Indirect Model-Based Speech Enhancement

Le Roux, J.; Hershey, J.R.

TR2012-016 March 2012

Abstract

Model-based speech enhancement methods, such as vector-Taylor series-based methods (VTS) [1, 2], share a common methodology: they estimate speech using the expected value of the clean speech given the noisy speech under a statistical model. We show that it may be better to use the expected value of the noise under the model and subtract it from the noisy observation to form an indirect estimate of the speech. Interestingly, for VTS, this methodology turns out to be related to the application of an SNR-dependent gain to the direct VTS speech estimate. In results obtained on an automotive noise task, this methodology produces an average improvement of 1.6 dB signal-to-noise ratio (SNR), relative to conventional methods.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

INDIRECT MODEL-BASED SPEECH ENHANCEMENT

Jonathan Le Roux, John R. Hershey

Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, Cambridge, MA 02139, USA

ABSTRACT

Model-based speech enhancement methods, such as vector-Taylor series-based methods (VTS) [1, 2], share a common methodology: they estimate speech using the expected value of the clean speech given the noisy speech under a statistical model. We show that it may be better to use the expected value of the noise under the model and subtract it from the noisy observation to form an *indirect* estimate of the speech. Interestingly, for VTS, this methodology turns out to be related to the application of an SNR-dependent gain to the direct VTS speech estimate. In results obtained on an automotive noise task, this methodology produces an average improvement of 1.6 dB signal-to-noise ratio (SNR), relative to conventional methods.

Index Terms— Speech enhancement, vector Taylor series, VTS, Algonquin, log power spectrum

1. INTRODUCTION

Model-based speech enhancement methods, such as vector-Taylor series-based methods (VTS) [1, 2, 3], use statistical models of both speech and noise to produce estimates of clean speech from noisy observations. In model-based methods, typically the clean speech is estimated directly by computing its expected value under the model, given the noisy observation. In this paper we argue that this is not necessarily the best approach. In fact, it may be better to do the reverse: use the expected value of the noise given the observation and subtract to *indirectly* estimate the clean speech.

Historically, different methods have been used for speech enhancement in quasi-stationary noise [4, 5, 6, 7] and feature-based noise compensation for automatic speech recognition (ASR) [1, 2, 8]. The latter methods are based on speech and noise models in log-spectrum-based feature domains. High-resolution versions of these methods [9] produce a reconstruction of the original speech signal, and their performance can thus be compared with traditional speech enhancement methods in quasi-stationary noise conditions. In feature-based noise compensation, the usual practice is to directly estimate the speech features, and use them to reconstruct the time domain estimate of speech. We show empirically that better performance can be obtained by reversing the process to first estimate the noise signal, and then subtract this estimate from the noisy speech signal. Surprisingly, this simple reversal of the estimation process yields an average improvement of 1.6 dB in signal-to-noise ratio (SNR) on an automotive noise task, relative to directly using the expected value of the speech.

2. VECTOR-TAYLOR SERIES-BASED METHODS

In *high-resolution* noise compensation techniques [9], the speech and noise are modeled by Gaussians or Gaussian mixture models in the short-time log-spectral domain, rather than in a feature domain having reduced spectral resolution, such as the mel spectrum typically used for speech recognition. This is done, along with using the appropriate complementary analysis and synthesis windows, for the sake of *perfect reconstruction* of the signal from the spectrum, which is impossible in a reduced feature set. Here we condition the short-time speech log power spectrum $\mathbf{x}_t$ at frame t on a discrete state s_t. We assume that the noise is quasi-stationary, so we posit only a single Gaussian for the noise log power spectrum $\mathbf{n}_t$:

$$p(\mathbf{x}_t, s_t) = p(s_t)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathsf{x}|s_t}, \boldsymbol{\Sigma}_{\mathsf{x}|s_t}), \quad p(\mathbf{n}_t) = \mathcal{N}(\mathbf{n}_t|\boldsymbol{\mu}_{\mathsf{n}}, \boldsymbol{\Sigma}_{\mathsf{n}}). \tag{1}$$

The *log-sum approximation*, used in [2, 9], uses the log of the expected value (with respect to the phase) in the power domain to define an interaction distribution over the observed noisy spectrum $y_{f,t}$ in frequency f and frame t:

$$p(y_{f,t}|x_{f,t}, n_{f,t}) \overset{\text{def}}{=} \mathcal{N}(y_{f,t} \mid \log(\mathrm{e}^{x_{f,t}} + \mathrm{e}^{n_{f,t}}), \psi_f), \tag{2}$$

where ψ_f is a variance intended to handle the effects of phase.

To perform inference in this model requires computing the following likelihood and posterior integrals

$$p(\mathbf{y}_t|s_t) = \int p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{n}_t)p(\mathbf{n}_t)p(\mathbf{x}_t|s_t)\ \mathrm{d}\mathbf{x}_t\ \mathrm{d}\mathbf{n}_t, \tag{3}$$

$$\mathrm{E}(\mathbf{x}_t|s_t) = \int \mathbf{x}_t\ p(\mathbf{x}_t, \mathbf{n}_t|\mathbf{y}_t, s_t)\ \mathrm{d}\mathbf{x}_t\ \mathrm{d}\mathbf{n}_t \tag{4}$$

$$= \int \mathbf{x}_t\ \frac{p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{n}_t)p(\mathbf{n}_t)p(\mathbf{x}_t|s_t)}{p(\mathbf{y}_t|s_t)}\ \mathrm{d}\mathbf{x}_t\ \mathrm{d}\mathbf{n}_t. \tag{5}$$

These integrals are intractable due to the nonlinear interaction function in (2). In iterative vector Taylor series, also known as Algonquin [10, 11], this limitation is overcome by linearizing the interaction function at the current posterior and iteratively refining the posterior. We describe this procedure here briefly, omitting the time index in cases when we are dealing with a single frame.

To simplify the notation, we concatenate $\mathbf{x}$ and $\mathbf{n}$ to form the joint vector $\mathbf{z} = [\mathbf{x}; \mathbf{n}]$, where ; indicates vertical concatenation. We define the prior $p(\mathbf{z}|s) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\mathsf{z}|s}, \boldsymbol{\Sigma}_{\mathsf{z}|s})$, where

$$\boldsymbol{\mu}_{\mathsf{z}|s} = \begin{bmatrix} \boldsymbol{\mu}_{\mathsf{x}|s} \\ \boldsymbol{\mu}_{\mathsf{n}} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_{\mathsf{z}|s} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathsf{x}|s} & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathsf{n}} \end{bmatrix}. \tag{6}$$

We also define the interaction function $g(\mathbf{z}) = \log(\mathrm{e}^{\mathbf{x}} + \mathrm{e}^{\mathbf{n}})$, where the logarithm and exponents operate element-wise on $\mathbf{x}$ and $\mathbf{n}$.

The interaction function is linearized at $\tilde{\mathbf{z}}_s$, for each state s, yielding:

$$p_{\text{linear}}(\mathbf{y}|\mathbf{z};\tilde{\mathbf{z}}_s) = \mathcal{N}(\mathbf{y}; g(\tilde{\mathbf{z}}_s) + \mathbf{J}_g(\tilde{\mathbf{z}}_s)(\mathbf{z} - \tilde{\mathbf{z}}_s), \Psi) \tag{7}$$

where $\mathbf{\Psi}$ is a diagonal covariance matrix with $\mathbf{\Psi}_{f,f} = \psi_f$ and $\mathbf{J}_g(\tilde{\mathbf{z}}_s)$ is the Jacobian matrix of g, evaluated at $\tilde{\mathbf{z}}_s$:

$$\mathbf{J}_g(\tilde{\mathbf{z}}_s) = \left.\frac{\partial g}{\partial \mathbf{z}}\right|_{\tilde{\mathbf{z}}_s} = \left[\mathrm{diag}(\tfrac{1}{1+\mathrm{e}^{\tilde{\mathbf{n}}_s - \tilde{\mathbf{x}}_s}}) \quad \mathrm{diag}(\tfrac{1}{1+\mathrm{e}^{\tilde{\mathbf{x}}_s - \tilde{\mathbf{n}}_s}})\right]. \quad (8)$$

It is then straightforward to derive the likelihood,

$$p(\mathbf{y}|s;\tilde{\mathbf{z}}_s) = \mathcal{N}(\boldsymbol{\mu}_{\mathsf{y}|s;\tilde{\mathbf{z}}_s}, \mathbf{\Sigma}_{\mathsf{y}|s;\tilde{\mathbf{z}}_s}), \quad (9)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\mathsf{y}|s;\tilde{\mathbf{z}}_s} &= g(\tilde{\mathbf{z}}_s) + \mathbf{J}_g(\tilde{\mathbf{z}}_s)(\boldsymbol{\mu}_{\mathsf{z}|s} - \tilde{\mathbf{z}}_s), \\ \mathbf{\Sigma}_{\mathsf{y}|s;\tilde{\mathbf{z}}_s} &= \mathbf{\Psi} + \mathbf{J}_g(\tilde{\mathbf{z}}_s)\mathbf{\Sigma}_{\mathsf{z}|s}\mathbf{J}_g(\tilde{\mathbf{z}}_s)^\top. \end{aligned} \quad (10)$$

The posterior state probabilities are then given by

$$p(s|\mathbf{y};(\tilde{\mathbf{z}}_{s'})_{s'}) = \frac{p(\mathbf{y}|s;\tilde{\mathbf{z}}_s)}{\sum_{s'} p(\mathbf{y}|s';\tilde{\mathbf{z}}_{s'})}. \quad (11)$$

The posterior mean and covariance of the speech and noise are

$$\begin{aligned} &\boldsymbol{\mu}_{\mathsf{z}|\mathbf{y},s;\tilde{\mathbf{z}}_s} = \\ &\quad \boldsymbol{\mu}_{\mathsf{z}|s} + \mathbf{\Sigma}_{\mathsf{z}|s}\mathbf{J}_g(\tilde{\mathbf{z}}_s)^\top \mathbf{\Sigma}^{-1}_{\mathsf{y}|s;\tilde{\mathbf{z}}_s}\Big(\mathbf{y} - g(\tilde{\mathbf{z}}_s) - \mathbf{J}_g(\tilde{\mathbf{z}}_s)(\boldsymbol{\mu}_{\mathsf{z}|s} - \tilde{\mathbf{z}}_s)\Big) \\ &\mathbf{\Sigma}_{\mathsf{z}|\mathbf{y},s;\tilde{\mathbf{z}}_s} = \Big[\mathbf{\Sigma}^{-1}_{\mathsf{z}|s} + \mathbf{J}_g(\tilde{\mathbf{z}}_s)^\top \mathbf{\Psi}^{-1}\mathbf{J}_g(\tilde{\mathbf{z}}_s)\Big]^{-1}. \end{aligned} \quad (12)$$

Iterative VTS updates the expansion point $\tilde{\mathbf{z}}_{s,k}$ in each iteration k as follows. The expansion point is initialized to the prior mean, $\tilde{\mathbf{z}}_{s,1} = \boldsymbol{\mu}_{\mathsf{z}|s}$, and it is subsequently updated to the posterior mean of the previous iteration, $\tilde{\mathbf{z}}_{s,k} = \boldsymbol{\mu}_{\mathsf{z}|\mathbf{y},s;\tilde{\mathbf{z}}_{s,k-1}}$. Because of this, although $p(\mathbf{y}|s;\tilde{\mathbf{z}}_{s,k})$ is Gaussian for a given expansion point, the value of $\tilde{\mathbf{z}}_{s,k}$ is the result of iterating and depends on $\mathbf{y}$ in a nonlinear way, so that the overall likelihood is non-Gaussian as a function of $\mathbf{y}$.

The posterior means of the speech and noise components are subvectors of $\boldsymbol{\mu}_{\mathsf{z}|\mathbf{y},s;\tilde{\mathbf{z}}_s} = [\boldsymbol{\mu}_{\mathsf{x}|\mathbf{y},s;\tilde{\mathbf{z}}_s}; \boldsymbol{\mu}_{\mathsf{n}|\mathbf{y},s;\tilde{\mathbf{z}}_s}]$. The conventional method uses the speech posterior expected value to form a *minimum mean-squared error* (MMSE) estimate of the log power spectrum:

$$\hat{\mathbf{x}} = \sum_s p(s|\mathbf{y};(\tilde{\mathbf{z}}_{s'})_{s'})\boldsymbol{\mu}_{\mathsf{x}|\mathbf{y},s;\tilde{\mathbf{z}}_s}. \quad (13)$$

For each frame t, the MMSE speech estimate is combined with the phase $\boldsymbol{\theta}_t$ of the noisy spectrum to produce the complex spectral estimate,

$$\hat{X}_t = \mathrm{e}^{\frac{\hat{\mathbf{x}}_t}{2} + \mathrm{i}\boldsymbol{\theta}_t}. \quad (14)$$

We shall refer to this estimate as the VTS MMSE.

3. PROPOSED METHOD

Model-based approaches typically combine noisy phases with estimated speech energies. This is problematic in situations where speech has significant energy but is still masked by noise. In these situations, the noisy phases are more appropriately combined with estimated noise energies. Model-based estimates of the noise accomplish precisely that. Thus, an interesting approach may be to indirectly compute the speech by subtracting the noise estimate from the noisy speech. In methods based on the log-sum approximation, such as VTS, the MMSE estimates of the signal and noise are indeed not symmetric, in the sense that they do not necessarily add up to the original signal. Therefore, an indirect approach will lead to a non-trivially different speech estimate than standard VTS.

3.1. Indirect VTS

We can write the noise MMSE estimate as

$$\hat{\mathbf{n}} = \sum_s p(s|\mathbf{y};(\tilde{\mathbf{z}}_{s'})_{s'})\boldsymbol{\mu}_{\mathsf{n}|\mathbf{y},s;\tilde{\mathbf{z}}_s}. \quad (15)$$

We can then subtract it from the observed speech to estimate the complex spectra:

$$\begin{aligned} \check{X}_t &= Y_t - \mathrm{e}^{\frac{\hat{\mathbf{n}}_t}{2} + \mathrm{i}\boldsymbol{\theta}_t} \\ &= \left(\mathrm{e}^{\frac{\mathbf{y}_t}{2}} - \mathrm{e}^{\frac{\hat{\mathbf{n}}_t}{2}}\right)\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}_t}, \end{aligned} \quad (16)$$

which we shall refer to as the indirect VTS log-spectral estimator. The latter expression is reminiscent of spectral subtraction, but it is more sophisticated: unlike spectral subtraction, the noise estimate being subtracted in a given time-frequency bin is estimated under statistical models of speech and noise, given the observation. In fact, it can be shown that, for small ψ_f, indirect VTS is approximately equivalent to an SNR-dependent suppression rule applied to the VTS estimate $\hat{X}_t$, with gain $g = \sqrt{r}/(\sqrt{1+r}+1)$, where $r = e^{\hat{\mathbf{x}}_t - \hat{\mathbf{n}}_t}$ is the VTS estimate of the SNR, and we neglect the influence of overlap-add in the resynthesis.

3.2. Acoustic model weights

In addition to the proposed estimation process, we investigated three other factors, each of which independently helps increase the average signal-to-distortion ratio (SDR) improvement in empirical evaluation. The first is to impose acoustic model weights α_f for each frequency f. These weights differentially emphasize the acoustic-likelihood scores as compared to the state priors. This only affects estimation of the speech-state posterior, which becomes:

$$p(s|\mathbf{y};(\tilde{\mathbf{z}}_{s'})_{s'}) = \frac{\prod_f p(\mathbf{y}_f|s;\tilde{\mathbf{z}}_{f,s})^{\alpha_f}}{\sum_{s'}\prod_f p(\mathbf{y}_f|s';\tilde{\mathbf{z}}_{f,s'})^{\alpha_f}}. \quad (17)$$

The weights α_f we used were inspired by the use in speech recognition of both pre-emphasis to remove low-frequency information and the mel-scale, which among other things de-emphasizes the weight of higher frequency components by differentially reducing their dimensionality. Thus the weights were chosen to follow a Gamma distribution over frequency with its mode at 1875 Hz, and a shape parameter of 37 Hz, such that the distribution decays to low values at 0 Hz and at the Nyquist frequency (8000 Hz).

3.3. MMSE Truncation

A second factor is the use of truncation to the region of feasibility to address errors in the VTS iterations. The exact log-sum model does not allow MMSE estimates of the speech or noise that are greater than the observation by any significant margin. However, in the VTS approximation, the speech and/or noise estimates can be much greater than the observation, depending on the linearization point. A simple remedy for this is to truncate the speech and noise estimates so that they do not exceed the observation. This has the effect of prompting a faster recovery from VTS optimization errors.

3.4. Noise estimation

A third factor investigated here concerns the estimation of the noise model's mean from a non-speech segment of data, assumed to occur in the first few frames. The conventional method is to estimate the noise model using the mean of the first few frames (assumed to be non-speech) in the log-spectral domain. Instead we investigated taking the mean in the power domain, so that

$$\boldsymbol{\mu}_{\mathrm{n}} = \log\left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{e}^{\mathbf{y}_t}\right). \qquad (18)$$

This has the benefit of reducing the influence of small outliers, and thus providing a smoother estimate. The variance about the mean was calculated in the usual way.

4. EVALUATION

4.1. Experimental conditions

The sampling rate was 16 kHz. Time-frequency analysis was performed using a frame length of 640 samples, 50% overlap and a sine window for analysis and re-synthesis.

The noisy speech data was obtained by synthetically mixing clean speech with car noise at various signal-to-noise ratio (SNR) levels. The speech data consisted of 4620 training utterances and 1676 test utterances by male and female speakers taken from the TIMIT training and test sets. The training and test sets had disjoint sets of speakers. The car noise data was randomly extracted from the CU-Move corpus [12], and additively mixed with the speech data at a random SNR uniformly sampled between -5 dB and 30 dB, taking into account speech activity. The speech model GMM consisted of 256 components which were trained on the clean speech training data.

4.2. Results

The outputs of the algorithms were quantitatively evaluated using the `bss_eval` toolbox [13]. The results are given in terms of signal-to-distortion ratio, signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR). Note that these quantities do not take into account speech activity, and the value of the SNR of the input computed with this toolbox is different (generally lower) than the value used to build the data. For consistency, the initial SNR used in the results is that of the `bss_eval` toolbox, and is thus not uniformly sampled between -5 dB and 30 dB, but roughly uniformly sampled between -10 dB and around 25 dB. For comparison, we show results for four classical speech enhancement algorithms: spectral subtraction ('SS') [14], Ephraim and Malah's amplitude estimator ('E&M MMSE') [4], Ephraim and Malah's log-amplitude estimator ('E&M log-MMSE') [5], and a more sophisticated algorithm combining Optimally-Modified Log Spectral Amplitude Estimator and Improved Minima Controlled Recursive Averaging ('OMLSA-IMCRA') [6, 7].

We first look at the behavior of the speech MMSE, referred to as 'VTS MMSE', and the speech obtained from the noise MMSE, refered to as indirect VTS speech estimate or simply 'indirect VTS'. The evolution of the SDR improvements depending on the VTS iteration number (1 meaning no re-estimation of the expansion point) is shown in Fig. 1. Focusing first on the red dashed curve, we see that, when the speech and noise posteriors are not truncated to the observation, the VTS MMSE can suffer unless at least two VTS iterations are performed. It is not clear that further improvements can be gained beyond the second iteration. Using the truncation technique described in Section 3.3 on the posteriors leads to an increase in SDR improvement from $+6.3$ dB to $+8.0$ dB for the VTS MMSE without iteration, and VTS iterations lead to no improvements in our setup. The VTS MMSE performances with and without truncation become very similar after VTS re-estimation.

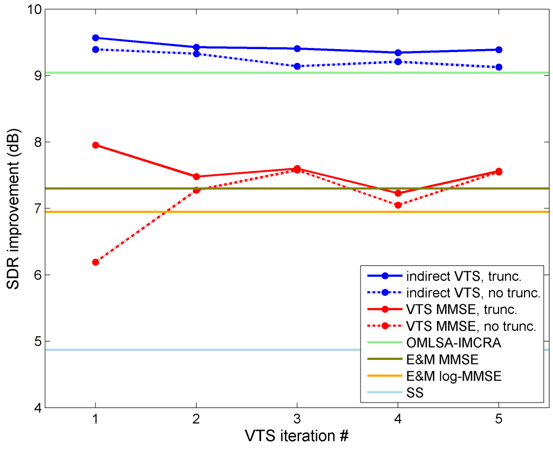

Fig. 1. Evolution of the SDR improvement depending on the VTS iteration number for the VTS MMSE and the speech obtained from the noise MMSE (indirect VTS), with and without truncation to the interaction function. Average SDR improvements for classical algorithms are shown for comparison.

Table 1. *Comparison of the mean SDR, mean SIR and mean SAR for four existing algorithms, VTS MMSE and the proposed indirect VTS method.*

Algorithm	SDR	SIR	SAR
No Processing	9.0	9.0	57.6
SS	13.9	18.3	17.3
E&M MMSE	16.3	20.8	19.1
E&M log-MMSE	16.0	19.3	19.6
OMLSA-IMCRA	18.1	22.9	20.5
VTS MMSE	17.0	19.3	21.6
indirect VTS	18.6	23.0	21.2

On the other hand, the proposed indirect VTS method, in blue, shows consistently high performance, and does not gain from VTS iterations. Whereas the VTS MMSE's performance is close to Ephraim and Malah's amplitude and log-amplitude estimators, we see that our method outperforms OMLSA-IMCRA on the task. Numerical results are presented in Table 1 as absolute values, and in Table 2 in terms of SDR improvements with respect to the initial SDR ('No Processing' in Table 1). Note that the SAR of the input is not infinite in practice due to numerical precision. We also show in Fig. 2 a comparison of the SDR improvements depending on the initial SNR range. In informal listening tests, the proposed approach led to better-sounding signals overall, more effectively cancelling the noise at the cost of a slight thinning of the sound.

We now consider the other experimental factors: the use of acoustic model weights in the likelihood, *aw* (Section 3.2), use of truncation of the posteriors to the observation, *tr* (Section 3.3), and estimation of the noise mean in the power domain, *pm* (Section 3.4). We show in Table 3 the SDR improvements obtained for the VTS MMSE and the indirect VTS when all three of these factors are used, *all*, when one of them is discarded, and when all three of them are discarded. We can see that each of them contributed significantly to improve the performance of both the VTS MMSE and indirect VTS. While indirect VTS seems less sensitive to the use of these factors, they

Table 2. *Comparison of the mean, median, minimum and maximum SDR improvements for four existing algorithms and the proposed indirect VTS method.*

Algorithm	mean	median	min	max
SS	4.9	4.7	-3.2	13.6
E&M MMSE	7.3	7.1	-3.4	17.8
E&M log-MMSE	6.9	6.9	-3.0	16.5
OMLSA-IMCRA	9.0	9.0	0.2	19.6
VTS MMSE	8.0	8.0	-10.7	17.7
indirect VTS	9.6	9.5	-3.1	19.7

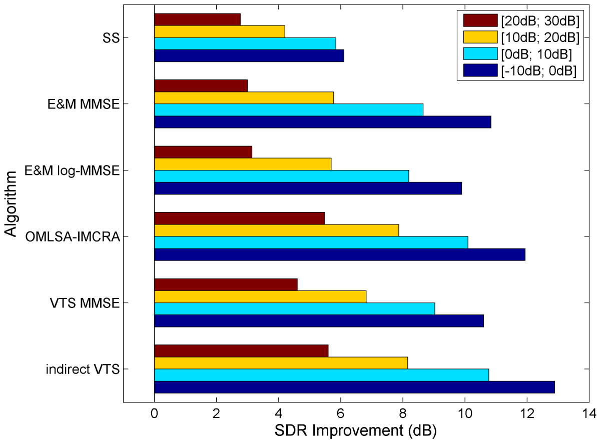

Fig. 2. Comparison of the average SDR improvement depending on the initial SNR range for the proposed indirect VTS method, VTS MMSE and four classical speech enhancement methods, including the state-of-the-art OMLSA-IMCRA.

each provided roughly an increase in average SDR improvement of $+0.2$ dB, altogether providing a $+0.6$ dB improvement.

We finally note that, although it may not result in an improvement of performance in terms of SDR, adding a small bias to the noise mean can lead to different trade-offs between SIR and SAR, which might be desirable depending on the application.

5. CONCLUSION

In this paper, we investigated an alternative to traditional model-based speech enhancement methods. Whereas these methods focus on reconstruction of the expected value of the speech given the observation, here we show improvements obtained via the expected value of the noise given the observation. Although conceptually the difference is subtle, the gains in enhancement performance on a simple VTS-based model are significant enough to warrant further investigation into this methodology for other model-based approaches.

6. REFERENCES

Table 3. *Influence of various factors on the performance in terms of SDR improvement for the VTS MMSE and the indirect VTS. $-$pm: no power-domain mean for the noise and use of log-domain mean instead; $-$tr: no truncation on the speech and noise MMSE; $-$aw: no acoustic model weights; $-$all: $-\{pm, tr, aw\}$; all: $\{pm, tr, aw\}$.*

Algorithm	all	$-$pm	$-$tr	$-$aw	$-$all
VTS MMSE	8.0	7.5	6.2	7.4	3.1
indirect VTS	9.6	9.4	9.4	9.3	9.0

[1] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proc. ICASSP*, May 1996, vol. 2, pp. 733–736.

[2] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero, “ALGONQUIN – learning dynamic noise models from noisy speech for robust speech recognition,” in *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pp. 1165–1171. 2002.

[3] T. Yoshioka and T. Nakatani, “Speech enhancement based on log spectral envelope model and harmonicity-derived spectral mask, and its coupling with feature compensation,” in *Proc. ICASSP*, May 2011, pp. 5064–5067.

[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 443–445, Apr. 1985.

[6] I. Cohen, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, 2002.

[7] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] S. J. Rennie, T. T. Kristjansson, P. A. Olsen, and R. Gopinath, “Dynamic noise adaptation,” *Proc. ICASSP*, May 2006.

[9] T. T. Kristjansson and J. R. Hershey, “High resolution signal reconstruction,” in *Proc. ASRU*, 2003, pp. 291–296.

[10] B. J. Frey, L. Deng, A. Acero, and T. T. Kristjansson, “ALGONQUIN: Iterating Laplace’s method to remove multiple types of acoustic distortion for robust speech recognition,” in *Proc. Eurospeech*, Sep. 2001, pp. 901–904.

[11] T. T. Kristjansson, H. Attias, and J. Hershey, “Single microphone source separation using high resolution signal reconstruction,” in *Proc. ICASSP*, May 2004, pp. 817–820.

[12] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, “”CU-Move”: Analysis & corpus development for interactive in-vehicle speech systems,” in *Proc. Eurospeech*, 2001.

[13] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] S. F. Boll, “Suppression of acousic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 3, pp. 113–120, Apr. 1979.