Speech Enhancement by Indirect VTS

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

TR2012-011 March 2012

Abstract

Model-based speech enhancement methods, such as vector-Taylor series-based methods (VTS), 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, relative to conventional methods.

Acoustical Society of Japan (ASJ) Spring Meeting

Speech enhancement by indirect VTS [*]

• •Jonathan Le Roux, John R. Hershey (MERL)

1 Introduction

Model-based speech enhancement methods, such as vector-Taylor series-based methods (VTS), 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, relative to conventional methods.

2 VTS-Based Methods

In high-resolution noise compensation techniques [1], the speech and noise are modeled by Gaussians or Gaussian mixture models in the short-time log-spectral domain for the sake of perfect reconstruction of the signal from the spectrum. Here we condition the short-time speech log 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 spectrum $\mathbf{n}_t$:

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

$$p(\mathbf{n}_t) = \mathcal{N}(\mathbf{n}_t|\boldsymbol{\mu}_{\mathsf{n}}, \boldsymbol{\Sigma}_{\mathsf{n}}). \tag{2}$$

The *log-sum approximation* [1] 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:

$$p(y_{f,t}|x_{f,t}, n_{f,t}) \overset{\text{def}}{=} \mathcal{N}(y_{f,t} \mid \log(\mathrm{e}^{x_{f,t}} + \mathrm{e}^{n_{f,t}}), \psi_f), \tag{3}$$

where ψ_f is a variance that handles phase effects.

The likelihood and posterior integrals required to perform inference in this model are intractable due to the nonlinear interaction function in (3). In iterative vector Taylor series, also known as Algonquin [2], this limitation is overcome by linearizing the interaction function at the current posterior and iteratively refining the posterior. At each step, this linearization leads to a linear Gaussian model. Denoting by $\tilde{\mathbf{z}}_s = [\tilde{\mathbf{x}}_s; \tilde{\mathbf{n}}_s]$ the linearization point for state s, we can easily obtain the posterior state probabilities $p(s|\mathbf{y}; (\tilde{\mathbf{z}}_{s'})_{s'})$, the posterior mean of the speech $\boldsymbol{\mu}_{\mathsf{x}|\mathbf{y},s;\tilde{\mathbf{z}}_s}$ and that of the noise $\boldsymbol{\mu}_{\mathsf{n}|\mathbf{y},s;\tilde{\mathbf{z}}_s}$, as well as their joint posterior covariance.

The conventional method uses the speech posterior expected value to form a *minimum mean-squared error* (MMSE) estimate of the log spectrum:

$$\hat{\mathbf{x}} = \sum_s p(s|\mathbf{y}; (\tilde{\mathbf{z}}_{s'})_{s'})\boldsymbol{\mu}_{\mathsf{x}|\mathbf{y},s;\tilde{\mathbf{z}}_s}. \tag{4}$$

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,

$$\hat{X}_t = \mathrm{e}^{\hat{\mathbf{x}}_t + \mathrm{i}\boldsymbol{\theta}_t}. \tag{5}$$

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. We can write the noise MMSE estimate as

$$\hat{\mathbf{n}} = \sum_s p(s|\mathbf{y}; (\tilde{\mathbf{z}}_{s'})_{s'})\boldsymbol{\mu}_{\mathsf{n}|\mathbf{y},s;\tilde{\mathbf{z}}_s}. \tag{6}$$

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

$$\check{X}_t = Y_t - \mathrm{e}^{\hat{\mathbf{n}}_t + \mathrm{i}\boldsymbol{\theta}_t} = \left(\mathrm{e}^{\mathbf{y}_t} - \mathrm{e}^{\hat{\mathbf{n}}_t}\right)\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}_t}, \tag{7}$$

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.

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:

$$p(s|\mathbf{y}; (\tilde{\mathbf{z}}_{s'})_{s'}) = \frac{\prod_f p(\mathbf{y}_f|s; \tilde{\mathbf{z}}_{f,s})^{\alpha_f}}{\sum_{s'}\prod_f p(\mathbf{y}_f|s'; \tilde{\mathbf{z}}_{f,s'})^{\alpha_f}}. \tag{8}$$

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).

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.

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.

[*] • • •VTS •(MERL) •

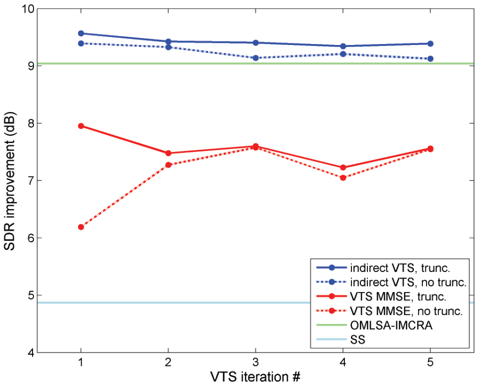

Fig. 1 *Evolution of the SDR improvement depending on the VTS iteration number for the VTS MMSE and the speech obtained from the noise MMSE (indirect VTS), with and without truncation to the interaction function. Average SDR improvements for classical algorithms are shown for comparison.*

The conventional method is to estimate the noise model using the mean of the non-speech frames in the log-spectral domain. Instead we investigated taking the mean in the power domain. 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

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 from the TIMIT database with car noise randomly extracted from the CU-Move corpus, at various randomly sampled signal-to-noise ratio (SNR) levels. The speech model GMM consisted of 256 components which were trained on the clean speech training data.

The results are given in terms of signal-to-distortion ratio, signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR). For comparison, we show results for two classical speech enhancement algorithms: spectral subtraction ('SS') [3] and the state-of-the-art algorithm combining Optimally-Modified Log Spectral Amplitude Estimator and Improved Minima Controlled Recursive Averaging ('OMLSA-IMCRA') [4, 5].

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 above 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. On the other hand, the proposed indirect VTS method, in blue, shows consistently high performance, outperforming OMLSA-IMCRA on the task, and does not gain from VTS iterations. Numerical results are presented in Table 1.

Table 1 *Comparison of the mean SDR, mean SIR and mean SAR for two 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
OMLSA-IMCRA	18.1	22.9	20.5
VTS MMSE	17.0	19.3	21.6
indirect VTS	18.6	23.0	21.2

Table 2 *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

We now consider the other experimental factors: the use of acoustic model weights in the likelihood, *aw*, use of truncation of the posteriors to the observation, *tr*, and estimation of the noise mean in the power domain, *pm*. We show in Table 2 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 each provided roughly an increase in average SDR improvement of +0.2 dB, altogether providing a +0.6 dB improvement.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] ——, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.