A SUPERVISED SIGNAL-TO-NOISE RATIO ESTIMATION OF SPEECH SIGNALS

Pavlos Papadopoulos, Andreas Tsiartas, James Gibson, and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA 90089

ppapadop@usc.edu, tsiartas@usc.edu, jjgibson@usc.edu, shri@sipi.usc.edu

ABSTRACT

This paper introduces a supervised statistical framework for estimating the signal-to-noise (SNR) ratio of speech signals. Information on how noise corrupts a signal can help us compensate for its effects, especially in real life applications where the usual assumption of white Gaussian noise does not hold and speech boundaries in the signal are not known. We use features from which we can detect speech regions in a signal, without using Voice Activity Detection, and estimate the energies of those regions. Then we use these features to train ordinary least squares regression models for various noise types. We compare this supervised method with state-of-the-art SNR estimation algorithms and show its superior performance with respect to the tested noise types.

Index Terms— signal-to-noise ratio estimation, speech signal processing, supervised learning

1. INTRODUCTION AND RELATED WORK

Signal to noise ratio (SNR) is one of the most fundamental metrics used in signal processing. It is defined as the ratio of signal power to noise power expressed in decibels (dB), and gives information about the level of background noise present in a speech (or other) signal. Its estimation in practice is however challenged by the diversity in the types and manner in which a signal can get corrupted. Moreover, the inherent variability in the signal itself (e.g., speech) adds an additional layer of challenge to SNR computation. Therefore, it is vitally important to study and estimate the effect of noise on the original signal in meaningful ways.

Speech processing in real life is challenged by a variety of environment and channel noise conditions making the design of robust applications an ongoing quest. For example, there is a renewed effort on robust Voice Activity Detection under the DARPA RATS program wherein the speech signal is degraded by a variety of, possibly unknown, channel conditions. This paper focuses on improved SNR computation especially targeting noisy speech signals.

Robust estimation of speech signal's SNR in turn can help guide the design of robust applications including Automatic Speech Recognition (e.g. [1], [2]), speech enhancement (e.g. [3], [4], [5]), and noise suppression [6].

Many methods have been proposed in literature for speech SNR estimation. In [7] the authors employ Voice Activity Detection (VAD) techniques to separate speech and noise regions and estimate SNR from the respective power in those regions. Ephraim and Malah in [3] derived a short-term spectral amplitude (STSA) estimator which minimizes the mean-square error of the spectral magnitude to estimate the *a-priori* SNR. This work has been the foundation for many subsequent research efforts (e.g. [4],[8], [9], [10]) and has resulted in many variations and improvements of the original algorithm.

The NIST SNR measurement ([11]) uses a method based on sequential Gaussian mixture estimation to model the noise. It then creates a short-time energy histogram which is used to estimate the energy distributions of the signal and noise from which SNR is estimated.

Other approaches rely on estimation of the speech and noise spectra (e.g. [1]), or track spectral minima in frequency bands which are used for optimal smoothing of the power spectral density (PSD) of the noisy speech signal, and use the estimated PSD and statistics of the spectral minima for a noise estimator (e.g. [12], [13]).

Finally, there are methods that make assumptions about the distribution of the signal, noise, or both in order to estimate the relative energy of each (e.g. [14]). While others use statistics from waveform samples, i.e. in [15] kurtosis values are used to estimate SNR in each frequency band.

Our proposed method is based on features that capture the presence of speech in the noisy signal and formulates a regression model, estimating its coefficients with ordinary least squares. It should be noted that our scheme does not require a Voice Activity Detection step. Our system supports two functionalities. First, we assume that we already know what kind of noise corrupts the signal and we use the the appropriate regression model. In the second case, we have no prior knowledge about the kind of noise that corrupts the signal .We use a classifier to identify the kind of noise and use the appropriate regression model. We compare our method with other state-of-the-art estimation algorithms such as the NIST SNR measurement ([11]) and the Waveform Amplitude Distribution Analysis (WADA) presented in [14]. Our experiments demonstrate that the proposed method outperforms these state-of-the-art systems.

In section 2 we present the features we use as well as the formulation of our algorithm. In section 3 we describe our experimental setup and how we chose the various parameters of our model. In section 4 we show the results of our SNR estimation method and compare it with other SNR estimation methods. Finally in section 5 we present our conclusions and discuss future work directions for the SNR estimation task.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. METHODOLOGY

In this work, our goal is to estimate the SNR of spontaneous speech signals or signals where speech boundaries are not available to us. Although, there are different kinds of SNR criteria, such as Global SNR, Local SNR, Segmental SNR ([7]), we focus on the estimation of Global SNR. Global SNR gives us information about the effect of noise on the whole signal and is defined as:

$$\mathrm{SNR} = 10 \cdot \log_{10} \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i)}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} n^2(i)}} \tag{1}$$

where the numerator is the root-mean square of the speech signal and the denominator is the root-mean square of the noise signal, expressing their respective energies $\mathbf{P}(S)$ and $\mathbf{P}(N)$.

Assuming that the noise is additive, the observed signal $x(i)$ is a sum of the speech signal $s(i)$ and the noise signal $n(i)$, $x(i) = s(i) + n(i)$, i being the time index. Furthermore, if the speech and noise signals are independent and zero-mean we can rewrite equation (1) as:

$$SNR = 10 \cdot \log_{10} \frac{\mathbf{P}(X) - \mathbf{P}(N)}{\mathbf{P}(N)} \tag{2}$$

which will be the basis of our estimation formulae.

Our approach focuses on finding regions of speech presence (and absense) in the signal without requiring VAD. We measure the respective energies of these regions, and create SNR estimators based on the formula of equation (2). Afterwards, we create a regression model, which we train with ordinary least squares and get our final SNR estimation.

To distinguish the regions of speech presence and absence in the signal we use a variety of features such as long-term energy, variability, pitch, and voicing probability. We take percentile windows of those features and calculate the energies $\mathbf{P}(X)$ and $\mathbf{P}(N)$ corresponding to those windows. The bands of high and low energies offer a reasonable approximation for representing speech from noisy speech regions. Such an estimate can be expressed as:

$$E_{a-b}^{c-d} = 10 \cdot \log_{10} \frac{\mathbf{P}(X_c^d) - \mathbf{P}(X_a^b)}{\mathbf{P}(X_a^b)} \tag{3}$$

where the values a, b, c, d correspond to percentile values where energy is concentrated. For example, if $a = 90\%$ and $b = 95\%$ then the expression $\mathbf{P}(X_a^b)$ is the average energy of the region where 90% to 95% of energy is concentrated. Since signals can be of arbitrary length and speech boundaries are unknown we make these estimates by using different empirical choices for windows defined by the values of a, b, c, d.

Moreover, since the transitions of both energy and feature values are abrupt we apply smoothing to increase the robustness of the estimates. However, since smoothing also alters the original values we use different smoothing window lengths in an attempt to both balance the robustness of the estimates and retain the original feature and energy values. In the following sections, we examine the features we used in more detail.

2.1. Long-Term Energy

Since SNR is the ration of energies, we first calculate the long-term energy in each frame from the spectrogram (the average energy in each frame). Then we apply different smoothing windows, using the moving average smoothing method.

For every case of smoothing window length, we estimate $\mathbf{P}(X)$ and $\mathbf{P}(N)$ by taking percentile windows on the long-term energy and substitute those values in (3). So, for different smoothing windows and energy regions we have different features.

2.2. Long-Term Signal Variability (LTSV)

Long-Term Signal Variability (LTSV) was proposed in [16] and is a way of measuring the degree of non-stationarity in a signal. Since speech is non-stationary, we can use LTSV to identify speech regions in a signal. Hence, we can make estimates of $\mathbf{P}(X)$ and $\mathbf{P}(N)$ based on percentage regions of variability and measure the respective energies of those regions. For example, when noise is stationary we can deduce that speech is present in the region where 85% to 90% of LTSV is concentrated. On the other hand, in the region 10% to 15% where LTSV is concentrated only noise is present.

An estimate based on variability is similar to the one of equation (3), where the windows of energy used for the estimates correspond to regions of the LTSV. However, before we compute those estimates we first apply smoothing windows on LTSV and median filtering on the corresponding energy regions.

2.3. Pitch

Another measure we can use to identify speech regions is through pitch detection. We use the openSMILE software, [17], to extract pitch information from the signal. Since pitch transitions are abrupt, and speech exists in the neighbour of pitch regions we apply smoothing on the outcome of pitch detection. Afterwards, we estimate $\mathbf{P}(X)$ and $\mathbf{P}(N)$ based on percentage regions of pitch presence in the signal in a similar fashion as in equation (3).

2.4. Voicing Probability

The final measure we employ to identify speech regions is the voicing probability. We use the openSMILE software ([17]) to calculate the voicing probability in each frame. Higher values of voicing indicate speech presence while lower indicate speech absence.

2.5. System Description

Based on the features described we created regression models for different types of noise (white, pink, car interior, machine gun, and babble speech noise). We chose these types of noises to test how our methods performs under both stationary and nonstationary noise conditions.

Our system supports two use cases. In the first case, we assume that we already know what kind of noise corrupts the signal and we use a linear regression model for every noise kind. The SNR estimation is based on the features we described and is given by:

$$\widehat{SNR} = \sum_{i=1}^{M} a_i \cdot f_i + \epsilon \tag{4}$$

where M is the number of features, ϵ is the disturbance term, a_i and f_i are the regression coefficients and the regressors respectively.

In the second case, we have no prior knowledge about the kind of noise that corrupts the signal. Instead, we use a classification scheme to identify the noise type and use the appropriate regression model. n [18], the authors use a K-Nearest Neighbour Classifier (KNN) classifier based on Bark scale features to classify noise types. In our work we have used a KNN classifier on 13 MFCCs.

3. EXPERIMENTAL SETUP

The total number of regressors we used in our models is 312 (24 from long-term energy, 216 from LTSV, 36 from pitch and 36 from voicing) and we estimate the features' coefficients with ordinary least squares. The regressors result from a combination of smoothing window lengths and regions of the features from which we make energy estimations according to the formula 3.

In the case of Long Term Energy and LTSV the window length ranges from 0.3ms to 1.8ms with a 0.3ms step, while in Pitch and Voicing Probability the window lengths are 0.9ms, 1.6ms, 2.2ms, 2.8ms, 3.4ms, and 4.1ms. The value pairs a, b, c, d in 3 we used to estimate the energies are shown in table 1

a	b	c	d
85%	95%	5%	15%
80%	90%	10%	20%
75%	85%	15%	25%
5%	15%	85%	95%
10%	20%	80%	90%
15%	25%	75%	85%

Table 1. Percentile Pair values of pitch windows from which we calculate the average energy

These values where the result of experimental procedure. Our experiments showed that adding more features (i.e. more smoothing windows, etc) boosts the performance of the estimation. Since this is a work in progress, in the future we plan to provide detailed analysis of the impact each feature has on the model.

For every noise type we used 1680 clean speech files from the TIMIT Database sampled at 16KHz in which we introduced silence periods randomly selected between 3 and 10 seconds to create signals with unknown speech boundaries. Then we added noise at six SNR levels (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB), resulting in a total of 10080 training samples per regression model.

For the KNN classifier we used 20 nearest neighbors (K=20) based on 13 MFCCs. We used the same set of 1680 files (adding noise for every SNR level) to train the KNN classifier. The final decision is made by calculating the probability of each class in every frame and then follows a majority vote.

4. EXPERIMENTAL RESULTS

We have tested our system for five different noise types. We randomly selected 150 files from the TIMIT database (there was no overlap between the training and testing files). In each file we introduced 3 to 10 seconds silence regions and then added noise at 6 different SNR levels. We compared our method with the WADA and NIST SNR estimation methods using the mean absolute error metric. In all cases we found that our method outperforms the other methods.

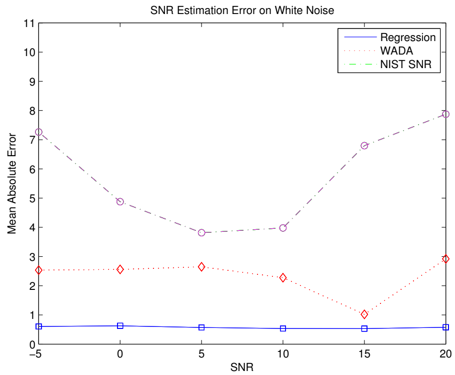

Fig. 1. Mean absolute error for White Noise.

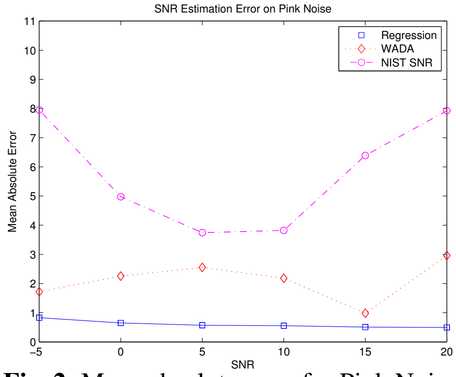

Fig. 2. Mean absolute error for Pink Noise.

In figures 1, 2,3 the results of white, pink and car interior noise are presented. By comparing the mean absolute error of our method and the WADA and NIST SNR method for 6 different SNR levels,it is clear that our method provides better estimates for every SNR level (difference in error ranges from 0.3db to 7db).

In the case of machine gun noise (figure 4) our method greatly outperforms the other methods (difference in mean absolute error is about 30db). Both WADA and NIST SNR fail to provide accurate estimates as shown from their mean

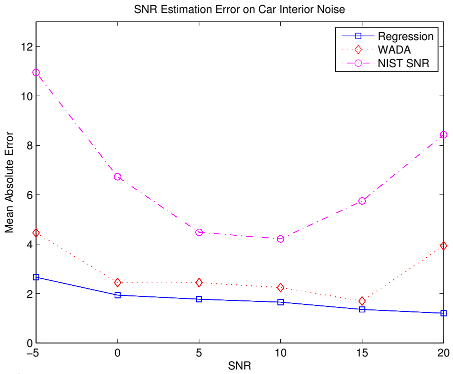

Fig. 3. Mean absolute error for Car Interior Noise.

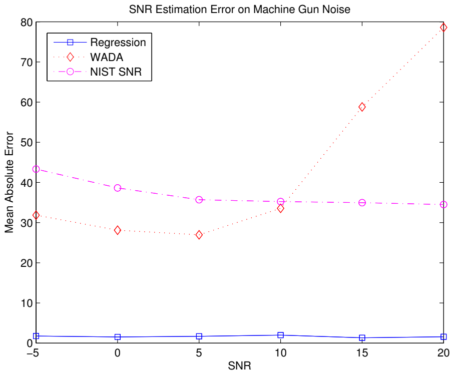

Fig. 4. Mean absolute error for Machine Gun Noise.

absolute error values. The reason for this is that our method does not make any assumptions about stationarity. Also this indicates that our method can perform well across different noise types with different characteristics.

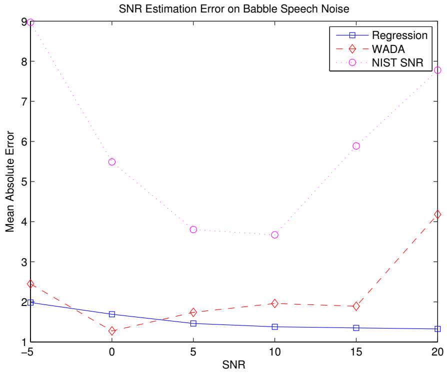

Fig. 5. Mean absolute error for Babble Speech Noise.

Finally, in the case of Babble Speech Noise (figure 5) we can see that only for 0dB the WADA method performs better. Since babble speech noise is similar to speech some of our features(i.e. pitch,voicing) fail at same energy levels. However, our method gives better estimates overall.

The above results refer to the case where we know the type of noise that corrupts the signal and we choose the appropriate regression model. In the second set of experiments we used the same test set of files. In every case we corrupted a signal with a noise that was used for training the KNN classifier, the signal was correctly classified and the appropriate regression model was used. Since our classifier achieved perfect accuracy for the given set of noises, we tried to corrupt a signal with high frequency Noise (which was not used for training the classifier). The classifier chose the regression model for white noise. In figure 6 we can see the results when we corrupted signals with high frequency noise and used the white noise regression model to estimate the SNR.

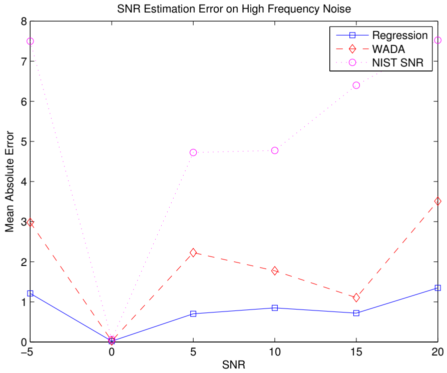

Fig. 6. Mean absolute error for High Frequency Noise by using the regression model of white noise.

In all the cases we examined our method outperforms other state-of-the-art methods, especially when the kind of noise that corrupts the signal is known. When the noise is unknown the performance of our method depends on the outcome of the KNN classifier, for instance in the example of high frequency noise if the classifier chose the regression model of machine gun noise we would have failed to provide accurate SNR estimates.

5. CONCLUSIONS AND FUTURE WORK

We have presented a novel method for Global SNR estimation using regression models which are trained on features that can be ranked by presence of speech. We tested our method for various noise types with different statistical properties and demonstrated that it successfully provides an accurate SNR estimation. Furthermore, we compared our work with two other SNR estimation algorithms (WADA, NIST SNR) and the proposed method in general outperforms across all experimental conditions.

Finally, we plan to attempt to generalize across noise types. Moreover, we want to improve our channel classification by employing features that can capture noise characteristics, since it is well known that MFCCs are not very robust under noise conditions. We also plan to test more advanced classifiers (e.g. DBN-DNN, SVMs, etc) as well as adaptive schemes and soft assignment approaches that will generalize better for unseen noise conditions.

6. REFERENCES

[1] H. G. Hirsch and C. Ehricher, “Noise estimation techniques for robust speech recognition,” in *Proc. IEEE ICASSP*, 1995.

[2] J. Morales-Cordovilla, N. Ma, V. Sanchez, J. Carmona, A. Peinado, and J. Barker, “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*, 2011, pp. 4808–4811.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] C. Plapous, C. Marro, and P. Scalart, “Improved Signal to Noise Ratio Estimation for Speech Enhancement.” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[5] Y. Ren and M. T. Johnson, “An improved SNR estimator for speech enhancement.” in *ICASSP*. IEEE, 2008, pp. 4901–4904.

[6] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression.” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 184–192, 2003.

[7] M. Vondrášek and P. Pollák, “Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency,” *Radioengineering*, vol. 14, pp. 6–11, 2005.

[8] I. Cohen, “Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[9] S. Suhadi, C. Last, and T. Fingscheidt, “A data-driven approach to a priori snr estimation.” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 186–195, 2011.

[10] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, pp. 443–445, 2003.

[11] “The NIST Speech SNR Measurement,” http://www.nist.gov/smartspace/nist_speech_snr_measurement.html.

[12] M. Rainer, “An efficient algorithm to estimate the instantaneous snr of speech signals,” in *Third European Conference on Speech Communication and Technology, EUROSPEECH*, 1993.

[13] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[14] C. Kim and R. M. Stern, “Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis,” in *Proc. Interspeech*, 2008, pp. 2598–2601.

[15] E. Nemer, R. Goubran, and S. Mahmoud, “SNR estimation of speech signals using subbands and fourth-order statistics,” *Signal Processing Letters, IEEE*, vol. 6, no. 7, pp. 171–174, 1999.

[16] P. Ghosh, A. Tsiartas, and S. Narayanan, “Robust Voice Activity Detection Using Long-Term Signal Variability,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[17] “openSMILE,” http://opensmile.sourceforge.net/.

[18] C. Eamdeelerd and K. Songwatana, “Audio noise classification using bark scale features and k-nn technique,” in *International Symposium on Communications and Information Technologies, ISCIT 2008.*, pp. 131–134.