An ERB Loudness Pattern Based Objective Speech Quality Measure

Guo Chen[1], *Vijay Parsa*[1,2], *and Susan Scollie*[2]

[1]Department of Electrical & Computer Engineering and [2]National Centre for Audiology
University of Western Ontario, London, Ontario, Canada
E-mails: guo.chen@nca.uwo.ca, parsa@nca.uwo.ca, scollie@nca.uwo.ca

Abstract

This paper presents an objective speech quality measure which is based on loudness patterns using the equivalent rectangular bandwidth (ERB) scale. The proposed measure, called the loudness pattern distortion (LPD), is computed from the differences between the loudness patterns of the original and processed speech. The LPD measure takes into account the transmission through the outer and middle ear, the calculation of an excitation pattern from the physical spectrum, and the transformation of an excitation pattern to a loudness pattern. The effectiveness of the proposed measure was demonstrated by experimental evaluations in comparison with the standard ITU-T P.862 (PESQ) using three coded speech database of the ITU-T P-series Supplementary 23.

Index Terms: speech quality, objective measure, loudness pattern

1. Introduction

A speech quality measure is highly desirable and beneficial in the field of speech processing, especially as a valuable assessment tool for the development of speech coding and enhancing techniques. During the course of designing a speech compression system, it is desirable to have a speech quality measure for indicating the amount of distortion introduced by the compression algorithm and for optimizing the system structure. For example, in mobile communications and voice over internet protocol (VoIP), the speech signal is compressed into a compact representation before transmission and is reconstructed at the other end. The speech encoding/decoding process invariably introduces distortion and a speech quality measure allows for the relative comparison of various speech coding techniques. Commonly, two approaches, subjective and objective, are used for measuring the speech quality. Subjective measures are based on the perceptual ratings by a group of listeners, who subjectively rank the quality of speech. The most widely used subjective test is the absolute category rating (ACR) method [1] which results in a mean opinion score (MOS). In the ACR test, listeners rate the speech quality by using a five-point scale, in which the quality is represented by five grades - excellent(5), good(4), fair(3), poor(2), and bad(1). Typically, the ratings are collected from a pool of listeners and the arithmetic mean of their ratings forms the MOS ratings. While subjective opinions of speech quality are preferred as the most trustworthy criterion for speech quality, they are also time-consuming, expensive and not reproducible. In contrast, objective measures, which assess speech quality by using the extracted physical parameters, are less expensive to administer, save time, and give more consistent results. Also, the results conducted at different times and with different testing facilities can be directly compared. Thus, good objective measures are highly desirable in practical applications.

In the past decades, objective speech quality measures have received considerable attention [2, 3]. In general, objective measures can be divided into two groups. One is intrusive evaluation, which assesses speech quality by measuring the "distortion" between the input and output signals, and mapping the distortion values to the predicted quality metric [4, 5, 6, 7]. The other is non-intrusive evaluation, which assesses speech quality only based on the output speech signal of a system under test [8, 9, 10]. According to the physical parameters exploited, objective measures can be classified into four groups: (i) time domain measures, (ii) linear predictive coefficient (LPC) based measures, (iii) frequency domain measures, and (iv) perception-based measures [2, 3]. Based on previous research, it has been found that perception-based speech quality measures exhibit higher correlations with subjective quality ratings [3, 4, 6, 7, 9]. It is well known that that the peripheral auditory system of humans, which is highly consistent from one person to another, plays an important role during speech quality rating. Normally, in a listening test, the auditory information conveyed by speech signal is first preprocessed by the peripheral auditory system and the highly compacted data obtained are then sent to the high-level brain function. The subjective quality rating is finally performed based on these data. Therefore, it is intuitive for us to design an objective quality measure by emulating this biological preprocessing and by comparing the reduced representations of the original and processed speech.

Currently almost all perception-based objective measures, such as techniques reported in [4, 6, 7, 9, 11, 12], are based on Zwicker's auditory model [13], which has lead to the definition of the Bark scale. Recently, more accurate psycho-acoustical experiments have lead to a revised Zwicker's model, i.e. the Moore and Glasberg model (the M-G model) [14, 15, 16]. In the M-G model, the notched noise method [14] was utilized to measure the auditory filter bandwidth rather than the classical masking methods involving a narrow-band masker and probe tone [13]. The M-G model has lead to the equivalent rectangular bandwidth (ERB) scale. In general, on the ERB scale the auditory-filter bandwidth is smaller than on the Bark scale, a difference which becomes larger for lower frequencies. Moreover, the M-G model can be used to better explain how equal-loudness contours change as a function of level, why loudness remains constant as the bandwidth of a fixed-intensity sound increases up to the critical bandwidth, and the loudness of partially masked sounds [17].

In this paper, we propose a novel objective speech quality measure, which is based on the loudness patterns of the M-G auditory model for normal hearing. The proposed loudness pattern distortion (LPD) measure emulates several properties of human auditory system, such as the transmission through the outer and middle ear, the calculation of an excitation pattern from the input physi-

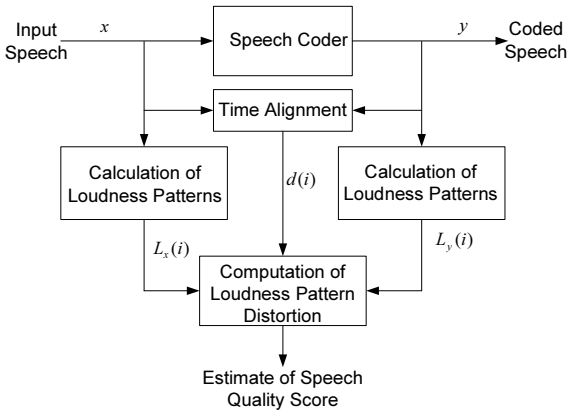

Figure 1: The block diagram of the LPD measure

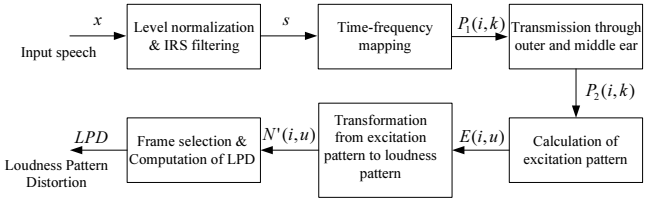

Figure 2: The calculation of loudness pattern

cal spectrum, and the transformation of an excitation pattern to a loudness pattern. The auditory properties of the M-G model will lead us to a new objective quality measure in the perceptual space.

2. The loudness pattern distortion measure

The block diagram of the LPD measure is shown in Fig.1. The time delay information is first estimated by using the cross-correlation method as stated in [5]. After the time delay information is calculated, both the original speech x and its coded version y are then separately processed by identical operations, leading to what we shall refer to as the loudness patterns, L_x and L_y, respectively. The quality measure is then defined by an appropriate distance between these two specific loudness patterns. The calculation of loudness pattern is shown in Fig. 2, which is based on the M-G model reported in [14, 15, 16]. The procedure is briefly formulated as follows.

2.1. Level normalization & IRS filtering

The level of the input speech is first normalized to -26 dBov and the receive-side modified intermediate reference system (IRS) filter is then applied to reflect the characteristics of handsets used in subjective listening tests[18].

2.2. Time-frequency mapping

The input speech signal with 8 kHz sampling rate were segmented into frames of 32 ms with an overlap of 50%, denoted by $s(i), i = 1, ..., I$. Each frame was transformed to the frequency domain using a Hamming window and a short-time FFT. The real and imaginary components were squared and added to obtain the short-time power spectrum, $P_1(i,k)$, where k represents the frequency scale.

2.3. Transmission through the outer and middle ear

The frequency response of the outer and middle ear was modelled by a frequency dependent weighting function as below [12]

$$H(k) = 10^{\frac{-2.184\left(\frac{f(k)}{1000}\right)^{-0.8}+6.5e^{\left(-0.6\left(\frac{f(k)}{1000}-3.3\right)^2\right)}-10^{-3}\left(\frac{f(k)}{1000}\right)^{3.6}}{20}}, \quad (1)$$

where $f(k) = k \cdot \frac{8000}{256}$ is in Hertz at the DFT bin k. The weighted power spectrum becomes $P_2(i,k) = H(k) \cdot P_1(i,k)$.

2.4. Calculation of excitation pattern

The excitation pattern represents the output level of successive auditory filters as a function of their center frequencies. Each auditory filter represents the frequency selectivity of the inner ear at a particular frequency, where the filter shape varies with the input level. In the LPD measure, we used the ERB scale to represent the bandwidth of the auditory filters. While the ERB is a measure for the bandwidth of auditory filters, the ERB-rate is a value on the ERB scale which is conceptually related to the Bark scale [13, 17]. The ERB and ERB-rate are approximated by the formulae: ERB/Hz $= \quad 24.7(4.37f/\text{kHz} + 1)$ and ERB-rate/ERB $= 21.4log_{10}(4.37f/\text{kHz} + 1)$.

In the M-G model, instead of partitioning the excitation transformation into two distinct steps of the spectrum integration over unit Bark-lengths and the frequency spreading as in Zwicker's model [13], a more compact representation of the excitation pattern $E(f_c)$ at frequency f_c is given by $E(f_c) = \int_0^\infty W(f, f_c, P_2)P_2(f)df$, where $W(f, f_c, P_2)$ is the auditory filter and $P_2(f)$ is the input power spectrum. The auditory filters are level-dependent on the argument P_2 and the shapes are modelled by the so-called roex-filter shapes ("rounded exponentials"). The shape of a roex-filter is defined by $W(f, f_c, P_2) = \left(1 + p\frac{|f-f_c|}{f_c}\right)\exp\left(-\frac{|f-f_c|}{f_c}\right)$.

The parameter p controls the slopes of the auditory filters. A more precise modelling [14, 15, 16] assumes asymmetric auditory filters parameterized by upper and lower slopes values p_u and p_l, respectively. Correspondingly, in calculating the output of a given filter arising from a given component, the p value is computed depending on whether the filter is centered above or below the frequency of that component, and the total input level of the auditory filter. For the slope of the low frequency skirt of the auditory filter, the variation can be described in terms of the parameter p_l. Let Q denote that input level in dB/ERB, and let $p_{l(Q)}$ denote the value of p_l at level Q, then $p_{l(Q)} = p_{(51)} - \frac{0.35p_{(51)}(Q-51)}{p_{(51,1k)}}$, where $p_{(51)}$ is the value of p at that center frequency for the input level of 51 dB/ERB and $p_{(51,1k)}$ stands for the value of p_l at 1 kHz for a input level of 51 dB/ERB. The value of $p_{(51)}$ can be calculated by $p_{(51)} = 4f_c/\text{ERB}(f_c)$. On the other hand, the changes in slope of the high-frequency skirt of the auditory filter with level tend to be rather small and correspondingly the value of p_u is set to be the same as p_{51}. More details on the calculation of the slope p can be found in [14, 15, 16].

2.5. Transformation from excitation patterns to loudness patterns

A loudness pattern (loudness density) in sone per ERB-rate can be calculated from the associated excitation pattern. While the excitation pattern represents the distribution of excitation along the basilar membrane, the loudness per ERB-rate (the specific loudness pattern) corresponds more closely to the distribution of neural activity. Accordingly, the specific loudness pattern is closely related to the subjective perception of speech signal. To calculate the loudness pattern it requires the availability of a computational procedure such as the one given by Zwicker's model [13]. Recently a modified version of Zwicker's model incorporating a more analytical formulation was introduced by Moore and Glasberg [16]. This revised model has been shown to account more accurately for various subjective loudness data. From the excitation patterns, the loudness patterns are calculated for three different cases as follows [16].

Case 1: IF $(10^9 \geq E(f_c) \geq E_{THRQ}(f_c))$ THEN

$$N'(f_c) = C[(G(f_c)E(f_c) + A(f_c))^{\alpha(f_c)} - A(f_c)^{\alpha(f_c)}] \quad (2)$$

Case 2: IF $(E(f_c) > 10^9)$ THEN

$$N'(f_c) = C\left(\frac{E(f_c)}{1.115}\right)^{0.2} \quad (3)$$

Case 3: IF $(E(f_c) < E_{THRQ}(f_c))$ THEN

$$N'(f_c) = C\left(\frac{2E(f_c)}{E(f_c) + E_{THRQ}(f_c)}\right)^{1.5} [(G(f_c)E(f_c) + A(f_c))^{\alpha(f_c)} - A(f_c)^{\alpha(f_c)}], \quad (4)$$

where C is a constant with the value of 0.047 and here the frame index i is omitted for simplification. The frequency dependent constants used are approximated by the following equations as stated in [11],

$$G(f) = \frac{E_{THRQ}(500Hz)}{E_{THRQ}(f)} \quad (5)$$

$$E_{THRQ}(f) = 1.4 + 0.4 \times 10^{0.3(f/kHz)^{-0.8}} \quad (6)$$

$$\alpha(f) = 0.171 + \frac{0.032085}{0.1 + G(f)^{0.25}} \quad (7)$$

$$A(f) = 2.8 + \frac{2}{0.1 + G(f)^{0.25}}. \quad (8)$$

Next, the obtained specific loudness pattern across frequency is summed to form the overall loudness pattern of a speech frame. In practice, to achieve greater accuracy, the specific loudness is calculated at 0.1-ERB intervals, and the sum is then divided by 10 as below,

$$L_x(i) = \frac{1}{10}\sum_{u_{f_c}=1}^{U} N'_x(i, u_{f_c}), \quad L_y(i) = \frac{1}{10}\sum_{u_{f_c}=1}^{U} N'_y(i, u_{f_c}), \quad (9)$$

where U is the total number of auditory filters on the ERB scale and u_{f_c} is the u-th auditory filter with the center frequency f_c.

2.6. Frame selection and Computation of LPD

In the calculation of the LPD measure, only frames that meet or exceed the set thresholds (in sone) in both $L_x(i)$ and $L_y(i)$ are used. For the original speech, the threshold was set to 15 sones below the loudness of the peak frame in L_x. For the coded speech, the threshold was set to 20 sones below the loudness of the peak frame in L_y. Only frames that meet or exceed both of these sone thresholds are retained. Assuming that there are I' speech frames retained, the LPD measure is calculated by:

$$LPD = \frac{1}{I'}\sum_{i=1}^{I'}[L_x(i) - L_y(i)] + L_{offset}, \quad (10)$$

where L_{offset} is an offset constant.

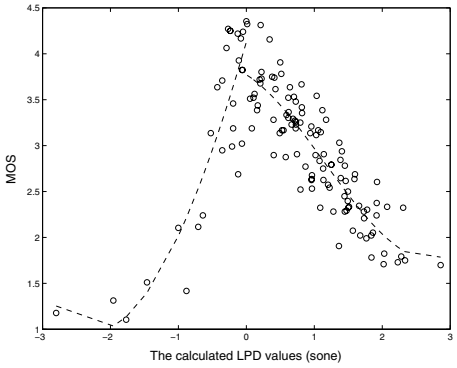

Figure 3: The plots of the computed LPD values versus the condition-averaged MOS ratings. The dash lines are the fitting of two third-order polynomials. The left curve:$EMOS = -0.0570t^3 + 0.3856t^2 + 2.5521t + 4.1268$; The right curve:$EMOS = 0.1431t^3 - 0.4946t^2 - 0.4503t + 3.7708$.

3. Experimental results

As a preliminary evaluation of the proposed LPD measure, we performed a comparison test with the ITU-T standard P.862 (PESQ) [6]. The experimental data consist of 528 subjective MOS ratings which include three subjective MOS databases (English, French, Japanese) obtained in listening opinion tests as described in Experiment One of the ITU-T P-Series Supplement 23 [19]. Each of these databases contains 44 sentence pairs spoken by four talkers (two female and two male) and each sentence pair stands for one condition under test. The correlation coefficient (denoted by ρ) and standard error of estimate (denoted by ϵ), defined in [2], were used to evaluate the performance. In order to predict MOS ratings, we used a third-order polynomial to fit the various scatter plots as suggested by [10], i.e.

$$EMOS = at^3 + bt^2 + ct + d. \quad (11)$$

Here t is the LPD values and $EMOS$ is the corresponding estimated MOS values as "predicted" by the fitted function.

Fig. 3 presents the computed LPD values versus the condition-averaged MOS ratings (132 subjective scores in total). From Fig. 3, it can be observed that the values of MOS ratings exhibited an asymmetrical distribution with respect to the LPD values. In order to reflect this asymmetrical property, we used two third-order polynomials to fit the scatter plots as shown in Fig. 3. One is for the plots with negative LPD values while the other is for those with non-negative LPD values.

Table 1: Performance of the LPD measure compared with the PESQ with condition-averaged MOS ratings

Speech Database	ρ		ϵ	
	LPD	PESQ	LPD	PESQ
P.Sup23 Exp1A(French)	0.940	0.918	0.268	0.421
P.Sup23 Exp1D(Japanese)	0.915	0.937	0.277	0.251
P.Sup23 Exp1O(English)	0.945	0.941	0.266	0.316
Total (3 Databases)	0.895	0.869	0.352	0.403

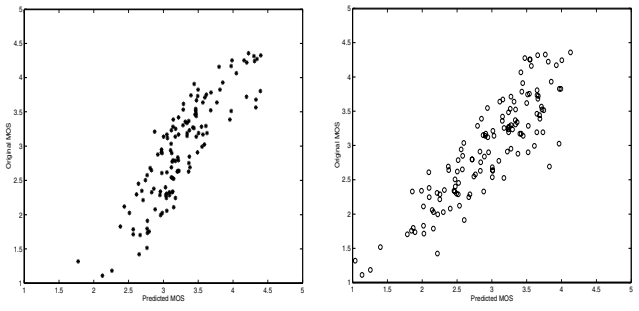

Figure 4: The predicted condition-averaged MOSs for the whole total databases (three databases). *: the results of the PESQ (ρ=0.8691,ϵ=0.4026); o: the results of the LPD measure (ρ=0.8946,ϵ=0.3522).

The performance of the LPD measure was evaluated in terms of the correlation coefficient and standard error of estimate between the actual MOS ratings and the output values of the LPD measure, which were obtained after the third-order polynomial regression analysis. Table II presents the experimental results of the ITU-T standard P.862 (PESQ) and the LPD measure. Fig. 4. shows the plots of the predicted results of the PESQ and the LPD measure. As shown in Table II, the LPD measure outperformed the P.862 for the subjective databases Exp1A and Exp1O but slightly lowered for the database Exp1D. For the entire MOS database, the correlation of the proposed method attained 0.8946 with a standard error of 0.3522, which was better than that of the PESQ.

4. Conclusions

In this paper we presented a new objective speech quality measure which is based on the loudness patterns of the Moore and Glasberg's auditory model. The effectiveness of the proposed loudness pattern distortion (LPD) measure was demonstrated by experimental evaluations in comparison with the standard ITU-T P.862 (PESQ). The experimental results show that the correlation of the proposed measure reached 0.8946 across the entire databases which is better than that of the standardized PESQ.

5. Acknowledgements

We gratefully acknowledge the financial support by the Oticon Foundation, Denmark, the Ontario Rehabilitation Technology Consortium, Canada, and the NSERC, Canada.

6. References

[1] ITU, "Methods for subjective determination of transmission quality," *ITU-T P.800 Recommendation*, Aug. 1996.

[2] S.R.Quackenbush, T.P.Barnwell-III, and M.A.Clements, *Objective Measures of Speech Qaulity*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] A.Rix, "Perceptual speech quality assessment - a review," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 3, pp. 1056–1059.

[4] J.G.Beerends and J.A.Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, March 1994.

[5] S.Voran, "Objective estimation of perceived speech quality - part i. development of the measuring normalizing block technique," *IEEE Transactions on speech and audio processing*, vol. 7, no. 4, pp. 371–382, July 1999.

[6] ITU, "Perceptual evaluation of speech quality," *ITU-T P.862*, 2001.

[7] G.Chen, S.Koh, and I.Soon, "Enhanced itakura measure incorporating masking properties of human auditory system," *Signal Processing*, vol. 83, pp. 1445–1456, 2003.

[8] G.Chen and V.Parsa, "Non-intrusive speech quality evaluation using an adaptive neuro-fuzzy inference system," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 403–406, May 2005.

[9] G.Chen and V.Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Philadelphia,USA, March 2005, vol. 1, pp. 385–388.

[10] ITU, "Single ended method for objective speech quality assessment in narrow-band telephony applicaitons," *ITU-T P.563*, 2004.

[11] T.Thiede, *Perceptual audio quality assessment using a non-linear filter bank*, Ph.D. thesis, Technical University of Berlin, Berlin, 1999.

[12] ITU, "Method for objective measurements of perceived audio quality," *ITU-R BS.1387-1*, Nov 2001.

[13] E.Zwicker and H.Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, Germany, 1990.

[14] B.R.Glasberg and B.C.J.Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[15] B.C.J.Moore, B.R. Glasberg, and T.Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–239, Apr. 1997.

[16] B.C.J.Moore and B.R. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hearing Research*, vol. 188, pp. 70–88, 2004.

[17] B.C.J. Moore, *An introduction to the psychology of hearing*, Academic Press, Boston,MT, 2003.

[18] ITU, "Subjective performance assessment of telephone-band and wideband digital codecs," *ITU-T P.830*, 1996.

[19] ITU, "Itu-t coded-speech database," *ITU-T P-series supplementary 23*, 1998.