RESEARCH ARTICLE OPEN ACCESS

The Short-Time Silence of Speech Signal as Signal-To-Noise Ratio Estimator

Azhar S. Abdulaziz[1, 2], Veton Z. Këpuska[3]
[1]*Department of Electrical and Computer Engineering, Florida Institute of Technology, Florida, USA*
[2]*Department of Computer and Information, College of Electronic Engineering, Ninevah University, Mosul, Iraq.*
[3]*Department of Electrical and Computer Engineering, Florida Institute of Technology, Florida, USA*

ABSTRACT
It is proposed in this paper to use a small portion of the audio speech signal to estimate Signal-to-Noise Ratio (SNR). It is found that, the first 30 ms duration has enough information about the SNR in advance. The first 30 ms of a recorded speech usually comes from the silence rather than speech. This is because the speaker usually starts the recording process or wait for it before he/she can deliver the utterance. For testing and comparing the proposed estimator, different noisy corpora are built upon the TIMIT data. The average estimation of the suggested algorithm proves to get better results as compared to the Waveform Amplitude Distribution Analysis (WADA) and the National Institute of Standard and Technology (NIST) SNR estimators. The complexity of the STS-SNR estimator is less than both as it only processes a small portion of the audio samples.
Keywords: SNR Estimation, Noisy Speech, Signal-to-Noise Ratio, Short-Time Silence.

I. INTRODUCTION

The signal-to-noise ratio (SNR) estimation algorithms has been investigated deeply and used for different applications. They could be used to improve speech enhancement, detection and recognition algorithms in different ways [1]. Those estimators usually use all signal samples to compute the SNR. For some applications, it isenough to know if the signal has high or low SNR in order to do further operations on the speech signal for better recognition accuracy as in [2], [3] and [4]. Many of the SNR estimation approaches are based on either a pre-specified weighting factor or preceding assumptions of some parameters in the signal model [5]. The spoken utterance SNR is a ratio between the power of two random signals, the speech and the noise. This phenomenon of variability and randomness behavior makes the SNR estimation more difficult to investigate [6].

The National Institute of Standard and Technology NIST developed a well-known SNR estimator NIST-STNR [7]. It uses the entire speech signal by framing it to 20 ms frames to estimate the histogram of the root-mean square (RMS) of the power. Those frames are 50% overlapped among each other, then used for histogram updates. When the audio is finished, the resulting power histogram will be analyzed so that the 15% of the total area from the left represents the noise while the 85% of the area from the left of the histogram represents the signal [8].

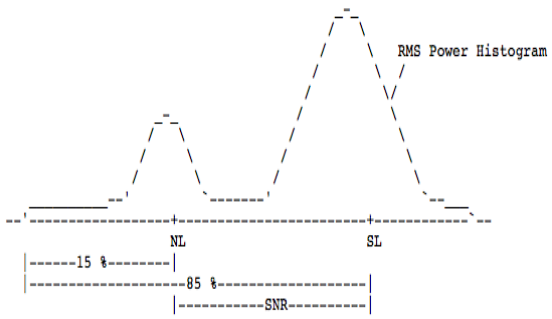


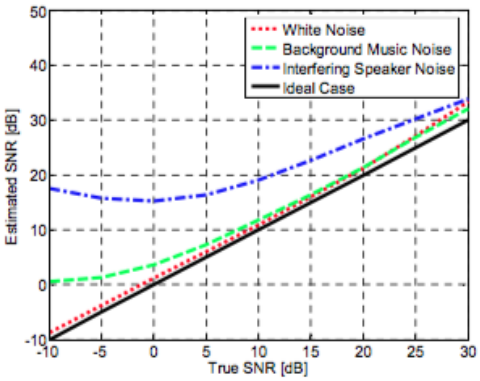
Figure 1: NIST-STNR Estimation using RMS power Histogram [8]

The NIST algorithm tries to separate the high and the low power of the audio signal using the its probability density function (PDF). It is supposed that, the noise and signal PDF boundaries lies in the selected regions with higher probability.

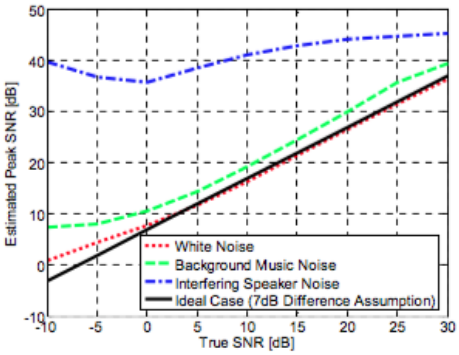
While the power PDF is the key for NIST SNR estimator, there is another approach of examining the amplitude PDF as in the Waveform Amplitude Distribution Analysis (WADA-SNR) algorithm in [1]. This approach is based on the assumption that the clean speech amplitude has proximately a symmetrical Gamma distribution function. According to Kim and Stern in [1], when a white Gaussian noise is added to the clean speech there will be a unique parameter G_z that can determine SNR.

The researchers used a pre-calculated lookup table to store the G_z parameterfor different SNRs. It is important to note that although the WADA-SNR approach assumes that the noise has

Gaussian distribution, the empirical results shows that it has superiority over the NIST-STNR even for other types of noise like background speech and music as shown in the figure 2[1].



a- Results with WADA-SNR



b- Results with NIST-SNR

Figure 2: Comparison of the average SNR estimates between WADA and NIST SNR algorithms on DARPA-RM database [1]

The rest of the paper will explain the proposed Short-Time Silence SNR estimator (STS-SNR). A comparison of the simulation results between the suggested STS-SNR algorithm with the WADA and NIST-STNR approaches is in section 3, while section 4 will is dedicated for the conclusion and future works.

II. THE SHORT-TIME SILENCE SNR ESTIMATOR

Based on what has been discussed in the previous section, an SNR estimator cannot give its results unless all the audio signal samples are processed. Those full-audio length approaches would seize the resources for some real-time live audio applications, like Automatic Speech Recognition (ASR). The suggested algorithm is to process only a small amount of samples from the audio signal to guess the SNR. The proposed algorithm is refereed as the Short-Time Silence (STS-SNR) estimator, which will consider only the first 30 ms at the beginning of the audio. This estimator assumes that the first 30 ms of the tested audio represents silence rather than speech. It is more realistic to have this consideration as the speaker cannot deliver his/her utterance within the first 30 ms. It is also assumed in this research that the SNR is not changing during the time of interest.

The proposed Short-Time Silence SNR (STS-SNR) estimator is described by the following steps:

1. Take the first 30 ms duration directly after the microphone is on, which is denoted here as the Noise Frame N_{Frame}.
2. Subtract the mean of the N_{Frame}
3. Estimate the Power Spectral Density (PSD) of the N_{Frame} using Fast-Fourier Transform (FFT) of 512 points and taking only 0 to 8 kHz band in consideration. Where PSD for the 30 ms N_{PSD} is:

$$N_{PSD} = |N_{Frame}(\omega)|^2 \quad (1)$$

Where $N_{Frame}(\omega)$ is the spectrum of the audio frame.

4. Reform the PSD by taking the absolute difference of its flipped version to produce a white-like PSD using the following process:

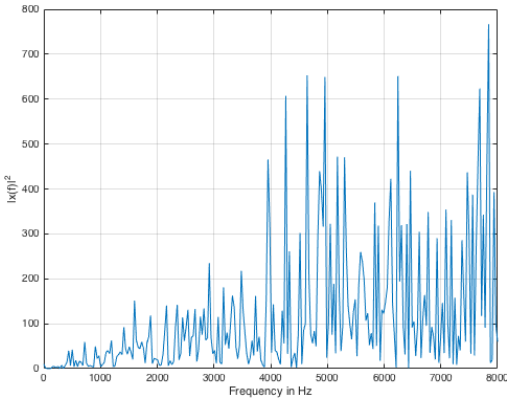
$$N_{Reformed} = |N_{PSD} - N_{PSD}^T| \quad (2)$$

here, N_{PSD}^T is the flipped version of the noise power spectral density vector N_{PSD}.

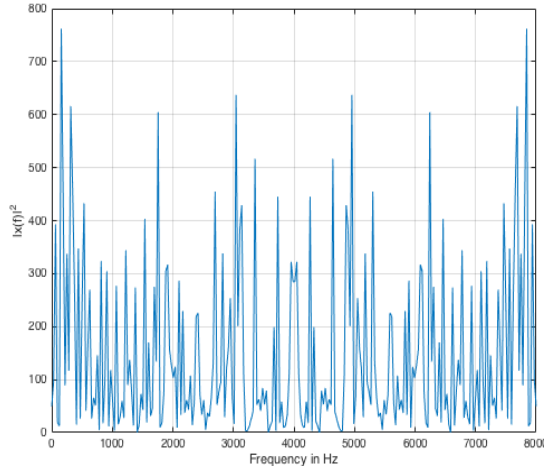
5. Take the average of the first and last quarter of the 8 kHz band of the $N_{Reformed}$. The average of those quarters is then considered as the estimated noise power spectral density $\widehat{N}_{PSD}$.
6. The estimated SNR in dB is:

$$SNR_{dB} = offset - 10 \times \log_{10}(\widehat{N}_{PSD}) \quad (3)$$

As non-white noise is also expected, the N_{PSD} is reformed in step 4 to produce a white-like PSD. This step will eliminate the effect of some colored noise that might appear while preserve the white noise PSD. In figure 3, the $N_{Reformed}$ is shown for a speech that was sampled by 16KHz sampling rate and has an additive blue noise that made the SNR 10 dB.



(a) PSD before reforming



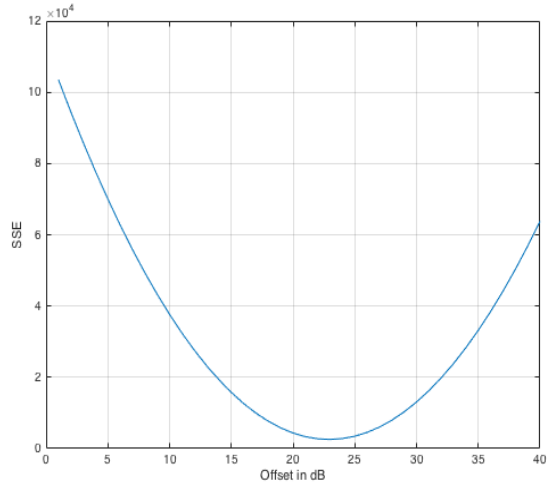
(b) PSD after reforming

Figure 3: The reforming operation in step 4

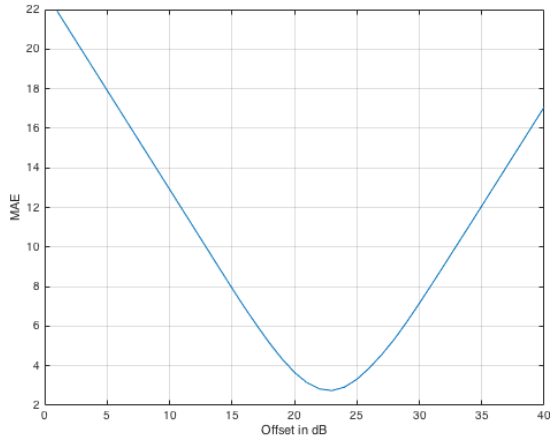
The measured pauses in a music concert are an ambient noise of 30 dB in advance as each frequency band contributes a different amount of dB power [9]. Hence, step 5 is designed to estimate the maximum contribution of each frequencyband by averaging the reformed spectrum from step 4. The N_{PSD} in the final stepshows to change linearly with the SNR of the speech utterance. However, there is an offset of the estimator mean from the real mean.

Experiments on different SNR noisy speech showed that the offset in equation (3) is not varying with the SNR. It was found that use of value 23 as an offset to get minimum error for different kinds and levels of noise. This exact value of the offset is estimated by the minimum sum of square error (SSE) and the minimum mean absolute error (MAE) as depicted in figure 4.

Experiments on the TIMIT test data and the NOIZEUS [10] show that this offset in equation 3 would be 23 dB as it gave the minimum for both the Sum of Square Error (SSE) and the Mean Absolute Error (MAE) as shown in figure 4 below.



(a) SSE for different Offsets



(b) MAE for different offsets

Figure 4: Testing different offsets for N_{Frame} of NOIZEUS noisy speech at 15 dB SNR

The offset value in equation (3) represents the minimum power gap between the moderate speech and the faint noise. So that the speech is differentiated from the noise by the speaker and his audience. It is common for a speaker to raise his vocal power to compete the ambient noise energy at least to hear himself, and by doing so allowing the target audience to hear clearly. According to Gordon J. King in [9], the Faint Noise starts by the 20 dB Sound Pressure Level (SPL) for whisper and goes up to 40 dB SPL for public library noise. The moderate speech lies in a pressure region of what he called Moderate Noise which is between 40 and 60 dB SPL [9]. Any increase in the faint noise will make the speaker, who aims to speak moderately, to increase his utterance power to compete the additional noise. Therefore, the constant offset of 23 in equation 3 seems to be reasonable and close to theory.

III. SIMULATION RESULTS

The STS-SNR estimator was evaluated using the DARPA TIMIT [11] and the noisy speech corpus NOIZEUS that is described in [10]. However, as the latter corpus speech and noise signals were filtered by the modified Intermediate Reference System (IRS) filters, an additional pre-processing was needed to be applied for getting the un-filtered spectrum.

By comparing the average estimated SNR of the WADA, NIST-STNR and the proposed STS-SNR, it seems that the latter has a better response. Figure 5 explains that the proposed STS-SNR is closer to the ideal case in average for NOIZEUS corpus.

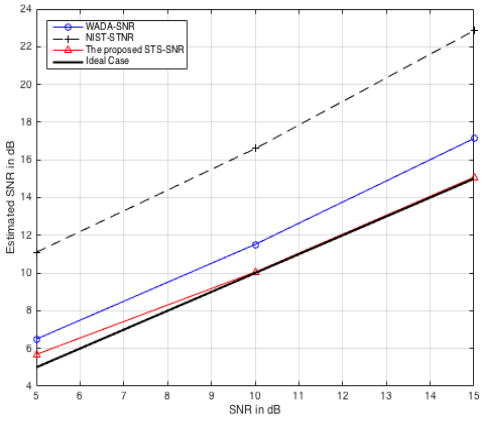


Figure 5: The average estimated SNR comparison for NOIZEUS corpus.

The audio of the TIMIT corpus is recorded in a quiet room environment [12], therefore different types of noise are added to it throughout this work. Additive White Gaussian Noise (AWGN), blue and pink are artificially computed and added to the TIMIT audio. Besides, crowd babble noise is also considered, where a crowd of people speaking randomly together. Both artificial and crowd babble noise are added to make noisy TIMIT of SNR ranges from 5 to 50 dB with 5 dB steps.

For white, pink and blue artificial additive noise, the STS-SNR shows to be a better SNR estimator than the other two approaches. Figure 6 below shows the comparison for the TIMIT corpus with white noise only as the responses due to pink and blue noise types have the same pattern.

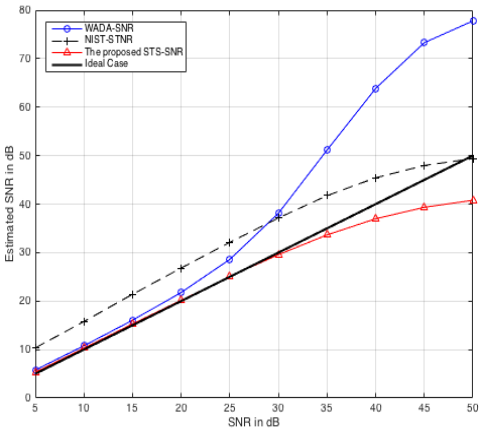


Figure 6: White Noise estimated SNR using TIMIT corpus

For babble crowd TIMIT the mean estimated SNR was almost better using the proposed STS-SNR compared with the WADA and the NIST-STNR, as shown in figure 7.

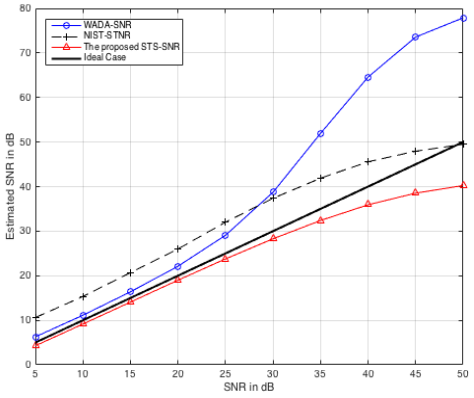


Figure 7: Crowd Babble Noise estimated SNR using TIMIT corpus.

The WADA-SNR has a good estimation when the SNR is less than 20 dB. However, when the test is run over higher noise levels, the proposed STS-SNR shows an advantage over the WADA approach. The results in figures 5 and 6 shows that the proposed STS-SNR estimator has an effective true mean detection in average.

Meanwhile, for the test cases below 20 dB SNR, the STS-SNR has shown to give a higher deviation from the mean, when it is compared with the WADA algorithm. As shown in figure 8 below, for SNR below 20 dB, the Mean Absolute Error (MAE) is higher for NIST-STNR and the proposed STS than the WADA approach. However, the MAE for STS-SNR is still less than both WADA and NIST-STNR in average for the whole range.

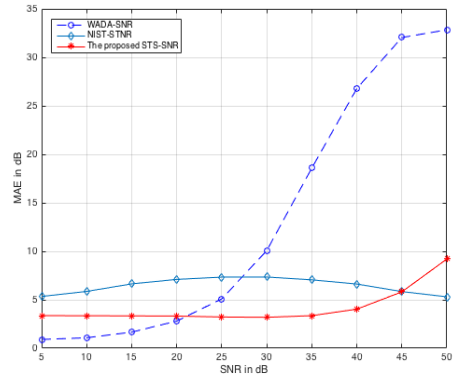


Figure 8: The MAE Comparison for NIST, WADA and STS Estimators.

For the complexity comparison among those SNR estimators, the STS-SNR is considered less expensive. As the proposed STS-SNR uses only 30 ms N_{Frame} window of the signal, it is expected to perform faster than the NIST and WADA SNR approaches. The computational expenses of the three estimators were test edusing the Real Time ratio (RT) as in the following equation:

$$RT = \frac{Processing\ \ Time}{Audio\ \ Duration} \quad (4)$$

Table 1 shows a comparison among the three SNR estimators, NIST, WADA and the proposed STS approach.

Table 1: Average Real-time Ratio for SNR Estimators

SNR Estimator	Average RT
NIST-STNR	0.0182
WADA-SNR	0.0033
STS-SNR	0.0026

Those measurements are the average RT's of when the algorithms were tested using the TIMIT test corpus of 1680 speech audio files.

IV. CONCLUSION

The proposed algorithm for the SNR estimation has shown to give a close guess to the ideal SNR for audio speech. It is less complicated and more efficient to expect the SNR in advance. The suggested SNR estimator in this paper shows to have a higher error for SNR less than 20 dB, as compared to the WADA approach. Nevertheless, for higher SNRs, the proposed STS algorithm estimates the SNR more accurately than the WADA estimator. In general, the STS algorithm has an advantage over the WADA and the NIST-STNR approaches. Even though the WADA estimator is based on extensive integration, its complexity is reduced by using the offline pre-calculated table [1]. However, as it is noted in table 1, the proposed STS-SNR estimator is still less expensive than the WADA-SNR.

The drawbacks of the proposed STS-SNR are there should be a 30 ms of silence at the beginning and the SNR is not changing. If the noise level is changed later, another silence period should be considered to re-estimate the new SNR ratio. In this case, a Voice Activity Detector (VAD) should be used to pick silencesthat occur within the speech as well. For Automatic Speech Recognizers (ASR) applications, there is another solution. Some of modern ASRs, like CMU Sphinx, can detect a silence and its samples could be easily extracted. This is because silence is treated as a word that represent a non-speech event and stored in a filler dictionary [13]. In future works, a continuous feedback from within-speech silences will be added allowing the STS to update the SNR continuously.

REFERENCES

[1]. C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis." in *INTERSPEECH*, 2008, pp. 2598–2601.

[2]. H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 153–156.

[3]. J. Morales-Cordovilla, N. Ma, V. Sa´nchez, J. L. Carmona, A. M. Peinado, J. Barker *et al.*, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4808– 4811.

[4]. H.-W. Park, A.-R. Khil, and M.-J. Bae, "Pitch detection based on signal- to-noise-ratio estimation and compensation for continuous speech signal," in *Convergence and Hybrid Information Technology*. Springer, 2012, pp. 767–774.

[5]. T. Moazzeni, A. Amei, J. Ma, and Y. Jiang, "Statistical model based snr estimation method for speech signals," *Electronics letters*, vol. 48, no. 12, pp. 727–729, 2012.

[6]. P. Papadopoulos, A. Tsiartas, J. Gibson, and S. Narayanan, "A supervised signal-to-noise ratio estimation of speech signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 8237–8241.

[7]. National Institute of Standards and Technology, "NIST speech signal to noise ratio measurements," 2008.

[8]. D. Ellis. (2011) Objective measures of speech qual- ity (snr).the laboratory for the recognition and organiza- tion of speech and audio (labrosa). [Online]. Available: http://labrosa.ee.columbia.edu/projects/snreval/

[9]. G. J. King, *The audio handbook*. Newnes-Butterworths Group, J. W. Ar- rowsmith Ltd., 1975.

[10]. Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007.

[11]. W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[12]. C. Becchetti and K. P. Ricotti, *Speech Recognition: Theory and C++ Implementation (With CD)*. John Wiley & Sons, 2008.

[13]. Carnegie Mellon University. (2000) Manual for the sphinx-iii recognition system,[Online]. Available at: http://www.speech.cs.cmu.edu/sphinxman/. http://www.speech.cs.cmu.edu/sphinxman/scriptman1.html#00