Unified Framework for Single Channel Speech Enhancement

Ivan Tashev[1)], Andrew Lovitt[2)], and Alex Acero[1)]
[1)] Microsoft Research, [2)] Microsoft Corporation
{ivantash, anlovitt, alexac}@microsoft.com

Abstract

In this paper we describe a generic architecture for single channel speech enhancement. We assume processing in frequency domain and suppression based speech enhancement methods. The framework consists of a two stage voice activity detector, noise variance estimator, a suppression rule, and an uncertain presence of the speech signal modifier. The evaluation corpus is a synthetic mixture of a clean speech (TIMIT database) and in-car recorded noises. Using the framework multiple speech enhancement algorithms are tuned for maximum performance. We propose a formalized procedure for automated tuning of these algorithms. The optimization criterion is a weighted sum of the mean opinion score (PESQ-MOS), signal-to-noise-ratio (SNR), log-spectral distance (LSD), and mean square error (MSE). The proposed framework provides a complete speech enhancement chain and can be used for evaluation and tuning of other suppression rules and voice activity detector algorithms.

1. Introduction

Aside from capturing the desired speech signal, the microphones in telecommunication systems capture ambient noise and the speech signal corrupted by room reverberation. As a result the captured speech signal has a decreased quality and thus the understanding requires higher cognitive load. Speech enhancement is a general term for signal processing algorithms aiming to improve various properties of the captured speech signal. Noise reduction (suppression and cancellation), echo reduction (cancellation and suppression) and de-reverberation are typical speech enhancement applications. The noise suppression, echo residual suppression, and some de-reverberation algorithms involve the application of a real valued, time varying filter to the frequency-domain transform of a captured signal.

In this paper we propose an architecture for a single channel speech enhancement framework for suppression based algorithms, and present a formalized procedure for the evaluation and the tuning of noise suppressors for achieving maximum performance. The optimality criterion utilized is a weighted sum of several evaluation parameters with major contribution from the MOS computed using the PESQ algorithm. As an illustration of the framework and the evaluation procedure, we provide a comparison of the most commonly used suppression rules where each one is optimized separately.

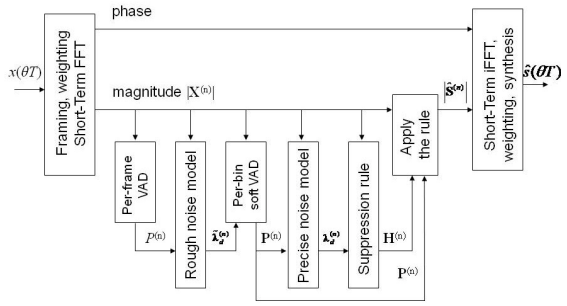

Figure 1. System design for the speech enhancement framework proposed in this paper.

2. Speech enhancement framework

The overall architecture of the speech enhancement framework is shown in Figure 1. Let $s(n\theta)$ represent values from a finite-duration analog signal sampled at a regular interval θ. The corrupted sequence is represented by the additive observation model $x(n\theta) = s(n\theta) + d(n\theta)$, where $x(n\theta)$ represents the observed signal at time index n, and $d(n\theta)$ is additive random noise, which is uncorrelated with the original signal. The goal of signal enhancement is to form an estimate $\hat{s}(n\theta)$ of the underlying signal $s(n\theta)$ based only on the observed signal $x(n\theta)$. In many implementations where efficient real-time performance is required, the set of observations is filtered using the overlap-add method of short-time Fourier analysis and synthesis. Taking the discrete Fourier transform on windowed frames yields K complex frequency bins

per frame: $X_k^{(n)} = S_k^{(n)} + D_k^{(n)}$. In this paper the frame index $^{(n)}$ is omitted wherever possible. Noise reduction in this framework may be viewed as the application of a suppression rule, or nonnegative real-valued gain $H_k^{(n)}$, to each bin k of the observed signal spectrum $X_k^{(n)}$, in order to form an estimate $\hat{S}_k^{(n)}$ of the original signal spectrum. In most cases the suppression rule is a function of the *a priori* and *a posteriori* SNRs. They are defined [5] respectively as:

$$\xi_k \triangleq \frac{\lambda_s(k)}{\lambda_d(k)},\ \gamma_k \triangleq \frac{\left|X_k\right|^2}{\lambda_d(k)} \tag{1}$$

Here $\lambda_d(k) \triangleq E\left\{\left|D_k\right|^2\right\}$ and $\lambda_s(k) \triangleq E\left\{\left|S_k\right|^2\right\}$ are the power spectra of the noise and speech signals.

The estimation of the suppression rule assumes a mixture of speech and noise, which is not always the case as pauses are an integral part of human speech. Therefore the suppression rule is modified to reflect this mixture with the probability of the speech signal presence: $\hat{S}_k^{(n)} = P_k^{(n)} H_k^{(n)} X_k^{(n)}$. This spectral estimate is then inverse-transformed to obtain the time-domain signal reconstruction. In a real valued suppression rule the phase of the estimated signal is the same as that of the input signal [2].

3. Voice activity detectors and noise models

The suppression rule estimation is based on statistical models of the noise and speech signal. It is assumed that the noise changes slower than the speech signal and can be modeled as a zero mean Gaussian process. Thus the model consists of one parameter per frequency bin – the noise variance $\boldsymbol{\lambda}_d^{(n)}$. The noise variances are estimated and updated during speech pauses, which are determined by a voice activity detector (VAD). The most commonly used algorithms use statistical methods and need a noise model to distinguish the signal from the noise. Due to this dependency a dual stage voice activity detector is proposed.

The first stage classifies audio frames by estimating the probability of the speech signal presence $\tilde{P}^{(n)}$ in the current frame. In many cases the first stage is signal energy based, and gives a binary decision: speech or noise. The signal energy is estimated after applying a filter $\mathbf{W}$ as $L^{(n)} = \sum_{k=0}^{K-1}\left(W_k . \left|X_k^{(n)}\right|\right)^2$. The filter is optimal in the sense that it maximizes the difference between the noise and speech segments:

$$\mathbf{W} = \arg\max_{\mathbf{w}}\left[\sum_{k=0}^{K-1}\left(W_k . \left|\overline{S}_k^{(n)}\right|\right)^2 - \sum_{k=0}^{K-1}\left(W_k . \left|\overline{D}_k^{(n)}\right|\right)^2\right], \tag{2}$$

where $\left|\overline{\mathbf{S}}\right|$ and $\left|\overline{\mathbf{D}}\right|$ are the average speech and noise magnitude spectra respectively. For typical speech and noise spectra this filter converges to a high pass filter with a cut-off frequency around 150-300 Hz. Some of the standardized weighting functions, such as [12], can be substituted as well. The noise floor can be adaptively tracked with two different time constants: one larger when the current level is higher than the estimate and one smaller, when the level is lower than the estimate. Thus the minimum level will follow changes down quickly and up slowly:

$$L_{\min}^{(n)} = \begin{cases} \left(1 - \frac{T}{\tau_{up}}\right) L_{\min}^{(n-1)} + \frac{T}{\tau_{up}} L^{(n)} & L^{(n)} > L_{\min}^{(n-1)} \\ \left(1 - \frac{T}{\tau_{down}}\right) L_{\min}^{(n-1)} + \frac{T}{\tau_{down}} L^{(n)} & L^{(n)} \le L_{\min}^{(n-1)} \end{cases}. \tag{3}$$

Here T is the frame duration, τ_{up} and τ_{down} are the two time constants, and $L_{\min}^{(n)}$ is the estimated noise floor. The decision for switching the state $V^{(n)}$ (speech $\tilde{\mathrm{H}}_1$ or noise $\tilde{\mathrm{H}}_0$) is threshold based, using two thresholds η_{down} and η_{up} ($\eta_{up} > \eta_{down}$):

$$V^{(n)} = \begin{cases} \tilde{\mathrm{H}}_0 & if \quad L^{(n)} / L_{\min}^{(n)} < \eta_{down} \\ \tilde{\mathrm{H}}_1 & if \quad L^{(n)} / L_{\min}^{(n)} > \eta_{up} \\ V^{(n-1)} & \text{otherwise} \end{cases} \tag{4}$$

Regardless of its simplicity, this binary decision VAD classifies the signal well. Utilizing the per-frame VAD decision a rough noise model $\tilde{\boldsymbol{\lambda}}_d^{(n)}$ is built and updated only during noise frames $\left(V^{(n)} = \tilde{\mathrm{H}}_0\right)$:

$$\tilde{\lambda}_d^{(n)}(k) = \tilde{\lambda}_d^{(n-1)}(k) + \frac{T}{\tau_r}\left(\left|\overline{X}_k^{(n)}\right|^2 - \tilde{\lambda}_d^{(n-1)}(k)\right). \tag{5}$$

Here τ_r is the updating time constant, $\left|\overline{X}_k^{(n)}\right|$ is the averaged magnitude of the input signal across several neighboring bins. The rough noise model is used by the secondary VAD, which provides the speech signal presence probability for each frequency bin $\mathbf{P}^{(n)}\left(\mathrm{H}_1 \mid \left|\mathbf{X}\right|^2\right)$. A statistically based VAD is published in [9]. Assuming a Gaussian distribution of the noise and speech signal, with variances $\tilde{\lambda}_{d,k}^{(n)}$ and $\tilde{\lambda}_{s,k}^{(n)}$ respectively, we consider two hypotheses: H_1 (speech plus noise) and H_2 (noise only) with probability density functions:

$$p\left(X_k \mid \mathrm{H}_0\right) = \frac{1}{\pi\lambda_{d,k}} \exp\left(-\frac{\left|X_k\right|^2}{\lambda_{d,k}}\right) \tag{6}$$

Table 1. Suppression rules optimized in the framework.

Rule	Ref.	Formula
MMMSE	[1]	$H_k = \frac{\lvert X_k\rvert^2 - \lambda_d(k)}{\lvert X_k\rvert^2}$
MMSE with DDA	[1], [5]	$H_k = \frac{\xi_k}{1+\xi_k}$
Maximum Likelihood	[4]	$H_k = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{\xi_k}{1+\xi_k}}$
Spectral Subtraction	[3]	$H_k = \sqrt{\frac{\xi_k}{1+\xi_k}}$
Short Term MMMSE	[5]	$H_k = \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\left[(1+\nu_k)\mathrm{I}_0\left(\frac{\nu_k}{2}\right)+\nu_k\mathrm{I}_1\left(\frac{\nu_k}{2}\right)\right]\exp\left(\frac{\nu_k}{2}\right)$
Short Term log-MMSE	[6]	$H_k = \frac{\xi_k}{1+\xi_k}\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{\exp(-t)}{t}dt\right\}$
JMAP SAP	[7]	$H_k = \frac{1}{2(1+\xi_k)}\left(\xi_k + \sqrt{\xi_k^2 + 2(1+\xi_k)\frac{\xi_k}{\gamma_k}}\right)$
MAP SAE	[7]	$H_k = \frac{1}{2(1+\xi_k)}\left(\xi_k + \sqrt{\xi_k^2 + (1+\xi_k)\frac{\xi_k}{\gamma_k}}\right)$
MMSE SP	[7]	$H_k = \sqrt{\frac{\xi_k}{1+\xi_k}\left(\frac{1+\nu_k}{\lambda_k}\right)}$

$$p(X_k \mid \mathrm{H}_1) = \frac{1}{\pi(\lambda_{d,k}+\lambda_{s,k})}\exp\left(-\frac{\lvert X_k\rvert^2}{\lambda_{d,k}+\lambda_{s,k}}\right) \quad (7)$$

which yields a likelihood ratio

$$\Lambda_k \triangleq \frac{p(X_k \mid \mathrm{H}_1)}{p(X_k \mid \mathrm{H}_0)} = \frac{1}{1+\xi_k}\exp\left(\frac{\gamma_k\xi_k}{1+\xi_k}\right). \quad (8)$$

Using a first order HMM the authors apply a "hangover scheme" for smoothing the likelihood in time:

$$\Gamma_k^{(n)} = \frac{a_{01}+a_{11}\Gamma_k^{(n-1)}}{a_{00}+a_{10}\Gamma_k^{(n-1)}}\Lambda_k^{(n)}, \quad (9)$$

where a_{ij} is the probability for switching from state H_i to state H_j. The smoothed likelihood for speech signal presence is

$$\hat{\Lambda}_k^{(n)} = \frac{P(\mathrm{H}_0)}{P(\mathrm{H}_1)}.\Gamma_k^{(n)} \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \eta \quad (10)$$

which is either compared with certain threshold η for binary decision or used to estimate the probability:

$$P_k^{(n)} = \frac{\hat{\Lambda}_k^{(n)}}{1+\hat{\Lambda}_k^{(n)}}. \quad (11)$$

The speech presence likelihood for the entire frame is:

$$\Lambda^{(n)} = \exp\left(\frac{1}{K}\sum_k \log\Lambda_k^{(n)}\right). \quad (12)$$

Then is applied a smoothing procedure as in equations (9) and (10) with state change probabilities b_{01} and b_{10}.

The speech presence probability for the entire frame $P^{(n)}$ is estimated according to (11).

The presented soft VAD is just an example as any published VADs which provide a speech presence probability per frequency bin can be used and evaluated in the presented framework.

The speech signal is sparse in frequency domain and even with speech present in the current frame many frequency bins contain only noise. We can therefore update the noise variance in the precise noise model $\boldsymbol{\lambda}_d^{(n)}$:

$$\lambda_{d,k}^{(n)} = \lambda_{d,k}^{(n-1)} + \left(1-P^{(n)}\right)\left(1-P_k^{(n)}\right)\frac{T}{\tau_p}\left(\left\lvert X_k^{(n)}\right\rvert^2 - \lambda_{d,k}^{(n-1)}\right). \quad (13)$$

4. Suppression rules

Utilizing the known *a priori* and *a posteriori* SNRs defined in (1) and the magnitude $\left\lvert X_k^{(n)}\right\rvert$, the goal is to estimate the suppression rule which is optimal based on the criteria. A selection of the most commonly used suppression rules are presented in Table 1.

For a Gaussian distribution of the noise and speech signals the optimal suppression rule in the MMSE sense is the well known Wiener filter [1]:

$$H_k = \frac{\lambda_s(k)}{\lambda_s(k)+\lambda_d(k)} = \frac{\xi_k}{1+\xi_k}. \quad (14)$$

The estimation of the prior SNR is not trivial, using the ML estimator $\xi_k \approx \gamma_k - 1$ leads to:

$$H_k \approx \frac{\gamma_k - 1}{\gamma_k} = \frac{\max\left[0, \left(\lvert X_k\rvert^2 - \lambda_d(k)\right)\right]}{\lvert X_k\rvert^2}. \quad (15)$$

In the widely used form of equation (15), the Wiener suppression rule requires only the noise variance λ_d. This Wiener estimator introduces a distortion to the estimated signal which is called musical noise. This is the audible bubbling heard during the pauses, which is caused by the approximation in (15). To reduce the distortions in [3] Boll proposed the spectral subtraction rule, which is less aggressive and introduces less distortion but suppresses less noise.

Later McAulay and Malpass [4] derived a maximum-likelihood (ML) spectral amplitude estimator under the assumption of a 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.

As an extension of the underlying model, Ephraim and Malah [5] derive a minimum mean-square error short-time spectral amplitude estimator (ST MMSE)

assuming that the Fourier expansion coefficients of the original signal and the noise may be modeled as statistically independent, zero-mean, Gaussian random variables. Their suppression rule is a function of both the *a priori* and the *a posteriori* SNRs. In Table 1 $I_0(\cdot)$ and $I_1(\cdot)$ denote modified Bessel functions of zero and first order and $\nu_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k$. This spectral magnitude estimator provides noise suppression while maintaining lower distortions and fewer artifacts. The shape of this suppression rule is shown in Figure 2.

The fact that humans hear in a logarithmic scale of the sound pressure level is used in [6] where Ephraim and Malah derive a suppression rule which is optimal in the MMSE log-spectral amplitude sense (ST log-MMSE). Regardless of the quite different criterion for optimality, this suppression rule is surprisingly similar to the ST MMSE suppression rule. The mean of the difference between the two 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.

In [7] Wolfe and Godsil derive three more suppression rules, Joint Maximum A Posteriori Spectral Amplitude and Phase (JMAP SAP) Estimator, Maximum A Posteriori Spectral Amplitude (MAP SA) Estimator, and MMSE Spectral Power (MMSE SP) Estimator. The derived suppression rules are efficient to implement for real-time execution as they do not contain Bessel functions and exponents. While the derivation is interesting, the suppression rules have quite similar to ST-MMSE shape.

5. Prior SNR estimation and uncertain presence of the speech signal

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

$$\xi_k^{(n)} = \alpha \frac{\left|\hat{X}_k^{(n-1)}\right|^2}{\lambda_d^{(n)}} + (1-\alpha).\max\left[0,\left(\gamma_k^{(n)}-1\right)\right]. \quad (16)$$

The ability to estimate the *a priori* SNR enables the usage of equation (14), maximum likelihood, and spectral subtraction rules in their more precise forms.

All suppression rules so far were derived under the assumption that both speech and noise signals are present. The speech signal is sparse in both time and frequency, and in their paper McAulay and Malpass [4] use a modifier for the suppression rule:

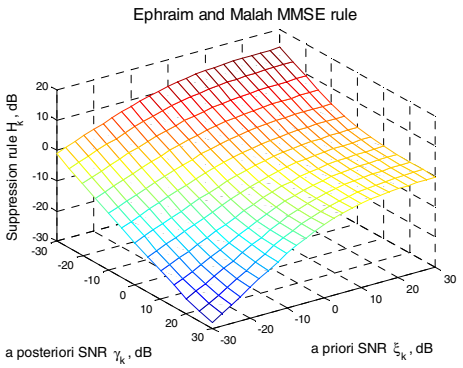

Figure 2. ST-MMSE suppression rule gain plotted against the *a priori* and *a posteriori* values.

$$\tilde{H}_k^{(n)} = P_k^{(n)}.H_k^{(n)}, \quad (17)$$

where $\tilde{H}_k^{(n)}$ is the modified suppression rule, $H_k^{(n)}$ is the estimated suppression rule under the assumption of speech signal presence and $P_k^{(n)}$ is the probability of a speech signal in the current frame. The derivation of (17) is generic and does not depend on any assumptions for specific distributions of the speech and noise.

Ephraim and Malah in [5] use an earlier work [8] to modify the suppression rule under the uncertain presence of a speech signal:

$$\tilde{H}_k = \frac{\Lambda(X_k,q_k)}{1+\Lambda(X_k,q_k)}.H_k \quad (18)$$

where $\Lambda(Y_k,q_k)$ is the generalized likelihood ratio:

$$\Lambda(Y_k,q_k) = \mu_k \frac{P(X_k \mid \mathrm{H}_1)}{P(X_k \mid \mathrm{H}_0)} \quad (19)$$

with $\mu_k \triangleq (1-q_k)/q_k$ and q_k is the probability of only noise signal in the k-th frequency bin. From equations (18) and (19) it is seen that from the detection and estimation standpoint the suppression rule modifier is just the probability of a speech signal presence, which is generalized in (17).

6. Practical aspects

Proper measures should be taken to prevent division by zero; usually this is implemented by adding a small real number to the denominator. In most runtime libraries the exponents or logarithms have their limitations for input parameters and provide invalid result beyond them. In such cases the argument value should be limited to meet these limitations. For example: the exp() argument in a double precision implementation should be limited to values below ~700.

Estimation of the *a priori* SNR using DDA converts the noise suppressor into a system with feedback. This raises stability concerns when the suppression gain has a value above one. Limiting the suppression rule to one guarantees the noise suppressor stability.

When the suppression gain goes to zero it causes musical noise. To reduce this effect in practice the suppression gains are limited to a certain minimal values and thus equation (17) becomes:

$$\tilde{H}_k^{(n)} = \left[G_{Um} + (1 - G_{Um}) P_k^{(n)} \right] \left[G_m + (1 - G_m) H_k^{(n)} \right] \quad (20)$$

Here G_m is the minimal gain for the suppression rule and G_{Um} is the minimal gain for applying the uncertain presence of the speech signal. These parameters are usually frequency independent.

7. Evaluation and tuning

The goal of speech enhancement and noise suppression is not actually to suppress noise. We do not want MMSE estimation, ML, or ST log-MMSE estimation of the speech signal. The actual goal is for humans to perceive the output as having a better quality, i.e. we want to maximize the perceptual sound quality.

The most commonly used criterion for perceptual speech quality is the Mean Opinion Score (MOS) [10]. This method involves real humans and is thus slow and expensive. The Perceptual Evaluation of Sound Quality (PESQ), standardized in [11], is a signal processing algorithm which provides an estimation of MOS in a fast and efficient way. Recently published results show high correlation between the PESQ and speech recognition rate [13]. Among the most common evaluation parameters for noise suppressors is the improvement in signal-to-noise-ratio (SNR), defined as the proportion of the average power of the signal and noise frames. It is usually measured in dB or in dBC if it is C-weighted. When the clean speech signal is available (either when using synthetically corrupted signals, or recorded with a close talk microphone) other commonly used evaluation parameters are the mean square error (MSE) and the log-spectral distance (LSD).

It is important that the tuning and evaluation of statistically based speech enhancement algorithms happen against a large corpus of data.

The speech enhancement framework described has a set of parameters that cannot be estimated directly. These parameters are the adaptation time constants $(\tau_{down}, \tau_{up}, \tau_r, \tau_p)$, thresholds (η_{down}, η_{up}), probabilities $(a_{01}, a_{10}, b_{01}, b_{10})$, geometry weightings (α_{VAD}, α), and limiting gains (G_m, G_{Um}). These parameters convert the evaluation and tuning process of each speech enhancement algorithm into a multidimensional optimization problem to find the best values of these parameters. The optimization metric can be the weighted sum of several optimization criteria:

$$Q = \sum_i w_i Q_i. \quad (21)$$

Once we have defined the optimization criterion the tuning procedure can use most of the algorithms for mathematical optimization, such as the simplex method and the entire variety of gradient based approaches.

It is common practice to split the evaluation corpus in three parts: the first is used with the optimization algorithm, while the second is evaluated after every iteration. The optimization procedure stops if several consecutive iterations show worse results against the second set. The optimal solution is the one which was best with regards to the second set of data. The final evaluation is conducted with the third set, which the optimization process did not use.

Table 2. Optimization results for rules in Table 1.

Rule	PESQ	SNR	MSE	LSD	G_m	G_{Um}
Base line	2.449	7.62	0.718820	7.654		
MMSE	2.839	22.11	0.434427	9.410	0.329	0.092
MMSE DDA	3.020	31.58	0.419704	10.256	0.020	0.100
ML	2.795	19.24	0.457038	8.188	0.001	0.018
SS	2.983	25.02	0.421912	9.394	0.001	0.083
ST-MMSE	3.003	28.15	0.419987	9.316	0.024	0.100
ST log-MMSE	3.018	30.49	0.420202	9.797	0.001	0.103
JMAP SAP	3.011	28.18	0.419490	9.392	0.020	0.100
MAP SAE	3.021	29.42	0.419297	9.640	0.020	0.100
MMSE SP	3.011	28.70	0.419637	9.468	0.001	0.097

8. Experimental results

The presented framework was used to tune and evaluate a speech enhancement algorithm for a communication system integrated into a car.

The evaluation corpus was synthetically generated. We measured the impulse responses between several potential positions of the driver's mouth and a set of microphone positions. The measurements were done by playing a chirp signal through a head and torso simulator and recording the response with the microphone. The same microphone was used for recording noises in various driving conditions such as a parked car, side street driving, busy street driving, and highway driving. These four basic scenarios were modified by turning the air conditioner on at various levels. As a clean speech source we used the 6300 utterances from TIMIT database [14]. Each utterance was convolved with a randomly selected impulse response and randomly selected noise segment was added. Proper correction for the Lombard effect was done on energy level based on the SNR. The corpus was divided in three parts with the same distribution of the SNRs. For tuning the parameters the steepest gradient descent

Table 3. Optimal parameter values for the VAD

τ_{down}	τ_{up}	η_{down}	η_{up}	τ_r	τ_p	a_{01}	a_{10}	b_{01}	b_{10}	α_{VAD}	α
0.035	20.0	1.1	2.9	1.39	0.825	0.37	0.11	0.41	0.73	0.940	0.974

algorithm was used. The optimization criterion Q_1 was PESQ with weight $w_1 = 1.0$. The other three criteria (SNR in dBC Q_2, MSE – Q_3, and LSD on dBC – Q_4) had weights $w_2 = 0.001$, $w_3 = -0.01$, and $w_4 = -0.001$. The optimization criterion is computed as the average result from all files in the set.

Each of the presented suppression rules was tuned separately using sets one and two. The results from the evaluation of the suppression rules with the third set are presented in Table 2. The SNRs and LSDs are measured in dBC. Overall the parameters of the two stage VAD converged to approximately the same values and the parameters that varied the most were the minimal gains. Table 3 records the optimal values of the VAD parameters, while the gain parameters for each suppression rule are shown in Table 2. All time constants in Table 3 are in seconds.

9. Discussion and conclusions

The presented single channel framework for suppression rule based speech enhancement is a flexible tool for evaluation and tuning of these algorithms.

The proposed formalized optimization procedure and evaluation criterion is applicable for many speech enhancement algorithms and their components. The VAD used is just an example and can be replaced with almost any of the published algorithms. It is difficult to determine how much improvement the optimization process brings, as it depends on the initial condition. In all cases the procedure saves time and resources over manual evaluation and tuning.

Using a proper evaluation corpus is critical for the successful tuning. The distribution of the SNRs and the noise characteristics should be as close as possible to the real conditions.

Overall the best performing algorithms with the corpus we used are MAP SAE and MMSE DDA which improve PESQ with 0.57 points. Among the best performers are JMAP SAP, MMSE SP, and ST log-MMSE. Increasing in LSD shows that all algorithms do not deal well with lower magnitudes and SNRs. We evaluated several other industrial noise suppression systems against the same corpus and they provided improvement in the range of 0.17–0.28 PESQ points, i.e. the proposed architecture and optimization methodology bring an audible improvement in perceptual sound quality.

10. References

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

[2] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[3] 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.

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

[5] 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, pp. 1109–1121, Dec. 1984.

[6] 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.

[7] P. Wolfe, S. Godsill. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. Proc. of 11th IEEE Workshop on Statistical Signal Processing, 496–499, 2001.

[8] D. Middleton, R. Esposito. "Simultaneous Optimum Detection and Estimation of Signals in Noise." *IEEE Trans. on Information Theory*, vol. IT-14, No. 3, May 1968.

[9] J. Sohn, N. Kim, W. Sung. "A statistical model based voice activity detector." *IEEE Signal Processing Letters*, vol. 6, No. 1, pp. 1-3, January 1999.

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

[11] 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.

[12] ITU-T Recommendation O.41. "Psophometer for use on telephone-type circuits". Geneva, Switzerland, 1994.

[13] P. Ding, J. Hao. "Assessment of Correlation between Objective Measures and Speech Recognition Performance in the Evaluation of Speech Enhancement". Proc. of Interspeech 2008, Brisbane, Australia.

[14] John S. Garofolo, et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, 1993.