SNR-BASED ASSESSMENT OF QUALITY OF SPEECH ENHANCEMENT USING SINGLE-CHANNEL METHODS

Zdenek Smékal[*] *and Robert Vích*[**]

[*]*Institute of Telecommunications, Brno University of Technology*
[**]*Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic*
smekal@feec.vutbr.cz

Abstract: In the paper, the authors discuss methods for assessing the quality of noise-degraded speech using the signal-to-noise ratio (SNR). This objective assessment is of advantage in the evaluation of speech encoding and also in the rough assessment of the effectiveness of single-channel methods of speech enhancement. The application of listening tests is costly and therefore subjective assessment methods are only used for optimized enhancement methods. First, the classical definitions of SNR are given, where the knowledge of undisturbed speech signals is assumed. Besides the classical definitions, the definitions of SNR for averaging finite signal frames of different energy levels are considered, as is the case of speech. The case is also discussed when only a mixture of useful signal and noise is available. Here it is no longer possible to use the classical definitions; it is necessary to identify noise in speech pauses. For this purpose, a suitable voice activity decoder (VAD) is necessary. The classical methods for assessing the quality of speech were mainly proposed and used to assess vocoders, in which it can be assumed that non-distorted speech signal is available. In the case of single-channel enhancement methods no uncorrupted speech signal is available and therefore the definition of SNR must be slightly adapted or new approaches must be sought.

1 Introduction

There are two large groups of methods for testing the quality, acceptability and intelligibility of speech [1]. The subjective methods include various types of listening tests. These tests have for long been performed in industries or in the army, in particular for speech coding, speech synthesis, and enhancement of noise-degraded speech [2, 3, and 4]. A group of listeners assess the respective recordings, using standard assessment forms or tables. They can assess, for example, intelligibility by means of rhyme tests, speech acceptability via assessing the signal quality, the quality of background, and overall quality. A disadvantage of these tests lies in their time demands, and much depends on the different hearing sensitivities of listeners (their hearing should be examined prior to the test, and their mother tongue should be the same as the language of tests), the tests are language-dependent, assessment conditions may differ, the statistical evaluation of tests is demanding, etc. Subjective tests can be used with advantage when comparing and assessing submitted prototypes that are to be used in practice.

There is also another group of tests, which employ objective methods. When reporting about new methods or enhancement of existing methods in renowned journals, numerous authors often point out that their method enhances the signal-to-noise ratio (SNR). They consider this parameter an objective assessment of the quality of the method that enhances the noise-degraded speech signal. Speech itself contains noise components (in particular in unvoiced sounds of speech), and it is difficult to distinguish whether this is undesirable noise that does not correlate with speech or noise that belongs to speech. There are a number of definitions of SNR [1, 5], which describe the signal as a whole or in partial sub-bands, and define SNR as a ratio of energies or as a ratio of the largest absolute values, etc. There are also recommendations of the International Telecommunication Union-Telecommunication

Standardization Sector (ITU-T) how to proceed in subjective and objective tests, and also other literature [6, 7, 8, and 9].

2 Definition of signal-to-noise ratio

2.1 Intrusive Approach

SNR is a frequently used parameter, for example, in the assessment of coding and compression systems. Assume we have at our disposal the digitized signal of a mixture of speech and noise without convolution distortion $y[n]$ and the original non-disturbed speech signal $s[n]$ and enhanced speech signal $\hat{s}[n]$, which results from processing the mixture of speech and noise by a suitable enhancement method [1, 10, 11]. The difference between the waveforms of the speech signal $s[n]$ and enhanced speech $\hat{s}[n]$ determines the error $\varepsilon[n]$:

$$\varepsilon[n] = s[n] - \hat{s}[n] \quad . \tag{1}$$

Now we determine the square of this difference and denote it as error energy:

$$E_\varepsilon = \sum_{n=-\infty}^{\infty} |\varepsilon[n]|^2 = \sum_{n=-\infty}^{\infty} |s[n] - \hat{s}[n]|^2 \quad . \tag{2}$$

The speech signal itself has the energy:

$$E_s = \sum_{n=-\infty}^{\infty} |s[n]|^2 \quad . \tag{3}$$

The resultant SNR in the decibel scale (dB) R is then equal to:

$$R = 10\log_{10}\frac{E_s}{E_\varepsilon} = 10\log_{10}\frac{\sum_{n=-\infty}^{\infty} |s[n]|^2}{\sum_{n=-\infty}^{\infty} |s[n] - \hat{s}[n]|^2} \quad . \tag{4}$$

This classical definition has its drawbacks in that if the undistorted signal $s[n]$ is equal to the enhanced signal $\hat{s}[n]$, then R is given by an indeterminate form and increases beyond all bounds. This sets the upper limit that SNR can reach. The lower limit is given by the condition $\hat{s}[n] = 0$ for all n, and then it holds $R = 0$ dB. It can roughly be said that according to (4) the range of R is $(0, \infty)$. The more similar the undistorted signal and the enhanced signal are, the greater the value of the signal-to-noise R ratio. Similar to the time domain, signals in the frequency domain can be compared. We know that for discrete signals the Parseval theorem holds [12]:

$$E_s = \sum_{n=-\infty}^{\infty} |s[n]|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(\omega)|^2 \mathrm{d}\omega \quad , \tag{5}$$

$$E_\varepsilon = \sum_{n=-\infty}^{\infty} |\varepsilon[n]|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(|S(\omega)| - |\hat{S}(\omega)|\right)^2 \mathrm{d}\omega \quad , \tag{6}$$

where $S(\omega)$ is the Fourier transform of discrete signal $s[n]$, i.e. it holds $S(\omega) \Leftrightarrow s[n]$ and $\hat{S}(\omega) \Leftrightarrow \hat{s}[n]$. The signal-to-noise ratio in the frequency domain is thus equal to:

$$R = 10\log_{10} \frac{\sum_{n=-\infty}^{\infty} |S(\omega)|^2}{\sum_{n=-\infty}^{\infty} \left(|S(\omega)| - |\hat{S}(\omega)| \right)^2} \quad . \tag{7}$$

These classical definitions (4) and (7) are not very satisfactory in practice. An infinitely long signal is never available-in practice, all signals are finite. In the frequency domain the range is limited, but the calculation cannot be performed continuously, and discrete calculation will limit the precision. Much always depends on what signal the SNR is calculated for. Speech is very time-variant and there are considerable magnitude differences between, for example, voiced and unvoiced frames. The two definitions, (4) and (7), are in practice not very suitable for speech. It will be more opportune to use a method that will average finite frames of speech signal; in this way the considerable energy difference between various speech frames can be suppressed a bit. We will select a suitable time frame where the speech signal parameters are approximately stationary (between 10ms and 30ms, depending on the type of speaker). Then the frame based SNR can be defined [1]:

$$R_{\text{seg}} = \frac{1}{M} \sum_{i=0}^{M-1} 10\log_{10} \left(\frac{\sum_{n=m_i-N+1}^{m_i} |s[n]|^2}{\sum_{n=m_i-N+1}^{m_i} |s[n] - \hat{s}[n]|^2} \right) \quad , \tag{8}$$

where we assume that we have M frames of length N, m_i, $i = 0,1,\cdots,M-1$. Thus the average value of SNR R_{seg} for the whole signal is obtained. A problem may arise when the speech includes pauses where there is no signal. Incorrect results can then be obtained. This can be prevented either by identifying the pauses and excluding them from processing or by setting the lower limit of SNR to, say, 0dB and substituting this limit for lower values. Another extreme can be seen in values exceeding 35dB, where listeners are no longer capable of perceiving any major magnitude differences. The two limits guarantee that the resultant signal will not be shifted in either direction.

In [13] and in other references the perception qualities of human hearing are considered in the calculation of SNR. The human ear introduces linear and non-linear distortion into the transfer and processing of audio information, and performs the masking. Critical frequency bands were therefore created in the form of band passes (Bark, Mel scale, etc. [14]) for the masking effects to play their role. The definition of SNR is then equal to:

$$R_{\text{freqseg}} = \frac{1}{M} \sum_{i=0}^{M-1} \frac{\sum_{k=1}^{K} w_{i,k} 10\log_{10} \frac{E_{s,k}(m_i)}{E_{\varepsilon,k}(m_i)}}{\sum_{k=1}^{K} w_{i,k}} \quad , \tag{9}$$

where M is the number of frames denoted $m_0, m_1, m_2, \cdots, m_{M-1}$. Each speech frame is divided in the spectrum into K frequency bands. The value is the energy in the k-th frequency band of the m_i-th frame of undisturbed speech, and $E_{\varepsilon,k}(m_j)$ is the energy in the k-th frequency band of the m_i-th frame of error signal $\varepsilon[n]$ according to (1). The weighting coefficients $w_{i,k}$ are used to set the importance of individual frequency bands of the given frame. It is known, for example, that human hearing is most sensitive in the range from 2 kHz to 4 kHz. The weight in these frequency bands will certainly be higher than in the neighbouring bands.

2.2 Non-Intrusive Approach

If recordings of pure speech (without noise) and recordings of noise alone are available, there is no problem in calculating the SNR: the classical definition from the preceding chapter will be used. A worse case will occur if we have a recording of speech with noise detected by one microphone. This case is typical of assessing the single-channel methods of speech enhancement [15]. Here the voice activity detector (VAD) must be used, which will estimate frames of speech activity and frames that do not contain speech (pauses), i.e. frames that only contain noise.

Assume we have at our disposal only the digitized signal of a mixture of speech and noise $y[n]$ without convolution distortion. Assume that all frames are of the same length of N samples, and the utterance is divided into M frames. The average energy of speech with noise will be determined according to the relation:

$$E_y[i] = \frac{1}{N} \sum_{n=iN}^{(i+1)N-1} |y[n]|^2 \quad , \quad i=0, 1, 2, \quad M-1 \; . \tag{10}$$

Assume further that the speech activity is monitored using the f_{VAD} parameter, for which it holds $f_{\mathrm{VAD}}=1$ when there is speech in the frame, and $f_{\mathrm{VAD}}=0$ when there is only noise in the frame. The average energy of frames that were detected by the VAD as mere noise, i.e. the average energy is speech pauses ($f_{\mathrm{VAD}}=0$), will be obtained using:

$$E_\eta = \frac{\sum_{i=0}^{M-1} \left(1 - f_{\mathrm{VAD}}[i]\right) E_y[i]}{\sum_{i=0}^{M-1} \left(1 - f_{\mathrm{VAD}}[i]\right)} \tag{11}$$

The average energy of the other frames of speech with noise is:

$$E_x = \frac{\sum_{i=0}^{M-1} f_{\mathrm{VAD}}[i] E_y[i]}{\sum_{i=0}^{M-1} f_{\mathrm{VAD}}[i]} \quad . \tag{12}$$

Assuming that the speech is degraded only by additive noise, the average energy of speech can be estimated approximately as:

$$E_s = E_x - E_\eta \quad . \tag{13}$$

The SNR of a mixture of speech and noise can then be estimated using the logarithm of the ratio of energy (13) to energy (11):

$$R_{\mathrm{VAD}} = 10 \log_{10} \frac{E_s}{E_\eta} \quad . \tag{14}$$

To assess enhanced speech, a procedure similar to that in the preceding case of additive noise can be used. Again, speech and pause frames are determined, and relation (11) is used to calculate the energy of residual noise. The energy of enhanced speech can be determined according to relation (12). The effectiveness of the enhancement method can then be estimated since the SNR of a mixture of speech and noise and the SNR of enhanced speech are available.

There are also methods for estimating noise in the spectral domain (e.g. [16]). In the spectrum, the frequency region with minimum magnitude is estimated. From each frequency vector of a given time frame these minimum values are summarized into one vector. From these values the average value is calculated, converted into energy, and regarded as the value E_η, which is then substituted into relation (14).

3 Conclusion

In the paper, different definitions of SNR are summarized which are used in practice to compare different methods of compression, coding and enhancement, and to test and compare prototypes. Much depends on what assessment the definition of SNR should be used for. In the case of compression and coding methods the initial undisturbed speech signal is available together with the signal that has been degraded one way or another by the compression algorithm or the coding procedure. This is logically reflected in the SNR definitions, which employ pure and distorted speech signals. In the case of speech recordings made in a noisy environment or degraded by transfer through the communication channel, signal contaminated by noise is only available. Classical definitions cannot be applied and distinction must be made between speech with noise and noise, the latter being identified in speech pauses. To identify noise, the voice activity detector is used, on which the precision of determining speech pauses and also the precision of calculating SNR depend. Besides assessing the signal quality by means of SNR there are other objective methods but they are often mathematically more complicated and more computation-demanding. SNR is therefore often used, which is simple to implement.

Another objective method for assessing speech enhanced by diverse methods can be seen in the application of the automatic speech recognition system for continuous speech with a large vocabulary [17].

Objective parameters serve for rough estimation of the applicability of the proposed algorithms. After selecting the optimum algorithm, subjective tests are used, but they require considerably more time and are more costly.

Acknowledgements

This work was supported within the framework of project No 102/07/1303 of the Grant Agency of the Czech Republic and the National Research Program of the Academy of Sciences of the Czech Republic "Information Society" No 1ET301710509.

References

[1] Deller, J.R., Jr., Hansen, J.H.L., Proakis, J.G.: *Discrete-Time Processing of Speech Signals*. The IEEE, Inc. New York, 1993.

[2] Voiers, W.D.: Diagnostic Evaluation of Speech Intelligibility. In: M.E. Hawley (ed.): *Speech Intelligibility and Speaker Recognition*. Stroudsburg, Pa.: Dowden, Hutchenson, and Ross, 1977 pp. 374-387.

[3] Flanagan, J.L.: Speech Coding. *IEEE Trans. on Communication Theory*, Vol. 27, April 1979, pp.710-736.

[4] Mahdi, A.E.: Voice Quality Measurement in Modern Telecommunication Networks. In: *CD Proceedings of the 6th EURASIP Conference Focused on Speech & Image Processing, Multimedia Communications & Services* (EC-SIPMCS), Maribor, Slovenja, June 27-30, 2007, pp.29-36.

[5] Deng, L., O'Shaughnessy, D.: *Speech Processing-A Dynamic and Optimization-*

Oriented Approach. Marcel Dekker, Inc., New York, USA, 2003.
[6] ITU-T Recommendation P.56, *Objective Measurement of Active Speech Level*. ITU-T, Geneva, 1993.
[7] ITU-T Recommendation P.800, *Methods for Subjective Determination of Speech Quality*. ITU-T, Geneva, 1996.
[8] ITU-T Recommendation P.835, *Subjective Test Methodology for Evaluating Speech Communication System that include Noise Suppression Algorithm*. ITU-T, Geneva, 2003.
[9] Jekosh, U.: *Voice and Speech Quality Perception-Assessment and Evaluation*. Springer-Verlag, Berlin, 2005.
[10] Smékal, Z., Sysel. P.: Single-Channel Noise Suppression by Wavelets in Spectral Domain. In: *Proceedings of the International Workshop COST 2102 "Verbal and Nonverbal Communication Behaviour"*, March 29-31, 2007, Vietri Sul Mare, Italy.
[11] Vondra, M. Vích, R.: Adaptive Comb Filtering in Speech Enhancement by Spectral Subtraction. In: *Proceedings of the 18th International Conference "Elektronische Signalverarbeitung ESSV 2007"*, September 10-12, 2007, Cottbus, Germany.
[12] Proakis, J.G., Manolakis, D.G.: *Digital Signal Processing-Principles, Algorithms, and Applications*. Third Edition. Prentice Hall, Upper Saddle River, New Jersey, 1996.
[13] Tribolet, J.M., Noll, P., McDermot, B.J., et al: A Study of Complexity and Quality of Speech Waveform Coders. In: *Proc. of the IEEE International Conf. on Acoustics, Speech, and Signal Processing*, Tulsa, Okla., 1978, pp.586-590.
[14] Zwicker, E., Fastl, H.: *Psychoacoustics, Facts and Models*. Springer Verlag, Heidelberg, Germany, 1998.
[15] Sysel, P.: *One-Channel Speech Enhancement Method based on Wavelet Transform in Spectral Domain*. PhD Thesis, Institute of Telecommunications, Brno University of Technology, Czech Republic, 2007. (In Czech)
[16] Martin, R.: Noise Power Spectral Density Estimation based on Optimal Smoothing and Minimum Statistics. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, July 2001, pp. 504-512.
[17] Nouza, J., Žďánský, J., Červa, P., Kolorenč, J.: Continual On-line Monitoring of Czech Spoken Broadcast Stream. In: *Proceedings of the International Conference on Spoken Language Processing-Interspeech 2006*, ICSLP 2006, September 2006, Pittsburgh, USA, pp. 1650-1653.