ON THE USE OF ARTIFICIAL NEURAL NETWORK TO PREDICT DENOISED SPEECH QUALITY

Anis Ben Aicha

Laboratoire de recherche COSIM
Ecole Supérieure des Communications de Tunis
Université de Carthage, Tunisie
anis_ben_aicha@yahoo.fr

ABSTRACT

Existing objective criteria for denoised speech assessment have as output one score indicating the quality of processed speech. Even it is well useful when it is about comparing denoised techniques between each others, they failed to give with enough accuracy an idea about the real corresponding Mean Opinion Score rate (MOS). In this paper, we propose a new methodology to estimate MOS score of denoised speech. Firstly, a statistical study of existed criteria based on boxplot and Principal Component Analysis (PCA) analysis yields to select the most relevant criteria. Then, an Artificial Neural Network (ANN) trained in selected objective criteria applied on the denoised speech is used. Unlike traditional criteria, the proposed method can give a significant objective score directly interpreted as an estimation of real MOS score. Experimental results show that the proposed method leads to more accurate estimation of the MOS score of the denoised speech.

Index Terms— speech enhancement, speech assessment, MOS, ANN

1. INTRODUCTION

Measuring speech quality constitutes an important task for evaluating many recent speech applications such as telephony, telephony over IP, coding, watermarking, speech enhancement, etc. Traditionally, user's opinions are measured using slow and costly subjective listening tests [1, 2]. In this test, listeners rate the speech they heard on a five-point opinion scale, ranging from 'bad' to 'excellent'. The ratings are unsigned integer scores ranging from 1 for 'bad' to 5 for 'excellent'. Then an average of these scores is computed and defined as the well-known Mean Opinion Score (MOS). It is widely used to characterize the quality of the speech. As an alternative to subjective measurement, an automated 'objective' criterion provides a rapid and economical way to estimate user opinion and makes it possible to perform real-time speech quality measurement. Many objective criteria are developed in the literature. They can be classified into three groups according to the domain in which they operate. We refer to temporal measures [3], spectral measures [4–8] and perceptual measures [9–11].

Most of objective criteria are mainly developed to assess speech signal for specific context. We relate for example PESQ criterion which is developed to evaluate speech over telecommunication systems [11]. For the specific case of speech enhancement context, only few attempts are conducted in the literature such as composite criteria [12] or perceptual audible degradation [13]. However, there is not yet a standard for denoised speech assessment. To choose which criterion is most suitable for speech enhancement context, it is mandatory to examine the correlation of objective measure with subjective one (MOS) [12]. It is found that existed criteria are not well correlated with subjective tests [12, 13]. Proposed measures in [12] and [13] show a better correlation to subjective tests which are about 0.68 and 0.79 for composite criteria and perceptual audible degradation respectively. However, it is still not sufficient to estimate with enough accuracy subjective evaluation.

Existing objective criteria compute one score for the degraded speech. The range and signification of the score depends on the theory on which objective criterion is based. Hence, it is possible to compare denoising techniques or algorithms between them. However, the main difficulty is to get a subjective interpretation of the obtained score. In other word the corresponding MOS score. It is found that the errors related to the correlation between objective and subjective measure are not negligible even for the latest proposed criteria in the area (about 0.46 for composite criteria and 0.37 for perceptual audible degradation) [13]. This means that when we assess a denoised signal using an objective criterion, we are not really sure about its correspond subjective note.

In this paper, we propose a novel method of denoised speech quality estimation. An Artificial Neural Network (ANN) trained in selected objective criteria applied on the denoised speech is used. Unlike traditional criteria, the proposed method can give a significant objective score directly interpreted as an estimation of real MOS note. A detailed description of the algorithm's functional blocks is presented in section 5.

The remainder of the paper is organized as following. In the second section, we present briefly an overview of objective criteria followed by the used corpus description. Section three, is reserved to analyze and explain how traditional criteria failed to estimate accurately subjective quality of denoised speech. In the forth section, we present selected objective criteria using boxplot and PCA analysis. In section five, we present the novel methodology to estimate subjective score according to MOS recommendation. Section six is reserved to experimental results and discussions. Finally, Section eight is devoted to the conclusion.

2. CRITERIA OVERVIEW AND SPEECH CORPUS DESCRIPTION

2.1. Criteria overview

Many objective criteria, well correlated with MOS, were proposed to estimate speech quality with low cost. Among a long list, the following most relevant objective criteria are used in this paper:

- temporal domain criterion:segmental SNR SNR_{seg} [3].
- spectral domain criteria: Log Likelihood Ratio LLR [4], Log-Area Ratio LAR [4], Itakura-Saito distortion measure IS [5], Cepstral distance CEP [6], Weighted-Slope Spectral distance WSS [7] and frequency SNR $fwSNR$ [8].
- Perceptual criteria: Modified Bark Spectral Distortion $MBSD$ [9], Perceptual Speech Quality Measurement $PSQM$ [10], Perceptual Evaluation of Speech Quality $PESQ$ [11], composite criteria $Covl$ [12] and Perceptual Signal to Audible Noise and Distortion Ratio $PSANDR$ [13].

2.2. Speech corpus description

The used corpus was designed to evaluate speech enhancement algorithms [12, 13]. A total number of 570 sentences is obtained from noisy signals corrupted by several kind of noises (white, babble, car, factory and f16) with input SNR range from -5 dB to 25 dB. The used denoising methods encompass four different classes of algorithms [12]:

- Spectral subtractive: multiband spectral subtraction, and spectral subtraction using reduced delay convolution and adaptive averaging.
- Subspace: generalized subspace approach, and perceptually based subspace approach.
- Statistical-model-based on minimum mean square error: mmse, log-mmse, and log-mmse under signal presence uncertainty.
- Wiener-filtering type algorithms: the *a priori* SNR estimation based method, the audible-noise suppression method, and the method based on wavelet thresholding the multitaper spectrum.

All these files are subjectively evaluated using standardized MOS methodology to get correspond subjective scores. Then, the files are classified into five classes according to their subjective MOS scores.

3. FAILURE OF OBJECTIVE CRITERIA TO ESTIMATE SUBJECTIVE QUALITY ACCURATELY

According to our database, we have for each 570 speech signal the corresponding MOS score. We have assessed all these signals using mentioned objective criteria in section 2. To display how well the MOS test and the objective criterion are correlated, we use scatter plot [14]. It is a plot representing in ‘x’ axis the computed scores obtained by objective criterion and in ‘y’ axis the real MOS scores [14]. To predict MOS scores from objective scores, a mapped function is used. Generally, a polynomial of second degree is used. The more the scatter plot is close to mapping function, the well the objective criteria is correlated with MOS test.

Without lost of generality and because of the lack of space, we present in Fig. 1 the scatter plot relative to the PESQ measure. Though, same remarks can be conducted for the remainder of the objective criteria. Firstly, we can remark that the points are very scattered around the mapping function. This means that even for PESQ, which known to have a high correlation with MOS test, it is difficult to predict subjective evaluation from the objective scores. This is an expected result, since as it is said in the introduction, most of objective criteria are developed for different tasks but not for the denoised speech assessment [12]. To show how it is no accurate to estimate MOS scores from objective ones, we have picket in the figure Fig.1 two signals with the same objective PESQ (PESQ=2) scores but with completely different real MOS scores (MOS=1.3 and MOS=3.6) (picket signals are noticed with boxes).

To show how it is difficult to estimate MOS score from any objective criterion, we present in Table 1, the Pearson correlation coefficients and correspond mean square error. We have to notice that we are interesting here on the mean square error which as depicted with bold values, still not negligible even for the most correlated criterion PSANDR (0.39). This is catch up with remarks of scatter plot analysis. Hence, we conclude that when denoised speech is assessed with objective criteria, we are not really sure about its correspond subjective evaluation.

4. NON REDUNDANT CRITERIA SELECTION

The purpose of this section is to select the most relevant criteria from the set of listed 11 objective criteria in section 2. Let us formulate the problem in two points as following. Firstly,

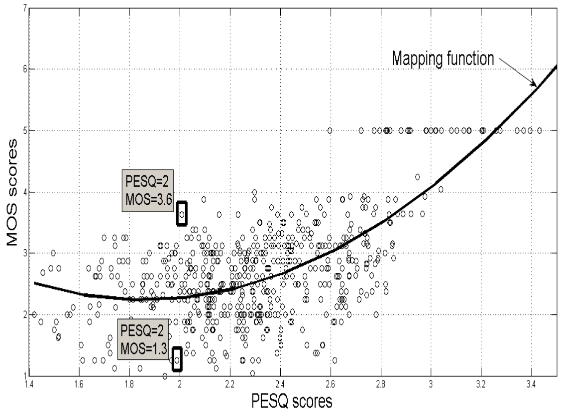

Fig. 1. Scatter plot of MOS scores versus PESQ scores.

Table 1. Correlation coefficient with MOS and mean square error of objective criteria.

Criteria	Correlation coefficient	mean square error
SNR	0.76	**0.41**
fwSNR	0.59	**0.51**
LLR	-0.53	**0.53**
IS	0.42	**0.60**
WSS	0.53	**0.57**
PESQ	0.67	**0.47**
Covl	0.68	**0.46**
PSANDR	0.78	**0.39**

from the large set of objective criteria, which ones are complementary and not redundant such as they can be used to extract the maximum information to determine to which MOS scale level belongs a given denoised speech? Secondly, which criteria are less confusing about the MOS scale level to which belongs a denoised speech when an objective score is computed? To answer these questions, we have proceed into two steps using boxplot analysis and Principal Component Analysis (PCA) [15–17]. In the following, we present a recap of the used experimental protocol [17].

- **Experimental protocol:** Each audio file of the database is evaluated using all previously defined objective criteria and the subjective score MOS. The subjective class to which it belongs is then identified. For each criterion, we group together the values which give the same subjective class. We hence obtain five sets of values which are analyzed using the boxplot toolbox.
- **Boxplot analysis:** We have computed and represented for each criterion the parallel boxplots versus subjective classes. Because the lack of space and without lost of generality, we have presented the computed boxplots of two criteria: $fwSNR$ and CEP. Fig. 2 shows that the different boxplots of $fwSNR$ haven't the same centrality. Moreover, centrality increase with the subjective classes, which means that the objective score obtained by $fwSNR$ can give an idea about the subjective class to which the processed signal belongs. It is retained as a candidate for relevant criteria selection. However, we notice that the values range of parallel boxplots overlap which means that there is some values for which the subjective class assignment is still confused. In the case of CEP criterion, the centrality for different classes have the same value range which means that this objective criteria cannot give an idea about the subjective class to which the processed speech belongs. Hence, CEP criterion is discarded from the set of criteria to be used for powerful evaluation. We have studied parallel boxplots of all objective criteria using the same methodology. We have discarded $CEP, LAR, IS, MBSD$ and $PSQM$ criteria and kept $SNR, fwSNR, WSS, LLR, PESQ, PSANDR$ as candidates for further steps of relevant criteria selection.
- **PCA analysis:** After discarding features which are source of confusion during t he procedure of relevant criteria selection, we propose to optimize the use of retained ones. For such purpose, Principal Component Analysis (PCA) is used [16]. It is found that the first and the second components represent about 80 % of the total variance. Projected criteria in the space generated by the first and the second components leads to find the most relevant criteria. Tab. 2 represents the ordered classical objective criteria according to their relative relevance. PSANDR seems to be the best one relevant to represent the total variation.

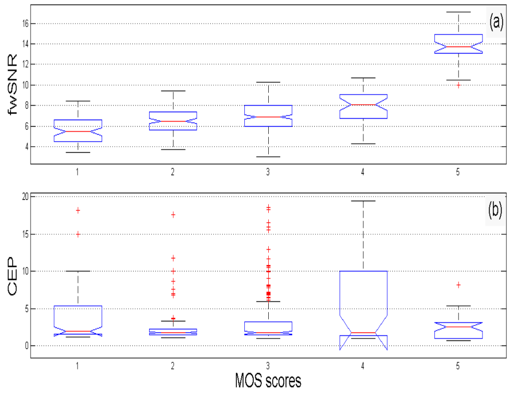

Fig. 2. Boxplot corresponding to (a) $fwSNR$ and (b) CEP.

5. NEW STRATEGY TO ESTIMATE MOS SCORE

5.1. Motivation and proposed idea

The proposed idea for estimating MOS scores of denoised signal is depicted in detail in Fig.3. We prose to use chosen

Table 2. Relevance of objective criteria in the principal components.

Objective criteria	$PSANDR$	SNR	LLR
Norm	0.73	0.66	0.54
	fw_{SNR}	$PESQ$	WSS
	0.52	0.51	0.43

criteria from previous section as attribute of the denoised signal. Hence, for each denoised signal $\hat{S^n}$, we compute scores c_1^n, c_2^n, c_3^n, c_4^n, c_5^n and c_6^n obtained by PSANDR, SNR, LLR, fwSNR, PESQ and WSS respectively to form a descriptor vector $A^n = [c_1^n, c_2^n, c_3^n, c_4^n, c_5^n, c_6^n]^T$. We want now to find the MOS score of the denoised signal S^n using its relative descriptor A^n. This can be easily identified as classical pattern recognition. Perhaps, one of the most successful algorithms in the field is the Artificial Neural Network (ANN) [18]. It can be seen as a statistical model which handles the data of non linear relationship based on linear combination of fixed non linear basis function known as activation functions. This feature may justify more our use of ANN as a predictor of MOS scores. In fact, previous attempts [12,13] combine linearly existed criteria under an implicitly assumption that there is some kind of linearity relationship between classical criteria. This assumption seems to be a hard one. This fact explains limits of composites criteria. By using ANN, we profit from the non linearity process handling non linear data. In this paper, we propose to use one of the most successful models of this type in the context of pattern recognition which is the feed-forward neural network, also known as the multilayer perceptron.

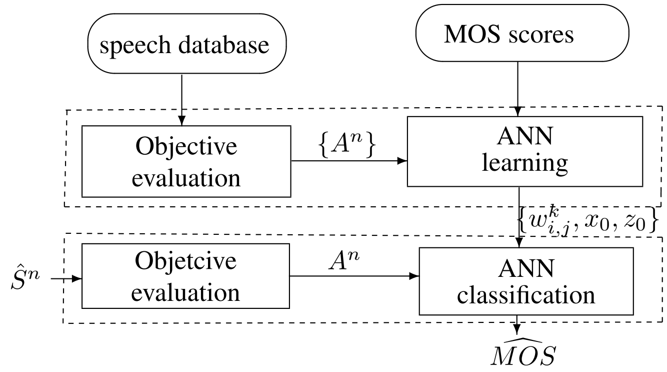

Fig. 3. Flowchart of denoised speech quality estimation using ANN.

5.2. Neural network design

The used network diagram is shown in Fig.4. It is a two layered ANN network. The input variable are $\{c_1^n, c_2^n, c_3^n, c_4^n, c_5^n, c_6^n\}$ obtained by classical criteria. The hidden layer contains 10 neurons and the output layer represents 5 MOS scales. $\{w_{i,j}^k\}$ are the weight parameters. Hidden variable x_0 and z_0 represent the bias parameters. The used activation function is hyperbolic tangent sigmoid.

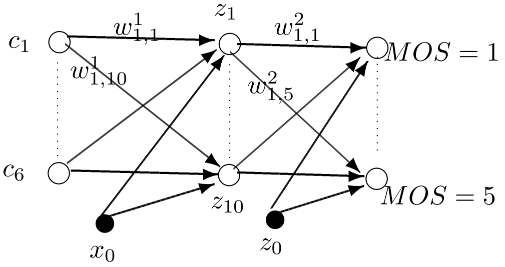

Fig. 4. ANN network diagram.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

The database has been divided into two subsets. First one contains 80% of denoised signal with known MOS scores. It is used for learning phase. The remainder of the database is used to test the proposed ANN. We have computed the error rate which gives an idea about the correctly identification of MOS scores. The error rate is computed for the two subsets: learning and testing. Tab.3 summarize experiments results. Table rows present the used criteria for the learning and testing process. We have tested 6 cases. Firstly, we have considered only PSANDR as attribute for ANN learning and testing. Then, we have added progressively the selected criteria according to their relevance (see previous section). As it is presented in Table 3, we can notice that the lower learning and testing error rate are obtained when all selected criteria are used. A second experience consists of testing separately classical criteria as ANN attributes. In Table 4 we can see clearly that classical criteria and even composite ones cannot estimate MOS scores with acceptable accuracy. However, when selected criteria are used the error rate is reduced to 0.29 which is about 0.3 less than error rate obtained when using separately classical criteria. This fact proves the validity of using selected criteria and ANN for MOS estimation.

Table 3. Evaluation of ANN for all selected criteria.

Criteria	learning error	test error
{PSANDR}	0.61	0.63
{PSANDR,SNR}	0.48	0.51
{PSANDR,SNR,LLR}	0.45	0.46
{PSANDR,SNR,LLR,fwSNR}	0.32	0.37
{PSANDR,SNR,LLR,fwSNR,PESQ}	0.30	0.35
{PSANDR,SNR,LLR,fwSNR,PESQ,WSS}	**0.23**	**0.29**

7. CONCLUSION

Statistical analyses of objective criteria are presented to select the more relevant ones for the specific case of denoised

Table 4. Evaluation of ANN for separated criteria.

Criteria	learning error	test error
SNR	0.53	0.55
LLR	0.60	0.65
fwSNR	0.48	0.51
PESQ	0.43	0.49
WSS	0.48	0.50
PSANDR	0.61	0.63
Covl	0.71	0.73

speech assessment. Objective measures PSANDR, SNR, LLR, fwSNR, PESQ and WSS are kept from a set of 11 studied criteria. The novel proposed methodology to assess denoised speech consists of estimating the subjective MOS score using an Artificial Neural Network as a pattern recognition tools. For each assessed speech signal, scores obtained from selected objective criteria are used as a set of labels for the designed ANN. Experimental results show that the designed ANN using selected criteria leads to improve the ability of objective criterion to estimate subjective score.

REFERENCES

[1] ITU-T P.800, *Subjective assessement methods of the transmission quality,* ITU-T Recommendation P.800, 1996.

[2] ITU-T P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,* ITU-T Recommendation P.835, 2003.

[3] S.R. Quackenbush, T.P. Barnawell and M.A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, 1988.

[4] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech Signal Process*, 1:67-72, 1975.

[5] F. Itakura, S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. Int. Congr. Acoust.,* pp. 17-20, 1978.

[6] N. Kitawaki, H. Nagabuchi and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Select. Areas Commun.,* 6:242248, 1988.

[7] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7, pp. 1278–1281, 1982.

[8] J. Tribolet, P. Noll, B. McDermott and R. E. Crochiere, "A study of complexity and quality of speech waveform coders, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 586-590, 1978.

[9] W. Yang, M. Benbouchta and R. Yantorno, "Performance of a modified bark spectral distortion measure as an objective speech quality measure," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Process., ICASSP*, pp. 541-544, 1998.

[10] ITU-T P.861, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs* $(PSQM)$, ITU-T Recommendation P.861, 1996.

[11] ITU-T P.862, *Perceptual evvaluation of speech quality* $(PESQ)$*, and objective method for end-to-end speech quality assessment of nerrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, 2000.

[12] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, 16(1):229–238, 2008.

[13] A. Ben Aicha and S. Ben Jebara, "Perceptual speech quality measures separating speech distortion and additive noise degradations," *Speech Communication*, 54(4):517–528, 2012.

[14] S. Wang, A. Sekey and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. on Select. Areas in Commun.*, vol. 10, pp. 819-829, 1992.

[15] P. J. Rousseuw, I. Ruts, and J. W. Tukey, "The boxplot: A bivariate boxplot," *The American Statistician*, 53: 382-387, 1999.

[16] J. F. Hair and *all*, *Multivariate Data Analysis*, Prentice Hall, 2006.

[17] A. Ben Aicha, S. Ben Jebara, "Statistical selection of relevent objective criteria for speech enhancement assessement," submitted in *Proc. Advanced Technologies for Signal and Image Processing, ATSIP*, 2014.

[18] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, John Wiley and Sons, 2001.