A comparative study on acoustic and modulation domain speech enhancement algorithms for improving noise robustness in speech recognition

Belinda Schwerin and Stephen So

School of Engineering and Built Environment,
Gold Coast Campus, Griffith University, QLD, 4222.
{b.schwerin, s.so}@griffith.edu.au

Abstract

This paper investigates whether modulation domain speech enhancement methods are better than corresponding acoustic domain methods when used as a preprocessor to automatic speech recognition. It is well known that linguistic information of speech is contained not only in the short-time magnitude spectrum but also in its temporal evolution. In addition, this study investigates whether popular metrics used in speech enhancement (such as PESQ, segmental SNR, STOI) are indicative of ASR performance. ASR experiments on the TIMIT speech corpus corrupted by various noises were performed to compare recent modulation domain methods with their acoustic domain variants.

Index Terms: modulation domain, robust speech recognition, speech enhancement

1. Introduction

It is commonly known that automatic speech recognition (or ASR) systems trained on clean speech will perform poorly when applied to speech that has been corrupted by environmental noise, in so-called mismatched conditions. A number of approaches have been investigated to mitigate the degrading effects of noise in ASR systems and these have been reported widely in the speech literature [1]. One popular approach is to apply a speech enhancement-based preprocessor on the noisy speech before it is passed to the ASR system. The premise for this approach is to use a speech enhancement algorithm to reduce the level of noise in order to improve the quality and intelligibility of the speech, which should assist in improving the ASR performance.

A large number of speech enhancement algorithms have been reported in the literature. We can classify all of these algorithms into two categories, based on the domain they process in. Typically, the short-time Fourier transform (or STFT) of the speech signal is processed, where the speech is windowed into short frames and a discrete Fourier transform is computed for each frame. In *acoustic frequency domain methods*, such as spectral subtraction [2], Wiener filtering [3] and MMSE-STSA (minimum mean squared error-short time spectral amplitude) [4, 5], the estimation is performed on the magnitude or power spectrum across all acoustic frequencies within each frame. A notable characteristic of acoustic frequency enhancement methods is that they generally process each short time frame independently without exploiting inter-frame dependencies that model the temporal dynamics of speech. For *modulation domain methods*, the estimation is performed on the modulation frequencies within the time trajectory of the magnitude [6] or the real/imaginary parts [7] of the STFT at each acoustic frequency; they are able to enhance the temporal dynamics of the power spectrum of the speech. Several modulation domain speech enhancement algorithms, such as modulation domain spectral subtraction [8], MMSE modulation magnitude estimation [9], and the modulation domain Kalman filter [10] have been reported to outperform their acoustic frequency domain analogues in terms of the quality of the enhanced speech.

In this paper, we investigate whether the advantage offered by recent modulation domain speech enhancement algorithms that are tuned for human listening, also results in improved performance in hidden Markov model (HMM)-based speech recognition. The feature vectors used in typical ASR systems comprise a parametric representation of the power spectrum, such as Mel-frequency cepstral coefficients (MFCCs) [11], as well as their first and second derivatives (also known as delta and delta-delta coefficients) [12], in order to exploit the temporal movements of the vocal tract. These derivatives have been shown in [13] to be equivalent to applying bandpass filters on the time sequences of spectral parameters (or TSSPs). Therefore, modulation domain enhancement methods have the potential to provide a better set of feature vectors for the ASR system.

Previous studies on modulation domain-based preprocessing in ASR, such as RASTA IIR filtering [6] and FIR-Slepian bandpass filtering [13], have demonstrated improvements in ASR accuracy when using basic filtering techniques in the modulation frequency domain. Therefore, this study examines the ASR performance when the noisy speech is preprocessed by recent and more sophisticated modulation domain speech enhancement algorithms. ASR experiments were performed on the TIMIT speech corpus [14] using the HMM Toolkit (or HTK) [15], that compare the performance across several acoustic and modulation domain enhancement methods for the white, F16 and babble noises. Phoneme correctness results from these experiments are presented along with speech quality (PESQ and segmental SNR) and intelligibility metric (STOI).

2. Method

In this paper we aim to evaluate the effect of applying modulation domain and RI (real and imaginary)-modulation domain based speech enhancement methods in the preprocessing stage, on the recognition rates of ASR. For this purpose, noisy stimuli were processed using various modulation domain, RI-modulation domain, and (for comparison) acoustic domain enhancement algorithms, then ASR experiments were conducted on these preprocessed speech stimuli. Details of these experiments are described below.

2.1. Speech corpus

The TIMIT speech corpus [14] was used for the ASR experiments. This corpus consists of 6300 utterances recorded from 630 different male and female speakers. The dataset is sampled at 16 kHz, and divided into training and testing sets. The training set consists of 3696 clean utterances from 462 speakers. The core test set, consists of 192 utterances from 24 speakers. Clean stimuli of the test set were corrupted with various noise types at input SNRs ranging from 0 dB to 20 dB. Noise types investigated include white (AGWN), babble, F16 and factory noises, and were generated with use of noise samples from the NOISEX-92 noise corpus [16]. Noisy test stimuli were processed by each speech enhancement method before their use in the ASR experiments.

2.2. Speech enhancement algorithms

The speech enhancement algorithms that were investigated include spectral subtraction, minimum mean-square error amplitude estimation, and Kalman filtering. Methods were implemented in the modulation domain, the RI-modulation domain, and for comparison in the acoustic domain. A total of 8 different enhancement methods were considered. Table 1 summarises each speech enhancement method evaluated, along with key parameters used in their implementation. Parameters applied for each are consistent with those given by the cited reference work and/or implementation.

2.3. ASR experiments

Automatic speech recognition (ASR) experiments made use of the TIMIT speech corpus (see section 2.1). The ASR model was generated using clean stimuli from the training set only. To prevent the biasing of the results, we removed the *sa** utterances from both the training and testing sets, as was done in [1]. For testing, noise corrupted stimuli from the core test set were first processed by each enhancement method described in section 2.2. Recognition tests were then conducted for each noise type, input SNR, and enhancement method type.

ASR experiments were conducted using an HTK-based triphone recogniser. Three states per HMM and 8 Gaussian mixtures per state were used. Consistent with [21], the set of 48 phonemes was reduced to 39 for testing. A frame size of 25 ms, and frame shift of 10 ms was used. MFCC features, energy coefficients, and first and second order derivatives were used, to give 39 coefficients in total. Cepstral mean subtraction was also applied. The bigram language model was used. Recognition used the Viterbi decoder, with no pruning factor, likelihood scaling factor of 8 and a penalty of 0. Recognition rates of phonemes were determined for each noise, input SNR, and treatment type in terms of correctness (Corr %).

Table 1: *Enhancement methods evaluated for preprocessing stage of ASR. Important parameters used for each method are also given. These include Acoustic frame duration (AFD), Acoustic frame shift (AFS), Modulation domain frame durati on (MFD), Modulation frame shift (MFS), Smoothing factor α, number of Linear Prediction Coefficients (LPCs) used to model speech p , and the number of LPCs used to model noise q.*

Method	Implementation details
AcSSUB	Acoustic domain spectral subtraction [2] AFD = 20 ms, AFS = 10 ms, Power spectral subtraction.[17]
MdSSUB	Modulation domain spectral subtraction [8] AFD = 32 ms, AFS = 8 ms, MFD = 256 ms, MFS = 32 ms, Power spectral subtraction
RISSUb	RI-modulation spectral subtraction [18] AFD = 25 ms, AFS = 2.5 ms, MFD = 120 ms, MFS = 15 ms, Magnitude spectral subtraction
LOGAcMME	Acoustic MMSE Log-amplitude estimation [5] AFD = 20 ms, AFS = 10 ms, $\alpha = 0.98$ [17]
LOGMME	Modulation MMSE Log-amplitude estimation [9] AFD = 32 ms, AFS = 1 ms, MFD = 32 ms, MFS = 2 ms, $\alpha = 0.996$
LOGRIMME	RI-modulation MMSE Log-amplitude estimation [19] AFD = 32 ms, AFS = 1 ms, MFD = 32 ms, MFS = 2 ms, $\alpha = 0.996$
AcKal [20]	Acoustic domain Kalman filtering AFD = 50 ms, AFS = 6.25 ms, $p = 20$, $q = 10$.
MdKal [10]	Modulation domain Kalman filtering AFD = 32 ms, AFS = 4 ms, MFD = 40 ms, MFS = 40 ms, $p = 4$, $q = 8$.

3. Results and discussion

Results for ASR experiments for each noise type are shown in Table 2. Objective evaluation of the enhanced utterances used in each experiment are also shown in terms of mean PESQ score. Table 3 shows mean segmental SNR and STOI for each treatment type.

For higher SNRs, LOGAcMME was shown to be very effective in improving speech recognition rates. However, for lower SNRs, the results were less clear. When dealing with white noise, AcKal, a method originally designed for compensating AWGN corrupted stimuli, resulted in the highest recognition performance at all SNRs. RISSUB and ModSSUB outperformed AcSSUB, and LOGRIMME and LOGMME outperformed LOGAcMME at lower SNRs. For utterances corrupted with F16 noise at low SNRs, we similarly found that processing with LOGRIMME resulted in better recognition rates than when LOGAcMME was used. When considering babble noise, LOGAcMME was in general found to be the highest performing method. However, for the spectral subtraction-based methods, MdSSUB and RISSUB were found to outperform AcSSUB. This trend was also noticed in the Kalman filters, where MdKal outperformed AcKal. The recognition accuracies for babble noise indicated some degree of consistency with PESQ scores shown in Table 2. However, in general for the other noise types, it was found that high PESQ was not indicative of good recognition accuracy. This was expected since PESQ was originally developed for measuring perceptual speech quality in speech coding applications for human listeners.

The intelligibility metric, short-time objective intelligibility measure (or STOI) [22], on the other hand, consistently gave preference to MdSSUB. On the other hand, segmental SNR gave preference to LOGMME for babble noise, and either LOGMME or LOGRIMME for white and F16 noise types. These results highlight the difference between various metrics and ASR recognition rates, and the difference between methods yielding improved human listener preference and those yielding better ASR rates.

Considering the results reported for LOGMME and LO-

Table 2: *TIMIT experimental results: mean PESQ scores and phoneme correctness (%) scores for babble, F16, factory, and white noises (clean corr = 75.82%). Highest scores are in bold.*

Algorithm	Mean PESQ					Corr (%)				
SNR (dB)	0	5	10	15	20	0	5	10	15	20
Noisy (babble)	1.75	2.10	2.45	2.79	3.13	24.64	34.81	46.16	57.66	66.24
AcSSUB	1.74	2.21	2.64	3.06	**3.44**	26.01	35.76	47.18	54.21	62.44
MdSSUB	1.99	**2.38**	2.73	3.04	3.33	33.58	43.05	53.65	62.79	69.08
RISSUB	1.88	2.31	2.70	3.05	3.37	31.21	40.48	51.52	61.43	67.43
LOGAcMME	**2.01**	**2.38**	**2.74**	3.08	3.39	**34.46**	**45.13**	**55.04**	**63.39**	**69.32**
LOGMME	1.87	2.31	2.72	**3.11**	**3.44**	32.48	39.80	49.33	58.15	65.29
LOGRIMME	1.90	2.33	2.73	**3.11**	**3.44**	32.03	40.10	49.67	58.62	65.70
AcKal	1.92	2.22	2.51	2.83	3.13	31.46	37.49	43.95	50.31	55.29
MdKal	1.88	2.26	2.62	2.95	3.25	31.58	40.23	49.76	58.29	64.06
Noisy (F16)	1.64	2.01	2.37	2.73	3.08	16.48	27.65	39.53	53.27	63.11
AcSSUB	1.91	2.38	2.84	**3.28**	**3.66**	28.52	40.02	50.75	58.76	64.98
MdSSUB	**2.32**	**2.63**	2.91	3.17	3.41	38.47	47.86	57.43	63.94	68.34
RISSUB	2.15	2.52	2.85	3.15	3.43	38.74	46.41	56.36	61.75	67.08
LOGAcMME	2.24	2.61	**2.94**	3.23	3.51	35.56	48.38	**59.30**	**66.10**	**69.27**
LOGMME	2.05	2.45	2.87	3.24	3.56	38.21	48.47	57.59	62.35	65.42
LOGRIMME	2.14	2.54	2.92	3.26	3.56	**40.68**	**50.70**	58.66	62.51	65.77
AcKal	2.13	2.51	2.88	3.20	3.47	39.34	49.00	56.13	62.49	65.86
MdKal	2.10	2.45	2.75	3.01	3.27	35.54	45.57	53.29	58.37	62.93
Noisy (white)	1.37	1.71	2.09	2.46	2.83	11.39	20.57	30.58	43.24	54.65
AcSSUB	1.69	2.19	2.66	3.11	**3.52**	24.51	36.98	47.06	57.00	64.90
MdSSUB	**2.20**	**2.53**	2.79	3.04	3.29	31.91	41.75	50.93	59.07	64.98
RISSUB	2.08	2.41	2.71	2.99	3.27	33.20	41.03	49.32	56.01	62.35
LOGAcMME	2.01	2.43	2.79	3.09	3.38	27.39	38.63	50.27	60.64	66.62
LOGMME	1.92	2.32	2.69	3.03	3.37	31.80	41.90	52.31	57.97	62.36
LOGRIMME	1.97	2.37	2.72	3.05	3.38	32.69	42.69	53.32	59.38	63.27
AcKal	2.14	2.51	**2.83**	**3.15**	3.43	**36.32**	**46.77**	**56.40**	**62.96**	**67.00**
MdKal	1.94	2.33	2.64	2.91	3.16	30.13	40.80	50.23	57.00	61.10

GRIMME, it was noted that these methods incorporated the use of a smoothing parameter α which provided a trade-off between musical type noise distortion and slurring in the resulting reconstructed speech. The value of α applied was determined experimentally using human listening tests. Therefore, a preliminary investigation was made to determine if this parameter might significantly impact on the resulting recognition rates. Results for experiments were conducted utilising various α values between 0.96 and 0.998 and the results are shown in Table 4. The recognition rates shown are for the first 110 utterances of the test corpus, with the stimuli being corrupted with 5 dB of babble noise. Results show that reducing the value of α improved recognition rates, particularly for LOGMME. This suggests that further improvement could be attained by further tuning of the modulation domain based algorithms so that they are optimised for speech recognition.

4. Conclusions

In this study, the ASR performance of modulation domain-based speech enhancement methods was compared with that of their acoustic domain counterparts, when used as a preprocessor of speech prior to the ASR feature extraction. The aim was to determine if recently reported modulation domain methods, which were tuned for human listening, would also offer additional advantages when coupled with an ASR system. In the ASR experiments performed on the TIMIT speech corpus, it was found that certain methods had performed differently and for different noises. For the spectral-subtraction algorithms, the modulation domain methods were found to be universally better than their acoustic-based ones in phoneme correctness for all noises. For the logMMSE-based methods, the acoustic domain variant was found to be more effective for all noises except for white noise. To explain why the modulation-domain MME methods were under-performing, preliminary tests were performed to determine if better tuning parameters could improve the ASR performance. The results demonstrated that further tuning have the potential to improve their competitiveness in ASR performance. Lastly, common speech enhancement metrics for speech quality and intelligibility were not found to be reliably indicative of ASR accuracy.

5. References

[1] K. Paliwal, J. Lyons, S. So, A. Stark, and K. Wójcicki, “Comparative evaluation of speech enhancement methods for robust automatic speech recognition,” in *International Conference of Signal Processing and Communication Systems (ICSPCS)*. Gold Coast, Australia: IEEE, Dec 2010, pp. 1–5.

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

[3] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley, 1949.

[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. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[5] ——, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr 1985.

[6] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 578–589, Oct 1994.

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

Table 3: *TIMIT enhancement evaluation in terms of mean STOI scores and segmental SNRs for babble, F16, factory, and white noises (Highest scores are in bold).*

Algorithm		Mean STOI					Mean Segmental SNR				
	SNR (dB)	0	5	10	15	20	0	5	10	15	20
Noisy (babble)		-	-	-	-	-	-6.871	-5.997	-5.395	-5.016	-4.778
AcSSUB		0.620	0.761	0.866	0.934	0.969	-3.058	1.090	5.147	9.294	13.530
MdSSUB		**0.670**	**0.793**	**0.884**	**0.944**	**0.975**	-3.436	0.784	4.839	8.946	13.174
RISSUB		0.667	0.792	0.883	0.941	0.971	-3.085	1.033	4.702	7.989	10.821
LOGAcMME		0.656	0.772	0.863	0.925	0.962	-4.015	0.216	4.284	8.350	12.429
LOGMME		0.643	0.763	0.862	0.933	0.971	**-1.954**	**1.951**	**5.868**	**9.963**	**14.103**
LOGRIMME		0.651	0.776	0.874	0.940	0.974	-2.226	1.785	5.691	9.573	13.255
AcKal		0.600	0.700	0.776	0.834	0.876	-7.252	-4.155	-1.643	0.415	2.128
MdKal		0.658	0.774	0.862	0.923	0.969	-3.306	0.998	4.947	8.571	11.797
Noisy (F16)		-	-	-	-	-	-6.934	-6.042	-5.423	-5.032	-4.788
AcSSUB		0.647	0.789	0.890	0.950	0.978	-2.281	1.898	5.954	10.086	14.203
MdSSUB		0.751	**0.856**	**0.924**	0.**964**	**0.984**	0.063	3.559	6.996	10.635	14.498
RISSUB		**0.754**	0.854	0.920	0.959	0.979	**0.821**	3.912	6.765	9.418	11.766
LOGAcMME		0.711	0.818	0.895	0.944	0.972	-1.573	2.266	5.962	9.658	13.485
LOGMME		0.687	0.800	0.893	0.951	0.979	-0.107	3.465	7.237	**11.013**	**14.866**
LOGRIMME		0.707	0.823	0.907	0.957	0.981	0.372	**3.922**	**7.429**	10.824	14.126
AcKal		0.729	0.817	0.882	0.926	0.950	-1.563	1.607	4.261	6.615	8.761
MdKal		0.704	0.811	0.889	0.938	0.968	-1.185	2.656	6.164	9.402	12.299
Noisy (white)		-	-	-	-	-	-6.981	-6.073	-5.441	-5.042	-4.795
AcSSUB		0.601	0.762	0.880	0.948	0.979	-2.449	1.788	5.874	9.971	**14.027**
MdSSUB		0.718	**0.833**	**0.910**	**0.955**	**0.980**	0.421	3.541	6.715	10.080	13.691
RISSUB		0.706	0.822	0.901	0.948	0.974	0.926	3.697	6.340	8.849	11.140
LOGAcMME		0.686	0.797	0.881	0.935	0.967	-1.136	2.379	5.798	9.268	12.886
LOGMME		0.668	0.792	0.887	0.941	0.972	**0.234**	3.587	6.942	**10.250**	13.743
LOGRIMME		0.679	0.804	0.893	0.944	0.973	0.488	**3.783**	**6.976**	10.058	13.156
AcKal		**0.743**	0.831	0.900	0.945	0.970	-0.030	3.142	6.201	9.207	12.035
MdKal		0.685	0.798	0.884	0.939	0.968	-1.070	2.680	6.115	9.299	12.126

Table 4: *Effect of α on the recognition rates of LogMME. Test evaluated using model generated from training set, and reduced test set (dr1) containing 110 utterances and input SNR of 5 dB of babble noise (Highest scores are in bold)*

	LogMME$_M$	LogRIMME
α value	Corr (%)	Corr (%)
0.998	40.92	41.94
0.996	43.52	42.73
0.990	44.42	44.98
0.980	44.98	44.76
0.970	**47.13**	**45.43**
0.960	46.67	45.55

[8] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.

[9] K. Paliwal, B. Schwerin, and K. Wójcicki, "Single channel speech enhancement using MMSE estimation of short-time modulation magnitude spectrum," in *Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Florence, Italy, Aug 2011, pp. 1209–1212.

[10] S. So and K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, July 2011.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[12] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, Feb 1986.

[13] C. Nadeu, P. Pachés-Leal, and B.-H. Juang, "Filtering the time sequences of spectral parameters for speech recognition,," *Speech Communication*, vol. 22, no. 4, pp. 315–332, Sep 1997.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Engineering Department, Cambridge University, 2006.

[16] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul 1993.

[17] P. Loizou, *Speech Enhancement: Theory and Practice.* Boca Raton, FL: Taylor and Francis, 2007.

[18] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, no. 4, pp. 509–522, May 2013.

[19] B. Schwerin and K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Speech Comm.*, vol. 58, pp. 49–68, Mar 2014.

[20] S. So, A. George, R. Ghosh, and K. Paliwal, "A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement," *International Journal of Signal Processing Systems*, vol. 4, no. 4, pp. 263–268, Aug 2016.

[21] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[22] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar 2010, pp. 4214–4217.