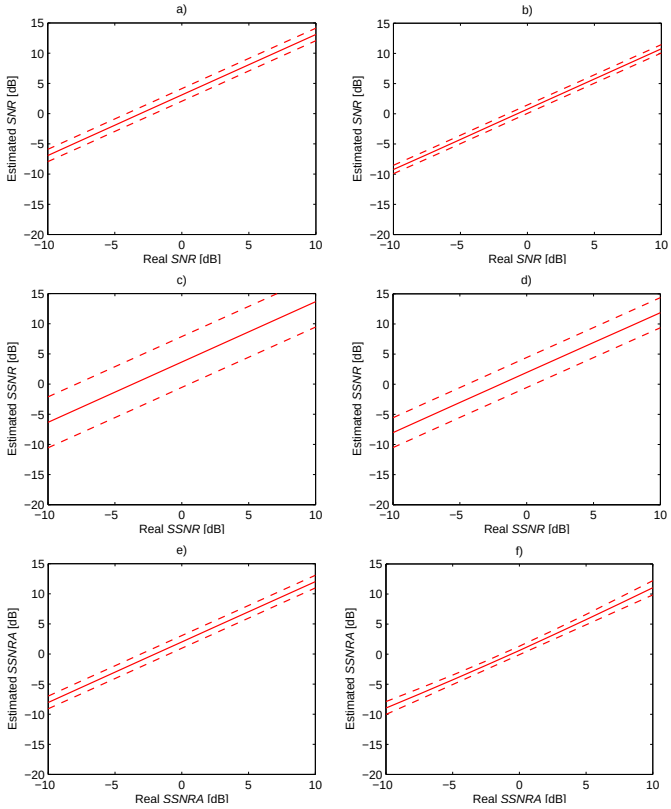

Fig. 4. Impact of voice activity detectors for *SNR*, *SSNR* and *SSNRA*: a), c) and e) energy detector, b), d) and f) cepstral detector.

4.3 Results on Database CAR2ECS

The first experiment [9] was realized with the CAR2ECS database. It contained clean speech recorded without background and also pure noise backgrounds without any utterance, typical for this environment. We created artificial noisy speech with three types of noise backgrounds, i.e. stationary noise, the background with slow non-stationarity, and highly non-stationary environment. Target SNRs of these mixtures were set in the range of $-10 \div 10$ dB. Ideal VAD was always used for creating of noisy speech. We obtained the set of 477 signals for further analysis from 53 speakers, both males and females. Utterances were different, so we tested fluent speech with small pauses as isolated words sequences with longer pauses, typical e.g. for digit strings.

4.3.1 VAD Influence in Criteria Definitions

Firstly, we have analyzed the results of SNR evaluation using the reference clean speech. VAD information was required to select the part of the signal used for SNR evaluation, both for global and segmental criteria. Possible error degraded evaluated criteria as the consequence of incorrectly selected signal parts.

The obtained results for evaluation with energy and cepstral VAD are shown in Fig. 4. The results with ideal detection are not presented because they are very close to the results with cepstral detector. The average value of evaluated criterion is always drawn by solid line, dashed ones represent range of the majority of variation given by the interval mean value with

plus-minus standard deviation. Thus we can assume that *SSNR* is sensitive to VAD accuracy which means in the consequence higher variance of target evaluated *SSNR*. On the other hand *SNR* and *SSNRA* give the results which are similar with respect to the average value and the variance.

4.3.2 VAD Influence in SNR Estimation

In this part we analyzed further degradation of the results, the error of VAD influences and also the estimation of noise power which is done during speech pauses. The estimations of *SNR*, *SSNR*, and *SSNRA* have slightly different numerical properties [9] which are described below.

The results of *SNR estimations* are shown in Fig. 5. When energy detector is used, see Fig. 5a), the estimation starts failing below 0 dB. The reference estimation is indicated by solid straight line. The usage of cepstral detector gives better results in mean value of estimated *SNR*, but stochastic error of the estimation increases.

Next Figure 6 shows the result of *SSNR estimation* [9]. Energy detector gives unusable results, see Fig. 6a), which are acceptable just for the highest *SSNR*. Figures 6b), 6c) and 6d) show the result with cepstral detector with impact of noise background to *SSNR* estimation.

In the case of *SSNRA* estimation, both energy and cepstral detector could be used with acceptable results. Figure 7 shows obtained results again in different noise backgrounds Energy detector gives very good results, if *SSNRA* is bigger than 0 dB. Cepstral detector keeps mean value of the estimation also below 0 dB, on the other hand, the variance of the estimation is slightly higher. For highly non-stationary noises, see Fig. 7e, f), the estimation is a general problem.

4.4 Results on Database SPEECON

We analyzed also signals from SPEECON database and the results were compared with the *SNR* evaluated by used recording platform. Both algorithms provided similar estimations of SNR using speech activity detector. SPEECON platform uses speech activity detector based on log-energy thresholding. Algorithms presented in this paper with cepstral detector can work with more precise speech/non-speech resolution which may yield to more realistic estimation under higher noise background. SPEECON platform gives in principle slightly optimistic estimation for low SNRs. We think that this hypothesis is confirmed by the results of our experiments.

We were comparing SPEECON SNR and SSNRA estimated with cepstral and energy detector. Under higher SNR all algorithms gave similar results, see Fig. 8. We can see similar histograms for all evaluated SNRs (SPEECON SNR is given by the curve) and also the correspondence between SPEECON SNR and our algorithm is very good. Fig. 9 gives similar results for two different sessions in car environment. We can see mixtures of two histograms. For higher SNR histograms fit quite well again. For lower SNR we can see that the estimation with cepstral detector gives according our opinion more realistic lower SNR. Estimation with energy algorithm gives result similar as SPEECON algorithm. Detailed analysis for different backgrounds was described in [9].

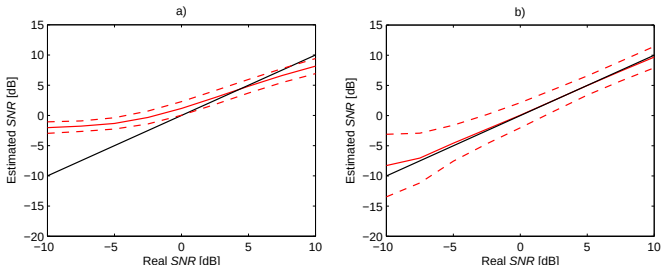

Fig. 5. Estimation of *SNR*: a) energy detector, b) cepstral detector

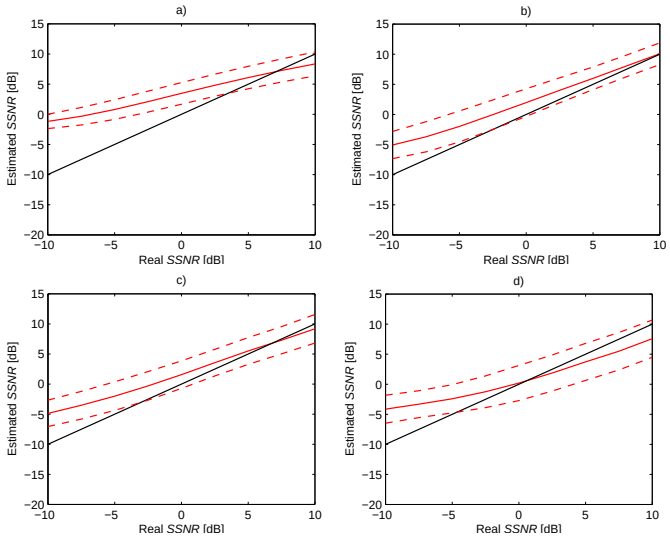

Fig. 6. Estimation of *SSNR*: a) usage of energy detector for stationary noise, b) cepstral detector for stationary noise, c) non-stationary noise with slower changes and d) non-stationary noise with fast changes.

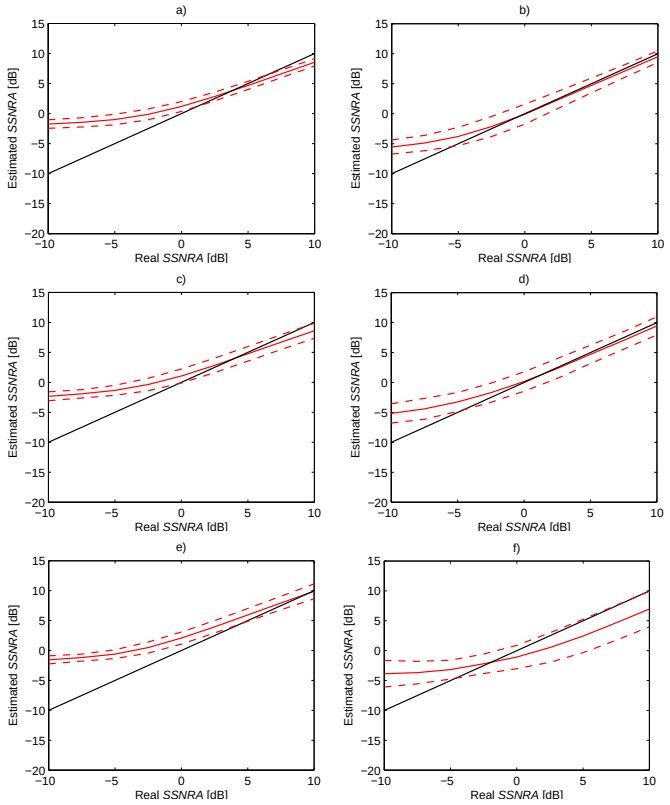

Fig. 7. Estimation of *SSNRA* for stationary noise, non-stationary noise with slower changes, non-stationary noise with fast changes: a) , c), and e) energy detector, b), d) and f) cepstral detector

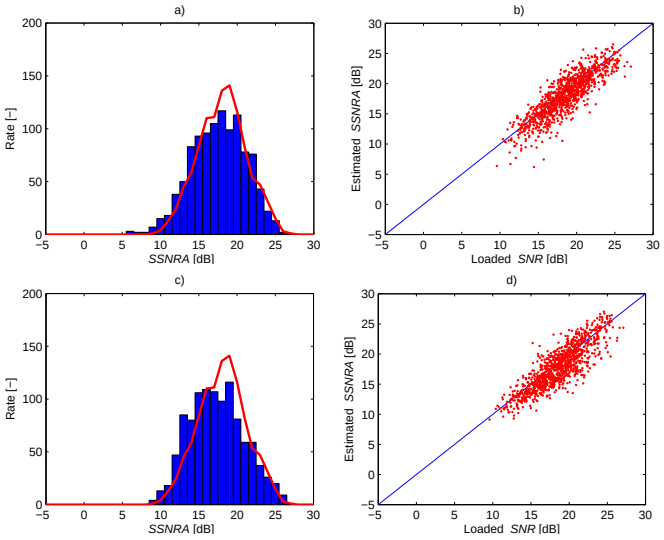

Fig. 8. Estimation of *SSNRA*: a), b) cepstral detector, c), d) energy detector in standard office environment.

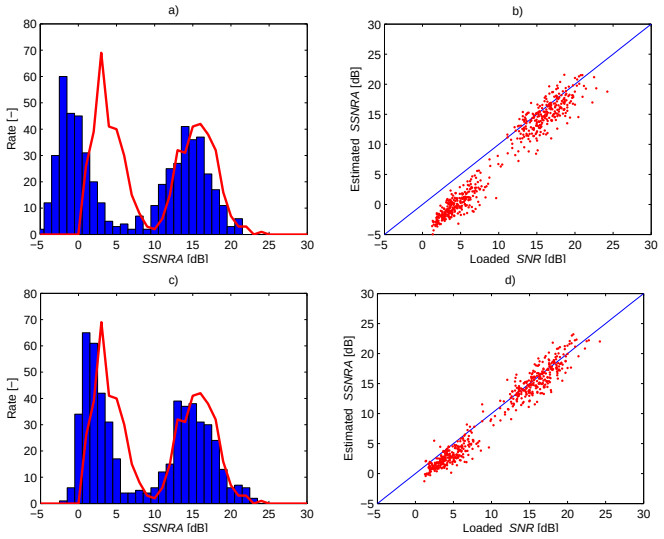

Fig. 9. Estimation of *SSNRA*: a), b) cepstral detector, c), d) energy detector in car environment.

5. Conclusions

This paper described the criteria for speech SNR measurement and the tool providing their estimation. Criteria giving the information independent on variability in speech pauses utterance were summarized. Estimation algorithms based on noise power estimation during non-speech part of the signal were described, commonly with suitable Voice Activity Detectors (VAD). The most important conclusions are:

- *SSNRA* seems to be an optimal criterion because of numerical stability and very low stochastic error in the case of the estimation from one signal without reference. The results are very good especially when the background is stationary or just slowly varying in its characteristics.
- Comparing the usage of simple energy and cepstral VADs, they both have advantages and disadvantages. Energy detector gives generally higher error of estimation, but it is easy to implement and sufficiently reliable for higher SNR. Cepstral detector is more precise for estimation of lower SNR with just slightly higher stochastic error, but the algorithm is more complex.
- All described criteria are implemented in the ***snr*** tool which is available with a source code and simple documentation at WEB-site noel.feld.cvut.cz.

Acknowledgment

The paper was partially supported by GACR 102/05/0278 ”New Trends in Research and Application of Voice Technology”, GACR 102/03/H085 ”Biological and Speech Signals Modeling”, and research activity MSM 6840770014 ”Research in the Area of the Prospective Information and Navigation Technologies”.

References

[1] HAIGH J. A., MASON. S. A voice activity detector based on cepstral analysis. In *Eurospeech 93*. Berlin (Germany), 1993.

[2] JELÍNEK, T. *Differential Cepstral Detector of Voice Activity*. Diploma theses CTU-FEE, 2004 (in Czech).

[3] JUNQUA, J.-C., HATON, J.-P. *Robustness in Automatic Speech Processing*. Kluwer Academic Publishers, 1996.

[4] KORTHAUER, A. Robust estimation of SNR of noisy speech signals for the quality evaluation of speech databases. In *Proc. Robust Methods for Speech Recognition in Adverse Conditions*. Tampere (Finland), 1999.

[5] MARTIN, R. An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Eurospeech 93*. Berlin (Germany), 1993, pp. 1093 - 1096.

[6] POLLÁK, P. Efficent and reliable measurements and simulation of noisy speech background. In *EUSIPCO 2002*. Toulouse (France), 2002.

[7] POLLÁK, P. Estimation methods of speech signal-to-noise ratio. *Acoustic Sheets, č.7*, 2001 (in Czech).

[8] RIS, Ch., DUPONT, S. Assessing local noise level estimation methods: Application to noise robust ASR. *Speech Communication*, 2001, pp. 141-158.

[9] VONDRÁŠEK, M. *Estimation of Speech SNR in Signal from Noisy Environment*. Diploma theses CTU-FEE, 2004 (in Czech).

[10] BAGWELL, Ch. SoX - Sound eXchange. *http://www.sox.com*: - Sox software WEB page.

About Authors...

Martin VONDRÁŠEK was born in Borovany, Czechoslovakia in 1977. He received the M.Sc. degree at the Faculty of Electrical Engineering of the Czech Technical University in Prague (FEE CTU) in 2004. Currently, he is a Ph.D. student at FEE CTU. His current research interests include speech enhancement, speech pre-processing for cochlear implants, and others.

Petr POLLÁK was born in 1966 in Ústí nad Orlicí (Czechoslovakia). After the graduation (ing. 1989) he joined Czech Technical University in Prague, Faculty of Electrical Engineering (CTU FEE), where he has received also further degrees (CSc. 1994, Doc. 2003). He works as a teacher at Department of Circuit Theory and as researcher in Speech Processing Group. His most important activities are in speech enhancement, robust speech recognition, speech database collection, and joined activities. He was the responsible person of several EC project aiming at speech database collection (SpeechDat, SPEECON, and other projects with European industrial partners).