Quantitative perceptual separation of two kinds of degradation in speech denoising applications

Anis Ben Aicha and Sofia Ben Jebara

Unité de recherche TECHTRA
Ecole Supérieure des Communications de Tunis, 2083 Cité El-Ghazala/Ariana, TUNISIE
anis_ben_aicha@yahoo.fr,sofia.benjebara@supcom.rnu.tn

Abstract

Classical objective criteria evaluate speech quality using one quantity which embed all possible kind of degradation. For speech denoising applications, there is a great need to determine with accuracy the kind of the degradation (residual background noise, speech distortion or both). In this work, we propose two perceptual bounds UBPE and LBPE defining regions where original and denoised signals are perceptually equivalent or different. Next, two quantitative criteria PSANR and PSADR are developed to quantify separately the two kinds of degradation. Some simulation results for speech denoising using different approaches show the usefulness of proposed criteria.

1. Introduction

Evaluation of denoised speech quality can be done using subjective criteria such as MOS (Mean Opinion Score) or DMOS (Degradation MOS) [1]. However, such evaluation is expensive and time consuming so that, there is an increasing interest in the development of robust quantitative speech quality measures that correlate well with subjective tests. Objective criteria can be classified according to the domain in which they operate. We relate for example the Signal to Noise Ratio (SNR) and segmental SNR operating in time domain [2], the Cepstral Distance (CD) and Weighted Slope Spectral distance (WSS) operating in frequency domain [2] and Modified Bark Spectral Distortion (MBSD) operating in perceptual domain [3]. Perceptual measures are shown to have the best chance of predicting subjective quality of speech and other audio signals since they are based on human auditory perception models.

The common point of all objective criteria is their ability of evaluating speech quality using a single parameter which embed all kind of degradations after any processing. Indeed, speech quality measures are basing their evaluation on both original and degraded speeches according to the following application C:

$$C: \begin{array}{l} \mathbb{E}^2 \longrightarrow \mathbb{R} \\ (x,y) \longmapsto c \end{array} \quad (1)$$

where $\mathbb{E}$ denotes the time, frequency or perceptual domain. x (resp. y) denotes original speech (resp. observed speech altered by noise or denoised speech after processing) and c is the score of the objective measure.

Mathematically, C is not a bijection from $\mathbb{E}^2$ to $\mathbb{R}$. It means that it is possible to find a signal y' which is perceptually different from y but has the same score than the one obtained with y ($c(x,y) = c(x,y')$). We relate for example the case of an original signal x which is corrupted by an additive noise to construct the signal y. Then, x is coded and decoded using a CELP coder to obtain the signal y'. It is obviously that the degradation noticed in both y and y' are not the same. Degradation of y is heard as a background noise and the degradation of y' is perceptually heard as distortion of original signal. However, in a previous work, we show that they have the same SNR [4].

In this paper, we aim improving speech quality evaluation by separating two kinds of degradation which are the additive residual noise and the speech distortion. Each degradation will be evaluated using its adequate criterion so that the non bijection C will be avoided and replaced by a bijection one characterized by a couple of outputs instead of a single output. Moreover, thanks to the advantage of perceptual tools in the evaluation of speech quality, the new couple of criteria will be based on auditor properties of human ear.

2. Study context: speech denoising

Before defining novel criteria of speech quality evaluation, let's define the different kinds of degradation altering speech. Without loss of generality, we consider the speech denoising application and we use spectral denoising approaches. They are viewed as a multiplication of noisy speech spectrum $Y(m,k)$ by a real positive coefficient filter $H(m,k)$ (see for example [5]). The estimated spectrum of clean speech is written

$$\hat{S}(m,k) = H(m,k)Y(m,k), \quad (2)$$

where m (resp. k) denotes frame index (resp. frequency index).

The estimation error spectrum $\xi(m,k)$ is given by

$$\xi(m,k) = S(m,k) - \hat{S}(m,k). \quad (3)$$

We assume that speech and noise are uncorrelated. Thus, the estimated error power spectrum is given by

$$\begin{array}{rl} E\{|\xi(m,k)|^2\} = & [H(m,f)-1]^2 E\{|S(m,k)|^2\} \\ & +H(m,k)^2 E\{|N(m,k)|^2\}, \end{array} \quad (4)$$

where $|N(m,k)|^2$ denotes the noise power spectrum.

Since $0 < H(m,k) < 1$, the first term of Eq. 4 expresses the 'attenuation' of clean speech frequency components. Such degradation is perceptually heard as a distortion of clean speech. However, the second term expresses the residual noise which is perceptually heard as a background noise. Since, it is additive, it is possible to formulate it as an 'accentuation' of clean speech frequency components.

3. Proposed perceptual characterization of audible degradation

We aim to perceptually characterize the degradation altering denoised speech. Hence, auditory properties of human ear are

considered. More precisely, the masking concept is used: a masked signal is made inaudible by a masker if the masked signal magnitude is below the perceptual masking threshold MT. In our case, both degradation can be audible or inaudible according to their position regarding the masking threshold. We propose to find decision rules to decide on the audibility of residual noise and speech distortion by using the masking threshold concept. If they are audible, the audibility rate will be quantified according to the proposed criterion. There are many techniques to compute masking threshold MT, we use in this paper Johnston model well known for its simplicity and well used in coding context [6].

3.1. Perceptual characterization of audible noise

According to MT definition, it is possible to add to the clean speech power spectrum, the MT curve (considered as a 'certain signal') so that the resulting signal (obtained by inverse FFT) has the same audible quality than the clean one. The resulting spectrum is called *Upper Bound of Perceptual Equivalence* "$UBPE$" and is defined as follows

$$UBPE(m,k) = \Gamma_s(m,k) + MT(m,k), \tag{5}$$

where $\Gamma_s(m,k)$ is the clean speech power spectrum.

When some frequency components of the denoised speech are above $UBPE$, the resulting additive noise is heard.

3.2. Perceptual characterization of audible distortion

By duality, some attenuations of frequency components can be heard as speech distortion. Thus, by analogy to $UBPE$, we propose to calculate a second curve which expresses the lower bound under which any attenuation of frequency components is heard as a distortion. We call it *Lower Bound of Perceptual Equivalence* "$LBPE$". To compute $LBPE$, we used the audible spectrum introduced by Tsoukalas *and al* for audio signal enhancement [7]. In such case, audible spectrum is calculated by considering the maximum between clean speech spectrum and masking threshold.

When speech components are under MT, they are not heard and we can replace them by a chosen threshold $\sigma(m,k)$.
The proposed $LBPE$ is defined as follows

$$LBPE(m,k) = \begin{cases} \Gamma_s(m,k) & \text{if } \Gamma_s(m,k) \geq MT(m,k) \\ \sigma(m,k) & \text{otherwise .} \end{cases} \tag{6}$$

The choice of $\sigma(m,k)$ obeys only one condition $\sigma(m,k) < MT(m,k)$. During this work, we choose it equal 0 dB.

3.3. Usefulness of $UBPE$ and $LBPE$

Using $UBPE$ and $LBPE$, we can define three regions characterizing the perceptual quantity of denoised speech: frequency components between $UBPE$ and $LBPE$ are perceptually equivalent to the original speech components, frequency components above $UBPE$ contain a background noise and frequency components under $LBPE$ are characterized by speech distortion. This characterization constitutes our idea to identify and detect audible additive noise and audible distortion. As illustration, we present in Fig. 1 an example of speech frame power spectrum and its related curves $UBPE$ (upper curve in bold line) and $LBPE$ (bottom curve in dash line). The clean speech power spectrum is, for all frequencies index, between the two curves $UBPE$ and $LBPE$. We remark that the two curves are the same for most peaks. It means that for these frequency intervals, any kind of degradation altering speech will be audible. If it quite over $UBPE$, it will be heard as background noise. In the opposite case, it will be heard as speech distortion.

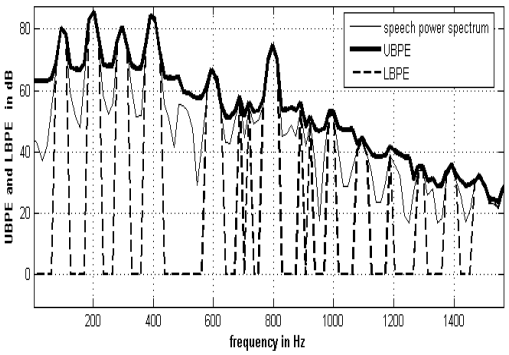

Figure 1: An example of $UBPE$ and $LBPE$ in dB of clean speech frame.

4. Audible degradation estimation

4.1. Audible additive noise PSD estimation

Once $UBPE$ calculated, the superposition of denoised signal power spectrum and $UBPE$ leads to separate two cases. The First one corresponds to the regions of denoised speech power spectrum which are under $UBPE$. In such case, there is no audible residual noise. In the second case, some denoised speech frequency components are above $UBPE$, the amount above $UBPE$ constitutes the audible residual noise. As illustration, we represent in Fig.2 an example of denoised speech power spectrum and its related $UBPE$ curve calculated from clean speech. The used denoising approach is spectral subtraction [5]. From Fig.2, we notice that frequency regions between 1 kHz and 2 kHz are above $UBPE$, they hence contain residual audible noise. In term of listening tests, such residual noise is annoying and constitutes in some cases the musical noise. Such musical noise is well popular and constitutes the main drawback of spectral subtraction.

Once the $UBPE$ is calculated, it is possible to estimate the audible power spectrum density of residual noise using a simple subtraction when it exists. Hence, the residual noise power spectrum density PSD is written

$$\Gamma_n^p(m,k) = \begin{cases} \Gamma_{\hat{s}}(m,k) - UBPE(m,k) \\ \text{if } \Gamma_{\hat{s}}(m,k) > UBPE(m,k) \\ 0 \qquad \text{otherwise} \end{cases} \tag{7}$$

where $\Gamma_{\hat{s}}(m,k)$ denotes the PSD of denoised speech and the suffix p designs the perceptually sense of the PSD.

4.2. Audible speech distortion PSD estimation

We use the same methodology as the one used for residual background noise. We represent in Fig.3 an example of denoised speech power spectrum and its related curve $LBPE$ calculated from the clean speech. We notice that some regions are under $LBPE$ (for example regions between 1.5 kHz and 2 kHz), they

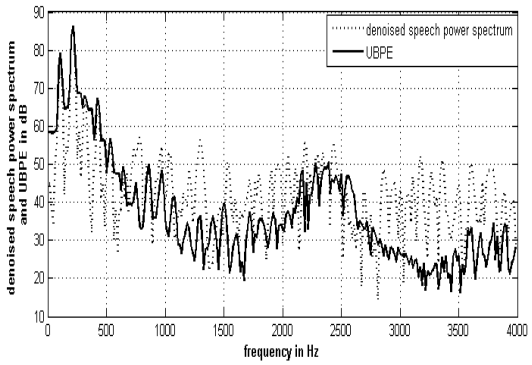

Figure 2: Superposition of a denoised speech power spectrum and its related clean speech $UBPE$.

hence constitute the audible distortion of the clean speech. In term of listening tests, they are completely different from residual background noise. They are heard as a loss of speech tonality.
It is possible to estimate the audible distortion PSD Γ_d^p as follows

$$\Gamma_d^p(m,k) = \begin{cases} LBPE(m,k) - \Gamma_{\hat{s}}(m,k) \\ \text{if } \Gamma_{\hat{s}}(m,k) < LBPE(m,k) \\ 0 \qquad \text{otherwise .} \end{cases} \tag{8}$$

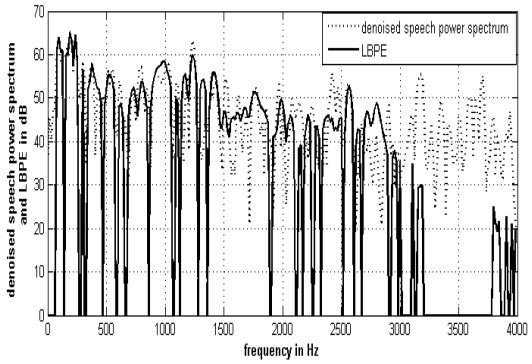

Figure 3: Superposition of a denoised speech frame and its related clean speech $LBPE$.

5. Audible degradation evaluation

In this section, we detail the proposed approach to quantify separately the two kinds of degradation. The assessment of the denoised speech quality by means of two parameters permits to overcome the problem of non bijection of classic objective evaluation and to better characterize each kind of speech degradation. Hence, instead of the application defined in Eq. 1, we develop a novel application from perceptual domain to $\mathbb{R}^2$

$$\begin{aligned} C: \quad & \mathbb{E}^2 \longrightarrow \mathbb{R}^2 \\ & (x,y) \longmapsto (PSANR, PSADR) \end{aligned} \tag{9}$$

where $PSANR$ and $PSADR$ are two parameters related respectively to the residual noise and the distortion.

The definition of $PSANR$ and $PSADR$ is inspired from the SNR definition which is the ratio between signal energy and noise energy. Thanks to Parseval theorem it can be calculated in frequency domain. Moreover, since the $UBPE$ and $LBPE$ are perceptually equivalent to the original signal, the proposed definition uses the energy of $UBPE$ and $LBPE$ instead of the energy of the clean speech. The time domain signal related to $UBPE$ is called "upper effective signal" whereas the time domain signal related to $LBPE$ is called "lower effective signal". In the following subsection, we define the proposed criteria.

5.1. Perceptual noise criterion PSANR

The perceptual residual noise criterion is defined as the ratio between the upper effective signal which is the $UBPE$ and the audible residual noise. The Perceptual Signal to Audible Noise Ratio $PSANR(m)$ of frame m is calculated in frequency domain (due to the Parseval theorem) and it is formulated as follows

$$PSANR(m) = \frac{\sum_{k=1}^{N} UBPE(m,k)}{\sum_{k=1}^{N} \Gamma_n^p(m,k)}. \tag{10}$$

5.2. Perceptual distortion criterion PSADR

By the same manner, we define the Perceptual Signal to Audible Distortion Ratio $PSADR(m)$ of frame m as a ratio between the lower effective signal which is $LBPE$ and the audible distortion. The $PSADR(m)$ is given by:

$$PSADR(m) = \frac{\sum_{k=1}^{N} LBPE(m,k)}{\sum_{k=1}^{N} \Gamma_d^p(m,k)}. \tag{11}$$

5.3. PSANDR criteria

to compute the global $PSANR$ and $PSADR$ of the total speech sequence, we are referred to the segmental SNR SNR_{seg} thanks to its better correlation with subjective tests when compared to the traditional SNR. The principle of segmental SNR consists on determining the SNR for each frame $SNR(m)$ and then calculating their geometric mean over the total number of frames $SNR_{seg} = \sqrt[N]{\prod_m^N SNR(m)}$ [2]. Moreover, since the SNR and SNR_{seg} are usually expressed in dB. The geometric mean is equivalent to the arithmetic mean in log domain.
Using this approach, we compute the global $PSANR$ and $PSADR$ for a given sequence of speech. Next, the couple ($PSANR$,$PSADR$) defines the new criterion to evaluate both kinds of degradation. We call it *Perceptual Signal to Audible Noise and Distortion Ratio* "$PSANDR$".

6. experimental results

6.1. Test signals

To show the ability of $PSANDR$ to take into account the perceptual effect of an additive noise, we add artificial noise, constructed from the masking threshold by multiplying it with a factor $\alpha \geq 0$ ($y(n) = s(n) + \alpha MT(n)$). In Fig.4, we represent the evolution of SNR_{seg}, $PSANR$ and $PSADR$ versus α. For the range of α between 0 and 1, SNR_{seg} decreases which means that there is a degradation of speech. This fact is true in term of signal to noise ratio but not true in term of perceptual sense, because the power of added artificial noise don't overtake MT. With $PSANR$, the amount of audible noise is

null (see Eq. 7) and the $PSANR$ is infinity which is truncated to 35 dB in our simulations. For $\alpha > 1$, the background noise becomes audible and the $PSANR$ decreases as α increases but remains above SNR_{seg}. This is explained by the ability of the clean speech to mask a certain portion of the added noise.

We notice that for any value of α, the second term $PSADR$ is still constant and is equal to 35 dB. In fact, there is no distortion of the clean speech and the only audible degradation is the background noise.

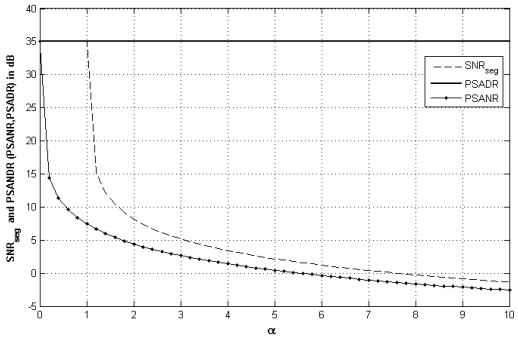

Figure 4: Evolution of SNR_{seg}, $PSANR$ and $PSADR$ versus α in case of additive noise.

6.2. real signals

Let's now compare some denoising techniques by means of the new objective criteria. We propose to denoise a corrupted signal, by gaussian noise with $SNR = 0$ dB, using the following techniques:

- Classical wiener filtering [5].
- Perceptual filtering proposed by Gustafsson *and al* in [8] which consists in masking the residual noise and allowing a variable speech distortion.
- Modified wiener technique [9]. In this technique, the shape of the tones is used as a selective parameter to detect and eliminate musical tones.

Evaluation of denoising quality is done using classic objective criteria (segmental SNR, WSS, MBSD) and the proposed PSANDR. Results are resumed in Tab.1. In term of SNR_{seg}, the used techniques are comparable even if there is a little improvement noticed with perceptual technique. But, subjective tests show that the denoised signals are completely different. Using WSS criterion, the best score is obtained with perceptual technique and it is nearly equal to the noisy speech score. Although, subjective tests show that the two signals are perceptually different. Indeed, the denoised speech using perceptual technique is heard as distorted version of clean speech and not as clean speech with background noise. In term of MBSD, the perceptual technique is also the best. However, this technique is characterized by a loss of the speech tonality comparing to wiener technique. Thus, we can see that classic evaluation tools don't give any idea of the kind and nature of the degradation of the signals. $PSANR$, giving idea about residual noise, shows that perceptual technique is the best one regarding noise attenuation. $PSADR$, determining the distortion of the denoised signals, shows that the important distortion is obtained using perceptual technique. These observations are confirmed by subjective tests.

Table 1: Evaluation of denoised signals.

	SNR_{seg} dB	WSS	MBSD	PSANR dB	PSADR dB
noisy speech	-4.30	46.07	2.32	-3.90	17.27
wiener technique	1.05	74.25	0.28	5.04	**7.53**
modified wiener	1.13	69.63	0.19	5.54	7.01
perceptual technique	**1.62**	**45.41**	**0.15**	**12.71**	6.93

7. Conclusion

The spectral and perceptual analysis of the degradation, in the case of denoised speech, imposes to separate between residual noise and signal distortion. We first propose two curves $UBPE$ and $LBPE$ to calculate the audible residual noise and audible distortion. Next, two parameters $PSANR$ and $PSADR$ characterizing the two kinds of degradation are developed. Simulation results comparing different denoising approaches and classical objective measures, show a better characterization of degradation nature of denoised signal. The calculation of the degree of correlation of the proposed criteria with MOS criterion constitutes the perspectives of our work.

8. References

[1] Recommandation UIT-T P.800. Mthodes d'valuation subjective de la qualité de transmission, 1996.

[2] J.H.L. Hansen and B.L. Pellom "An effective quality evaluation protocol for speech enhancement algorithms " Int. Conf. on Spoken Language Processing ICSLP, Austria 1998.

[3] W. Yang, M. Benbouchta and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech measure," Proc. Int. Conf. on Acoustics, Speech and Signal Processing ICASSP, vol 1, pp. 541-544, 1988.

[4] A. Ben aicha et S. Ben Jebara, "Caractérisation perceptuelle de la dégradation apportée par les techniques de débruitage de la parole," submited in Traitement et Analyse de l'Information Méthodes et Applications TAIMA, Tunisia 2007.

[5] J.S Lim and A.V Oppenheim, Enhancement and Bandwidth Compression of Noisy Speech, in Proc. IEEE, vol. 67, pp. 1586-1604, 1979.

[6] J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," IEEE J. Select. Areas Commun, vol 6, pp. 314-323, 1988.

[7] D. E. Tsoukalas, J. Mourjopoulos and G. Kokkinakis, "Speech enhancement based on audible noise suppression," IEEE Trans. Speech and Audio Processing, vol. 5, no. 6, pp. 497- 514, November 1997.

[8] S. Gustafsson, P. Jax and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving backround noise characteristics," in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Seattle, WA, pp. 397-400, May 1998

[9] A. Ben Aicha, S. Ben Jebara and D. Pastor "Speech denoising improvement by musical tones shape modification," International Symposium on Communication, Control and Signal Processing ISCCSP, Morocco 2006.