EFFICIENT AND RELIABLE MEASUREMENT AND SIMULATION OF NOISY SPEECH BACKGROUND

Petr Pollák
Czech Technical University in Prague,
ČVUT FEL K331, Technická, 166 27 Praha 6, Czech Republic
Tel: +420 2 2435 2049; fax: +420 2 3333 9805
`e-mail: pollak@feld.cvut.cz`

ABSTRACT

One of the most required tasks in different applications of noisy speech processing are the measurement of signal-to-noise ratio (SNR) and precise simulation of noisy background. In this paper different methods of SNR estimation were compared with the reliability of these estimations from the noisy speech. It was found that segmental SNR seems to be the optimal solution for the measurement of the level of noise in the speech from the point of view reliability of the estimation procedure. The results with arithmetical averaging in segmental SNR evaluation are presented. This algorithm gives values close to standard global SNR and its big advantage is less sensitivity to the error of local SNR estimation.

1 INTRODUCTION

Many applications in concurrent speech processing research are more frequently oriented to the processing of real life speech. It brings the requirement of techniques like noise suppression and other speech enhancement techniques, noise robust recognition, etc. Evaluating these techniques, the precise analysis of noisy speech signals starts being very important step of this research. The most frequent analysis is usually focused on the evaluation of speech quality with main interest in the measurement of the level of noise in speech signal.

Noise level in the signal is quantified by well known criterion Signal-to-Noise Ratio (SNR). It is used also for the measurement of noise level in speech, however, it may cause some difficulties given by natural character of speech signal. To eliminate the influence of varying pauses in different speech utterances, the SNR should be computed only over the speech activity parts of the signal [2].

When only real noisy speech is available, speech and noise variances must be estimated only from the one signal. The majority of algorithms estimate noise variance, typically from speech pauses [2], [3], [6], by low-variance envelope tracking [5], by the statistical analysis of signal variance envelope [9], etc. Nice overview of different noise estimation techniques can be found in [7].

Algorithms usually require the information about speech presence. From this points of view, Voice Activity Detector (VAD) starts being very important part of the whole algorithm and as it will be shown later, the failures of VAD yield to final failure of whole SNR estimation. Different VAD can be found in [8], [1], [4]. However, we will discuss the influence of VAD to SNR estimation, no details will be presented on mentioned VAD algorithms.

In this contribution several definitions of SNR are presented. It will be shown that they give different results for same noise level. Secondly, the analysis of estimation algorithms is presented. It will be shown that different algorithms are more or less sensitive to the errors in the estimation procedure. Finally, formulae for artificial modelling of different noise backgrounds according to discussed criteria are derived.

2 SNR DEFINITIONS

In all further text we will assume that $s[n]$ is speech signal, $n[n]$ the additive noise, and $x[n] = s[n] + n[n]$ its mixture. General SNR is given as

$$SNR = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2}\,, \tag{1}$$

where σ_s^2 and σ_n^2 are variances (powers) of above discussed signals. For basic definition of SNR for noisy speech, the formula is completed by the information about speech activity.

Basic (global) SNR criterion:

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

where $vad[n]$ gives the information about speech presence in current signal sample (1 - speech is present, 0 - speech pause). Variances of speech and noise are computed from the parts of same length (where $vad[n] = 1$), so the computation can be based on energy basis.

Segmental SNR:

$$SSNR = \frac{1}{K}\sum_{i=0}^{L-1} VAD_i \cdot 10\log_{10}\frac{\sum\limits_{n=0}^{M-1} s_i^2[n]}{\sum\limits_{n=0}^{M-1} n_i^2[n]}\,, \qquad (3)$$

Modified segmental SNR:

$$SSNRA = 10\log_{10}\left(\frac{1}{K}\sum_{i=0}^{L-1} VAD_i \cdot \frac{\sum\limits_{n=0}^{M-1} s_i^2[n]}{\sum\limits_{n=0}^{M-1} n_i^2[n]}\right)\,, \qquad (4)$$

where $s_i[n] = s[m \cdot i + n]$, $n_i[n] = n[m \cdot i + n]$, M is the length of processed frame, m is the segmentation step, VAD_i is the information about speech presence in i-th frame, L total number of frames in analyzed signal, and finally K is the number of the frames with speech activity.

Segmental SNR (*SSNR*) with VAD is the criterion which is also not influenced by the non-stationarity of the speech or by different length of speech pauses. Moreover, it can quantify more precisely real level of non-stationary noise and it is highly correlated with the perception of the noisy speech by human ear, see [2].

Arithmetical segmental SNR (*SSNRA*) is based on arithmetical averaging of linear signal-to-noise ratios, while the *SSNR* is based on geometrical average of linear signal-to-noise ratios because the equation (3) can be rewritten as

$$SSNRA = 10\log_{10}\left(\prod_{j=0}^{K-1} VAD_j \cdot \frac{\sum\limits_{n=0}^{M-1} s_j^2[n]}{\sum\limits_{n=0}^{M-1} n_j^2[n]}\right)^{\frac{1}{K}}\,. \qquad (5)$$

Index j in the equation (5) represents only of the frames with speech activity, where $VAD_j = 1$ must be fulfilled to avoid the saturation of geometrical average on zero value. The difference between type of averaging of local signal-to-noise ratios is very important from the point of view estimation and it will be discussed in the following section.

It can be easily demonstrated that *SSNR* gives approximately -5 dB lower values than global *SNR* while the values of *SSNRA* and *SNR* are very close. It is given by the fact that the averaging before the evaluation of logarithm gives very similar results as global computation of variances.

3 NOISE BACKGROUND SIMULATIONS

In many cases we would like to have artificially modelled noisy background in the speech signal. This part gives the procedure how to create mixtures according to above defined criteria *SNR*, *SSNR*, or *SSNRA*. When the mixture is created as $x[n] = s[n] + k \cdot n_o[n]$, its *SNR* is[1]

$$SNR = 10 \cdot \log \frac{\sigma_s^2}{k^2\sigma_{no}^2}\,. \qquad (6)$$

The noise scaling factor k can be then computed according to the following formulae.

Mixture according to *SNR*:

$$k \;=\; \sqrt{10^{-\frac{SNR}{10}} \cdot \frac{\sum\limits_{n=0}^{l-1} s^2[n]vad[n]}{\sum\limits_{n=0}^{l-1} n_o^2[n]vad[n]}} \qquad (7)$$

Mixture according to *SSNR*:

$$k \;=\; \sqrt{10^{-\frac{SSNR}{10}} \cdot \left(\prod_{j=0}^{K-1} \frac{\sum\limits_{n=0}^{M-1} s_j^2[n]}{\sum\limits_{n=0}^{M-1} n_{o,j}^2[n]}\right)^{\frac{1}{K}}} \qquad (8)$$

Mixture according to *SSNRA*:

$$k \;=\; \sqrt{10^{-\frac{SSNRA}{10}} \cdot \frac{1}{K}\sum_{i=0}^{L-1} VAD_i \cdot \frac{\sum\limits_{n=0}^{M-1} s_i^2[n]}{\sum\limits_{n=0}^{M-1} n_{o,i}^2[n]}} \qquad (9)$$

The meanings of the variables are same as in the section 2 and only frames with speech activity must be averaged in the computation of (8) again (that's the reason why index j is used instead of i).

4 ESTIMATION ALGORITHMS

Above described definitions can be easily used for the estimation of SNR only when a reference signal is available. Usually, speech signal (clean) is used as the reference. Noise can be then computed in discrete time-domain as $n[n] = x[n] - s[n]$ because we assume that the noise is additive.

Having the real life problem, we must estimate SNR from one signal without any reference. It may be very difficult problem especially in the case of similar spectral characteristics (typically in car environment). The estimation of noise in the time-domain cannot give satisfying results including the correct phase information which is necessary for application of above discussed definition formulae. That's the reason why the estimation are usually provided in the power/variance-domain.

If the noise is additive and uncorrelated with speech signal (it can be assumed for signals from different

[1] Similarly it can be written also for *SSNR* and *SSNRA*.

sources), the variance of its mixture is given as $\sigma_x^2 = \sigma_s^2 + \sigma_n^2$. SNR can be then estimated two ways as

$$\widehat{SNR} = 10 \log_{10} \frac{\sigma_s^2}{\sigma_x^2 - \sigma_s^2} \tag{10}$$

$$\widehat{SNR} = 10 \log_{10} \frac{\sigma_x^2 - \widehat{\sigma}_n^2}{\widehat{\sigma}_n^2} . \tag{11}$$

The second approach given by the formula (11) is used most frequently because we can estimate much more easily the variance of noise background than very rapidly changing variance of speech signal. The problem is then simplified to the estimation of noise variance and we can find frequently used two basically different approaches.

Estimation by averaging in speech pauses - [2], [3], [6]
This algorithm represents the most frequently used approach. Usually, the exponential averaging is preferred because its recursive formula is advantageous for the evaluation. This algorithm gives very good results for correct VAD information.

Low-variance tracking - [5], [7]
The evaluation of local estimation of noise variance without VAD information is the main advantage of this algorithm but the VAD information is needed in next step because computation of segmental SNR should be done only over speech frames.

5 ANALYSIS OF ESTIMATION ERRORS

The analysis of estimation errors were provided on artificially mixed speech without noise and real car noise on different levels. It gives the information about the value which should be then estimated. The results of the estimations on -5 dB and 0 dB for *SSNR* are presented (0 dB and 5 dB for *SNR* and *SSNRA* because it should represent equivalent level of noise, see section 2). Finally, the criteria were also tested on real noisy speech data recorded in the running car. Following estimation errors were studied.

Power subtraction - The formula $\sigma_x^2 = \sigma_s^2 + \sigma_n^2$ is fulfilled only in the limit case and this fact brings the stochastic error into SNR estimation. It was analyzed by evaluation of SNR criteria with reference signal according to (10). Generally, obtained values were biased less than 0.5 dB with standard deviation in the range $0.1 \div 0.5$. It means that any estimation algorithm cannot work with lower estimation error.

Consequently, the subtractions in (10) and (11) can give negative values. They must be eliminated before the evaluation of logarithm. Mainly, it appears in global *SNR* computation. We can see on fig. 4 that in many cases we obtain due this fact the limit value -30 dB.

VAD error - Errors in VAD detection influence the SNR estimation on two levels: on the level of noise variance because it is estimated in speech pauses and in the averaging of local SNR during the evaluation of segmental SNR over speech frames. It is demonstrated on fig. 1 and 2 - solid line histograms represent the estimations of SNR with ideal VAD information (computed on clean speech), bar histograms were obtained with cepstral VAD detection on noisy speech, and finally dashed histograms show higher failures of SNR estimation when energy VAD was used on noisy speech.

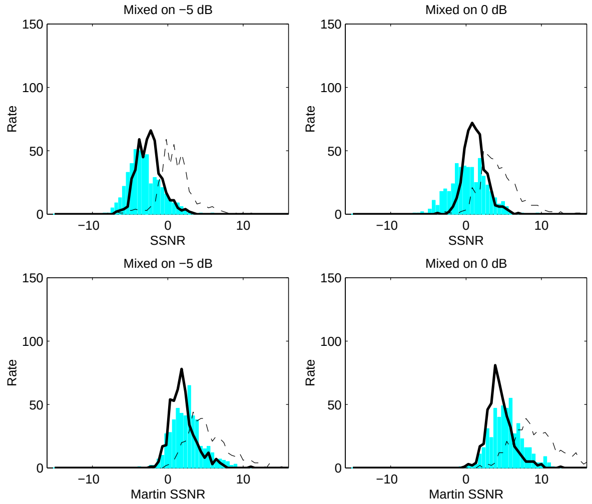

Figure 1: Histograms of SSNR estimation.

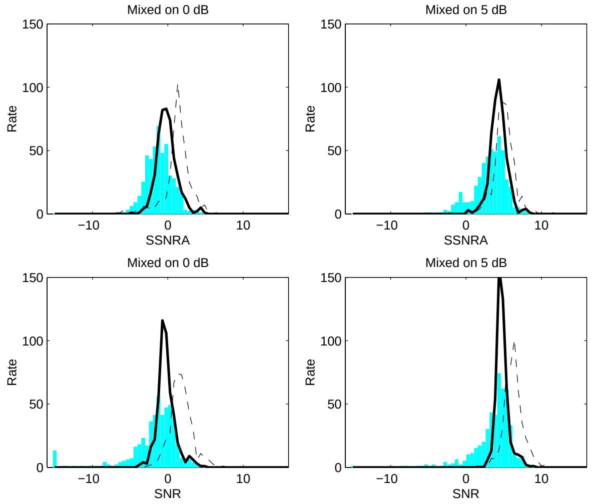

Figure 2: Histograms of SSNRA and SNR estimation.

Estimation of noise power - Having good VAD information, the noise power can be well estimated by the averaging in speech pauses. On fig. 1 and 2 we can see that we obtain very sharp histograms with relative small variance with ideal VAD. Martin's algorithm which works without VAD gives good results for higher SSNR while for lower values the estimated values start giving higher values, see fig. 1.

Error in local SNR - The last analysis was motivated by the comparison of *SSNR* and *SSNRA*. The influence of different values of one local SNR to the final *SSNR* and *SSNRA* was studied. The results are shown on fig.3 (lines with higher dynamics represent the averaging over

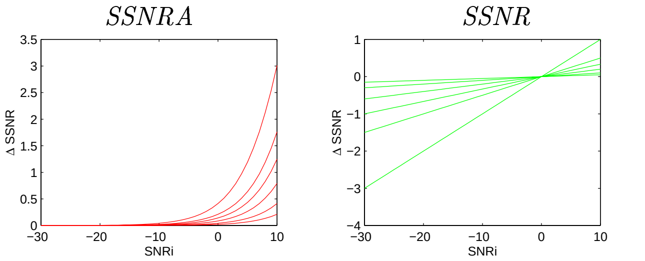

Figure 3: Influence of short-time SNR to SSNR and SSNRA (basic SSNR is 0 dB).

smaller number of frames or more frequent appearance of this value in the set of local SNRs respectively).

SSNRA is more influenced by higher values of local SNR while *SSNR* has linear relation on local SNR. From the estimation point of view the most important part is for lower local SNR because mainly these values may contain important estimation error (in high level noise case). It means that the estimation of *SSNRA* will be less sensitive to this error than the estimation of standard *SSNR*. It was also proved by the results on fig. 1 and 2 where the histograms for *SSNRA* are more narrow than for *SSNR*.

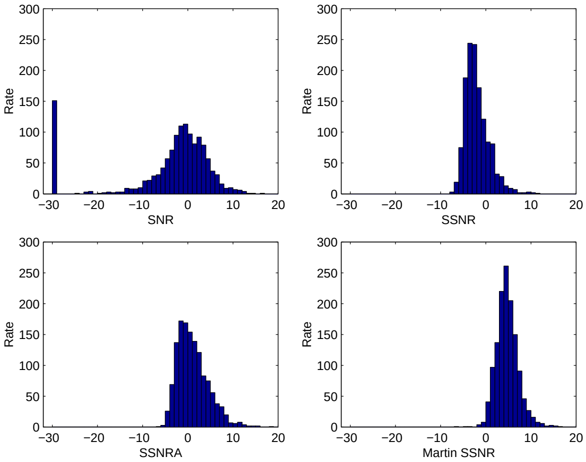

Figure 4: Histograms for SNR estimation of speech recorded in car.

6 CONCLUSIONS

Different methods of estimation of global and segmental SNR were compared. Main focus was given to the methods based on noise variance estimation during speech pauses and to the analysis of errors of the estimation algorithms.

Global SNR with VAD information (*SNR*) can give the information robust to speech pauses. Nevertheless, the estimation of this criterion can give low-saturated values when $\widehat{\sigma}_n^2 > \sigma_x^2$ (fig. 4). Mainly, it may be the consequence of VAD failure which may be quite frequent for low SNRs.

Segmental SNR (*SSNR*) is more robust criterion. In relation to *SNR* the value is approximately 5 dB less. The estimation gives quite satisfying results with reliable cepstral VAD. Low saturation cases appear in local SNRs and in comparison to *SNR* their influence is minimized during further averaging.

The best results were obtained for the estimation of Arithmetical SSNR (*SSNRA*). This criterion can be estimated with minimal standard deviation. In principle, the value is very close to standard global SNR but it is not influenced by the low-saturation effect. The main advantage is small sensitivity of this criterion to the estimation error on low local SNR.

Formulae for artificial mixing according to above described criteria were presented. They are very useful for the simulations of real noise environment during evaluation phase of different systems. The main advantage of these simulations is the availability of reference clean speech signal for further analysis.

Acknowledgement

This paper is the part of the research supported by grant “Theory and Applications of Voice Communication in Czech” with No. GACR-102/96/K087.

References

[1] J. A. Haigh and J. S. Mason. A voice activity detector based on cepstral analysis. In *Eurospeech’93*, pages 1103–1106, Berlin, September 1993.

[2] J.-C. Junqua and J.-P. Haton. *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1996.

[3] A. Korthauer. Robust estimation of the SNR of noisy spech signals for the quality evaluation of speech databases. In *Proc. of Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.

[4] A. Le Floc’h, R. Salami, B. Mouy, and J.-P. Adoul. Evaluation of linear and non-linear subtraction metods for enhancing noisy speech. In *Speech Processing in Adverse Conditions*, pages 131–134, Cannes-Mandelieu (France), November 1992.

[5] R. Martin. An efficient algorithm to estimate the instantaneous SNR of speech signals. In *Eurospeech’93*, pages 1093–1096, Berlin, Sep 1993.

[6] P. Pollák. Metody odhadu odstupu signálu od šumu v řečovém signálu. *Akustické listy*, 2001. In Czech language.

[7] Ch. Ris and S. Dupont. Assesing local noise level estimation methods: Application to noise robust asr. *Speech Communication*, pages 141–158, 2001.

[8] P. Sovka and P. Pollák. The study of speech/pause detectors for speech enhancements methods. In *Eurospeech’95*, pages 1575–1578, Madrid, Spain, September 1995.

[9] H. van den Heuvel. Validation criteria. SpeechDat(E) deliverable ED1.4.2, SPEX, May 1999.