Speech enhancement using STFT of real and imaginary parts of modulation signals

Belinda Schwerin[1], *Kuldip Paliwal*[2]

Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Nathan QLD 4111, Australia
[1] belinda.schwerin@griffithuni.edu.au,[2] k.paliwal@griffith.edu.au

Abstract

This paper investigates an alternate modulation (RI-modulation) AMS-based framework for speech enhancement, in which real and imaginary parts of the modulation signal are processed in secondary AMS procedures. We propose to apply MMSE magnitude estimation in this framework, and using subjective experiments, show that MMSE RI-modulation magnitude estimation produces stimuli which is preferred by listeners over RI-modulation spectral subtraction. Experiments presented also show that while this framework is suited to speech enhancement and offers theoretical advantages over the modulation AMS framework, resulting stimuli had similar quality to that produced by the corresponding modulation AMS-based method.

Index Terms: speech enhancement, MMSE short-time spectral magnitude estimator, modulation magnitude spectrum

1. Introduction

Many of the commonly used speech enhancement methods improve the quality of speech with use of an analysis-modification-synthesis (AMS) framework, where the magnitude spectrum is modified, and processing is done (traditionally) in the acoustic domain. Recent works have suggested the suitability of the modulation domain for speech enhancement. Early efforts assumed speech and noise to be stationary and applied fixed filtering of the trajectories of the acoustic magnitude spectrum. However, speech and noise are known to be nonstationary, so short-time processing of the trajectories of the acoustic magnitude spectrum, in a secondary modulation AMS procedure, was proposed in [1].

In many of the acoustic and modulation AMS-based methods, acoustic phase spectral information is left unmodified, justified by the assumption that the acoustic phase spectrum is less important for speech enhancement. However, for lower signal-to-noise ratios (SNRs), the use of noisy phase can introduce a type of roughness distortion to the speech [2]. What is more, the modulation AMS-based methods, such as that proposed in [1], make the assumption that noise is additive in the modulation signal.

In an alternative modulation framework proposed by Zhang and Zhao [3] (denoted the RI-modulation framework), the trajectories of the real and imaginary parts of the acoustic spectrum are used to construct the modulation signals that are processed in the secondary modulation AMS procedures. As before, each of the modulation magnitude spectra are modified using an algorithm such as spectral subtraction (as was done in [3]) then the signal is reconstructed to produce the enhanced speech signal. This framework offers the advantage that noisy acoustic phase spectra are not used to reconstruct the speech signal. Furthermore, noise in the modulation signal is in fact additive in this framework. While the modulation spectrum calculated in this way does not have the same physical interpretation, it does offer mathematical advantages, and resolves some of the assumptions constraining the modulation AMS framework.

In this work, we propose the use of minimum mean-square error (MMSE) magnitude estimation in the RI-modulation framework (denoted RIMME) to enhance speech quality. Results presented show that RIMME provides improved speech quality compared with RI-modulation spectral subtraction. To investigate the advantages of this alternate framework, we also compare, the quality of stimuli enhanced using the RI-modulation framework with those enhanced using the same approach in the modulation AMS and acoustic AMS frameworks. Results presented show that the quality of stimuli generated using RI-modulation based methods are generally comparable to those of modulation AMS-based methods.

The rest of this paper is organised as follows. Section 2 describes the RI-modulation framework and proposed RIMME method, then a comparison with spectral subtraction in that framework is given in Section 3. Section 4 then presents experiments comparing enhancement methods across different AMS frameworks and conclusions are drawn in Section 5.

2. RI-modulation MMSE method

2.1. RI-modulation AMS framework

In the RI-modulation framework, similar to that proposed in [3], the trajectories of the real and imaginary parts of the acoustic spectrum are used to construct modulation signals, then each signal is processed in secondary modulation AMS frameworks. The modulation magnitude spectra are processed and the signals reconstructed to produce the enhanced speech signal. A block diagram of this AMS-based framework for speech enhancement is shown in Fig. 1.

2.2. MMSE magnitude estimation in the RI-modulation framework

In this work we propose to apply the minimum mean square-error (MMSE) magnitude estimation approach to the estimation of the modulation magnitude spectrum in the RI-modulation framework. In this method, denoted RIMME, the modulation magnitude spectrum of clean speech (for both the real and imaginary trajectories) are estimated from noisy observations so as to minimise the mean-square error between the modulation magnitude spectra of the clean and estimated speech. Similar to [4], we assume that speech and noise are additive in the short-time acoustic domain, and individual short-time modulation spectral components of the clean speech ($\mathcal{S}(\ell, k, m)$) and the noise ($\mathcal{D}(\ell, k, m)$) are independent, identically distributed Gaussian random variables.

The modulation magnitude spectrum calculated from the real acoustic spectrum of clean speech is estimated from the noisy modulation spectrum under the MMSE criterion (following [5]) as

$$\left|\hat{\mathcal{S}}_R(\ell, k, m)\right| = \mathcal{G}_R(\ell, k, m)\left|\mathcal{X}_R(\ell, k, m)\right| \tag{1}$$

where $\mathcal{G}_R(\ell, k, m)$ is the RIMME spectral gain function given by

$$\mathcal{G}_R(\ell, k, m) = \frac{\sqrt{\pi}}{2}\,\frac{\sqrt{\nu_R(\ell, k, m)}}{\gamma_R(\ell, k, m)}\,\Lambda\Big[\nu_R(\ell, k, m)\Big] \tag{2}$$

in which $\nu_R(\ell, k, m)$ is defined as

$$\nu_R(\ell, k, m) \triangleq \frac{\xi_R(\ell, k, m)}{1 + \xi_R(\ell, k, m)}\,\gamma_R(\ell, k, m) \tag{3}$$

and $\Lambda\left[\cdot\right]$ is the function

$$\Lambda\big[\theta\big] = \exp\Big(-\frac{\theta}{2}\Big)\Big[(1{+}\theta)\,I_0\Big(\frac{\theta}{2}\Big) + \theta\,I_1\Big(\frac{\theta}{2}\Big)\Big] \tag{4}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. In Eq. 2, $\xi(\ell, k, m)$ and $\gamma(\ell, k, m)$ are interpreted as the *a priori* and *a posteriori* SNR. The *a posteriori* SNR is estimated as

$$\hat{\gamma}_R(\ell, k, m) = \frac{|\mathcal{X}_R(\ell, k, m)|^2}{\hat{\lambda}_R(\ell, k, m)}. \tag{5}$$

where $\hat{\lambda}_R(\ell, k, m)$ is an estimate of $\lambda_R(\ell, k, m) \triangleq \mathrm{E}\Big[\big|\mathcal{D}_R(\ell, k, m)\big|^2\Big]$. The decision-directed approach in the short-time spectral modulation domain is used to estimate the *a priori* SNR, with its minimum value is limited to lower bound ξ_{min} (to prevent excessively low SNRs adversely effecting the resulting estimate), and the decision-directed smoothing parameter α controlling the trade-off between noise reduction and transient distortion [6, 5].

3. RI-framework experiments

As previously mentioned, MMSE magnitude estimation has been shown in [4] to work well in the modulation AMS framework (denoted MME), addressing many of the problems introduced by the spectral subtraction approach. In this work, we have proposed to apply MMSE magnitude estimation in this RI-modulation framework (denoted RIMME), and in this section, we now evaluate the quality of speech enhanced using RIMME compared to RI-modulation spectral subtraction (as proposed in [3], via subjective experiments.

Spectral subtraction in the RI-modulation framework (denoted RISSub) type stimuli were enhanced using the RI-modulation framework described by Figure 1 and spectral subtraction approach similar to modulation spectral subtraction (ModSSub) [1]. Parameter values used for RISSub (acoustic frame duration AFD = 25 ms; acoustic frame shift AFS = 2.5 ms; modulation frame duration MFD = 120 ms; modulation frame shift MFS = 15 ms; magnitude is the spectral domain for

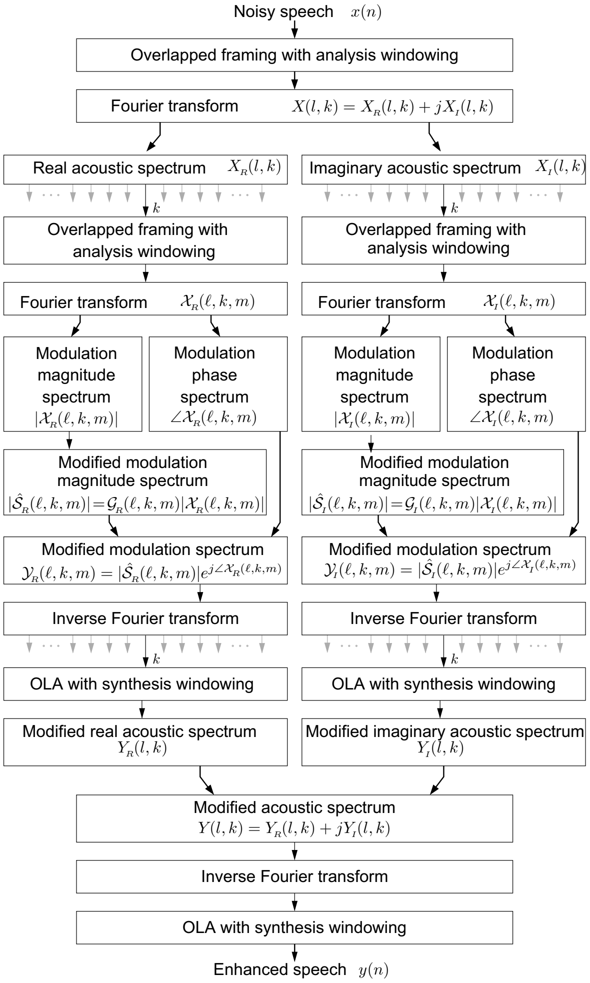
Noisy speech $x(n)$
Overlapped framing with analysis windowing
Fourier transform $X(l,k) = X_R(l,k) + jX_I(l,k)$
Real acoustic spectrum $X_R(l,k)$
Imaginary acoustic spectrum $X_I(l,k)$
k
k
Overlapped framing with analysis windowing
Overlapped framing with analysis windowing
Fourier transform $\mathcal{X}_R(\ell,k,m)$
Fourier transform $\mathcal{X}_I(\ell,k,m)$
Modulation magnitude spectrum $|\mathcal{X}_R(\ell,k,m)|$
Modulation phase spectrum $\angle\mathcal{X}_R(\ell,k,m)$
Modulation magnitude spectrum $|\mathcal{X}_I(\ell,k,m)|$
Modulation phase spectrum $\angle\mathcal{X}_I(\ell,k,m)$
Modified modulation magnitude spectrum $|\hat{\mathcal{S}}_R(\ell,k,m)| = \mathcal{G}_R(\ell,k,m)|\mathcal{X}_R(\ell,k,m)|$
Modified modulation magnitude spectrum $|\hat{\mathcal{S}}_I(\ell,k,m)| = \mathcal{G}_I(\ell,k,m)|\mathcal{X}_I(\ell,k,m)|$
Modified modulation spectrum $\mathcal{Y}_R(\ell,k,m) = |\hat{\mathcal{S}}_R(\ell,k,m)|e^{j\angle\mathcal{X}_R(\ell,k,m)}$
Modified modulation spectrum $\mathcal{Y}_I(\ell,k,m) = |\hat{\mathcal{S}}_I(\ell,k,m)|e^{j\angle\mathcal{X}_I(\ell,k,m)}$
Inverse Fourier transform
Inverse Fourier transform
k
k
OLA with synthesis windowing
OLA with synthesis windowing
Modified real acoustic spectrum $Y_R(l,k)$
Modified imaginary acoustic spectrum $Y_I(l,k)$
Modified acoustic spectrum $Y(l,k) = Y_R(l,k) + jY_I(l,k)$
Inverse Fourier transform
OLA with synthesis windowing
Enhanced speech $y(n)$

Figure 1: *Block diagram of the RI-modulation AMS-based framework for speech enhancement.*

subtraction; and spectral floor parameter β = 0.005), are those found to work best, and as reported in [3]. For consistency between methods, noise estimates for each approach were computed by recursive averaging of the relevant magnitude spectra in leading silence and updated during non-speech frames. For spectral subtraction, speech absences were determined using segmental SNR based voice activity detector (VAD). RIMME based stimuli were generated using an AFD of 32 ms, AFS of 1 ms, MFD of 32 ms, MFS of 2 ms, *a priori* SNR estimate lower bound of -25 dB, and decision-directed smoothing parameter of 0.998. Speech absence was determined using a log-likelihood based VAD.

For the purpose of this comparison, subjective experiments in the form of AB listening tests that determined listener preference, were conducted. For each stimuli pair, listeners were asked to make a subjective preference, with pair-wise scoring used to determine their preference score. Since use of the entire corpus was not feasible, experiments used four sentences (sp1, sp10, sp11 and sp27), belonging to two male and two female speakers from the Noizeus speech corpus [2]. Listening tests were conducted in two sessions, the first investigating additive white Gaussian noise (AWGN), and the second investi-

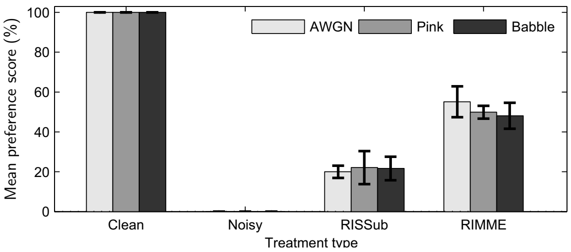

Figure 2: *Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded at 5 dB); and stimuli generated using the following treatment types: (c) RISSub [3]; and (d) RIMME.*

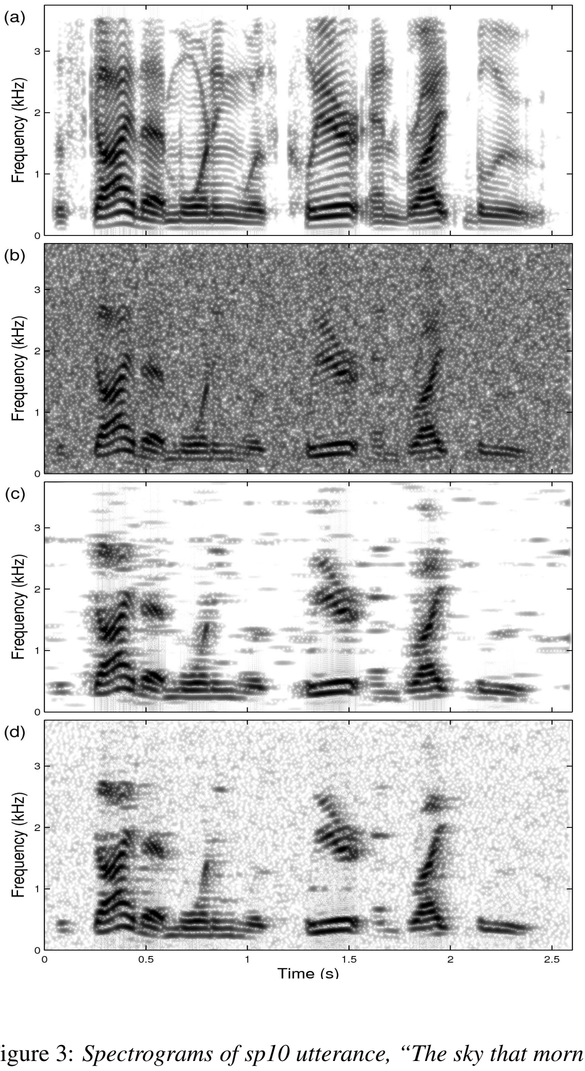

Figure 3: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) RISSub [3]; and (d) RIMME.*

gating pink and babble noise types. Eight listeners participated in each test.

Mean subjective preference scores for AWGN, pink and babble are shown in Figure 2. Results show that use of MMSE magnitude estimation in the RI-modulation framework provided considerable improvement compared to spectral subtraction in that same framework. Comparative scores were consistent across the noise types. These results are consistent with the improvement reported by MMSE modulation magnitude estimation over modulation spectral subtraction in [4].

Spectrograms of an example utterance and AWGN have also been included as Figure 3. Here we see that RISSub removes more background noise but also introduces a lot of perceptually annoying distortions seen as dark horizontal smears. These are not present in the RIMME stimuli, resulting in preference for RIMME over RISSub by listeners.

4. Comparison of AMS-based frameworks

In this section we compare the quality of stimuli enhanced using popular speech enhancement algorithms in the RI-modulation framework, with those enhanced using the modulation AMS framework [1] and the traditional acoustic AMS framework. The objective is to evaluate whether the theoretical benefits of the RI-modulation framework mentioned in the introduction, translate to improvements in quality compared to modulation AMS-based approaches.

4.1. Spectral subtraction experiments

In the first set of experiments we investigated the spectral subtraction approach. This is the method applied by [3] in the RI-modulation framework (RISSub). We compare it to the modulation spectral subtraction approach of [1] (ModSSub) and the acoustic spectral subtraction approach of [7] (SSub). SSub stimuli were generated using the reference implementation of [2]. ModSSub stimuli were enhanced using the modulation AMS framework and spectral subtraction algorithm as described in [1] (AFD = 32 ms; AFS = 8 ms; MFD = 220 ms; MFS = 27.5 ms; magnitude-squared spectral domain for subtraction and β = 0.002). RISSub stimuli were enhanced using the RI-modulation framework described by Figure 1 and spectral subtraction approach similar to ModSSub with parameters as given in the previous section. For consistency between methods, each used noise estimates computed by averaging the relevant magnitude spectra in the leading silence, updated during non-speech frames with speech absence determined using a segmental SNR based VAD.

Subjective experiments were used to compare the quality of processed stimuli. Again AB listening tests were used to determine listener preference. Eight listeners participated in each test. Experiments investigating stimuli corrupted with AGWN, then pink and babble noise types were conducted. Mean subjective preferences for each treatment and noise type are shown in Fig. 4. Subjective preference scores for AWGN show ModSSub to be preferred over RISSub, and both ModSSub and RISSub were preferred over SSub. Results for pink and babble noise indicate that, while still scoring ModSSub higher, there is little separation in the preference for ModSSub and RISSub stimuli.

For SSub, considerable musical noise can be heard in stimuli. In ModSSub and RISSub, on the other hand, musical noise is significantly reduced. However, both ModSSub and RISSub have other distortions introduced, but these were preferred by listeners over the musical noise of SSub. For ModSSub, there is a spectral smearing distortion that is heard as a type of reverberance. While speech is still intelligible, it is not as crisp as RISSub, which used a shorter MFD of 120 ms compared with

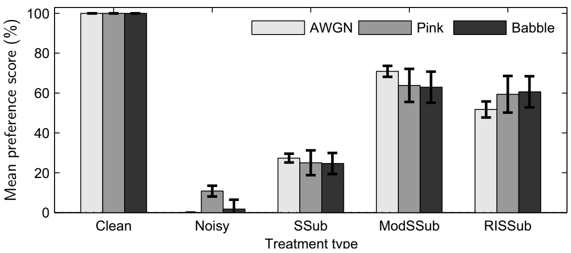

Figure 4: *Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded at 5 dB); and stimuli generated using the following treatment types: (c) SSub [7]; (d) ModSSub [1]; and (e) RISSub [3].*

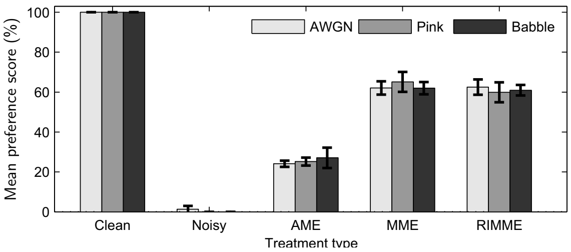

Figure 5: *Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded at 5 dB); and stimuli generated using the following treatment types: (c) AME [5]; (d) MME [4]; and (e) RIMME (proposed).*

220 ms used by ModSSub. However, RISSub had a ringing type of artifact which was disliked by listeners. Thus, while more background noise was removed by RISSub, the additional distortions introduced to RISSub made it less preferred. This was particularly the case for pink and AWGN stimuli.

From these results, we observe that while RISSub was a little more effective at removing noise than ModSSub, the distortion it introduced was disliked such that ModSSub was still generally preferred. Results also show that for less stationary noises such as babble, the difference in quality between RISSub and ModSSub stimuli was insignificant.

4.2. MMSE spectral estimation experiments

In both Section 3 and [4], MMSE magnitude estimation has been shown to improve speech quality compared to spectral subtraction in the respective AMS framework. In this section we now use subjective experiments to compare the effectiveness of the MMSE approach in each of the AMS frameworks being investigated.

Subjective experiments were conducted in the form of AB listening test, comparing the quality of clean, noisy, and noisy stimuli enhanced using MMSE-based enhancement methods: acoustic magnitude estimation [5] (AME), modulation magnitude estimation (MME) and RI-modulation magnitude estimation (RIMME). Type AME stimuli were generated using the reference implementation of [2]. Stimuli of type MME were generated using the modulation AMS framework as described in [4]. Stimuli of type RIMME were generated as described in Section 2. Both MME and RIMME stimuli were generated using an AFD of 32 ms, AFS of 1 ms, MFD of 32 ms, MFS of 2 ms, *a priori* SNR estimate lower bound (ξ_{min}) of -25 dB, and decision-directed smoothing parameter of 0.998. Each MMSE-based method calculates the initial noise estimate from the average relevant magnitude spectrum in leading silence, and is updated during non-speech frames, where speech absence is determined using a log-likelihood based VAD.

Mean subjective preference scores for each treatment and noise type are shown in Fig. 5. Subjective results for AWGN indicate MME and RIMME to be of similar quality and preferred over AME. Listening to stimuli, a clear improvement can be heard over AME, but the difference between MME and RIMME can not be distinguished. The residual noise in both is similar and non-distractive, and there is no musical noise apparent.

From these results, we therefore conclude that MMSE magnitude estimation performs equally well in the modulation and RI-modulation frameworks, with negligible difference between the resulting stimuli.

5. Conclusion

In this paper we have investigated the suitability of the RI-modulation framework for speech enhancement and proposed its use with MMSE magnitude estimation for improved speech quality. As found using the modulation AMS framework, the (proposed) MMSE RI-modulation magnitude estimation improved the quality of the processed stimuli compared to spectral subtraction in the RI-modulation framework. Results of subjective experiments evaluating spectral subtraction and MMSE magnitude estimation approaches in the RI-modulation framework, showed that stimuli had similar or reduced quality compared to those generated using the same approaches in the modulation AMS framework. These results suggest that while the RI-modulation framework did work well, the theoretical advantages have not resulted in an improvement in the overall quality of processed speech compared to those generated using the modulation AMS framework.

6. References

[1] Paliwal, K., Wojcicki, K. and Schwerin, B., “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain”, Speech Communication, 52(5), pp. 450–475, 2010.

[2] Loizou, P., “Speech Enhancement: Theory and Practice”, Boca Raton, FL: Taylor and Francis, 2007.

[3] Zhang, Y. and Zhao, Y., “Spectral subtraction on real and imaginary modulation spectra”, in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), Prague, pp. 4744–4747, 2011.

[4] Paliwal, K., Schwerin, B. and Wojcicki, K., “Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator”, Speech Communication, 54(2), pp. 282–305, 2012.

[5] Ephraim, Y. and Malah, D., “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”, IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no.6, pp. 1109–1121, 1984.

[6] Cappe, O., “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor”, IEEE Trans. Speech Audio Process., vol. 2, no. 2, pp. 345–349, 1994.

[7] Boll, S., “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113–120, 1979.