5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)
29-31 August 2016, Berlin, Germany

Objective Speech Quality Estimation of In-Ear Microphone Speech

João F. Santos[1,3], *Rachel E. Bouserhal*[2,3], *Jérémie Voix*[2,3], *Tiago H. Falk*[1,3]

[1]INRS-EMT, Université du Québec, Montréal, Canada
[2]ÉTS, Université du Québec, Montréal, Canada
[3]Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Canada

Abstract

Speech captured from an in-ear microphone (IEM) under an intra-aural device is beneficial in extremely noisy environments as it maintains a relatively high signal to noise ratio. Due to its limited bandwidth, speech enhancement is required in order to obtain a more natural speech. Consequently, quick and practical measurement of speech quality is important. In this paper, we compare the performance of the quality of intrusive and non-intrusive objective quality metrics on IEM speech, and propose an adaptation of a non-intrusive metric, the speech-to-reverberation modulation energy ratio (SRMR) to IEM speech signals. Changes are implemented to take into account the effect of the occluded ear on the recorded speech signals, which causes an amplification in the bone conduction sounds in the ear canal. We show that the updated SRMR metric, SRMR-IEM, significantly reduces the performance gap between non-intrusive and intrusive metrics.

Index Terms: speech quality, in-ear speech, speech enhancement

1. Introduction

In increasingly noisy work environments, solving the issue of good quality communication while maintaining proper hearing protection remains of significant interest. In particular, the use of in-ear microphones (IEM) to capture speech from occluded ears has gained attention [1, 2]. Unlike bone conduction microphones IEMs capture speech using air-conduction. This is beneficial as the IEM speech shares significant mutual information with speech captured in front of the mouth [3]. In noisy environments, it is advantageous to capture speech using an IEM as it is less susceptible to degradation from background noise [3]. Placed inside an occluded ear, after the passive attenuation of the earpiece, IEMs can capture a speech signal by way of the occlusion effect. The occlusion effect occurs when the ear canal is blocked allowing the energy from the ear canal wall vibrations to build up resulting in an amplification of the bone conduction sounds in the ear canal.

Although IEM speech can maintain a relatively high signal-to-noise ratio (SNR) in noisy environments, its quality suffers as a consequence of its limited frequency bandwidth. Originating from bone and tissue conduction and amplified via the occlusion effect, typically, the bandwidth of IEM speech is limited to 2 kHz [3]. In extremely noisy situations, even with a good passive attenuation from an earpiece noise can leak and is captured by the IEM in addition to the speech. To enhance the perceived quality of IEM speech, it is essential to first denoise the IEM speech from any residual noise and to extend its bandwidth in the high frequencies. Consequently, it is important to be able to quickly assess the quality of IEM speech and any subsequent enhancement. Such rapid quality measures could be considered and utilized to optimize speaker dependent factors in speech enhancement algorithms.

Objective speech quality metrics have been proposed as a way of making this process less time-consuming and laborious. Such metrics can be classified as intrusive, which require a clean reference signal that is compared to the distorted signal, and non-intrusive, which are reference-free. While both classes of metrics are useful, intrusive metrics are only suitable for offline system evaluation because of the reference signal requirement. Non-intrusive metrics, on the other hand, could be applied "in the loop" to fine-tune algorithm parameters on the fly. However, intrusive metrics usually have better performance than non-intrusive metrics as they are able to estimate distortions in a more direct way by comparing the corrupted/processed signal to the reference. Most speech quality metrics target speech in natural conditions or particular scenarios (such as telephony) and no metrics have been proposed to deal with IEM speech; however, some metrics have been adapted to hearing aid and cochlear implant users [4].

In this paper, we compare the performance of the quality of intrusive and non-intrusive objective quality metrics on IEM speech, and propose an adaptation of a non-intrusive metric, the speech to reverberation modulation energy ratio (SRMR) [5] to IEM speech signals. We show that the updated SRMR metric, SRMR-IEM, significantly reduces the performance gap between non-intrusive and intrusive metrics.

The remainder of this paper is organized as follows. In section 2, we describe the benchmark objective speech quality metrics, the proposed metric, and the in-ear speech quality dataset used for the experiments. In section 3, we present a comparison of the performance of the benchmark metrics and proposed metrics. In section 4, we discuss some of the limitations of the current study. Finally, in section 5 we conclude the paper and delineate some avenues for future work in this domain.

2. Materials and Methods

2.1. In-ear Speech Quality Dataset

Speech was recorded in an audiometric booth with an intra-aural communication headset containing an IEM, and an outer-ear microphone (OEM) as well as a digital audio recorder (Zoom®H4n) placed in front of the speaker's mouth (i.e ref signal). A female speaker read out the first ten lists of the Harvard phonetically balanced sentences [6] and speech was recorded at 8 kHz sampling rate and 16-bit resolution using all three microphones (IEM, OEM and ref) simultaneously. To simulate a condition where the environmental noise is high enough that residual noise "leaks" through the passive attenuation of the earplug, factory noise from the NOISEX-92 database [7] was mixed to the OEM speech at -5 dB (SNR) after recording. Using the mea-

10.21437/PQS.2016-15

sured attenuation of the earplug, the OEM noise was filtered and then added to the IEM speech (i.e IEM_N). The noisy signals were then denoised using an adaptive nLMS filtering process (i.e IEM_{NS}) and its bandwidth extended in the high frequencies (i.e IEM_{BWE}).

The quality of the four different IEM signals (IEM, IEM_N, IEM_{NS}, IEM_{BWE}) was assessed subjectively using an online MUlti Stimulus Test with Hidden Reference and Anchor (MUSHRA) [8] test. The speech signal captured in front of the mouth served as the reference signal while the noisy IEM signal served as the anchor as it is a bandlimited and noise-corrupted version of the reference signal. A total of 42 participants took part in the test. The test was performed using a web interface [9] which is freely available for other researchers.

2.2. Benchmark Objective Quality Measures

The speech to reverberation modulation energy ratio (SRMR) is a non-intrusive metric [5] that estimates speech quality as proportional to the ratio of energies in lower frequencies of the envelope modulation spectrum of a speech signal and energies in higher frequencies. The metric is computed as follows: first, the signal is processed by a 23 channel gammatone filterbank and the envelope of each band is computed using the Hilbert transform. Each of the band envelopes is then analysed by an 8 channel modulation filterbank, with center frequencies going from 4 Hz to 128 Hz. The energy of such bands is computed on a per-frame basis using 256 ms frames with 75% overlap, resulting in a $23 \times 8 \times$ number of frames tensor, which is averaged over the time dimension yielding the average modulation spectrum. The modulation bands corresponding to the lower ($<$20 Hz, dominated by the speech signal) and higher ($>$20 Hz, dominated by distortions) modulation frequencies are them summed and their ratio is computed. In [10], SRMRnorm, a metric based on SRMR with lower inter- and intraspeaker variability due to the use of a normalized modulation spectrum representation, is proposed. In SRMRnorm, the modulation filterbank center frequencies range from 4 to 40 Hz, which limits the influence of the pitch on the speech envelope, and the modulation spectrum dynamic range is limited to 30 dB, which was shown to reduce the effect of differences between different speakers and speech content in the speech material being evaluated.

As benchmark metrics, two intrusive speech quality metrics were considered in this study: the Perceptual Evaluation of Speech Quality (PESQ) [11] and Perceptual Objective Listening Quality Assessment (POLQA) [12]. Both metrics are ITU-T standards. PESQ was originally proposed for assessment of speech quality for users of telephony systems, while POLQA is extended to cover a broader range of scenarios, such as higher bandwidth signals, different types of speech processing (noise reduction, spectral shaping, voice quality enhancement), and recordings made using an ITU ear simulator.

We have also included two other non-intrusive metrics as benchmarks. ANIQUE+ [13], an ANSI standard, is another non-intrusive speech quality measure, which characterizes abnormal articulation-related distortions based on modulation spectral features. Finally, P.563 [14] is the ITU-T standard for non-intrusive objective speech quality metric for narrow-band telephony. P.563 computes mean opinion scores based on a number of internal features, which are aggregated to characterize five distortion classes: high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech. Once a major distortion class is detected, the intermediate scores are linearly combined with eleven other parameters to derive a final quality estimate.

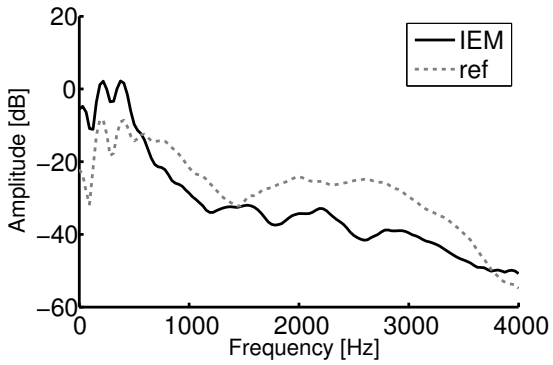

Figure 1: Long-term average spectra for in-ear speech vs. reference

It is important to note that none of these benchmark metrics has been designed or validated with IEM speech. However, since they are considered state-of-the-art in speech quality measurement, we used them as benchmarks in order to evaluate their performance and identify limitations. We expect intrusive metrics to have a smaller performance gap on IEM measured when compared to non-intrusive metrics, since they have access to a reference signal and can therefore estimate distortions more accurately.

2.3. Adapting SRMR for In-Ear Microphone Speech

Two adaptations were made to SRMRnorm in order to make it more suitable to IEM speech. The first adaptation was to apply a weighting function over the modulation spectrum to take the characteristics of the occluded ear canal into account. Figure 1 shows the long-term average spectra for speech recorded using the in-ear and the reference microphone (captured in front of the mouth) for 90 Harvard sentences recorded under the same conditions used for the subjective tests, but not used in the tests. The long-term average spectra was calculated by averaging spectrogram frames calculated with a window length of 20 ms and 50% overlap. We can see a boost of approximately 10 to 15 dB in lower frequencies, followed by an attenuation of less than 10 dB starting around 500 Hz and stronger attenuations as frequencies increase.

Based on these results, and since it has been previously reported in [3] that the occluded ear canal boosts frequencies below 2 kHz and attenuates higher frequencies, we decided to apply a simple inverted step function to cancel this effect: energies in channels with center frequency up to 2 kHz were attenuated by 15 dB, and energies in channels above 2 kHz amplified by 15 dB, with the 15 dB gain/attenuation being roughly the average gain/attenuation across all frequencies in the occluded signal. By using a rough average instead of optimizing the weighting function based on our data, we avoid overfitting the final metric to our currently available empirical results.

The second adaptation was to limit the modulation spectral energy range, as previously proposed in [10]. The procedure can be summarized in three steps:

1. Compute the peak energy, $\bar{E}_{peak}$, from the average modulation spectrum;
2. Clip each frame of the original modulation spectrum to the range $[\bar{E}_{peak} - E_{min}, \bar{E}_{peak}]$;
3. Recalculate the average modulation spectrum based on the clipped spectrum.

Table 1: Performance of the evaluated speech quality metrics. The values between parenthesis for SRMRnorm and SRMR-IEM represent the modulation energy range.

Metric	ρ	ρ_{sp}	ρ_{sig}	RMSE
PESQ	0.922	0.685	0.903	11.20
POLQA	0.906	0.814	0.894	11.65
P.563	0.561	0.522	0.585	21.11
ANIQUE+	0.531	0.502	0.535	21.98
SRMRnorm (30 dB)	0.536	0.408	0.528	22.10
SRMR-IEM (15 dB)	0.811	0.692	0.867	12.98
SRMR-IEM (30 dB)	0.727	0.533	0.728	17.84
SRMR-IEM (60 dB)	0.530	0.384	0.523	22.18

In [10], a threshold of $E_{min} = 30$ dB was used, as that value had been shown to significantly reduce inter- and intraspeaker variability for a dataset of consonant-vowel pairs recorded in an anechoic, noise-free condition. For speech recorded using an in-ear microphone, a similar reference dataset was not available so evaluating modulation spectrum variability across different speakers and speech content was not possible. Therefore, we tested the performance of the proposed metric using different threshold values, namely 15 dB, 30 dB, and 60 dB. The results for all threshold values are reported in the next section.

3. Results

Table 1 summarizes the performance of the evaluated metrics using four figures of merit: Pearson linear correlation (ρ), Spearman rank correlation (ρ_{sp}), Pearson correlation after a sigmoidal fit to the MUSHRA scores (ρ_{sig}) and root mean-squared error (RMSE) between the predicted scores and the average MUSHRA scores from the subjective test. The sigmoidal fit parameters are estimating by using a generalized linear model with a logit link function, after normalizing both the input and output variables to the interval [0, 1]. As can be seen, the intrusive metrics have significantly higher correlations than P.563, ANIQUE+ and SRMRnorm. SRMR-IEM using a modulation spectrum range of 15 dB and 30 dB show significant improvement compared to other non-intrusive metrics in all figures of merit, reducing the gap between intrusive and non-intrusive metrics. Limiting the range to 60 dB, however, shows similar performance to SRMRnorm, indicating that the full range of the signal is close to 60 dB. The RMSE of predictions with intrusive and non-intrusive methods follows a similar trend, with the best SRMR-IEM configuration reducing the error in almost 10 points in the MUSHRA scale when compared to SRMRnorm and only 1.8 points above PESQ, the best-performing metric.

Figure 2 shows scatterplots of the objective metric scores against the respective MUSHRA scores for all the tested conditions, with the sigmoidal fit overlaid. The updated SRMR-IEM, shown only for the best modulation spectrum range (15 dB), has reduced the variability of objective scores for all IEM scenarios, which results in a better discrimination between different conditions after the sigmoidal fit. We can see that predictions for the IEM scenarios overlap for most of the evaluated metrics, with the proposed metric giving slightly larger scores to the bandwidth-extended signals and PESQ having the opposite behavior. PESQ and POLQA, however, predict very low scores to the noisy IEM signal, going in line with subjective scores, while SRMR-IEM predicts these to be just slightly lower than for the other IEM scenarios.

4. Discussion

A metric designed to predict speech quality for IEM signals has to be able to reliably rate IEM signals recorded under different conditions and processed by noise reduction and bandwidth extension algorithms. We can see that even current standard intrusive metrics are not well-suited for this task. POLQA has a large variability for most IEM signals, with most of the conditions overlapping even though the subjective ratings for three of the four conditions evaluated do not overlap. The IEM-NS sentences, for example, span more than 50% of the whole metric range, while the subjective scores have very little variation (less than 10%). While PESQ is also not able to differentiate the bandwidth-extended signals from the IEM and IEM-NS ones, it shows the best discrimination capability for the IEM_N condition, giving significantly lower scores to that condition than to all the other IEM conditions.

Comparing SRMRnorm to the proposed metric, we can see that the dynamic range limitation strategy that was proposed in [10] alone does not account for the variability in IEM speech, as can be seen in Figure 2. We experimented with both the modulation frequency range in the modulation filterbank and the dynamic range limitation that were proposed in [10]. First, we tested frequencies from 30 Hz up to 128 Hz (the latter being the original range proposed in [5]) as the center frequencies for the last modulation filter. The modulation energy range limitation was kept at 30 dB. When we compared the correlations between the different configurations to the MUSHRA scores, we found out that similarly to normal speech, a range of 4 to 40 Hz was optimal. However, as detailed in the previous section, varying the modulation energy range led to significant improvements. Nonetheless, this variation alone is not responsible for all the improvements seen in the proposed version, as the acoustic frequency weighting scheme we propose here has also played an important role, as can be seen by comparing the performance results for SRMRnorm to the ones for SRMR-IEM with a 30 dB energy range; frequency weighting alone improved correlations from 0.536 to 0.727, and combining it with a reduction in modulation energy range further improved the performance.

While SRMR-IEM significantly reduces the variability of IEM measurements, even when compared to the intrusive metrics, it still does not discriminate the noisy IEM scenario from the other IEM scenarios. In the modulation spectrum representation we are using, the noise present in the noisy scenario does not change the modulation spectrum as much as expected by the model in clean speech conditions, as illustrated in Figure 3. We expect noise to significantly increase the amount of energy in modulation frequencies above 20 Hz. However, in the example illustrated by Figure 3, there are only small variations in the region between 15 and 25 Hz. Further investigation is required to find how to better represent IEM-related distortions in the modulation spectrum, or if another signal representation is needed to represent such distortions.

Our experimental setup has a few limitations that have to be taken into account. The weighting values chosen in this work are based on IEM speech captured for one female speaker. It is important to consider that these weighting values may vary with a more diverse range of speakers. For example, male speakers with more low frequency content may not benefit from a 15 dB attenuation in the low frequencies. Our proposed metric was evaluated on a very limited number of scenarios, including a single noisy environment and two speech enhancement algorithms (for noise reduction and bandwidth extension). In order to be able to use an objective metric to evaluate speech enhance-

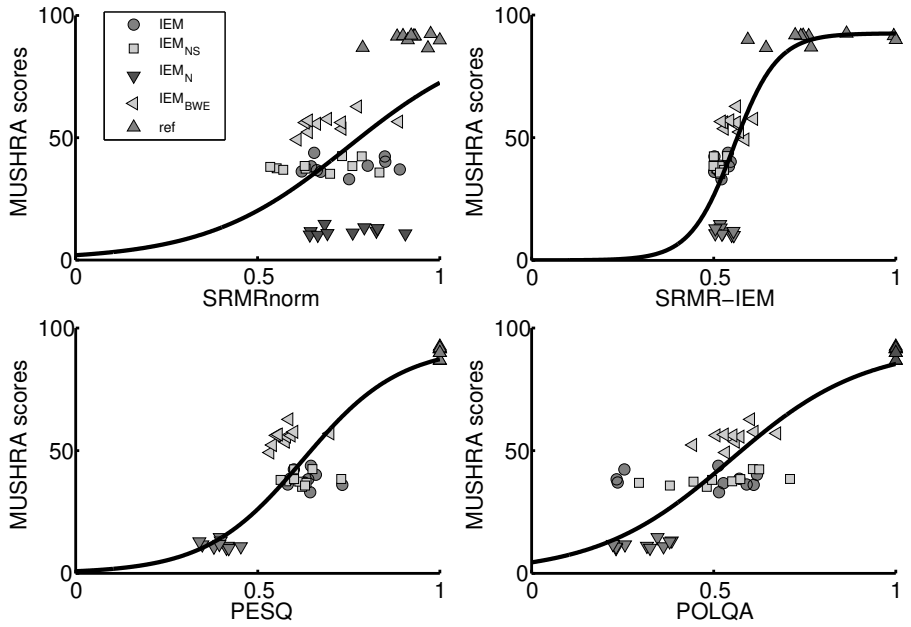

Figure 2: Scatterplots for SRMRnorm, SRMR-IEM (with 15 dB modulation range limitation), PESQ, and POLQA vs. MUSHRA scores

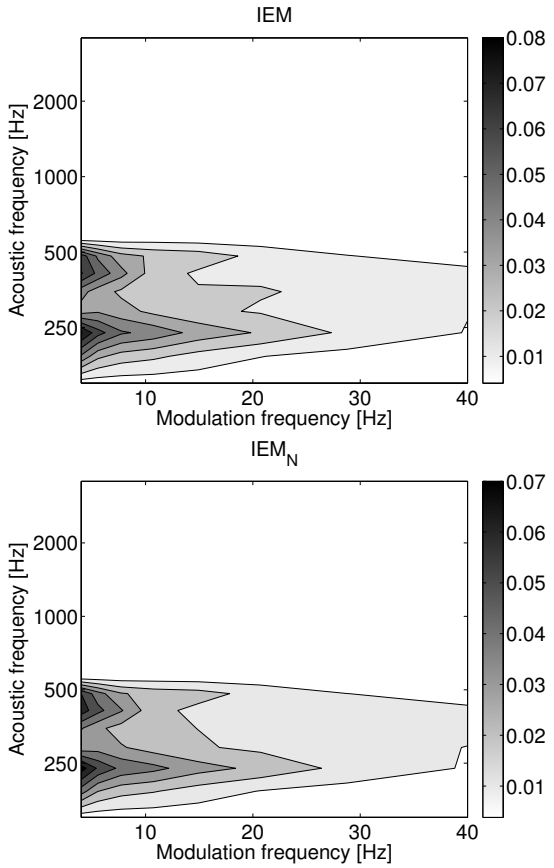

Figure 3: Modulation spectrums for a clean IEM signal (top) vs. a noisy one (bottom).

ment algorithms and speech produced in a variety of real-world environments, it would be relevant to evaluate and train mappings of the metric on a larger dataset, as well as cross-validate the results on an independent subset of data. Due to the limited availability of data collected with in-ear microphones, and especially results of subjective experiments using such data, we are currently not able to perform such analysis.

5. Conclusions

We evaluated intrusive and non-intrusive speech quality metrics with in-ear speech under different conditions. While most non-intrusive metrics showed poor performance compared to intrusive metrics, our proposed adaptation of the SRMRnorm metric for IEM speech, SRMR-IEM, significantly reduces this performance gap. As future work, we consider evaluating the proposed metric on a larger dataset, encompassing different scenarios and speech enhancement algorithms.

6. Acknowledgements

The authors would like to acknowledge funding from EERS Technologies Inc. and its Industrial Research Chair in In-ear Technologies (CRITIAS), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Fonds de recherche du Québec - Nature et technologies.

7. References

[1] R. E. Bou Serhal, T. H. Falk, and J. Voix, "Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones." in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 040013.

[2] J. Voix, "Did you say" bionic" ear?" *Canadian Acoustics*, vol. 42, no. 3, 2014.

[3] R. E. Bouserhal, T. H. Falk, and J. Voix, "On the potential for artificial bandwidth extension of bone and tissue conducted speech: a mutual information study," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5108–5112.

[4] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processsing Magazine*, March 2015.

[5] T. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766 –1774, Sep. 2010.

[6] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[7] A. Varga and H. J. 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, 1993.

[8] R. ITU-R, "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra)," *International Telecommunications Union, Geneva*, 2001.

[9] J. F. Santos, "mushra-ruby-server: Version used for the PQS 2016 paper," Jun. 2016. [Online]. Available: http://dx.doi.org/10.5281/zenodo.56061

[10] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014, pp. 55–59.

[11] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Intl. Telecom Union, Tech. Rep., 2001.

[12] ITU-T P. 863, "Perceptual Objective Listening Quality Assessment (POLQA)," ITU Telecommunication Standardization Sector (ITU-T), Tech. Rep., 2011.

[13] D.-S. Kim and A. Tarraf, "ANIQUE+: a new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, p. 221–236, 2007.

[14] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," Intl. Telecom Union, Tech. Rep., 2004.