Noise Update Modeling for Speech Enhancement: When do we do enough?

Nitish Krishnamurthy, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
n.krishnamurthy@student.utdallas.edu, john.hansen@utdallas.edu

Abstract

In speech enhancement, it is generally assumed that if you can update your noise estimate on a frame-by-frame basis, you should achieve the highest level of enhancement performance. However, for many noise types and environmental conditions, it is not necessary to perform an update on a frame-by-frame basis to achieve superior performance if the noise structure does not change rapidly. For applications where compute/memory resources are limited, better overall speech performance could be achieved if a more reasonable update rate is estimated so that available compute/memory resources could be made available to the enhancement algorithm itself. In this study, we propose a framework to model the noise structure with the goal of determining the best update rate required to achieve a given quality for speech enhancement. Speech systems generally develop specialized solutions for niose which are unique to each application (i.e., recognition, speaker ID, enhancement etc.). Here we propose a model to predict the noise update rate required to achieve a given quality for enhancement. We evaluate the algorithm across a corpus of four noise types under different levels of degradation. The error between the mean observed and the mean predicted Itakuta-Saito (IS) values of quality are typically between 0.06 to 1.78 IS for our model selected noise frame update rate of 1 frame every 5 frames using the Log-MMSE enhancement scheme. Finally we consider mobile and resource limited applications where such a framework would be useful

Index Terms: speech enhancement, noise modeling, noise update rate.

1. Introduction

The field of speech analysis and modeling has evolved dramatically over the past forty years, from early electronic models to source filter frameworks, to rapid DSP based strategies. Over the past decade, the use of speech technology in real environments has rapidly increased. This is due to the proliferation of mobile devices such as cell phone and PDA technology requiring speech recognition and interactive dialogue systems. Environmental noise has therefore become one of the primary causes that limit speech system performance in real world environments. Traditionally solutions for noise are addressed separately for every speech area (e.g., coding, recognition, enhancement, speech recognition, speaker ID, etc.). Current solutions to address noise fail in two areas which limit the interoperability of systems. These systems are usually customized to operate only under specific environments, but if the environment changes there is a rapid decrease in system performance. Systems that address this issue are computationally expensive in terms of resources that are expensive for mobile speech systems. It would stand to reason that a field of noise processing would evolve that would address a wider range of speech applications focused directly on noise analysis, modeling and processing in the real world. In this research, we propose a framework for noise modeling with specific application to determine the update noise rate required to achieve a given quality of speech enhancement in an environment. This would provide a means to better utilize the available computing resources of a mobile voice system. The main goals of this paper are (a) Using noise modeling to achieve effective speech enhancement performance with restricted computational resources. (b) To be able to redirect computational resources by limiting noise modeling requirements to improve overall speech system performance.

This work was supported by the U.S. Air Force Research Laborator, Rome NY, under contract F30602-03-1-0110, RADC under contract (FA8750-05-C-0029) and by University of Texas at Dallas under project EMMITT

2. Noise and Speech

Noise has been addressed differently for every speech area. For speech coding the issue of noise resistance is focused on strategies that make coefficients more robust to noise. Noise has been addressed for speech recognition and speaker identification systems under the title of environmental robustness. There are three basic approaches in the case of recognition systems, (a) Enhancing speech to obtain the estimate of clean speech and using that for recognition w purposes, (b) Model Adaptation under different noise conditions, (c) Retraining models or multiple models using noisy data. Noise is addressed differently across different areas of speech processing. One main problem to be addressed is that of speech communications in changing adverse conditions.

3. Noise and Speech Enhancement

Speech enhancement is the process of recovering the speech signal information from a received degraded feature vector. Speech enhancement is extremely important for speech systems because noise is ever present in communications [1]. (a) It is the process that aids speech communications under noisy conditions (voice communications such as telephony, hearing aids, etc). (b) It is a front-end for many speech systems such as speech recognition and speaker ID systems. The goals of speech enhancement are different under dif-

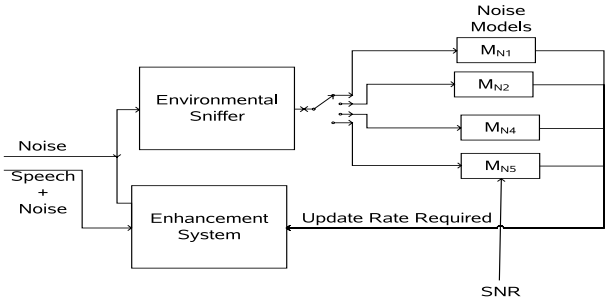

Figure 1: *The Update Rate Evaluation Algorithm*

ferent circumstances. Under extremely noisy conditions the main task of speech enhancement is to make speech communication possible by removing the noise from speech. Under mid to low noise conditions with intelligible speakers the speech enhancement focus shifts to reducing listener fatigue during speech communication.

The distortion at the output of the speech enhancement process is a function of mainly two variables: the environment (the power and stationarity of noise) and the algorithm itself (artifacts introduced). Here, the kind of environments which are to be addressed reflect to a large degree the output quality of the speech enhancement process. Most speech enhancement algorithms use a spectral estimate of the noise. In single channel systems, it is typical to use the first few frames of speech for noise estimation, or a voice activity detector is employed and noise is estimated at intervals where no speech is present. In dual channel systems, one channel is normally dedicated to noise estimates to enhance the speech plus the noise channel. Under non-stationary conditions we require regular noise updates to have a good estimate of the noise spectral structure. In this research, we focus on providing an estimate of the number of updates of noise required to obtain a given quality of enhancement, given the noise type and noise power in that environment. Our attempt here is not to formulate a new algorithm for enhancement or noise tracking methods but to provide a framework that would help the above algorithms perform better by reducing the computational load for a given task. If we know the number of noise updates required for a given quality of enhancement, we can utilize the computing resources in a better way. For example a stationary environment would require fewer updates to provide better enhancement relative to the clean speech than a non-stationary environment. So, under such environments the computational resources are therefore more effectively directed towards tasks other than noise estimation. If large amounts of computational resources are required to obtain a given quality of enhancement, algorithms that are more computationally intensive but provide better enhancement can be used.

4. Noise Processing Framework

We propose a framework towards achieving the above goal in Fig. 1.We construct a model for the noise update rate against the resultant quality measure for different noise types across different SNR values using speech degraded under pre-defined environmental conditions. We use these models to estimate the number of noise updates required for a given environment. In a real world scenario, the system deployment would be as follows: Once the noise type and power are determined [5], we use the pertinent model for that environment to estimate the update rate required for the required level of speech quality; conversely the output quality can be estimated given the update rate used for the environment.

4.1. The Itakura Saito Measure

We employ the Itakura-Saito measure [2] as an objective quality measure of the enhancement of the speech signal. This measure is based upon the dissimilarity between the all pole speech spectra of the original $1/A(\overrightarrow{a_s},\omega)$ and processed (degraded or enhanced) waveforms $1/A(\hat{\overrightarrow{a_s}},\omega)$ as,

$$d_j(\overrightarrow{a_s},\hat{\overrightarrow{a_s}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}[e^{v(\omega)} - v(\omega) - 1]d\omega \tag{1}$$

where,

$$v(\omega) = log\left(\frac{1}{\mid A(\hat{\overrightarrow{a_s}},\omega)\mid^2}\right) - log\left(\frac{1}{\mid A(\overrightarrow{a_s},\omega)\mid^2}\right) \tag{2}$$

where $\overrightarrow{a_s}$ and $\hat{\overrightarrow{a_s}}$ are the all pole model parameters from the gain normalized and the processed speech spectra. The symmetric IS measure is obtained as follows,

$$d_{IS(j)} = \frac{1}{2}\{d_j(\overrightarrow{a_s},\hat{\overrightarrow{a_s}}) + d_j(\hat{\overrightarrow{a_s}},\overrightarrow{a_s})\} \tag{3}$$

This measure assigns a large weight when the error is due to differences in shape or location of spectral peaks, and a smaller weight for error in spectral valleys. This is desirable because the human auditory system is more sensitive to errors in formant peaks, than to spectral valleys between peaks . If the value of $d_{IS(j)}$ is zero, then the speech signal $\hat{\overrightarrow{s}}$ at frame j is same as the noise free reference $\overrightarrow{s}$. The quality measure has an effect on the update rates and the performance of the noise modeling framework since different measures assign different weights across different attributes of speech signals. The IS measure is a non symmetric measure since the IS Measure between signals (a,b) is not the same as the IS Measure between signals (b,a), and the reason we use the symmetric version in Eq. 3. The IS score ranges from $0 - \infty$, with values between $0.0 - 0.75$ representing small levels of distortion typically associated with vocoder distortions, and values above 5.0 reflecting more severe background distortions. Average IS measures that are greater than 5 represent severe distortions and would be an upper bound across noise types.

4.2. The MMSE Algorithm

The enhancement algorithm we use is the minimum mean square error log-spectral amplitude estimator proposed by Ephraim and Malah [3]. We intensionally select a well established, traditional method in order to focus our efforts on noise modeling for enhancement. The log-MMSE algorithm uses a distortion measure based on the log spectra to estimate the speech signal, and assumes that over short intervals that speech and noise are essentially uncorrelated. The algorithm minimizes the following distortion measure,

$$E(\log A_K - \log \hat{A_K}) \tag{4}$$

The clean speech is estimated using the following expression

$$\hat{A_k} = e^{E\{\ln(A_k|Y_k)\}} \tag{5}$$

where $\hat{A_k}$,A_K ,Y_K denote the k^{th} Fourier expansion coefficient of the estimate of clean speech, the speech signal and the noisy observation. This estimator was used because of its relative simplicity over other enhancement algorithms.Other extensions to MMSE have been proposed which employ human auditory characteristics eg by the auditory masked threshold.

5. Algorithm

The overall process for noise update modeling is shown in Fig. 4. We assume a dual channel system with one channel that contains only noise and another that contains degraded speech. The type of noise is detected and the appropriate model for that noise is used. This model is constructed using the mean of the enhanced speech enhanced across different update-rates. When an unknown utterance arrives, the required model is selected and the models are interpolated to obtain the estimate of the IS measure (12th order LP model is used for IS). We use the 192 core set from TIMIT utterances to construct the models. The 192 sentence TIMIT set was degraded using four noise types: aircraft cockpit noise, multi-speaker babble noise, stationary car noise and white Gaussian noise (AIR,BAB,HWY,WGN). The four noise types were selected because of their varying degree of stationarity and their spectral properties. The degraded utterances are enhanced by the log-MMSE algorithm. The noise updates for the log-MMSE algorithm are performed using noise from the second channel (i.e., to ensure an even noise frame update process; future work will consider non-uniform update rates based on silence intervals for single-channel systems). The average IS measures are calculated for these utterances for different noise update rates and are recorded. The same procedure is carried out for a range of SNR values. This results in the final update rate model for the given environmental conditions.

While estimating the IS measure for the given update rate for an unknown utterance, the models corresponding to the detected noise environment are selected. These are interpolated to obtain an estimate of the model for the given SNR conditions. At given update rates, the respective IS measures can be estimated using this noise model. Alternatively, the update rate can be predicted using the model plots in Fig. 2 given that we select the particular IS score which is contained in the range (note: this is more appropriate for noise types such as BAB, which show a varying range of resulting IS speech quality as the update rate is reduced from 1:5 (one update each five frames) down to 1:50 because of the time varying nature of the noise; if there is no IS change versus noise frame update rate as in WGN, then it may not be possible to achieve a desired IS score by selecting the appropriate update rate since the model plot is flat).

6. Evaluation

The noise models for speech enhancement were obtained using the above outlined scheme. The flow diagram of the noise modeling process is summarized in Fig. 4, and the resulting models are illustrated in Fig. 2. The frame size used was 20 ms weighted using a Hamming window and the noise estimate was calculated as the magnitude square of the Fourier transform. The speech and noise signals were all sampled at 16kHz and PCM encoded. The Fig. 2(a) depicts the mean IS values after enhancement across all frames for different noise update rates with one noise update every $3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$ frames respectively. We represent this as (1:5) which corresponds to one update every 5 frames . The variance in IS measures across all frames for different update rates is shown in Fig. 2(b) and Fig. 2(c) shows the mean IS values after enhancement for the aircraft cockpit noise model for update rates at SNR values of 0 dB, 5dB and 10dB. A total of 24000 sentences were enhanced during the process of obtaining the update rate models in Fig. 2.

The **Stationary car noise** is low frequency restricted bandwidth noise from a Chevy Blazer traveling on a highway at 65 mph. As illustrated in the noise model plots the increase in the noise update rate does not result in an improved IS measure after enhancement. We note that update rates as low as one update every 1000ms (1:50 frames) gives the same resulting quality of enhancement as one update every 100ms (1:5 frames). This implies that an update rate of one update every 1000ms is sufficient to characterize the noise. **Aircraft cockpit noise** noise was recorded in a Lockheed C130 flying at 25,000 ft and is stationary in nature like the car noise. The major differences in the IS measure plots are due to the difference in noise bandwidths. The shape of the model plot is similar to that obtained for stationary car noise, but the level is elevated. This is due to higher degradation of speech spectral structure. **White Gaussian noise** is the most stationary of all the noise types considered. This noise type also degrades the spectral structure of speech the most due to its flat spectrum. This is reflected in the IS measure plots as they are essentially flat but shifted vertically, showing more degradation than AIR or HWY. The **Babble noise** is non-stationary and has spectral properties similar to speech. More spectral updates are required to obtain a reasonable enhancement under babble noise conditions. This noise type shows a general rising trend(i.e., as the update rate is reduced the resulting enhancement suffers).

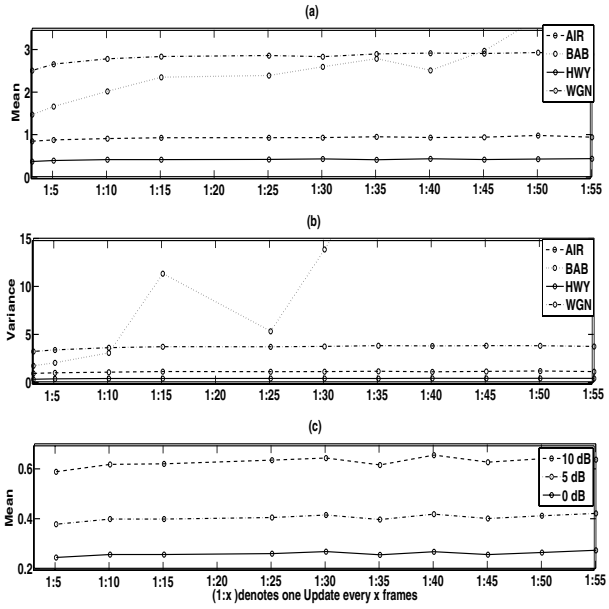

Figure 2: *(a) update rate model showing the mean for white Gaussian noise, babble noise, aircraft cockpit noise and stationary car noise (b) the variance of the IS measures for different update rates, and (c) the update rate Vs.. IS measure model for aircraft cockpit noise for different SNR values 0dB, 5dB and 10dB.*

The histogram of IS measures for speech degraded using different noise types is shown in Fig 3. The distributions are modeled using γ distributions since the IS measures cannot be negative. The relative effect of noise on the speech utterance can be gauged using the length of the tail of the γ distribution. As seen, the speech corrupted by the stationary car noise and aircraft noise have shorter tails than the speech corrupted by babble noise and white Gaussian noise. These models depend on the quality measure that is chosen.

To test the capability of these models to predict update rate vs. IS quality, a separate set of 192 TIMIT utterances were used. These were degraded at 2dB, 5dB, and 7dB SNR using the four

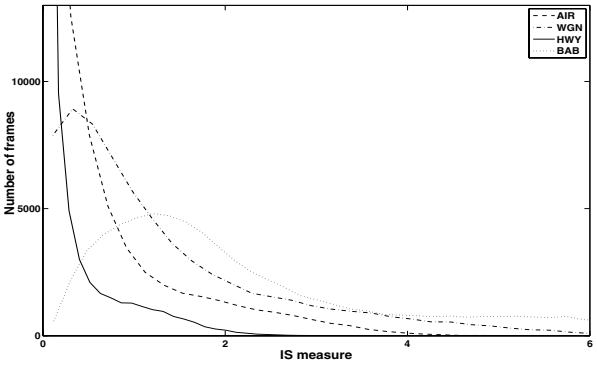

Figure 3: *distribution of IS measures for speech files degraded with different noise types*

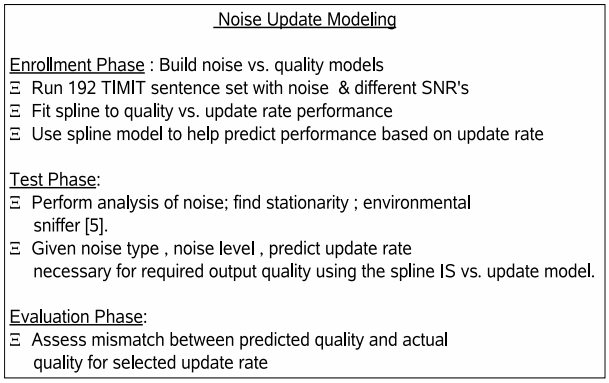

Figure 4: *Enrollment-Test-Evaluation flow our Noise Modeling framework*

noise types, and then enhanced using a noise update rate of one update every 100 ms or 5 frames (1:5). The average IS measure across frames was then compared to the estimated IS measure from the pre-trained models.

7. Results

The flow of the enrollment-test-evaluation (ETE) process is shown in Fig. 4, and all tests were performed according to this standard. The output from the evaluation phase is the error in the IS measures of the mean of the observed (from the experiments) and the estimated values(from the models). The models were trained at SNR values of 0 dB, 5dB and 10 dB. These were then tested for degradation levels of 2dB, 5dB and 7dB (in order to determine the ability of the noise update model to generalize for SNRs which were not used in training in order to predict accurate update rates. The results are shown in Table 1. The mismatch in mean IS values are typically of the order of coding distortions for the AIR, HWY, and BAB noise types. However, the error for WGN is more than 1.0, which is typically a little more than that caused by coding distortions this is due to the white spectral structure of WGN. Experiments were carried out using a noise update-rate of one in five frames. This update rate is approximately where plots for stationary noise types start to become flat and the decrease in update rate does not reflect on the IS measure plots. If a high noise update rate is required to achieve a required quality using spectral subtraction methods, it is possible to achieve better quality using other enhancement strategies such as iterative schemes (Auto-LSP [6]) that use more resources and therefore provide better overall enhancement.

Table 1: *Error between Predicted and Observed IS measures for noise types at SNR values at an update rate of 1 update every 5 frames*

SNR ↓ Noise types →	AIR	HWY	BAB	WGN
2dB	-0.068	-0.6361	0.7693	1.7853
5dB	-0.0515	- 0.5109	0.6693	1.6241
7dB	-0.2917	-0.683	0.3591	1.2683

8. Discussion

In this paper, we proposed a framework to model noise and its impact on the required update rate necessary to achieve a given enhancement quality. Furthermore, we used the resulting noise models to estimate the required update rate for different noise types and levels of degradation. This approach would be useful to improve performance of next generation mobile speech systems by allowing a more effective noise modeling solution to be established up-front, and re-directing available compute/memory resources to improve enhancement or other speech processing tasks, thereby improving overall system performance. In theory, the resulting speech quality could be reduced using the proposed scheme, however when compute/memory resources are limited for cellular platforms, PDAs, and other mobile systems, the trade-off is always in achieving the best overall solution. Future algorithm development could include:(i) integrating the proposed framework with voice activity detection system; (ii) exploring the impact of non-uniform noise frame estimation (here, we assumed a second channel was available to draw noise estimates from; further work could explore from a statistical perspective the potential improvement knowing that periods of silence are inserted between phrases in spontaneous speech) for single channel systems; (iii) considering alternative quality measures that have higher correlation with subjective quality for modeling the noise update rate. This research also provides insight into the enhancement process and its dependence on update rate. As seen from the models for stationary noise types, quality does not depend on the update rate if the update rate of noise is below a limit. Therefore, effective estimation of the degree of stationarity would also provide a useful metric for determining proper noise update rates. Finally, the resulting spline models could be employed in other speech applications that could benefir from front-end enhancement where computing resources are limited.

9. References

[1] Y.Ephraim,"Statistical-Model-Based Speech Enhancement Systems,"Proceeding of the IEEE, 80(10):1526-1555,1992.

[2] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, "Objective Measures of Speech Quality", Prentice-Hall, NJ, 1988.

[3] Y. Ephraim, D. Malah,"Speech enhancement using a minimum meansquare logspectral amplitude estimator," IEEE Trans. ASSP, Apr. 1985.

[5] M. Akbacak, J.H.L. Hansen, "General Issues in Environmental Noise Tracking for Robust In-Vehicle Speech Applications: Supervised vs. Unsupervised Acoustic Noise Analysis," DSP for Vehicular and Mobile Systems, paper M2-2, Sesimbra, Portugal, Sept. 3, 2005.

[6] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," IEEE Trans. Signal Processing, April 1991.