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 (SRMR) to IEM speech signals. 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]. In noisy environments, it is advantageous to capture speech using an IEM as it is less susceptible to degredation from background noise [3]. However, 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 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]. Consequently, it is important to be able to quickly assess the quality of IEM speech and any subsequent enhancement.

Objective speech quality metrics have been proposed as a way of making such measurements 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. Most of the speech quality metrics target speech in natural conditions 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 (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.

2. Materials and Methods

2.1. Benchmark Objective Quality Measures

Two intrusive speech quality metrics were considered in this study: the Perceptual Evaluation of Speech Quality (PESQ) [6] and Perceptual Objective Listening Quality Assessment (POLQA) [7]. Both metrics are ITU-T standards.

The Speech to Modulation Energy Ratio 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. In [8], an updated version of SRMR with lower variability is proposed. ANIQUE+ [9], an ANSI standard, is another non-intrusive speech quality measure whose internal model also considers modulation spectrum features. Finally, P.563 [10] is the ITU-T standard for non-intrusive objective speech quality metric for narrow-band telephony.

2.2. In-ear Speech Quality Dataset

Speech was recorded in an audiometric booth with an intra-aural communication headset containing an IEM 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 and speech was recorded at 8 kHz sampling rate and 16-bit resolution using both microphones simultaneously. After recording, factory noise from the NOISEX-92 database [11] was mixed to the IEM speech (i.e IEM_{N}) to simulate a condition where the environmental noise is high enough that residual noise "leaks" through the passive attenuation of the earplug. 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) [12] 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.

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 behavior of the occluded ear canal into account. Since the occluded ear canal boosts frequencies below 2 kHz and attenuates higher frequencies, we applied a simple inverted step function

Table 1: Performance of the evaluated speech quality metrics

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	0.520	0.396	0.513	22.3
SRMR-IEM	0.756	0.650	0.830	14.5

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. The second adaptation was to limit the modulation spectral energy range, as previously proposed in [8]. In [8], a range of 30 dB was used. For IEM speech, we found that a range of 15 dB was optimal.

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). As can be seen, the intrusive metrics have significantly higher correlations than P.563, ANIQUE+ and SRMRnorm. SRMR-IEM shows a significant improvement compared to other non-intrusive metrics in all figures of merit. The RMSE of predictions with intrusive and non-intrusive methods follows a similar trend.

Figure 1 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 has reduced the variability of objective scores for all IEM scenarios, which results in a better discrimination between different conditions after the sigmoidal fit.

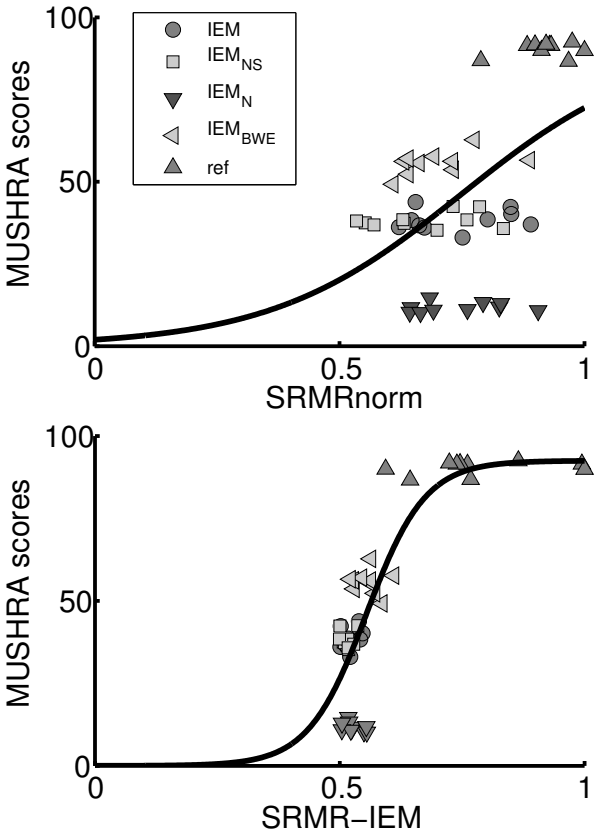

Figure 1: Scatterplots for SRMRnorm (top) and SRMR-IEM (bottom) vs. MUSHRA scores

4. 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 adapted for IEM speech, SRMR-IEM, significantly reduces this performance gap.

5. 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] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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