Information approach to signal-to-noise ratio estimation of the speech signal

Vasiliy Gai

Nizhny Novgorod State Technical Universtiy n.a. R.E. Alekseev, Nizhny Novgorod, Russian Federation
{vasiliy.gai@gmail.com}
http://www.nntu.nnov.ru/

Abstract. The article describes the method of signal-to-noise ratio estimation for speech signals. The proposed method is based on the theory of active perception. Within the scope of work assumes that the speech signal includes a desired signal (system formation) and noise. The conversions, which were described in the theory of active perception, allow allocating the desired signal and solving the problem of signal to noise ratio estimation. The work includes experimental data confirming workability of the proposed method.

Keywords: signal to noise ratio (SNR) estimation, speech signal, theory of active perception (TAP)

1 Introduction

Speech processing system (speaker identification, speech recognition), is working under noise conditions, must be possess stability to various distortion of input signal. Therefore, to adjust the signal processing algorithm by noise level, such systems must be possess the ability of level rating of signal's distortion (signal to noise ratio, SNR), and that kind of rating must be done only by a distorted signal.

Let us consider identity of existent methods of signal to noise ratio estimation:

1. by SNR estimation analyzable signal is divided into segments length by 20-30 ms., the overlap between segments is 50 percent, then the spectrum of each segment is calculated and by the spectra are done noise estimation [1];
2. methods of SNR estimation are developed with a glance that analyzable signal contain speech and pauses [2].

The analysis of works permits to sort out the following classes of methods of SNR estimation:

1. methods based on voice activity determination are to signal segmentation active speech and pauses, to signal and noise power estimate and to computing SNR [2];

2. methods are using controlled recursive averaging [3], in such methods estimation of noise is performed by averaging the previous values of the spectral power using a smoothing parameter, which depends on probability of the presence of the signal in different frequency band: shows that the presence of speech in some segment in a certain frequency band may be determined by relation to local energy of noisy speech to its minimum in that segment. If the value of this ratio is less than the threshold, we can conclude that no speech signal in segment;
3. methods of noise assessment based on the minimum statistic, methods which pertaining to this class are based on two assumptions [4]. First consists in independence of noise and speech, the second - the power spectrum of the noisy speech signal is similar to the noise power spectrum. Therefore, noise dispersion estimate is to calculate the minimum of the spectral concentration of noisy speech signal in fixed length segment. Disadvantages of the method are necessary to select the length of the segment. In this case the wrong choice can considerably affect on the assessment's result. Another disadvantage of the method is entering of a delay in the noise estimation parameters, as the length of the segment (1.6 - 2.8 s.) is chosen that to ensure include part of speech and pauses;
4. methods of noise parameters assessment based on the histogram is used observation that the most frequently occurring value of the energy in some frequency band corresponds to the noise level in this frequency band, i.e. the noise level corresponds to the maximum energy histogram [5].

In this work the solution of SNR estimation based on using a systematic approach to signal processing described in the theory of active perception [6].

Let us consider the basic propositions of the theory of active perception. From the point of view of an observer, sound signal contains a desired signal (information message) and a hindrance. Desired signal is the information, which is required for the observer to make decisions under the task, and the noise - all the other information. In this work, desired signal is considered as a system formation. In this case, it must contain the structural elements and connections.

Theory of Active Perception (TAP) contains description operations that allow to allocate the structural elements of the signal and links between them. For detection system elements in TAP is used integral conversion, and to identify links between elements - differentiation. The result of the identification of the differential structure is the signal's spectral description.

Conversion integration and differentiation together form a composition which is called U-transform: $U = d \circ \int$.

Transformations of integration and differentiation for one-dimensional signals realized with the help of four-base dimensional filter-coatings (F_0, F_1, F_2, F_3, see fig.1).

Let $f(t)$ – analyzable sound signal, observed on a finite time interval. Result of applying the U-transform to the signal f – multilevel (roughly exact) spectral representation $D = d_{ij}$, $i = \overline{1, K}$, $j = \overline{1, M_i}$, where K – is number of dissection level, M_i – number of signal's segments on the i-th dissection level,

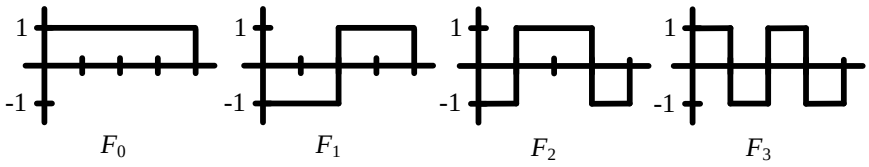

Fig. 1. Basis functions

d_{ij} – spectrum, which is included N spectrum factor (number of using filters), $d_{ij}\{k\}$ – k-th spectrum factor ($k = \overline{1, L}$), f_{ij} – signal's segment f, by which is calculated the spectrum d_{ij} (see fig.2). In calculating the spectral representation of the signal segments are not overlapped. Example U-transform computation is given in [7].

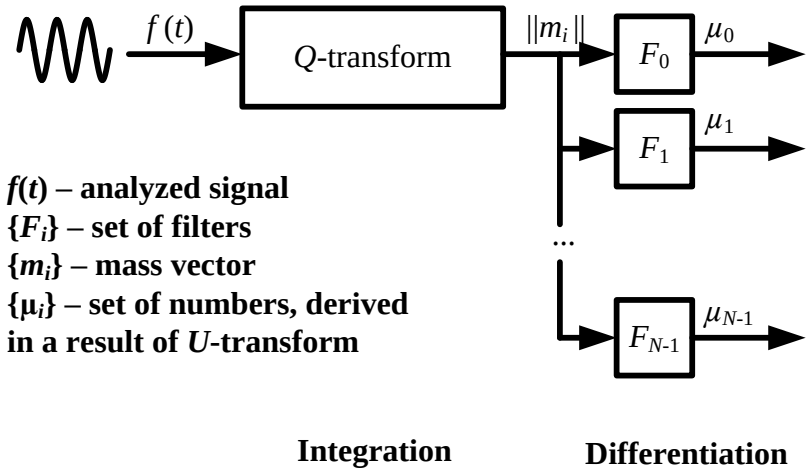

Fig. 2. U-transform circuit

Considered methods of SNR estimation based on use of the Fourier transform. Comparing the Fourier transform with U-transform (major transformation of TAP) can be noted following [6]:

1. Fourier coefficients, except for the certain their lack – complexity, there are integral characteristics that do not contain information about the structural properties of the signal;
2. filters are used in the U-transform, which is endowed with differentiating properties which allows to highlight structural elements of the signal and links between them.

1.1 Signal to Noise Ratio Estimation on the Basis of the Theory of Active Perception

Conducted researches have established properties of the spectra (within the U-transform) relating to the desired signal and pause. The desired signal can be represented as a set of voiced and unvoiced segments (see fig.3):

1. elements of signal spectrum, which is related to pause, close to zero values (section 1).
2. signal segment spectrum relating to voiced sound contains elements diverged considerably from each other by magnitude, but differ in a lesser degree than for a voiced segment (section 2);
3. signal segment spectrum relating to voiced sound contains elements diverged considerably from each other by magnitude (section 3).

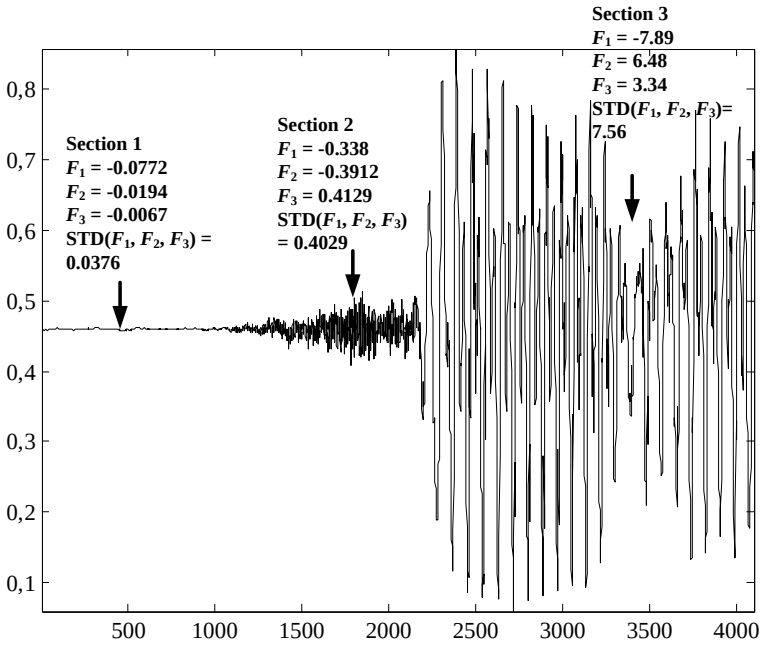

Fig. 3. Properties of Desired Signal And Pause

Let us consider the table (see tab. 1) in which gives values of the spectral coefficients for different sections of the signal and the standard deviation (STD). It may be noted that the result of exposure of noise on the signal:

1. on the set of spectra factors (under certain noise level) it becomes impossible to distinguish the unvoiced sections of the speech signal from the pause

sections (this is also corresponds to the subjective perception of the person of noisy speech signal);
2. values of spectra factors relating to voiced segments decrease with increasing noise.

Table 1. Influence of Noise on Speech

	Uniform Noise (SNR = 8 dB)			Normal Noise (SNR = 4 dB)		
Signal Section	1	2	3	1	2	3
Experiment N. 1 (segment size = 8 ms)						
STD(F_1, F_2, F_3)	0.4747	1.3124	13.0893	1.0637	1.2250	7.0924
F_1	0.6516	-1.1670	-3.3792	0.8600	0.5248	1.5579
F_2	0.8198	-0.0529	17.7059	1.6997	-1.9050	10.2402
F_3	-0.0735	1.4483	-6.2735	-0.4129	-0.4175	-3.8153
Experiment N. 2 (segment size = 8 ms)						
STD(F_1, F_2, F_3)	0.7981	0.5843	12.7010	1.3569	1.3808	5.7403
F_1	-0.7303	1.0742	-4.3398	0.7722	3.3808	0.2616
F_2	0.1213	0.9121	15.6942	-1.9404	1.2583	7.9577
F_3	-1.4737	1.9954	-7.8480	-0.5106	0.7893	-3.2680

It is known that an increase in the noise level rhythmic pattern of word or phrase is a parameter which is destroyed in the last turn [10]. Therefore, when SNR estimation we can use the signal segments for which the standard deviation of spectra factors max (desired signal) and the lowest (noise).

Let a finite time interval $[0; T]$ is taken sound signal $f(t)$, which is a function of speech signal $s(t, \lambda)$ and signal-independent noise $n(t)$:

$$f(t) = F(s(t, \lambda), n(t)), 0 \leq t \leq T, \tag{1}$$

where $\lambda = \lambda_1, \ldots, \lambda_m$ is vector of parameters of speech signal. It is assumed that direct observation is only the received signal $f(t)$ available. Let $s_a(t, \lambda)$ is active speech without pauses, then signal to noise ratio estimation (ξ) can be written as follows:

$$\xi = \frac{E\{s_a^2(t, \lambda)\}}{E\{n^2(t)\}}, \tag{2}$$

where E – calculating expectation's operator.

The proposed method of SNR estimation (at the i-th level of decomposition) includes the following operations:

1. the definition of the standard deviation (SD) of spectra factors relating to the signal: $\hat{\sigma} = \max[(\sigma(\{d_{ij}(t)\}))]$, $j = \overline{1, 2^{i-1}}$, $t = \overline{1, N}$;
2. definition STD of spectra factors relating to noise: $\hat{\sigma_n} = \min[(\sigma(\{d_{ij}(t)\}))]$, $j = \overline{1, 2^{i-1}}$, $t = \overline{1, N}$;
3. calculation of SNR estimation $\hat{\xi}$: $\hat{\xi} = 10 \log_{10} \frac{(\hat{\sigma_s} - \hat{\sigma_n})^2}{\hat{\sigma_n^2}}$.

The proposed method has two parameters:

1. signal segment length (in milliseconds) to compute the spectrum;
2. number of filters which is used in the calculation of the spectrum.

2 Computing Experiment

Let us consider results of SNR estimation based on proposed and existing methods. When testing was used database of records votes 100 speakers (audio sampling frequency – 16 kHz, depth coding – 8 bits). Computing experiment involved deformation records and signal to noise ratio estimation.

Table 2 - Table 3 shows the results of estimation of accuracy of calculation the SNR for existing algorithms.

Table 2. Results of SNR Estimation Method Proposed in [8]

Noise / SNR (in dB)	-2	4	10	16	22	28	34	40	Estimation error
Noise 1	-1.7	2.9	8.9	15.6	22.5	29.4	36.6	38.7	1.08
Noise 2	-1.35	3.11	9.11	15.9	22.9	30.2	37.3	38.6	1.29
Noise 3	-1.16	3.52	9.86	16.5	23.5	30.5	37.6	39.0	1.32
Noise 4	0.03	5.26	11.87	18.8	26.0	33.1	38.4	39.7	2.72

Table 3. Results of SNR Estimation Techniques

SNR (in dB) / Source	-10	-5	0	5	10	15	20	25	30	35	Error
[9], stationary noise	-	-6.1	-3.4	1.9	7.92	13.1	18.1	23.5	29.2	32.2	2.06
[10], white noise	-8.6	-3.7	1.2	6.01	11.02	16.02	21.57	26.85	32.96	-	1.48
[10], white noise	-0.2	-0.95	1.9	3.5	6.6	10	14.1	17.9	21.9	-	5.19
[2], stationary noise	-2	1.3	0.1	4.8	8.13	-	-	-	-	-	1.73

Table. 4 - Table. 7 shows the average test results of the proposed method (for two types of voices: male and female).

Table. 8 shows a calculation error signal / noise ratio under varying conditions. The table shows that the smallest error is achieved using 64 filters, with the duration of the analyzed signal has low effect on the accuracy of SNR estimation.

Analyzing the given table can be noted that the proposed method by accuracy of SNR estimation is not inferior to the existing methods, and in some cases - shows the best results. The advantages of the proposed method is also easy to implement and a wide range of SNR estimation: from -5 to +35 dB.

Table 4. Results of Accuracy of Method (Women's Voices, 8 Seconds)

	Algorithm Parameters / SNR (in dB)	-5	0	5	10	15	20	25	30	35
Uniform Noise	8 ms., 128 filters	-7.11	-2.85	4.01	10.15	15.93	21.07	26.30	31.50	36.11
Normal Noise		-5.82	-6.68	-1.55	4.52	10.07	15.57	20.55	25.49	31.03
Uniform Noise	4 ms., 64 filters	-2.87	0.20	5.29	11.28	17.03	22.24	27.22	32.76	37.65
Normal Noise		-1.68	0.38	2.49	6.76	11.12	16.68	21.93	26.62	31.97
Uniform Noise	2 ms., 32 filters	0.01	2.83	8.26	13.95	19.32	24.63	29.06	34.18	39.35
Normal Noise		5.03	5.79	8.48	9.36	14.00	20.24	24.41	29.99	34.93

Table 5. Results of Accuracy of Method, the Men's Voices, 4 Seconds

	Algorithm Parameters / SNR (in dB)	-5	0	5	10	15	20	25	30	35
Uniform Noise	8 ms., 128 filters	-9.59	-8.54	-2.07	4.17	10.47	16.31	21.72	27.17	31.98
Normal Noise		-7.36	-4.90	-6.13	-1.15	3.92	10.32	15.38	21.10	26.28
Uniform Noise	4 ms., 64 filters	-5.81	-3.20	0.28	5.37	12.23	18.22	22.89	28.89	33.41
Normal Noise		-1.87	-1.69	-0.26	2.04	7.11	11.69	17.34	22.39	28.36
Uniform Noise	2 ms., 32 filters	-0.45	0.48	3.60	9.68	14.86	20.34	25.51	30.47	35.80
Normal Noise		4.19	5.02	3.25	6.03	10.42	15.28	20.39	26.17	31.25

Table 6. Results of Accuracy of Method, Women's Voices, 8 Seconds

	Algorithm Parameters / SNR (in dB)	-5	0	5	10	15	20	25	30	35
Uniform Noise	8 ms., 128 filters	-7.34	-3.77	2.95	8.53	13.85	20.01	25.28	30.73	36.00
Normal Noise		-5.54	-6.38	-1.95	2.46	8.88	14.60	19.70	24.70	30.76
Uniform Noise	4 ms., 64 filters	-3.95	-0.42	5.18	10.77	15.30	21.21	26.58	31.31	37.02
Normal Noise		-0.34	-1.43	0.86	5.20	10.46	15.66	21.09	26.02	31.71
Uniform Noise	2 ms., 32 filters	1.41	4.86	9.04	15.37	19.14	24.40	31.39	34.15	39.96
Normal Noise		4.41	3.28	5.52	8.56	14.89	20.09	25.19	30.35	35.41

Table 7. Results of Accuracy of Method, Women's Voices, 4 Seconds

	Algorithm Parameters / SNR (in dB)	-5	0	5	10	15	20	25	30	35
Uniform Noise	8 ms., 128 filters	-8.35	-4.27	2.71	8.60	14.35	20.13	25.60	30.40	36.15
Normal Noise		-6.99	-6.57	-3.42	1.97	8.26	13.79	19.83	24.78	30.45
Uniform Noise	4 ms., 64 filters	-5.10	0.50	3.91	10.02	15.12	21.66	26.74	31.19	37.63
Normal Noise		-3.92	-1.11	0.01	4.79	9.88	15.81	20.96	27.10	31.46
Uniform Noise	2 ms., 32 filters	0.47	5.24	8.16	14.55	18.07	24.73	29.56	34.52	40.17
Normal Noise		3.39	5.00	6.01	8.75	13.99	18.99	24.98	29.97	35.75

Table 8. Calculation Error SNR (in dB)

	Men			Women		
Noise Type / filter amount	128	64	32	128	64	32
Uniform noise, 4 sec.	4.82	2.52	1.00	1.58	1.01	4.50
Normal noise, 4 sec.	8.62	6.24	4.60	5.88	3.58	2.05
Uniform noise, 8 sec.	1.33	1.76	4.07	1.42	0.98	4.97
Normal noise, 8 sec.	4.65	2.90	2.43	5.31	3.90	1.76

3 Conclusion

Method of signal/noise ratio estimation in the observed speech signal was devised. The proposed method is based on the theory of active perception. Using this theory reveals the structural elements of the signal and links between them, and thus to detect desired signal and hindrance.

Research of the algorithm on testing and real signals confirmed its efficiency and ability to use in speech signal processing, which requires adjustment to the quality of the analyzed signal.

References

1. Rangachari, S.: A noise-estimation algorithm for highly non-stationary environments. J. Speech Communication. 48, 220–231 (2006)
2. Vondrasek, M., Pollak, P.: Methods for Speech SNR Estimation: Evaluation Tool and Analysis of VAD Dependency. J. Radioengineering. 14, 6–11 (2005)
3. Cohen, I.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. J. IEEE Signal Processing Letters. 9, 12–15 (2002)
4. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. J. IEEE Transactions on Speech and Audio Processing, 9. 504–512 (2001)
5. Hirsch, H.-G., Ehrlicher, C.: Noise estimation techniques for robust speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 153–156 (1995)
6. Utrobin, V. A.: Physical interpretation of the elements of image algebra. J. Advances in Physical Sciences, 47, 1017–1032 (2004)
7. Gai, V. E.: Metod ocenki chastoty osnovnogo tona v uslovijah pomeh, 4, 65–71, (2013) (in Russian)
8. Stolbov, M.B.: Algoritm ocenki otnoshenija signal/shum rechevyh signalov. J. Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki, 82, 67–72 (2012)
9. Gerkmann, T.: Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. J. IEEE Transactions on Audio, Speech, and Language Processing. 20, 1383–1393, (2012)
10. Kim, C., Stern, R.M.: Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis. In: InterSpeech 2008, 2598–2601, Brisbane, Australia (2008)