A NON-INTRUSIVE QUALITY MEASURE OF DEREVERBERATED SPEECH

Tiago H. Falk and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada

ABSTRACT

A modulation spectral signal representation is investigated for non-intrusive quality measurement of reverberant and dereverberated speech. The representation is obtained by means of an auditory-inspired filterbank analysis of temporal envelopes of the speech signal. Modulation spectral cues are used to develop an adaptive measure which is shown to correlate well with subjective ratings of overall quality, colouration, and reverberation tail effects. The performance of the proposed measure is compared to that of four state-of-art quality measurement algorithms. Experiments show that substantial improvement is attained, in particular for reverberant speech enhanced by a delay-and-sum beamformer.

Index Terms— Objective quality measurement, modulation spectrum, non-intrusive, reverberation, dereverberation.

1. INTRODUCTION

With the advances in hands-free communication technologies, signal processing algorithms have been developed to combat unwanted reverberation effects. With reverberant speech, objective measures computed from the *measured* room impulse response (RIR), such as reverberation time and direct-to-reverberation energy ratio, are often used to characterize signal quality. With dereverberated speech, however, RIRs need to be *estimated* (e.g., via blind deconvolution) and often result in poor quality characterization. In addition, dereverberation algorithms can introduce audible artifacts to the speech signal; such artifacts are not captured by the estimated RIR. As a consequence, signal based quality measurement methods are needed. Today, subjective listening tests represent the most reliable method to quantify the perceived quality of dereverberated speech. Listening tests have also been used to characterize the subjective perception of colouration and reverberation decay tail effects [1]. Although reliable, subjective assessment tests are very expensive and time consuming, thus unsuitable for real-time processing.

For practical applications, signal-based objective quality measurement methods, which replace the listener panel with a computational algorithm, are needed. Objective quality measurement methods can be broadly classified as intrusive or non-intrusive. Intrusive measures depend on some form of distance metric between a clean reference signal (e.g., captured by a close-talking microphone) and the reverberant/dereverberated speech signal. Non-intrusive measures, on the other hand, do not depend on a reference signal and constitute a more challenging paradigm. In [1], several conventional *intrusive* measures, such as segmental signal-to-noise ratio, Bark spectral distortion, and cepstral distance, are tested as estimators of subjective perception of colouration, reverberation tail effects, and overall quality. It is reported that all measures attain poor correlation with subjective listening scores, thus signaling the need for more reliable estimators. In addition, since a clean reference signal is seldom available in practice, the development of a reliable *non-intrusive* measure is invaluable.

In this paper, reverberation cues obtained from the modulation spectrum are used to devise a non-intrusive signal based quality measurement tool. The proposed measure is tested on subjectively scored reverberant and dereverberated speech signals. Comparisons with several state-of-art intrusive and non-intrusive algorithms serve to demonstrate the gains in quality measurement performance obtained with the proposed measure. Experiments also suggest reliable estimation of subjective perception of colouration and reverberation tail effects.

2. MODULATION SPECTRAL REPRESENTATION

In this section, the motivation for using a modulation spectral representation is presented and the signal processing involved in the computation of the proposed measure is described.

2.1. Motivation

The motivation for using the modulation spectral representation originates from the fact that the diffuse reverberation tail can be modeled as an exponentially damped Gaussian white noise process. As the reverberation time increases, the signal attains more Gaussian white-noise like properties. In addition, it is known that the Hilbert temporal envelope can contain frequencies up to the bandwidth of its originating signal [2]. For clean (unreverberated) speech, Hilbert envelopes contain frequencies ranging from 2 Hz - 20 Hz with peaks at approximately 4 Hz, corresponding to the syllabic rate of spoken speech [3]. With reverberant speech, higher Hilbert envelope frequencies, henceforth referred to as modulation frequencies, are expected due to the "whitening" effect of the

Table 1. Modulation filter center frequencies (f_c) and bandwidths (BW) expressed in Hz.

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4.0	6.5	10.7	17.6	28.9	47.5	78.1	128.0
BW	2.4	3.9	6.5	11.0	18.2	29.1	47.6	78.8

reverberation tail. As such, features extracted from the modulation spectrum are expected to provide useful information for non-intrusive quality measurement and for estimation of colouration and reverberation tail effects.

2.2. Computation of Proposed Measure

To obtain the modulation spectral signal representation, the speech signal $s(n)$ is first filtered by a 23-filter critical-band gammatone filterbank to emulate cochlear signal processing. Filter center frequencies (termed *acoustic frequencies* to distinguish from modulation frequencies) range from 125 Hz to approximately half the sampling rate. Filter bandwidths are characterized by the equivalent rectangular bandwidth; in our simulations bandwidths range from 38 Hz to approximately 775 Hz (for 16 kHz sampling rate). The output signal of the j^{th} critical-band filter is denoted by $s_j(n) = s(n) * h_j(n)$, where $h_j(n)$ is the impulse response of the j^{th} filter. The Hilbert transform $\mathcal{H}\{\cdot\}$ is then used to obtain temporal envelopes $e_j(n)$ for each signal $s_j(n)$. Temporal envelopes are computed as $e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\{s_j(n)\}^2}$.

Temporal envelopes $e_j(n)$ are multiplied by a 256 ms Hamming window with 32 ms shifts; the envelope for frame m is represented as $e_j(m)$, where the time variable n is dropped for convenience. Here, 256 ms frames are used to obtain appropriate resolution for low modulation frequencies. The modulation spectrum for critical band j is obtained by taking the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m)$, i.e., $E_j(m; f) = |\mathcal{F}(e_j(m))|$ where f denotes modulation frequency bin. Modulation frequency bins are grouped into $\mathcal{K}$ bands in order to emulate an auditory-inspired modulation filterbank [4]. The k^{th} modulation band energy for frame m is denoted as $\mathcal{E}_{j,k}(m)$, $k = 1, \ldots, \mathcal{K}$. In our experiments, $\mathcal{K} = 8$ is used as it resulted in superior performance; center frequencies and bandwidths of the modulation filters are described in Table 1. Filters are second-order bandpass with quality factor $Q = 2$, as suggested in [4]. Modulation energy $\mathcal{E}_{j,k}(m)$ is then averaged over all active speech frames to obtain

$$\bar{\mathcal{E}}_{j,k} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \mathcal{E}_{j,k}^{act}(i), \tag{1}$$

where N_{act} denotes the number of active speech frames and $\mathcal{E}_{j,k}^{act}(i)$ the modulation energy of such frames.

As mentioned in Section 2.1, higher modulation frequencies are expected with reverberant speech due to reverberation tail effects. To verify this assumption, reverberant speech is generated by convolving 330 clean (anechoic) speech signals with room impulse responses measured by a linear microphone array in four different enclosures (reverberation time values of 274, 319, 422, and 533 ms). Additionally, a delay-and-sum beamformer is used to investigate the effects of multi-channel dereverberation on the modulation spectrum. The plots in Fig. 1 (a)-(b) depict average per-modulation band energy $\bar{\mathcal{E}}_k$ given by

$$\bar{\mathcal{E}}_k = \frac{1}{23} \sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k}, \tag{2}$$

averaged over all signals, for modulation bands $k = 1$ and $k = 6$, respectively. The plots depict modulation band energy of anechoic, reverberant, and dereverberated speech processed by the delay-and-sum beamformer (represented by DSB in the figure).

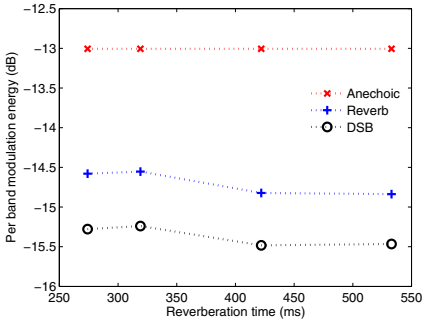
As seen from subplot (a), low-frequency modulation energy is reduced for reverberant and dereverberated speech signals. Such effects, however, are shown to be relatively *independent* of reverberation time and are likely due to early reflections. On the other hand, reverberation time dependency is observed for higher frequency modulation channels. From subplot (b), it can be seen that modulation energy increases almost linearly with reverberation time. Moreover, the delay-and-sum beamformer is shown to reduce high-frequency modulation energy by approximately 1 dB relative to reverberant speech. Such gains, however, are quite modest, as an approximate 6.5 dB difference remains between anechoic and dereverberated speech for a reverberation time of 533 ms.

Using this insight, an "adaptive" measure termed speech to reverberation modulation energy ratio (SRMR) is proposed for non-intrusive quality measurement of reverberant and dereverberated speech. The measure is given by

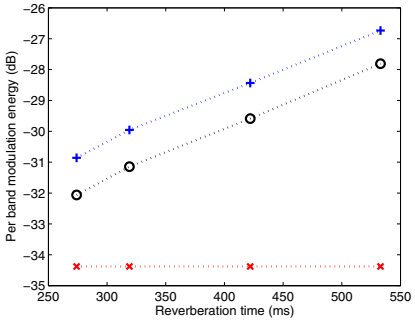
$$\text{SRMR} = \frac{\sum_{k=1}^{4} \bar{\mathcal{E}}_k}{\sum_{k=5}^{K^*} \bar{\mathcal{E}}_k} \tag{3}$$

and is adaptive as the upper summation bound K^* in the denominator is dependent on the speech signal under test. As mentioned previously, modulation frequency content for acoustic frequency band j is upper-bounded by the bandwidth of critical-band filter j. As such, speech signals with different acoustic frequency content, subjected to the same reverberation effects, should result in different modulation spectra.

Plots in Fig. 2 (a)-(b) illustrate one such example; subplots depict the percentage of modulation energy present per acoustic frequency band for speech signals from two different speakers with a reverberation time of 319 ms. As can be seen, for subplot (a), 90% of the total energy is obtained below 575 Hz; whereas for subplot (b), 90% of the total energy is obtained below 983 Hz. The bandwidths of the gammatone

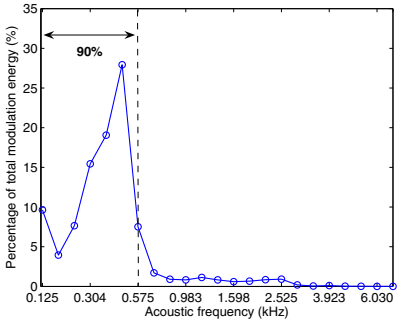


(a)

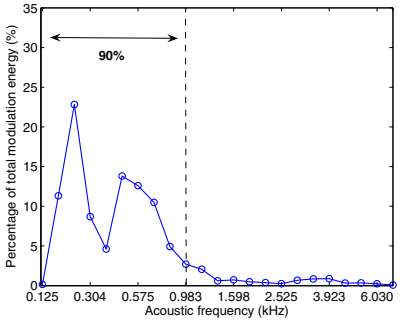


(b)

Fig. 1. Per-band modulation energy for modulation frequency band (a) $k = 1$, and (b) $k = 6$.



(a)



(b)

Fig. 2. Percentage of modulation energy, per acoustic frequency band, for speech signals from two different speakers.

filters centered at such frequencies are 86 Hz and 131 Hz, respectively. Hence, according to Table 1, negligible energy at modulation frequency band $k = 8$ is expected from the signal represented in subplot (a). In our simulations, K^* is chosen on a per-signal basis and depends on the bandwidth of the lowest gammatone filter for which 90% of the total energy is accounted for. As examples, for the speech signals represented in Fig. 2 (a)-(b), $K^* = 7$ and $K^* = 8$, would be used, respectively.

3. EXPERIMENTS

In this section, experiments with subjectively scored reverberant and dereverberated speech signals are described.

3.1. MARDY Database Description

A subjectively scored multi-channel acoustic reverberation database (MARDY) [1] is used in our experiments. The database uses room impulse responses which were collected with a linear microphone array in an anechoic chamber with reflective panels and with absorptive panels in place. Speaker to microphone distances varied between one to four meters (1-meter increments) and reverberation time values ranged from 291 ms to 447 ms. Reverberant speech was generated with the collected room impulse responses and anechoic speech from two speakers (one male and one female); additionally, three dereverberation algorithms were used. In our experiments, only reverberant speech and speech processed by a conventional delay-and-sum beamformer are used. Speech signals are digitized with 16-bit precision and stored with a 16 kHz sample rate. More detail regarding the development of the MARDY database can be found in [1].

A subjective listening test was performed following the guidelines described in [5]. In the test, 26 normal hearing listeners rated the subjective perception of colouration (COL), reverberation tail effect (RTE), and overall speech quality (OVL). Listeners used a 5-point scale where a rating of 5 indicated the best score and a rating of 1 the worst score for a given category. The individual ratings, averaged over all listeners, constitutes the widely used mean opinion score (MOS) [5]. It is noted that calibration speech examples were presented to the listeners in order to assist in identification and quantification of colouration and reverberation tail effects.

3.2. Experiment Results

The performance of the proposed measure is compared to that of four state-of-art algorithms: two intrusive, W-PESQ [6] and PEMO-Q [7], and two non-intrusive, P.563 [8] and ANIQUE+ [9]. For non-intrusive algorithms, a dowsampled (8 kHz) version of the MARDY database is required. Table 2 reports correlation values (ρ) attained between subjective scores and quality scores obtained with the four quality measurement algorithms and the proposed SRMR measure.

Table 2. Performance comparison between SRMR, SRMR*, W-PESQ, PEMO-Q, P.563, and ANIQUE+ on MARDY database. Average improvement is computed over the four quality measurement algorithms.

	Overall (reverberant + dereverberated)						Reverberant						Delay-and-sum					
Algorithm	COL	%↑	RTE	%↑	OVL	%↑	COL	%↑	RTE	%↑	OVL	%↑	COL	%↑	RTE	%↑	OVL	%↑
SRMR	0.82	–	0.83	–	0.80	–	0.81	–	0.84	–	0.81	–	0.85	–	0.83	–	0.79	–
SRMR*	0.73	36.2	0.80	16.6	0.77	12.1	0.73	28.8	0.83	5.9	0.81	0.0	0.72	45.8	0.75	33.7	0.72	22.1
W-PESQ	0.66	48.3	0.81	8.9	0.72	26.0	0.66	44.1	0.82	11.4	0.70	37.3	0.67	55.2	0.83	3.8	0.78	4.0
PEMO-Q	0.61	55.6	0.53	64.1	0.48	61.2	0.70	37.4	0.61	59.9	0.56	56.9	0.52	69.1	0.47	68.7	0.38	65.5
P.563	0.44	68.7	0.46	68.4	0.35	68.6	0.38	69.5	0.41	73.4	0.31	72.7	0.54	67.7	0.50	66.4	0.40	64.0
ANIQUE+	0.72	38.2	0.70	42.6	0.77	12.2	0.77	17.9	0.76	34.7	0.84	-15.3	0.67	54.7	0.57	61.5	0.67	34.5
Average	–	52.7	–	46.0	–	42.0	–	42.2	–	44.9	–	37.9	–	61.7	–	50.1	–	42.0

Additionally, to demonstrate the gains obtained with the adaptive SRMR measure, a comparison is also carried out with a non-adaptive measure. Denoted by SRMR* in the table, the non-adaptive version uses a fixed $K^* = 8$ value for all speech signals. The column labeled "%↑" lists the percentage improvement in correlation obtained by using SRMR relative to algorithm "X". The improvement is computed as

$$\% \uparrow = \frac{\rho_{SRMR} - \rho_X}{1 - \rho_X} \times 100\% \tag{4}$$

and indicates percentage reduction of the performance gap of algorithm X to perfect correlation.

As observed, the proposed measure is shown to reliably estimate the three quality dimensions for both reverberant and dereverberated speech. Overall, SRMR is shown to outperform intrusive and non-intrusive algorithms by an average 53%, 46%, and 42% for COL, RTE, and OVL, respectively. Additionally, improvements in performance of 36%, 17%, and 12% are attained relative to SRMR* for all data; more significant gains are obtained for dereverberated speech data. ANIQUE+ is shown to slightly outperform SRMR in OVL prediction for reverberant speech. Nonetheless, the capability of the proposed measure to reliably estimate colouration and reverberation tail effects, in addition to overall quality, make it a better candidate for non-intrusive evaluation of reverberant speech and of dereverberation algorithms.

4. CONCLUSION

Based on a modulation spectral signal representation, a speech to reverberation modulation energy ratio measure is proposed for *non-intrusive* quality measurement of reverberant and dereverberated speech. The performance of the proposed measure is compared to that of four state-of-art quality measurement algorithms and substantial improvement is reported. For reverberant speech, average improvements of 42%, 45%, and 38% are attained with the proposed estimators of colouration, reverberation tail effects, and overall quality, respectively. For dereverberated speech, the attained improvements are of 62%, 50%, and 42%, respectively.

5. ACKNOWLEDGEMENT

The authors would like to thank Mr. Petko Petkov for his assistance with running the ANIQUE+ algorithm and Mr. Jimi Wen for making the MARDY database available.

6. REFERENCES

[1] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. of the Intl. Workshop on Acoustic Echo and Noise Control*, 2006.

[2] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, pp. 87–90, March 2002.

[3] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Intl. Conf. Speech and Lang. Proc.*, Oct. 1996, pp. 2490–2493.

[4] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I – model structure," *Journal Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[5] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Intl. Telecom. Union, 2003.

[6] ITU-T P.862.2, "Wideband extension to Rec. P.862 for the assessment of wideband telephone networks and speech codecs," Intl. Telecom. Union, 2007.

[7] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[8] ITU-T P.563, "Single-ended method for objective speech quality assessment in narrowband telephony applications," Intl. Telecom. Union, 2004.

[9] ATIS-PP-0100005.2006, "Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," American National Standards Institute, 2006.