Reference-free SNR Measurement for Stationary Noises

Balázs Fodor, Tim Fingscheidt

Institut für Nachrichtentechnik, Technische Universität Braunschweig, Schleinitzstr. 22, 38106 Braunschweig, Deutschland

Email: {fodor,fingscheidt}@ifn.ing.tu-bs.de

Introduction

Signal-to-noise (SNR) measurement of noisy speech signals is an important topic in automotive environments, e. g., for investigating optimal microphone positions. To simulate a reference SNR condition, the desired SNR is adjusted by scaling the clean speech and noise signals separately according to the ITU-T Recommendation P.56 [1], followed by the addition of both signals. For this, the clean speech signal and the noise signal are measured by means of the active speech level and the root mean square (RMS) noise level, respectively.

The aim of this contribution is to measure the SNR of a noisy speech signal distorted by car noise as close as possible to the *reference* SNR from the setup above, however, without using any reference signals. This "reference-free" nature allows for a wide flexibility; the proposed method just has to be applied to the noisy speech signal. Moreover, it offers low complexity.

To measure the SNR of the noisy input signal in our approach, first the speech and noise power have to be estimated. The speech power is estimated by means of a noise tracking algorithm, and a voice activity detection (VAD) algorithm, the noise power is estimated by means of a speech pause detection (SPD) algorithm. Then, the measured SNR is the ratio of the speech and noise power estimates. The resulting raw SNR values have to be corrected by a mapping curve, in order to obtain unbiased measurements.

This contribution is organized as follows: After the introduction of the SNR measurement method, the applied VAD/SPD is described. This is followed by the evaluation of the proposal. The paper ends then with some concluding remarks.

Signal-to-Noise Ratio Measurement

The input signal $y(n)$ of the measurement system is assumed to consist of the clean speech signal $s(n)$ and an additive noise signal $n(n)$, with n being the discrete time index. In the following we assume a sampling rate of 16 kHz. After segmentation, windowing, and the discrete Fourier transform (DFT), the input signal can be rewritten as $Y(\ell,k) = S(\ell,k) + N(\ell,k)$, with ℓ being the analysis frame index, k being the frequency bin index. Then, the SNR is defined as

$$\mathrm{SNR} = \frac{\frac{1}{LK}\sum_{\ell}\sum_{k}|S(\ell,k)|^2}{\frac{1}{LK}\sum_{\ell}\sum_{k}|N(\ell,k)|^2} = \frac{P_S}{P_N}, \tag{1}$$

with L, K, and P_S, P_N being the the number of frames and frequency bins, as well as the speech and the noise power, respectively. P_S in (1) is estimated as $\widehat{P_S}(\ell,k) = \max\{P_Y(\ell,k) - \widehat{\sigma_N^2}(\ell,k), 0\}$, with $P_Y(\ell,k) = |Y(\ell,k)|^2$, and $\widehat{\sigma_N^2}(\ell,k)$ being a noise variance estimate which is determined by a noise variance tracking algorithm proposed in [2]. Since the tracking of noise variance during speech activity can lead to estimation errors, the estimation of P_S in (1) is only carried out in speech active frames as

$$\widehat{P_S} = \frac{1}{|\Lambda_1|\cdot K}\sum_{\ell\in\Lambda_1}\sum_{k}\max\left\{P_Y(\ell,k) - \widehat{\sigma_N^2}(\ell,k), 0\right\}, \tag{2}$$

with Λ_1 and $|\Lambda_1|$ being the set of active speech frames detected by a voice activity detection (VAD) described in the next section, and the number of its elements, respectively. For the estimation of P_N in (1), $P_Y(\ell,k)$ is averaged over all frequency bins k and those frames which belong to speech pause $\ell \in \Lambda_0$, with Λ_0 being the set of these frames as detected by a *separate* speech pause detection (SPD) described in the next section.

Voice Activity / Speech Pause Detection

This section describes the algorithms detecting frames with speech activity Λ_1 and speech pause Λ_0. Please note that for both the VAD and the SPD the same algorithm with different parameters is employed. The VAD/SPD is based on the frame energy of the noisy input signal spectrum calculated by means of the smoothed periodogram $\overline{|Y(\ell,k)|^2} = 0.5\cdot\overline{|Y(\ell-1,k)|^2} + 0.5\cdot|Y(\ell,k)|^2$ as follows

$$\overline{P_Y}(\ell) = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\overline{|Y(\ell,k)|^2}, \tag{3}$$

with $\mathcal{K}$ and $|\mathcal{K}|$ being a set of frequency bins containing relevant speech information covering frequencies between 500 Hz and 2500 Hz and the number of its elements, respectively. Based on an adaptive threshold $\Xi(\ell)$, the VAD/SPD schemes deliver hypotheses $H(\ell)$ being $H^{\mathrm{VAD}}(\ell)$ or $H^{\mathrm{SPD}}(\ell)$ based on the following 3 states:

H_{SP} : Speech *presence* is assumed if $\overline{P_Y}(\ell) > \Xi(\ell)$.
H_{ST} : Speech *transition* follows every H_{SP} decision with a duration of L_{tr} frames, unless $\overline{P_Y}(\ell) > \Xi(\ell)$.
H_{SA} : Speech *absence* is assumed in all other situations.

Frame ℓ is detected as *voice active* and becomes element of set Λ_1, if $H^{\mathrm{VAD}}(\ell) \in \{H_{\mathrm{SP}}, H_{\mathrm{ST}}\}$. Meanwhile, frame ℓ is considered as *speech absent* and becomes element of Λ_0, if $H^{\mathrm{SPD}}(\ell) = H_{\mathrm{SA}}$. The adaptive threshold $\Xi(\ell)$ is calculated as $\Xi(\ell) = \mu\cdot\Phi(\ell) + \alpha$, with the multiplicative factor μ and the additive term α. The VAD/SPD floor signal $\Phi(\ell)$ is updated depending on the VAD/SPD control parameter value of the last frame $\Upsilon(\ell-1)$ if this frame as follows

$$\Phi(\ell) = \begin{cases}\varepsilon(\ell)\cdot\Phi(\ell+1) + [1-\varepsilon(\ell)]\cdot\overline{P_Y}(\ell), & \text{for } \mathcal{A},\\ \Phi(\ell+1), & \text{else,}\end{cases} \tag{4}$$

with $\mathcal{A} = \left\{\overline{P_Y}(\ell-1) \le \nu\cdot\Upsilon(\ell-1) \wedge H(\ell-1)\in\{H_{\mathrm{SP}}, H_{\mathrm{SA}}\}\right\}$ with a threshold factor ν. The smoothing parameter $\varepsilon(\ell)$ is defined as

$$\varepsilon(\ell) = \begin{cases}\varepsilon_{\mathrm{up}}, & \text{if } \overline{P_Y}(\ell-1) > \Phi(\ell-1),\\ \varepsilon_{\mathrm{dn}}, & \text{if } \overline{P_Y}(\ell-1) \le \Phi(\ell-1).\end{cases} \tag{5}$$

Selecting $\varepsilon_{\mathrm{dn}} < \varepsilon_{\mathrm{up}}$ ensures that the floor signal is mostly updated during speech pause. The VAD/SPD control parameter $\Upsilon(\ell)$ is increased slightly under the hypothesis of speech presence and updated strongly under the hypothesis of speech absence as

$$\Upsilon(\ell) = \begin{cases}\delta\cdot\Upsilon(\ell-1), & \text{if } \overline{P_Y}(\ell-1) > \nu\cdot\Upsilon(\ell-1),\\ \overline{P_Y}(\ell), & \text{if } \overline{P_Y}(\ell-1) < \Upsilon(\ell-1),\\ \Upsilon(\ell-1), & \text{else,}\end{cases} \tag{6}$$

with the control update constant δ. We initiated the VAD/SPU control parameter as $\Upsilon(\ell=0) \to \infty$. The parameters of the VAD and the SPD are summarized in Table 2.

j	0	1	2	3	4	5
p_j	-7.25	4.394	-0.9265	0.1444	$-0.1133 \cdot 10^{-1}$	$0.3449 \cdot 10^{-3}$
q_j	9.277	-1.137	0.1686	$-0.5575 \cdot 10^{-2}$	$0.6753 \cdot 10^{-4}$	0

Table 1: Polynomial coefficients p_j and q_j of the mapping curve

Parameter	L_{tr}	μ	α	$\varepsilon_{\mathrm{up}}$	$\varepsilon_{\mathrm{dn}}$	δ	ν
VAD	5	3	$3 \cdot 10^8$	0.875	0.5	1.025	2
SPD	9		10^8			1.015	

Table 2: Parameters of the VAD and the SPD

Absolute estimation error	Relative frequency $T_m = 8$ s	$T_m = 80$ s	$T_m = 480$ s
≤ 2 dB	90.0 %	99.2 %	100 %
≤ 1.5 dB	84.4 %	97.0 %	99.8 %
≤ 1 dB	74.5 %	92.5 %	97.4 %
≤ 0.5 dB	47.6 %	76.9 %	85.4 %
≤ 0.25 dB	26.0 %	49.3 %	61.2 %

Table 3: Relative frequency of absolute measurement error for different measurement durations T_m

Evaluation

In order to evaluate the proposed method, we performed the following simulations on a *training* data set: 720 speech files (48 speech files in 15 different languages) were taken from the NTT Multi-Lingual Speech Database for Telephonometry [3], each with a length of 8 s. Car noise signals with the same length were randomly taken from the ETSI database [4]. Both the speech and noise signals were filtered according to the ITU-T P.341 filter mask [1]. A number of 51 reference input SNR values from -15 dB until 35 dB were employed in 1 dB steps; the desired SNRs were adjusted according to ITU-T Recommendation P.56 [1]. After the superposition of both the speech and noise signals, a frame-wise processing was carried out as follows: At a sampling frequency of 16 kHz, the segmentation was done by a Hann window, the analysis frame length was $L = 512$ samples, the analysis frame shift took 50 %. Within the frame-based processing, $P_Y(\ell, k)$, and the noise variance estimate $\widehat{\sigma_N^2}(\ell, k)$ were computed, as well as the VAD and the SPD were employed. Since these algorithms need some time to adapt, the first 15 frames were dropped. Then, the *raw* SNR value for the ith speech file was calculated as

$$\widehat{\mathrm{SNR}}_{\mathrm{raw}}(i) = 10 \log \frac{\widehat{P_S}(i)}{\widehat{P_N}(i)}. \tag{7}$$

This whole process was repeated for all speech files, resulting in a total of 720 raw SNR estimates per reference input SNR. Moreover, since the noise variance tracking, the VAD, and the SPD are quite challenging at low SNRs, a non-linear relationship between the reference SNR and the measured raw SNR values was observed. Therefore, the raw SNR measures have to be corrected by a mapping curve, in order to obtain unbiased results. The mapping curve is basically the inverse of the mean of the measured raw SNRs which we decided to approximate by a polynomial fit [5]. The mapping function $\widehat{\mathrm{SNR}} = f(\widehat{\mathrm{SNR}}_{\mathrm{raw}})$ is thus defined as

$$\widehat{\mathrm{SNR}} = \begin{cases} -11\mathrm{dB}, & \widehat{\mathrm{SNR}}_{\mathrm{raw}} < -0.7\mathrm{dB}, \\ \sum_{j=0}^{5} p_j \cdot \widehat{\mathrm{SNR}}_{\mathrm{raw}}^{\,j}, & -0.7\mathrm{dB} \leq \widehat{\mathrm{SNR}}_{\mathrm{raw}} < 11\mathrm{dB}, \\ \sum_{j=0}^{4} q_j \cdot \widehat{\mathrm{SNR}}_{\mathrm{raw}}^{\,j}, & \widehat{\mathrm{SNR}}_{\mathrm{raw}} \geq 11\mathrm{dB}, \end{cases} \tag{8}$$

with the coefficients p_j and q_j as shown in Table 1.

In order to evaluate our mapped SNR estimator, we performed the simulation steps described above for SNR values from -10 dB until 30 dB in 1 dB steps on our *test* speech and noise data set taken from the same databases and being of equal size as the training data set. The resulting $\widehat{\mathrm{SNR}}_{\mathrm{raw}}(i)$ value in (7) for the ith speech file was corrected by the mapping function (8). This whole process was repeated for all speech files; the bias could significantly be reduced due to the mapping.

Correlation coefficient ρ		
$T_m = 8$ s	$T_m = 80$ s	$T_m = 480$ s
0.9908	0.9977	0.9991

Table 4: Correlation coefficient between the measured SNRs and the reference SNR values for different measurement durations T_m

We evaluated the performance of the proposal w. r. t. the absolute measurement error which is shown in Table 3 for the corrected measurements. The values based on 8 s signals may be too high for some applications. Applying longer speech sequences, however, e. g., by averaging estimation results in groups, the absolute error can be reduced. This can be observed in Table 3 for groups of 10 files of 8 s each and 60 files of 8 s each, reflecting a measurement duration of $T_m = 80$ s and $T_m = 480$ s, respectively. As a conclusion, we recommend averaging SNR measurements based on groups of 8 s signals in order to reduce the absolute measurement error. In addition, the performance of the proposal was evaluated by calculating a correlation coefficient between the SNR measurement employing ITU-T P.56 reference levels and the proposal as [6]

$$\rho = \frac{\sum_i \left(\widehat{\mathrm{SNR}}(i) - \overline{\widehat{\mathrm{SNR}}}\right)\left(\mathrm{SNR}_{\mathrm{ref}}(i) - \overline{\mathrm{SNR}_{\mathrm{ref}}}\right)}{\sqrt{\sum_i \left(\widehat{\mathrm{SNR}}(i) - \overline{\widehat{\mathrm{SNR}}}\right)^2 \cdot \sum_i \left(\mathrm{SNR}_{\mathrm{ref}}(i) - \overline{\mathrm{SNR}_{\mathrm{ref}}}\right)^2}}, \tag{9}$$

for all speech signals and reference SNR values with $i \in [1\ 51 \times 720]$, $\mathrm{SNR}_{\mathrm{ref}}$ and $\overline{(\cdot)}$ being the speech file index, the reference SNR value from P.56 reference-based measurement, and the mean operator, respectively. The correlation coefficient for different measurement durations T_m (again, simulated by assembling groups of the 8 s database signals) can be seen in Table 4. The proposed method achieves a correlation coefficient larger than 0.99 in all cases.

Conclusions

This paper presents a reference-free SNR measurement method that requires only the noisy speech signal as input. It was thoroughly tested and evaluated for car noises and compared to ground truth by reference-based measurement of speech and noise separately using P.56. The resulting SNR estimates achieve a correlation coefficient of greater than 0.99 compared to the reference. Moreover, for a measurement length of 80 s, the absolute measurement error was below 1 dB in more than 92% of the cases.

References

[1] ITU, "ITU-T Recommendation G.191, Software Tools for Speech and Audio Coding Standardization," Mar. 2010.

[2] Fodor, B.; Scheler, D.; Suhadi, S.; Fingscheidt, T., "Talk-And-Push (TAP) - Towards More Natural Speech Dialog Initiation," in *36th AES Int. Conference*, Dearborn, MI, USA, June 2009.

[3] NTT, "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1994.

[4] ETSI EG 202 396-1, "Speech Processing, Transmission and Quality Aspects (STQ), Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation technique and Background Noise Database," 2008.

[5] Mathews, J. H.; Fink K. K., *Numerical Methods Using Matlab*, Prentice-Hall, Upper Saddle River, NJ, USA, 4th edition, 2004.

[6] Burington, R.S.; May, D.C., *Handbook of Probability and Statistics with Tables*, McGraw-Hill, New York, USA, 2nd edition, 1970.