Data Driven Suppression Rule for Speech Enhancement

Ivan Tashev and Malcolm Slaney
Microsoft Research, Redmond, WA 98052, USA
ivantash@microsoft.com, malcolm@ieee.org

***Abstract*—Audio signal enhancement often involves the application of a time-varying filter, or suppression rule, to the frequency-domain transform of a corrupted signal. Classic approaches use rules derived under Gaussian models and interpret them as spectral estimators in a Bayesian statistical framework. This mathematical approach provides rules that satisfy certain optimization criteria – maximum likelihood, mean square error, etc. In this paper we propose to learn the suppression rule from a representative training corpus and make it optimal in the sense of best perceived quality. This can be measured, for example, with the wideband PESQ algorithm, for which we cannot derive an analytic estimator. The proposed suppression rule is evaluated in controlled environment and shows improvements in the range of 0.1–0.2 PESQ points on a data corpus with SNRs ranging from -10 to +50 dB.**

Keywords—speech enhancement; noise suppression; suppression rule;

I. INTRODUCTION

This paper addresses an important problem in audio signal processing: How do we create a filter that enhances audio (i.e. noisy speech) so the result is optimum for human perception? The conventional approach is a short-time spectral attenuator, where we suppress broadband noise using a time-varying filter applied to the frequency-domain transform of the corrupted signal. Designing such a filter is a difficult problem because we don't have access to the original signal. We can measure the statistics of the received signal, and we can often estimate something about the added noise, but we don't know anything about the original speech.

Typically a derivation assumes the speech and noise signals have Gaussian distributions and the suppression rule is a function of the *a priori* and *a posteriori* signal-to-noise ratios (SNR) [1]. With various approximations there are simple estimates for the "optimal" filter, but the resulting systems do not perform as well as desired. Note, there are two halves of the enhancement problem. One must first estimate the statistics of the underlying signals, and then one can use these estimates to design a suppression filter. This paper only addresses the second problem: how do we design suppression filters with high perceptual quality, when we have exact statistical measures of the original signals?

In this paper we describe the speech-enhancement filter-design problem as an optimization problem. This avoids the need to fully characterize the signal's statistics, and the formal use of many approximations. Instead, we learn that in *this* condition the best filter has *this* gain. We derive a new human-perception-friendly *suppression rule* by learning from a representative corpus of noisy data. With the appropriate error metric, we can improve noisy speech by a significant measure on a perceptual scale.

By optimizing on real data, we address two problems. The first is that it is convenient to assume that the speech signal is well modeled by a Gaussian distribution, but this is not true in either the time [2] or frequency domains [3]. Several attempts to derive a suppression rule using a super-Gaussian distribution model for the speech signal lead to very complex derivations and marginal improvements. The second problem is that what we actually want is for humans, listening to the noise-suppressor output, to perceive its quality as better. Classically, we approximate the desired goal of maximizing perceived quality with an approach that is easier to describe mathematically, such as mean squared error, maximum likelihood, or log-mean squared error.

In signal processing and communications the gold standard for measuring the perceived quality of the speech signal uses a metric called the mean opinion score (MOS) [4]. The measurement process consists of asking multiple people to listen to audio files, rank the audio quality with a number between one and five, and then averaging for the final result. This is a time consuming and expensive process. Instead the International Telecommunication Union standardized several algorithms for objective measurement of the sound quality. The most commonly used is the Perceptual Evaluation of the Sound Quality (PESQ) algorithm [5]. It exists in narrowband (8 kHz sampling rate) and wideband (16 kHz sampling rate) versions and is a *de-facto* computational proxy for the MOS procedure. While more computationally expensive, and much harder to analyze, it gives an estimate that is a better fit to perception than conventional approaches such as MSE.

Our approach for addressing both problems above is to learn the suppression rule from a synthetic data corpus, for which we have perfect estimates of the *a priori* and *a posteriori* SNRs. As a form of regularization and interpolation we propose a two-dimensional sigmoid function for the suppression rule. We use optimization to estimate the parameters of this function for three different ways of measuring performance: magnitude and log-magnitude estimators (first issue), and maximum perceptual quality (second issue). We achieve improvements that are very audible and range from 0.1–0.2 PESQ points better than the best known suppression rule, over a broad range of SNRs (–10 to +50 dB).

II. MODELING AND SUPPRESION RULES

There are many ways to design a suppression rule. We want to highlight five styles of rules, both to show how our optimization approach can simulate the classic algorithms, and to show

that we can do better with the proper optimization criteria. We compare and contrast all these algorithms in Table 1.

A. The Data

Let $x_n = x(nT)$ represent values from a finite-duration analog signal sampled at a regular interval T. We represent a corrupted sequence with an additive observation model $y_n = x_n + d_n$, where y_n represents the observed signal at time index n, x_n is the original signal, and d_n is additive random noise, uncorrelated with the original signal. The goal of signal enhancement is then to form an estimate $\hat{x}_n$ of the original signal x_n based on the observed signal y_n. In many implementations where efficient on-line performance is required, the set of observations $\{y_n\}$ is filtered using the overlap-add method of short-time Fourier analysis and synthesis. Taking the discrete Fourier transform on N windowed intervals of length $2K$ yields K frequency bins per frame: $Y_k = X_k + D_k$, where all these quantities are complex. Noise reduction may be viewed as the application of a suppression rule, or nonnegative real-valued gain H_k, to each bin k of the observed signal spectrum Y_k, in order to form an estimate $\hat{X}_k$ of the original signal spectrum: $\hat{X}_k = H_k \cdot Y_k$. This spectral estimate is then inverse-transformed to obtain the time-domain signal reconstruction.

Within such a framework, a simple Gaussian model often proves effective [1]. In this case the elements of $\{X_k\}$ and $\{D_k\}$ are modeled as independent, zero-mean, complex Gaussian random variables with variances $\lambda_x(k)$ and $\lambda_d(k)$, respectively: $X_k \sim \mathrm{N}_2\left(0, \lambda_x(k)\right)$, $D_k \sim \mathrm{N}_2\left(0, \lambda_d(k)\right)$.

B. Minimimum Mean-Squared Error (MMSE)

A frequent goal in signal enhancement is to minimize the mean-squared error of an estimator. Within the framework of Bayesian risk theory, this MMSE criterion may be viewed as a squared-error cost function. Considering the corrupted signal model, Bayes' rule, and the prior distributions defined above, the optimal suppression rule in an MMSE sense is $H_k = \dfrac{\lambda_x}{\lambda_x + \lambda_d}$, which is recognizable as the well-known Wiener filter [6]. The estimation of the clean signal variance is not trivial. Using the ML estimator $\lambda_x \approx \lambda_y - \lambda_d$ leads to:

$$H_k \approx \frac{\lambda_y - \lambda_d}{\lambda_y} \approx \frac{\max\left[0, \left(|X_k|^2 - \lambda_d(k)\right)\right]}{|X_k|^2}. \qquad (1)$$

C. Spectral Subtraction

In the widely used form of Eq. (1), the Wiener suppression rule requires only the noise variance λ_d. The noise variance can often be estimated from periods when the speech signal is silent, between words, but estimating $\lambda_x(k)$ is more difficult. This Wiener estimator introduces a distortion to the estimated signal that is called musical noise. This is the audible bubbling heard during pauses, which is caused by the approximation in the design process. To reduce the distortions Boll [7] proposed the spectral subtraction rule, which is less aggressive and introduces less distortion but suppresses less noise.

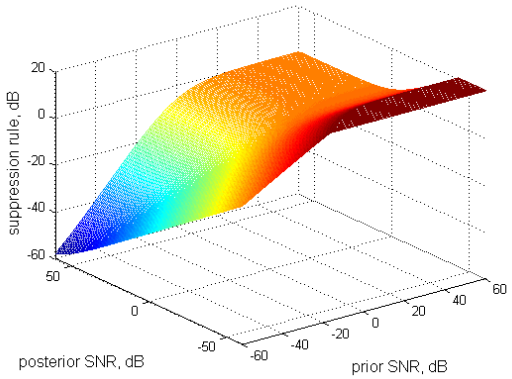

Figure 1. A typical suppression rule, in this case from ST MMSE, showing the "optimal" gain as a function of the prior and posterior SNR.

D. Maximum Likelihood

Later McAulay and Malpass [8] derive a maximum-likelihood (ML) spectral amplitude estimator under the assumption of Gaussian noise and an original signal characterized by a deterministic waveform of unknown amplitude and phase. This suppression rule is always greater than 0.5, which completely eliminates the musical noise, but reduces the amount of noise it suppresses.

E. Short-Time Minimum Mean-Squared Error (ST MMSE)

As an extension of the underlying model, Ephraim and Malah [9] derive a minimum mean-squared error short-time spectral-amplitude estimator based on the assumption that the Fourier expansion coefficients of the original signal and the noise may be modeled as statistically independent, zero-mean, Gaussian random variables. They introduce the *a priori* and *a posteriori* signal-to-noise ratios as:

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \quad \text{and} \quad \gamma_k \triangleq \frac{|Y_k|^2}{\lambda_d(k)} \qquad (2)$$

respectively. Their suppression rule is a function of these two SNRs: $H_k = f(\xi_k, \lambda_k)$. In the ST MMSE row of Table 1 $\mathrm{I}_0(\cdot)$ and $\mathrm{I}_1(\cdot)$ denote modified Bessel functions of zero and first orders, and $\nu_k \triangleq \dfrac{\xi_k}{1+\xi_k}\gamma_k$. This spectral magnitude estimator provides noise suppression while maintaining lower distortions and fewer artifacts. The shape of this suppression rule is shown in Figure 1.

F. Short-Term log-MMSE (ST log-MMSE)

Ephraim and Malah use the fact that humans hear sound pressure on a logarithmic scale to derive a suppression rule that is optimal in the MMSE log-spectral amplitude sense [10]. Regardless of the quite different criterion for optimality, the resulting suppression rule is surprisingly similar to the ST MMSE suppression rule. The mean of the difference between

the two gain rules is 1.12 dB, and the maximum difference is only 1.46 dB for $\xi_k \in [-30,+30]$ dB and $\gamma_k \in [-30,+30]$ dB.

The success of the Ephraim and Malah suppression rules is largely due to the authors' decision-directed approach (DDA) for estimating the *a priori* SNR ξ_k. For a given audio frame *n*, the decision-directed *a priori* SNR estimate $\hat{\xi}_k$ is given by a geometric weighting of the SNR in the previous and current frames:

$$\hat{\xi}_k = \alpha \frac{\left|\hat{X}_k(n-1)\right|}{\lambda_d(n-1,k)} + (1-\alpha)\max\left[0, \gamma_k(n)-1\right], \alpha \in [0,1) \quad (3)$$

Using the suppression rule to remove more noise often adds more distortion to the speech enhancement output. A way to mitigate this is using psychoacoustic-based speech enhancement algorithms. They estimate the masking threshold of human hearing and do not remove noise, or limit its removal to the level which humans can't hear. This estimation approach is applicable to any of the suppression rules and we are not going to discuss it further as it is not a suppression rule *per se*.

III. Model-Based suppression rules

Given a corpus of synthetic training data, which is a mixture of noise and speech signals in known proportions, we can measure the precise *a priori* and *a posteriori* SNRs of Eq. (2) for each audio frame and frequency bin, *k*. The perfect suppression rule has a desired gain of:

$$H_k^{(n)} = \frac{\left|X_k^{(n)}\right|}{\left|Y_k^{(n)}\right|}. \quad (4)$$

Given a sufficiently large training corpus, we have millions of data points as triplets $[\xi_i, \gamma_i, H_i]$: the *a priori, a posteriori* SNRs and desired gain. This converts the design problem into modeling problem $H(\xi,\gamma)$, which we can tackle with large-data machine-learning techniques.

Considering the shape of the known suppression rules, which largely varies between 0 and 1, we model the suppression rule with a logistic function [11] and define the following two-dimensional sigmoid function with 9 parameters:

$$\begin{aligned} H(\xi,\gamma,\mathbf{p}) &\triangleq \frac{P}{\left(1+\exp(-Q)\right)} + R, \text{ where} \\ P &\triangleq p_1 + p_2\xi + p_3\gamma \\ Q &\triangleq p_4 + p_5\xi + p_6\gamma \\ R &\triangleq p_7 + p_8\xi + p_9\gamma \\ \mathbf{p} &= [p_1, p_2, \ldots, p_9] \end{aligned} \quad (5)$$

The equations for *P*, *Q*, and *R* include linear terms of the prior and posterior probabilities to allow simple rotations/skews. The design problem then becomes one of finding the values of the parameters vector $\mathbf{p}$ that give a model with the closest fit to the desired suppression gain over all the training data. This optimization can be done in the linear domain:

$$\mathbf{p} = \arg\min_{\mathbf{p}} \left\{ \sum_n \left[x(nT) - \hat{x}(nT,\mathbf{p}) \right]^2 \right\}, \quad (6)$$

or in the logarithmic domain:

$$\mathbf{p} = \arg\min_{\mathbf{p}} \left\{ \sum_n \left| \log\left(x(nT)\right) - \log\left(\hat{x}(nT,\mathbf{p})\right) \right|^2 \right\}. \quad (7)$$

These two amplitude estimators are optimal in the same sense as the ST MMSE and ST log-MMSE suppression rules respectively. Note that the minimization process is naturally self-weighted – we weight those regions where we have more triplets $[\xi_i, \gamma_i, H_i]$ more heavily than the regions with less data.

In prior work [12] the suppression rule is represented as a 10x10 matrix and the values of the matrix elements are optimized directly using the recognition rate of a speech recognizer. This effort brought minimal improvements. The reason for this is the large number of optimization parameters and the fact that we may try to optimize points for which we not have sufficient data in the training corpus. The suppression model in Eq. (5) conveniently solves these problems – it has a small parameter space and the parameters are the same for the entire region of useful prior and posterior SNRs. For each vector of parameters we can process the entire training corpus and compute the values of the objective sound quality $PESQ_l$, for each of the *L* files in the data corpus. Then the optimization problem is defined as:

$$\mathbf{p} = \arg\max_{\mathbf{p}} \left\{ \frac{1}{L} \sum_l PESQ_l \right\}. \quad (8)$$

By defining the problem in such a way we are able to tailor the suppression rule for specific set of input data (SNRs, type of noise), stressing more or less different regions of the enhancement space. In all of these cases we find a suppression rule optimal for the specific problem. We have the flexibility to include additional components in the optimization criterion, such as MSE and/or log-MSE. The full optimization in this case looks like:

$$\mathbf{p} = \arg\max_{\mathbf{p}} \left\{ \frac{1}{L} \sum_l \left[w_1 PESQ_l - w_2 MSE_l - w_3 logMSE_l \right] \right\}. \quad (9)$$

With this additional flexibility we can change the weights of the different components of the optimization criterion. Theoretically $\mathbf{w} = [0,1,0]$ should be equivalent to the ST MSE suppression rule, and $\mathbf{w} = [0,0,1]$ should be equivalent to the ST log-MSE suppression rule. In practice the MSE error (6) has a large effect where the speech signal has higher amplitudes, while log-MSE (7) gives more weight to the regions with lower amplitudes. This allows fine tuning of the received result, still keeping PESQ as the component with the highest weight.

IV. Experimental Results

Table 1 summarizes the two types of experiments we performed to evaluate our approach. The first experiments measure the ability of our parameterized model to fit the existing suppression rules, and measure their performance. A second

TABLE I. EXPERIMENTAL RESULTS

Rule	Ref./ Eq.	Formula	PESQ	LSD	MSE	Interp. Error, dB
Do nothing		$H_k \equiv 1$	2.332	12.45	1.89E-03	0.000
MMSE with DDA	[6], [9]	$H_k = \frac{\xi_k}{1+\xi_k}$	3.393	9.64	2.75E-05	0.871
Maximum Likelihood	[8]	$H_k = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{\xi_k}{1+\xi_k}}$	2.558	11.04	5.37E-04	0.013
Spectral Subtraction	[7]	$H_k = \sqrt{\frac{\xi_k}{1+\xi_k}}$	3.310	7.28	3.64E-05	0.436
Short Term MMSE	[9]	$H_k = \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\left[(1+\nu_k)\mathrm{I}_0\left(\frac{\nu_k}{2}\right)+\nu_k\mathrm{I}_1\left(\frac{\nu_k}{2}\right)\right]\exp\left(\frac{\nu_k}{2}\right)$	3.468	7.02	3.16E-05	1.050
Short Term log-MMSE	[10]	$H_k = \frac{\xi_k}{1+\xi_k}\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{\exp(-t)}{t}dt\right\}$	3.508	6.93	3.00E-05	1.140
Optimal MSE (Eq. 6, eq. 9) w = [0,1,0]	(5)	p=[0.4095, 0.0274, -0.0642, -9.2203, -0.4936, 0.7046, 2.7159, -0.0343, -0.0198]	3.667	6.95	2.42E-05	
Optimal log-MSE (Eq. 7, eq. 9) w = [0,0,1]	(5)	p=[-0.6886, 0.05656, -0.05879 -10.0, -0.4723, 0.7856 4.571, -0.07352, -0.03235]	3.600	6.65	3.98E-05	
Optimal Quality (Eq. 8, eq. 9) w = [1,1,0.01]	(5)	p=[2.2136, 0.0427, -0.0558, -4.9285, -0.5288, 0.6952, 2.7026, -0.0368, -0.0210]	3.722	6.70	1.27E-03	

set of evaluations uses the full optimization criteria. But first we describe our experimental setup.

A. Data

We created a data corpus that consists of two parts, training and testing, created from the corresponding parts of the TIMIT database [13]. Each data file contains ten randomly selected utterances from different speakers. To the clean speech files we added stationary Hoth noise [14] (standardized model of the noise spectrum in living rooms and office spaces) with the right levels to achieve SNRs ranging from -10 to $+50$ dB with step size of 5 dB, i.e. $-10, -5, 0,$ etc. Having the clean speech, the noise, and the mixture allowed us to calculate the exact prior and posterior SNRs, used for optimization and evaluation of the suppression rules:

$$\xi_k^{(n)} \triangleq \frac{\left|X_k^{(n)}\right|^2}{\lambda_d(k)} \quad \text{and} \quad \gamma_k^{(n)} \triangleq \frac{\left|Y_k^{(n)}\right|^2}{\lambda_d(k)}. \tag{10}$$

To avoid glitches, we limited the measured prior and posterior SNRs to fall between $[0.001, 1000]$. In addition, we limited all of the estimated values of the suppression rules to fall between $[0.001, 10]$. We optimized all of the suppression rules using the training-data corpus and we evaluated using the testing-data corpus. We used PESQ as the main evaluation parameter, but we also computed the average MSE and LSD as follows:

$$MSE \triangleq \frac{1}{N}\sum_n \left[x(n) - \hat{x}(n)\right]^2 \tag{11}$$

$$LSD \triangleq \sqrt{\frac{1}{NK}\sum_n\sum_k \left|10\log_{10}\left(\frac{\hat{X}_k^{(n)}}{X_k^{(n)}}\right)\right|^2}. \tag{12}$$

The process of conversion to the frequency domain and back was performed using the scripts provided in Tashev's book [15]. The first row of Table 1, the "Do Nothing" row, shows the measured speech degradation for the noisy test files without processing.

B. Comparison to Existing Suppression Rules

Our first group of experiments verified how well the parameterized suppression model fits the existing approaches (described in Section II), and evaluates their enhancement performance on our database. For each suppression rule, we found the model parameters, **p**, that minimize the mean-squared difference in the log domain between the desired (classic) suppression rule and the parameterized suppression model (Eq. (5):

$$\mathbf{p} = \arg\min_{\mathbf{p}}\left\{\sum_i \left[\log\left(H\left(\xi_i,\gamma_i,\mathbf{p}\right)\right) - \log\left(H_{rule}\left(\xi_i,\gamma_i\right)\right)\right]^2\right\}. \tag{13}$$

The prior and posterior SNRs varied over the range of $[-60\text{ dB}, +60\text{ dB}]$. The column "Interp. Error" in Table 1 shows the difference between the rule-based suppression function and the parameterized models. The errors are small and this shows that our suppression model (Eq. 5) can adequately represent the known suppression rules.

In terms of the speech enhancement results, the ML suppression rule has the lowest performance, as expected, and the Spectral Subtraction and Wiener suppression rules perform

substantially better. The best performing rules from the standard set are ST MMSE and ST log-MMSE.

C. Optimized Suppression Rules

We can improve our enhancement ability using the parameterized model and the optimization criteria proposed in Section III. The rows "Optimal MSE" and "Optimal log-MSE" show the results from the optimization criteria in equations (6) and (7). As expected, the performance is close to ST MSE and ST log-MSE suppression rules. The main benefit from this model-based approach is that the calculation of equation (5) is much faster than calculation of either of these two suppression rules. Figure 2 shows the shape of the "Optimal log-MSE" suppression rule. It keeps the shape of ST log-MSE rule show in Figure 1. Furthermore, Figure 3 shows the number of points we have in each 1x1 dB square of the prior and posterior SNRs. While we have data to cover most of the SNR space, we note that there are areas for which we do not have sufficient information – this justifies using the parameterized model.

The final row of Table 1 ("Optimal quality") contains the result optimized using Eq. (9). In this particular case we used weight vector $\mathbf{w} = [1.0, 1.0, 0.01]$ which, considering the values range of each parameter, gives most of the weight to PESQ. This is why this suppression rule performs best in PESQ terms, while the performance in MSE and LSD terms is lower compared to the other suppression rules.

Figure 4 shows the PESQ as function of the input SNR for this and some of the other suppression rules from Table 1. It shows the expected trend of having higher PESQ when the input SNR is higher, and the different performance of the classic suppression rules. The "Optimal quality" suppression rule outperforms all suppression rules for the entire range of the input SNRs. Informal blind listening tests with several audio professionals confirmed that the signals processed with this suppression rule sound audibly better than any other suppression rule.

V. Discussion and Conclusions

In this paper we propose to learn, from training data, suppression rules for speech enhancement algorithms. This approach addresses two issues. The first is the classic assumption that the magnitude of the speech signal follows a Gaussian distribution. Using a training corpus we take advantage of the statistical properties of the signal and noise, and how they combine, to find the parameters of a suppression function that optimizes any one of several criteria. An interesting question is why do two of the rules we learn perform similar to their equivalents – ST MMSE and ST log-MMSE suppression rules. A potential reason is the different way the prior SNR is defined and estimated. In Eq. (2) the SNRs are defined as a long term statistical parameter – proportion of the variations of the speech and noise signals. In reality they are estimated as instantaneous SNRs as in Eqs. (3) and (10). A computationally-efficient way to estimate these suppression rules is the primary benefit from this effort.

The second issue we address is the optimality of the suppression rule. Conventional suppression rules are optimal in the MSE sense, the ML sense, the log-MSE sense, etc. What we actually want is for humans to perceive the quality of the output signal as better, compared to the non-processed signal. In Section III we propose a methodology to find a suppression rule, for specific data corpus, that is optimal in an objective sound quality measure – PESQ. The new suppression rule fits in the existing speech enhancement framework and can easily replace any of the other suppression rules.

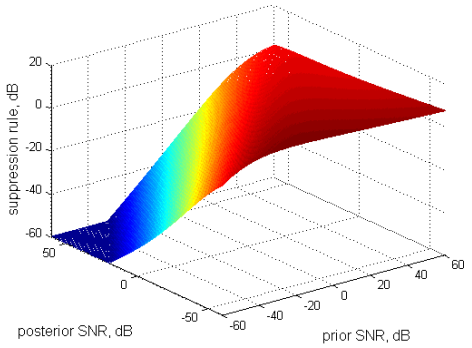

Figure 2. Optimal log-MSE suppression rule

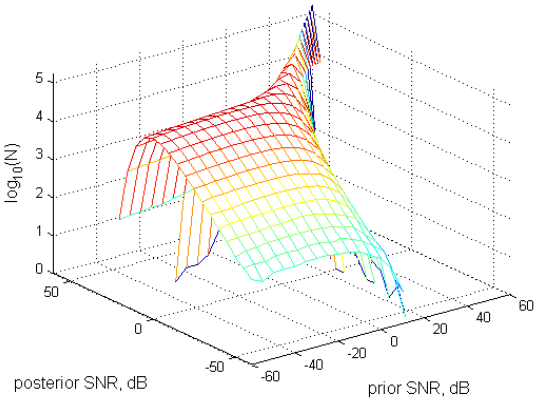

Figure 3. Number of points for each pair $[\xi, \gamma]$.

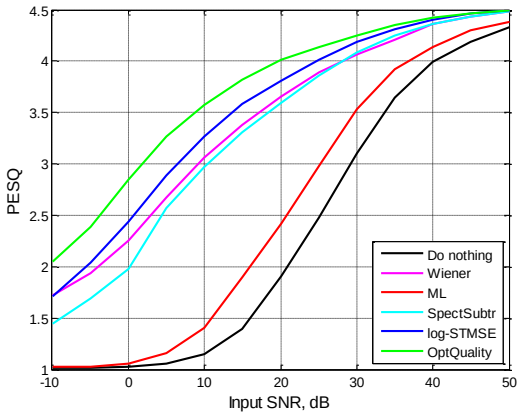

Figure 4. PESQ as function of the input SNR for various rules.

Overall, with direct optimization of the suppression rule using a perceptual quality measurement, we believe that the abilities of the existing speech enhancement framework to provide better sound quality are pretty much exhausted. To further improve the capabilities for speech enhancement we should utilize additional properties of the speech and noise signals. For example, the assumption that the frequency bins are statistically independent allows us to process the bins independently, but this is also not quite true. The speech signals in neighboring frequency bins are highly correlated, as are frequency bins containing the harmonics of the pitch signal. By assuming independence of the consecutive audio frames, we can process them efficiently, but the consecutive audio frames of the speech signal are also highly correlated. The existing framework only benefits from the temporal correlation when estimating the prior SNR (Eq. 3). Processing all frequency bins and several consecutive audio frames together to estimate the output audio frame provides a lot of opportunities for better quality enhancement. As modeling this process analytically is very complex, we believe that it can and should be approached using the science of machine learning from a large data corpus.

REFERENCES

[1] P. J. Wolfe and S. J. Godsill. "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement." In *Proceedings of the IEEE Workshop on Statistical Signal Processing*, pp. 496–499, 2001.

[2] S. Gazor, W. Zhang. "Speech probability distribution." *IEEE Signal Processing Letters*, vol. 10, No. 7, pp. 204–207, July 2003.

[3] I. Tashev and A. Acero. "Statistical Modeling of the Speech Signal." In *International Workshop on Acoustic, Echo, and Noise Control (IWAENC)*, Tel Aviv, Israel, 1 September 2010

[4] ITU-T Recommendation P.800. "Methods for subjective determination of transmission quality." Geneva, Switzerland, 1996.

[5] ITU-T Recommendation P.862. "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." Geneva, Switzerland, 2001.

[6] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Principles of Electrical Engineering Series. MIT Press, Cambridge, MA, 1949.

[7] S. Boll. "Suppression of acoustic noise in speech using spectral subtraction." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, pp. 113–120, 1975.

[8] R. J. McAulay and M. L. Malpass. "Speech enhancement using a soft-decision noise suppression filter." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, 1980.

[9] Y. Ephraim, D. Malah. "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator." *IEEE Trans. On Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, No. 6, December 1984.

[10] Y. Ephraim and D. Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[11] Pierre-François Verhulst. "Notice sur la loi que la population poursuit dans son accroissement." *Correspondance mathématique et physique* 10: 113–121, 1838.

[12] I. Tashev, J. Droppo, and A. Acero. "Suppression rule for speech recognition friendly noise suppressors." In *Proceedings of Eight International Conference Digital Signal Processing and Applications* DSPA'06, Moscow, Russia, March 2006

[13] John S. Garofolo, et al. "TIMIT acoustic-phonetic continuous speech corpus." Linguistic Data Consortium, Philadelphia, 1993.

[14] D. F. Hoth. "Room noise spectra at subscriber's telephone location." *Journal of the Acoustical Society of America*, vol. 12, pp. 499–504, 1941.

[15] I. J. Tashev. *Sound Capture and Processing: Practical Approaches*. pp. 388, Wiley, July 2009.