Interspeech 2018
2-6 September 2018, Hyderabad

Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition

Anderson R. Avila[1,2], *Jahangir Alam*[2], *Douglas O'Shaughnessy*[1], *Tiago H. Falk*[1]

[1]INRS-EMT, University of Quebec, Canada
[2]Computer Research Institute of Montreal (CRIM)

anderson.avila@emt.inrs.ca, jahangir.alam@crim.ca, doug@emt.inrs.ca, falk@emt.inrs.ca

Abstract

In this study, the performance of two enhancement algorithms is investigated in terms of perceptual quality as well as in respect to their impact on speech emotion recognition (SER). The SER system adopted is based on the same benchmark system provided for the AVEC Challenge 2016. The three objective measures adopted are the speech-to-reverberation modulation energy ratio (SRMR), the perceptual evaluation of speech quality (PESQ) and the perceptual objective listening quality assessment (POLQA). Evaluations are conducted on speech files from the RECOLA dataset, which provides spontaneous interactions in French of 27 subjects. Clean speech files are corrupted with different levels of background noise and reverberation. Results show that applying enhancement prior to the SER task can improve SER performance in more degraded scenarios. We also show that quality measures can be an important asset as indicator of enhancement algorithms performance towards SER, with SRMR and POLQA providing the most reliable results.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Nowadays, existing emotion recognition systems can achieve satisfactory performance in controlled environments [1]. In the wild, i.e., in real world scenarios where background noise is commonly found, performance is reduced due to the numerous circumstances unseen during training, such as illumination, occlusion, background noise and room reverberation. This has motivated an increasing effort by the research community towards providing more realistic settings (e.g. unposed data collected beyond laboratory environment) so that the behaviour of such systems can be tested in more naturalistic scenarios [2]. For instance, several Challenges have been organized over the last few years, most notably the INTERSPEECH 2009 Emotion Challenge [3], EmotiW (Emotion Recognition In The Wild Challenge) [1], and the 2016 Audio/Visual Emotion Challenge (AVEC) [4].

In the case of speech emotion recognition (SER), detrimental effects of environmental noise can be mitigated by the use of speech enhancement algorithms. Most of these methods, however, are primarily designed for improving perceived quality and intelligibility and may not be tailored for speech emotion recognition. A handful of studies have evaluated enhancement algorithms considering the results attained by the SER systems, neglecting information regarding perceptual quality of the processed speech itself. For example, in [5], a new set of wavelet based features was adopted for classifying emotions. Two enhancement methods were used to suppress residual noise with their performance evaluated only in respect to the SER task. The authors in [6] presented a speech enhancement algorithm based on adaptive noise cancelation for improving emotional speech classification. A set of features based on cepstral analysis of pitch and energy contours was also introduced. In [2], a feature enhancement method based on an autoencoder with Long Short-Term Memory (LSTM) neural networks is proposed towards robust emotion recognition from spontaneous speech. Both additive and convolutional noise were explored and results showed considerable gains compared to the baseline system where no feature enhancement was applied.

The primary concern of speech enhancement for SER is to improve speech quality and intelligibility of the corrupted speech signal without removing important emotional cues. Although the aforementioned studies have shown the benefit of applying denoising algorithms on SER, little or no information is given in respect to the quality of the processed speech and how it correlates with SER performance. Such evaluation could provide prior information about the performance of these algorithms, which could be used to leverage SER performance. With this paper, we intend to fill this gap. We investigate the performance of two enhancement algorithms under two criteria: considering the performance of a SER system and in respect to the scores reported from three objective quality measures. We also report the correlations between the estimated perceptual qualities and the performance of the SER system under different background noise levels. This will give us insights on whether prior information about speech quality as well as intelligibility can benefit SER systems, e.g., by identifying the most effective enhancement algorithm to be applied on the speech signal before the performing SER.

The remainder of this paper is organized as follows. Section II presents materials and methods used for the experiments performed in this study. In Section III, we present the experimental results and discussion. Section IV gives the conclusions.

2. Methods and Materials

2.1. Corrupted speech and enhancement methods

In many everyday situations, the speech signal is acquired in a noisy environment. Hence, it is desirable to enhance it in order to improve intelligibility and speech quality prior to performing any speech related task. Here, two noise suppression algorithms used to enhance SER performance are presented.

2.1.1. Single-channel spectral enhancement

The single-channel spectral enhancement method (SSE) is based on the estimation of a real-valued spectral gain, $\hat{G}(\omega)$, which represents the amount of attenuation to be applied to the corrupted signal to obtain the enhanced signal. This is performed at each frequency component, as shown below:

10.21437/Interspeech.2018-2350

$$\hat{S}(\omega) = \hat{G}(\omega)X(\omega) \tag{1}$$

where $\hat{X}(\omega)$ represents the corrupted signal and $\hat{S}(\omega)$ is the STFT of the estimated speech signal. The gain is attained using the minimum mean square error (MMSE) for estimating the spectral magnitude of the clean speech signal. This requires the computation of the power spectral density (PSD) for the clean, noise and reverberant components. The readers are referred to [7] for more details regarding this method.

2.1.2. Relative convolutive transfer function

In this method, the late reverberation is modeled by an STFT domain moving average (MA) model using a relative convolutive transfer function (RCTF). The RCTF coefficients are modeled as a first-order Markov process and estimated using a Kalman filter. After computing the RCTF coefficients, an estimate of the late reverberation PSD can be obtained which allow spectral enhancement to achieve dereverberation and noise reduction. The problem can be described as

$$X(\omega) = S(\omega)H(\omega) + V(\omega) \tag{2}$$

where $S(\omega)$ is the anechoic signal representation in the STFT domain and $H(\omega)$ is the time-varying convolutive filter coefficients, used to model the reverberant signal and $V(\omega)$ is the additive noise. The PSD of the reverberant signal is considered time-varying while the noise PSD is assumed stationary. The reader is referred to [8] for a detailed description on this method.

2.2. Instrumental Quality Measures

A listening test is the most accurate method for evaluating speech signal quality [9]. Nevertheless, it presents shortcomings as the process can be time-consuming, expensive and more importantly it cannot be done in real time [10]. To overcome this limitation, objective perceptual quality models can be used to attain subjective quality estimation. Here, we present a non-intrusive instrumental quality measure, namely speech-to-reverberation modulation energy ratio (SRMR) [11] and two intrusive measures (i.e., a reference signal is required): the perceptual evaluation of perceived quality (PESQ) [12] and the perceptual objective listening quality assessment (POLQA) [13].

3. Experimental Setup

3.1. Corpus Description

The results presented in this paper have been obtained using anechoic speech files from the REmote COllaborative and Affective interactions (RECOLA) database [14]. It features 27 French-speaking subjects, 16 females and 11 males, from 3 different nationalities (French, Italian and German). Spontaneous interactions were collected during a conference call while a collaborative task was performed by the subjects. Six French speaking annotators measured emotion continuously providing a value, chosen over a time-continuous emotional scale (ranging from -1 to 1, with 0.01 step), for two dimensions: arousal and valence [14]. This dataset was also used to assess the performance of emotion recognition systems at the Audio/Visual Emotion Challenge (AVEC 2016) [4]. The annotated data was binned with a frame rate of 40 ms. The ground truth was computed as the mean value of the annotations at every time step. The dataset is segmented into three parts, i.e., training, development and testing sets, each containing 9 speech samples of 5 minutes. As labels were available only for training and development, all our experiments are based on these two sets. Corrupted speech files were attained with a babble type of noise from an airport lobby, under five speech-to-noise ratio was considered: 0 dB, 5 dB, 10 dB, 15 dB, 20 dB.

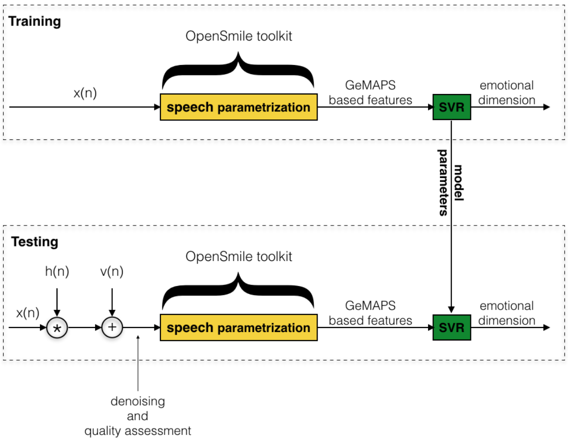

Figure 1: *Training setup.*

3.2. Benchmark and Performance Figures

The acoustic benchmark features adopted here are the same set of acoustic Low-Level Descriptors (LLD) used on the AVEC Challenge 2016 [4], which cover spectral, cepstral, prosodic and voice quality information. Such features are based on the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and extracted using the OpenSMILE feature extraction toolkit [15]. For all the LLD, arithmetic mean and the coefficient of variation are computed. Percentiles 20, 50 and 80, the range of percentiles 20-80 and the mean and standard deviation of the slope of rising/falling signal parts are applied to pitch and loudness. For more detailed information on these baseline features refer to [4].

As figure of merit, the concordance correlation coefficient (CCC) is used. The method computes the correlation considering the reproducibility and the level of agreement between two variables [16]. It combines Pearson's correlation coefficient (CC) and the square difference between the mean of the two samples, as follows

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x + \mu_y)^2} \tag{3}$$

where μ_x and μ_y represents the mean of each variable, σ_x and σ_y are their variances whereas ρ is the Pearson correlation coefficient.

3.3. Test Setup

As described in Figure 1, a front-end based on the OpenSMILE toolkit is used to extract GeMAPS features. The learning process is divided into two phases: training and testing. Different from the approach given in [2], where the authors used a small amount of noisy data during training, no compensation technique was considered in our experiments and only clean speech was used, characterizing more severe mismatch scenarios. Model development was based on the same steps taken for the AVEC Challenge 2016 [4]. That is, after training the support vector regressor (SVR) on the training set, the attained

Table 1: *Performance, in terms of CCC, of SER system after applying 3 enhancement algorithms and clean speech corrupted with additive noise.*

		Clean		0 dB		5 dB		10 dB		15 dB		20 dB		Average	
Algorithm	Features (Model)	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
Unprocessed	Benchmark (SVR)	0.786	0.461	0.222	0.048	0.326	0.047	0.455	0.033	0.542	**0.096**	0.594	**0.151**	0.427	0.075
SSE	Benchmark (SVR)	-	-	**0.439**	0.070	**0.573**	**0.106**	**0.653**	**0.059**	**0.670**	0.075	0.627	0.093	**0.592**	0.080
RCTF	Benchmark (SVR)	-	-	0.423	**0.130**	0.539	0.084	0.651	0.053	0.667	0.048	**0.662**	0.145	0.588	**0.092**

model and its respective parameters are validated on the development set. This process is performed several times until the optimal model is obtained. In the testing phase, a total of 9 speech samples per condition were used to test our model.

To evaluate the performance of each speech enhancement algorithm, all processed speech files are assessed using the instrumental quality measures described in Section 2. Each speech file contains two speakers: the interviewer in the background and the interviewed, who is the subject of the emotion recognition task. The development set has 4 males and 5 females with one subject per file. Each objective quality metric provides a score for each speech file, in total 9 scores are attained. The performance of each enhancement algorithms is evaluated based on the overall score of each condition, which is achieved by averaging over the 9 scores previously mentioned. We have three main goals with this experiment. First, a simple comparative analysis of these two enhancement algorithms. Second, comparison of the performance of each algorithm in respect to noisy (unprocessed) speech. And third is to investigate how the performance of each algorithm correlates with the performance of the SER system.

4. Experimental Results

In this section, the benchmark SER system performance is evaluated after applying speech enhancement on a speech signal corrupted with background noise. The performance of the two enhancement algorithms described in Section 2 is thus evaluated based on three instrumental quality measures. Results are presented in Table 1 in terms of the concordance correlation coefficient (CCC) and in Figure 2, which provides perceptual quality predictions of processed speech as well as the correlations between the predictions of each instrumental measure and the SER performance.

From Table 1, we can observe that the SER system is severely affected by background noise, especially for lower SNR, with both enhancement approaches being effective in mitigating this effect. For valence, for instance, results were improved for speech signals corrupted with 0, 5 and 10 dB, with SSE providing the best performance except at 0 dB. For arousal, enhancement helped to boost SER performance in every condition. At the most severe noise condition, 0 dB, CCC went from 0.222 to 0.439 after applying the SSE enhancement method. Similar improvement can be verified for the RCTF method. Overall, the SSE algorithm outperformed the RCTF.

In Figures 2-a, 2-b and 2-c, the performance of these two speech enhancement methods is compared also in terms of objective speech quality assessment. According to the scores obtained with POLQA, depicted in Figure 2-a, almost no improvement was found after applying speech enhancement on the corrupted speech signal, which does not relate with the results in Table 1. On the other hand, POLQA assessment shows that speech signals processed by the SSE method had better quality than the ones attained with the RCTF method. This finding relates to the SER performance after enhancement. For PESQ (see Figure 2-b), both methods improved the quality of the speech signal when compared to the scores attained from an unprocessed speech signal. Although this is congruent with the SER results, PESQ found the RCTF enhancement algorithm to outperform the SSE one. This is not in accordance with the performance reported in Table 1. Between these two methods, we can observe better results with the SSE method, which is supported by the SRMR measure in Figure 2-c. The measure was able not just to assess both enhancement methods, but also rank their results towards the benchmark SER performance.

Correlations between each objective measure and the SER performance for enhanced speech attained from SSE and RCTF, as well as for unprocessed speech are also given in Figure 2. For unprocessed speech (second row of Figure 2), the best correlations are achieved by POLQA, R^2=0.98, followed by SRMR, which provided R^2=0.97, both for arousal. PESQ achieved lower performance for arousal predictions compared to these measures, but provided higher correlation for valence. These results suggest that these measures can potentially be used to predict SER performance in noisy environments. The results for enhanced speech are found from Figure 2-g to Figure 2-i. Correlations with valence are low for both enhancement methods and for each measure. This is due to the fact that enhancement had no effect on valence performance, even worsening it in some cases. For arousal, on the other hand, correlations were much higher. For SSE, for instance, SRMR provided the best performance, R^2=0.84, against R^2=0.82 attained with POLQA.

5. Conclusion

In this paper, we proposed to investigate how the performance of speech enhancement algorithms for the SER task correlates with three instrumental quality measures. First, we compare these methods with respect to the performance of the benchmark speech emotion recognition. We found that both enhancement methods helped to boost performance of SER, with the SSE method achieving the best results. Second, we performed speech quality assessment using PESQ, POLQA and SRMR. We showed that these metrics were quite accurate in estimating quality across different SNR's, and also good indicators of SER performance, especially for arousal predictions. SRMR, for instance, was quite successful in ranking the adopted enhancement methods towards SER performance. For arousal, POLQA and SRMR seemed to correlate well with SER performance and could potentially be used for estimate performance in noisy environments.

6. Acknowledgement

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fonds de recherche du Québec - Nature et Technologies (FRQNT) and the Natural Sciences and Engineering Research Council of Canada for financial support.

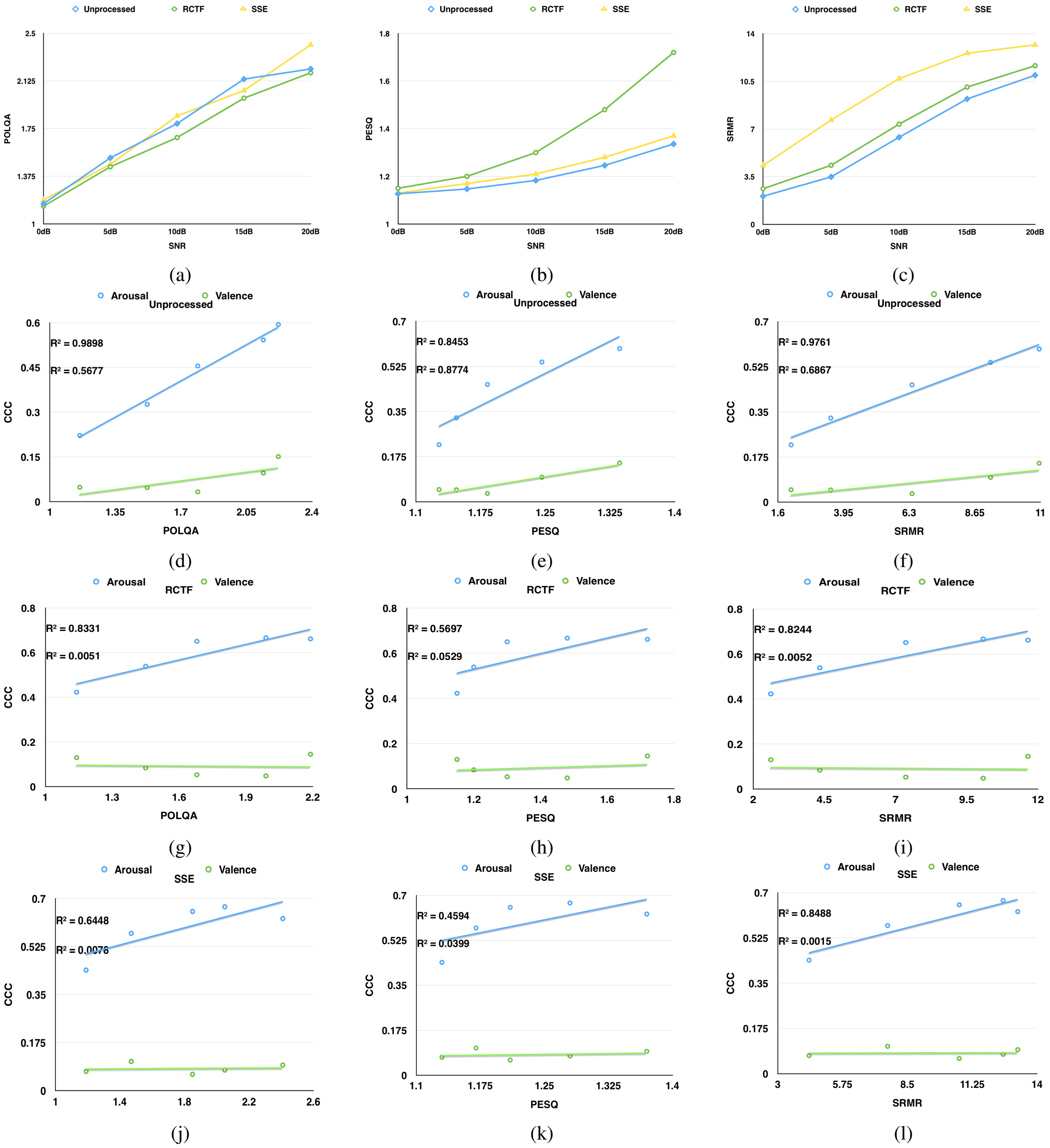

Figure 2: *Perceptual performance of two enhancement algorithms based on three perceptual measures. First row describes the the MOS estimated by (a) POLQA, (b) PESQ and (c) SRMR for five different noise levels. Graphs (d)-(l) report the correlation between the SER performance (CCC) and the MOS estimated by each measure.*

7. References

[1] A. Dhall and et al., “Video and image based emotion recognition challenges in the wild: Emotiw 2015,” in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2015, pp. 423–426.

[2] Z. Zhang and et al, “Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks,” in *Proceedings of the d17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 3593–3597.

[3] B. Schuller and et al., “The interspeech 2009 emotion challenge.” in *Interspeech*, 2009, pp. 312–315.

[4] M. Valstar and et al., “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[5] J. Vasquez-Correa and et al., “Emotion recognition from speech under environmental noise conditions using wavelet decomposition,” in *Proceedings of International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2015, pp. 247–252.

[6] A. Tawari and M. Trivedi, “Speech emotion analysis in noisy real-world environment,” in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 4605–4608.

[7] B. Cauchi and et al, “Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–12, Jul. 2015.

[8] S. Braun and et al, “Late reverberation psd estimation for single-channel dereverberation using relative convolutive transfer functions,” in *Proceedings of the IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[9] S. Möller and et al., “Speech quality estimation: Models and trends,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[10] J. You and et al., “Perceptual-based quality assessment for audiovisual services: A survey,” *Signal Processing: Image Communication*, 2010.

[11] T. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[12] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, Recommendation P.862, Feb. 2001.

[13] J. Beerends and et al., “Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part itemporal alignment,” *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[14] F. Ringeval and et al, “Introducing the recola multimodal corpus of remote collaborative and affective interactions,” in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[15] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[16] I. Lawrence and K. Lin, “A concordance correlation coefficient to evaluate reproducibility,” *Biometrics*, pp. 255–268, 1989.