



Enhanced Itakura measure incorporating masking properties of human auditory system

Guo Chen*, Soo Ngee Koh, Ing Yann Soon

Communication Research Laboratory, Nanyang Technological University, School of Electrical & Electronic Engineering, Block S2, S2-B4c-17, 40 Nanyang Avenue, Singapore 639798, Singapore

Received 2 May 2002; received in revised form 5 February 2003

Abstract

A new enhanced Itakura (E-Itakura) speech distortion measure is proposed in this paper. It incorporates masking properties of the human auditory system into the original Itakura measure. Inaudible noise components masked by speech signals are excluded from the calculation of the E-Itakura measure, while the intrinsic advantage of the Itakura measure is retained. The proposed new measure has been compared with the original Itakura distortion, frequency-weighted Itakura spectral distortion, cepstral distance and Bark spectral distortion measures. The comparison results show that the correlation between the original Itakura measure with speech quality has been improved from 0.73 to 0.89 with the incorporation of the enhancement feature, and that the E-Itakura measure offers a more consistent indication of the subjective quality of speech.

© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Speech distortion measure; Masking properties; Itakura measure

1. Introduction

An objective speech distortion measure that correlates well with a subjective speech quality measure is highly desirable and useful in the field of speech processing, especially as a valuable assessment tool for the development of speech coding and enhancing algorithms. The accuracy of a speech distortion measure is determined by correlating it with a known subjective measure of speech quality, such as the Mean Opinion Score (MOS), and the degree of correlation between the MOS data set and the distortion measure data set indicates the effectiveness of the objective measure. Detailed discussions on this issue can be found in [8].

The original Itakura measure [1] has been widely used in various research fields of speech processing. Though the Itakura measure can indicate the matching error between the clean reference spectrum and the noisy test spectrum well, its performance usually deteriorates when the processed speech signals are degraded heavily [13]. Soong and Li [6,13] improved the performance of the Itakura distortion measure by using a frequency-weighted Itakura spectral distortion (FWID) measure which weights the frequency components in the high SNR regions more than those in the low SNR regions.

The solution proposed in this paper is to incorporate the concept of the noise masking of the human auditory system into the original Itakura measure. Considering the fact that the perceptual distortion distance between two speech signals should not be solely based on the matching errors of the two speech spectra, and that

* Corresponding author.

E-mail addresses: egchen@ntu.edu.sg (G. Chen), esnkoh@ntu.edu.sg (S.N. Koh), eiyysoon@ntu.edu.sg (I.Y. Soon).

the perceptual response of the human auditory system to a speech signal is also a critical determining factor, the psychoacoustical properties of the human auditory system must therefore be factored into a speech distortion measure. Noise masking is a well-known psychoacoustical property of the human auditory system [9]. Masking is present because human auditory system is incapable of distinguishing two signals close in the time or frequency domain. The noise masking effects of the human auditory system have a direct influence on human perception of speech signals and they subsequently affect the correlation between the objective distortion measure and the subjective assessment of speech quality. It is therefore logical to postulate that the performance of a speech distortion measure could be improved if noise masking effects are considered in the process of calculating the distortion value. Noise masking effects are also used in audio and speech coding systems [3,5,11]. It has been shown that coding gain could be obtained with no loss of speech quality without transmitting spectral samples below the noise masking threshold. In a similar way, noise spectral components below the noise masking threshold are also excluded from the calculation of the proposed E-Itakura distortion measure.

The rest of the paper is organized as follows. In the next section, we outline the original Itakura speech distortion measure. In Section 3, we give the definition of the proposed E-Itakura measure and its computational procedure. Following that, we present our experimental results and discussion. Finally, in Section 5, we draw our conclusions.

2. The Itakura measure

Let $\sigma_x^2/|A_x(\omega)|^2$ and $\sigma_y^2/|A_y(\omega)|^2$ represent the power spectra of the reference autoregressive (AR) model $\sigma_x/A_x(z)$ and the test AR model $\sigma_y/A_y(z)$, respectively. The Itakura–Saito distortion measure is given by the following equation [2]:

$$d_{IS} \left(\frac{\sigma_x^2}{|A_x(\omega)|^2}, \frac{\sigma_y^2}{|A_y(\omega)|^2} \right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sigma_x^2/A_x(\omega)|^2}{\sigma_y^2/A_y(\omega)|^2} d\omega - \ln \left| \frac{\sigma_x^2}{\sigma_y^2} \right| - 1, \quad (1)$$

where

$$A_x(\omega) = A_x(z)|_{z=e^{j\omega}} = 1 + \sum_{i=1}^P a_x(i)e^{-ji\omega},$$

$$A_y(\omega) = A_y(z)|_{z=e^{j\omega}} = 1 + \sum_{i=1}^P a_y(i)e^{-ji\omega}.$$

σ_x^2 and σ_y^2 , $a_x(i)$ and $a_y(i)$ are the gains and i th LPC prediction coefficients of two p -order LPC models, respectively.

Since the d_{IS} measure is sensitive to the LPC gain terms, it is not very appropriate for calculating the distortion distance between two given speech LPC model spectra [7]. In order to minimize the gain sensitivity of the d_{IS} measure, the Itakura distortion measure, denoted by d_I , was derived by Itakura [1] as follows:

$$d_I = \log \int_{-\pi}^{\pi} \frac{|A_y(\omega)|^2}{|A_x(\omega)|^2} \frac{d\omega}{2\pi}. \quad (2)$$

3. The E-Itakura measure

The basic idea of the proposed E-Itakura measure involves the exclusion of noise spectral components below the noise masking threshold in the calculation of the distortion measure. First, the power spectra of the original speech and distorted speech are calculated. Second, based on the original speech spectrum, the noise masking threshold is calculated and then the masking matrix is obtained by comparing the power spectrum of the distorted speech with the noise masking threshold. Finally, the E-Itakura measure is calculated by using the formulation of the Itakura measure and the masking matrix.

The E-Itakura measure computes the distortion distance between the original speech and distorted speech frame by frame, with a 10th-order LPC analysis and the frame length set to 32 ms. The frame length is selected after examining the effect of frame length on the distortion measure, as discussed in the following sections. Each frame is weighted by a Hamming window, and consecutive frames are overlapped by 50%. A block diagram of the E-Itakura measure is shown in Fig. 1. The algorithm of the E-Itakura measure consists of the following main steps.

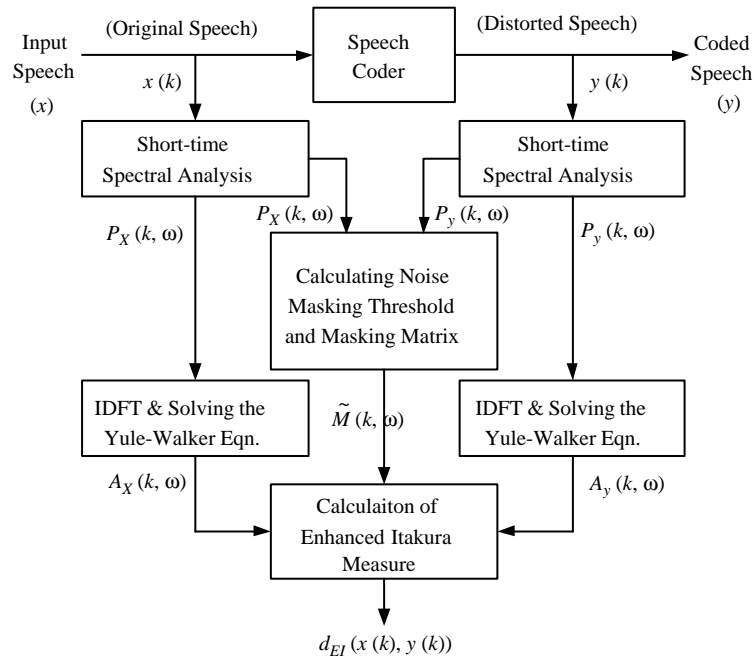


Fig. 1. Calculation of the E-Itakura measure.

3.1. Short-time spectral analysis

Assume two non-silence frames of speech samples, $x(k)$ and $y(k)$ ($k = 1, 2, \dots, K$), of the original speech x and distorted speech y , respectively. They are weighted by the Hamming window and discrete Fourier transformed (DFT). The real and imaginary components of the short-term speech spectrum are squared and added to obtain the short-term power spectrum, i.e.,

$$P(k, \omega) = \text{Re}^2(k, \omega) + \text{Im}^2(k, \omega), \quad (3)$$

where k is the ordinal number of non-silence speech frame and ω is the angular frequency in rad/s.

3.2. Calculation of the noise masking threshold

The noise masking threshold is obtained through modeling the frequency selectivity of the human ear and its masking properties. The computation steps are described in [3]. It is calculated based on the original speech and composed of the following steps, as illustrated in Fig. 2.

3.2.1. Critical band analysis

The short-term power spectrum $P(k, \omega)$ is partitioned into critical bands and the energies of each critical band are added up. Frequencies within the same critical band are equally perceived by the human ear. Since the bandwidth of the speech signal in our study is approximately 3.4 kHz, 18 critical bands are used for the E-Itakura measure calculation which covers 0–4 kHz. The energy in each critical band is summed as follows:

$$B(k, z) = \sum_{v=bl_z}^{bh_z} P(k, v), \quad z = 1, 2, 3, \dots, 18, \quad (4)$$

where bl_z and bh_z are the lower and upper boundaries of critical band z , respectively, $P(k, v)$ is the power spectrum, and $B(k, z)$ is the energy in critical band z . Table 1 shows the mapping from FFT bins to critical bands.

3.2.2. Masking across critical bands

A spreading function represented by a matrix $S(z_i, z_j)$ is used to estimate the effects of masking

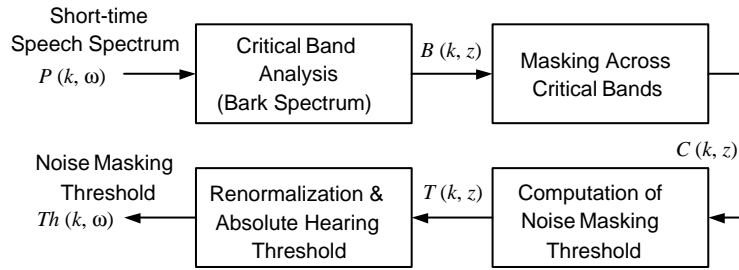


Fig. 2. Calculation of the masking threshold.

Table 1
Mapping from FFT bins to critical bands at a sampling frequency of 8 kHz and frame size $N = 256$ [3]

| Critical band number z | FFT bins | | Real frequencies (Hz) |
|--------------------------|-----------|--------|-----------------------|
| | Intervals | Number | |
| 1 | 1–3 | 3 | 0–94 |
| 2 | 4–6 | 3 | 94–187 |
| 3 | 7–10 | 4 | 187–312 |
| 4 | 11–13 | 3 | 312–406 |
| 5 | 14–16 | 3 | 406–500 |
| 6 | 17–20 | 4 | 500–625 |
| 7 | 21–25 | 5 | 625–781 |
| 8 | 26–29 | 4 | 781–906 |
| 9 | 30–35 | 6 | 906–1094 |
| 10 | 36–41 | 6 | 1094–1281 |
| 11 | 42–47 | 6 | 1281–1469 |
| 12 | 48–55 | 8 | 1469–1719 |
| 13 | 56–64 | 9 | 1719–2000 |
| 14 | 65–74 | 10 | 2000–2312 |
| 15 | 75–86 | 12 | 2312–2687 |
| 16 | 87–100 | 14 | 2687–3125 |
| 17 | 101–118 | 18 | 3125–3687 |
| 18 | 119–128 | 10 | 3687–4000 |

across the critical bands. The function used in this work has been proposed by Schroeder et al. in [10].

$$\begin{aligned}
 S(z_i, z_j) &= 15.81 + 7.5(z_i - z_j + 0.474) - 17.5 \\
 &\quad \times \sqrt{1 + (z_i - z_j + 0.474)^2}, \quad |z_j - z_i| < 25,
 \end{aligned}
 \tag{5}$$

where z_i is the Bark frequency of the masked signal, and z_j is the Bark frequency of the masking signal.

The critical band spectrum, $B(k, z_i)$, is then multiplied with $S(z_i, z_j)$ as follows:

$$\begin{aligned}
 C(k, z_i) &= \sum_{z_j=1}^{18} S(z_i, z_j)B(k, z_j), \\
 z_i &= 1, 2, 3, \dots, 18.
 \end{aligned}
 \tag{6}$$

The value of $C(k, z_i)$ denotes the spread critical band spectrum of the z_i th critical band of the k th speech frame.

3.2.3. Computation of the noise masking threshold

The noise masking threshold depends on the frequency and tonality (i.e., tone or noise like nature) of the signal. The computation of the threshold need consider tone masking noise and noise masking tone. The results of [3] suggest that, for tone like maskers, the masking threshold is estimated as $(14.5 + z_i)$ dB below $C(k, z_i)$ in dB, where z_i is the critical band index. On the other hand, for noise like maskers, the masking threshold is estimated as 5.5 dB below $C(k, z_i)$ uniformly across the spread critical spectrum. In order to determine the tonelike or noiselike nature of the signal, the spectral flatness measure (SFM) is used in [3]. In our study, the noise masking threshold $T(k, z_i)$ is computed by using the simplified method proposed by Sinha and Tewfik in [12], which avoids an accurate estimate of the signal tonality and therefore reduces the computation load. The relative offset $O(z_i)$ in decibels for the making energy in each critical band is given in [12] by a simple estimation, based on the fact that the speech signal has a tonelike nature in lower critical bands and a noiselike nature in higher bands. The resulting values for $O(z_i)$ are represented in

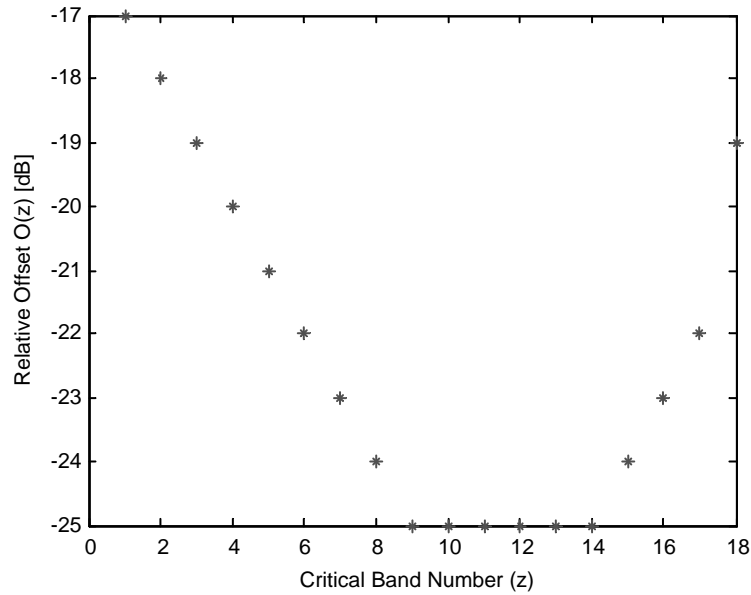


Fig. 3. The relative offset $O(z_i)$ used for noise masking threshold calculation.

Fig. 3. The noise masking threshold of the z_i th critical band of the k th speech frame, $T(k, z_i)$, is obtained by subtracting the offset $O(z_i)$ from the spread critical spectrum $C(k, z_i)$.

3.2.4. Renormalization and comparison with the absolute threshold of hearing

Since the energy estimates in each critical band are increased due to the effects of Eq. (6), the renormalization need be applied as described in [3], i.e., each $T(k, z_i)$ is multiplied with the inverse of the energy gain in each critical band. The renormalized $T(k, z_i)$ is then compared with the absolute threshold of hearing [3]. If any critical band has a calculated noise masking threshold lower than the absolute threshold, it is changed to the absolute threshold for that critical band. The final noise masking threshold $Th(k, z)$ in the Bark domain is obtained and it is further converted into the noise masking threshold $Th(k, \omega)$ in the frequency domain according to the mapping relation as shown in Table 1. An example of noise masking threshold for a given speech frame is represented in Fig. 4. The noise masking threshold is computed based on the clean speech signal and the noise spectral components of the distorted speech signal below the noise masking

threshold are excluded from the calculation of the distortion measure.

3.3. Calculation of E-Itakura measure

Based on the final noise masking threshold $Th(k, \omega)$, the masking matrix for the distorted speech frame $y(k)$, denoted $\tilde{M}(k, \omega)$, is constructed as follows:

$$\begin{aligned} \text{IF } P_y(k, \omega) < Th(k, \omega), \quad \tilde{M}(k, \omega) &= 0, \\ \text{IF } P_y(k, \omega) \geq Th(k, \omega), \quad \tilde{M}(k, \omega) &= 1. \end{aligned} \quad (7)$$

The inverse DFT (IDFT) is applied to $P_x(k, \omega)$ and $P_y(k, \omega)$ to yield the autocorrelation matrix $R_x(k)$ and $R_y(k)$. The LPC coefficients $a_x(k, i)$ and $a_y(k, i)$ ($i = 1, 2, \dots, p$) are then obtained by solving the Yule–Walker equation. The E-Itakura measure of the frame of original speech samples $x(k)$ and the distorted speech sample $y(k)$ can then be obtained as follows:

$$\begin{aligned} d_{\text{EI}}(x(k), y(k)) \\ = \log \int_{-\pi}^{\pi} \tilde{M}(k, \omega) \frac{|A_y(k, \omega)|^2}{|A_x(k, \omega)|^2} \frac{d\omega}{2\pi}. \end{aligned} \quad (8)$$

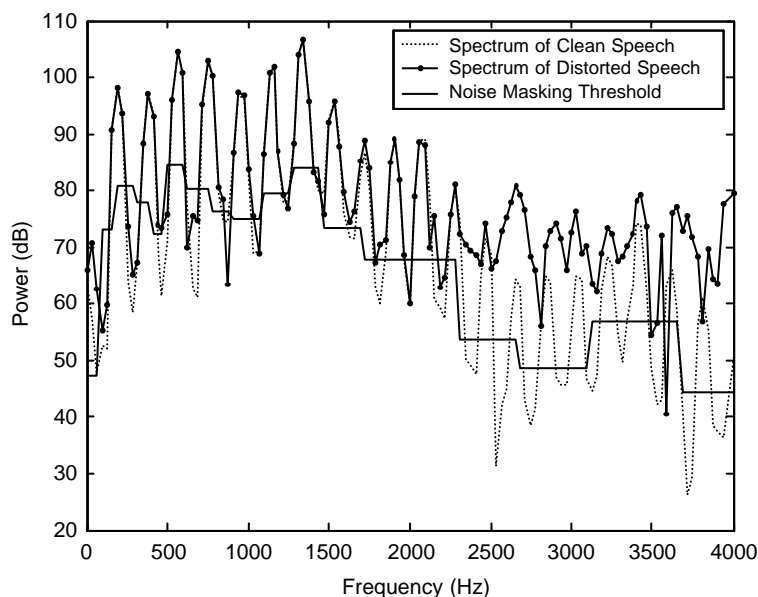


Fig. 4. An example of noise masking threshold for a 32 ms frame of a distorted speech (MNRU condition: SNR = 15 dB).

4. Experimental results and discussions

First, the performance of the E-Itakura measure with various speech frame sizes has been investigated in order to find the optimal frame length for the highest correlation coefficient. Second, in order to examine the performance of the E-Itakura measure, the comparisons between the performance of the E-Itakura measure with those of the original Itakura measure, the Bark spectral distortion measure (BSD) [14], the cepstral distance (CD) [4] and the FWISD [13] measure have also been carried out. In our work, we used a speech data set which included five Modulated Noise Reference Unit (MNRU) conditions (5–25 dB at 5 dB step) and various different types of speech coders such as ADPCM, GSM, IS54, FS1016, LD-CELP and CELP. A total of 136 MOS of the speech signals were obtained by Dynastat, Inc. of USA using the Absolute Category Rating method in a formal subjective test environment. Four panels of eight listeners were recruited to evaluate the materials. The speech materials were presented to listeners monaurally over earphones mounted in a circumaural cushion. Listeners were seated in a sound isolation chamber containing the listening stations. The audio

output of the DAT player was input to an audio distribution system for distribution to the listening stations. Each station in the booth was equipped with a personal computer system for rating scale presentation and data collection.

4.1. Evaluation method

Since the speech distortion measures are computed between two speech utterances, we evaluate the performance of different distortion measures by using the correlation coefficient, r , and standard error of estimate, s , between the distortion measures and MOS difference [8]. In our experiments, the 64-kbps PCM coded speech is regarded as the original speech. The evaluation method is described below.

Let x denote the original speech signal, and y^n ($n = 1, 2, 3, \dots, N$) denote the coded speech signals, where N is the total number of the different coded speech signals. Let M_x denote the MOS of x and M_y^n ($n = 1, 2, \dots, N$) denote the MOS of y^n . Assume two speech signals x and y^n are weighted by the Hamming window and separated into K frames, let the k th speech frames of x and y^n be denoted by $x(k)$ and $y^n(k)$ ($k = 1, 2, \dots, K$). The distortion distance between the

original and coded speech signals is then given by

$$\Gamma_i^n(x, y) = \frac{1}{K} \sum_{k=1}^K d_i(x(k), y^n(k)),$$

$$n = 1, 2, \dots, N, \quad (9)$$

where d_i indicates the i th distortion measure being studied.

Also, let MOSD^n denote the MOS difference of two speech signals x and y^n , i.e.,

$$\text{MOSD}^n = M_x - M_y^n, \quad n = 1, 2, \dots, N. \quad (10)$$

In order to assess the ability of the proposed new measure and that of others to predict MOSD ratings, we fitted second-order polynomial predictors to the various scatter plots by the following least-square linear regression equation:

$$\text{EMOSD}^n = \alpha + \beta \Gamma_i^n + \gamma (\Gamma_i^n)^2. \quad (11)$$

Subsequently, the correlation coefficient (r) and standard error of estimate (s) of EMOSD and MOSD are calculated by the following equations:

$$r = \frac{\sum_{n=1}^N (\text{EMOSD}^n - M_{\text{EMOSD}})(\text{MOSD}^n - M_{\text{MOSD}})}{\sqrt{\sum_{n=1}^N (\text{EMOSD}^n - M_{\text{EMOSD}})^2 \sum_{n=1}^N (\text{MOSD}^n - M_{\text{MOSD}})^2}} \quad (12)$$

$$s = \text{Var}_{\text{MOSD}} \sqrt{(1 - r^2)}, \quad (13)$$

where M_{MOSD} and Var_{MOSD} are the mean and variance of MOSD^n , respectively. r and s indicate the degree of consistency between the i th distortion measure and subjective assessment of speech quality in terms of MOS [8].

4.2. Selecting the optimal frame length

We have examined the effects of different frame lengths on the E-Itakura measure. With window lengths ranging from 8 to 40 ms, the highest correlation coefficient for mixed (male and female) speakers is obtained when the frame length is 32 ms (256 samples), as shown in Fig. 5. Each frame was weighted by a Hamming window, and consecutive frames were overlapped by 50%.

4.3. Comparison experiments

We compare the performance of the E-Itakura measure with those of the original Itakura, the FWISD, the CD and the BSD measure. In the FWISD measure, four bandwidth-broadening factors are set to be 0, 0.25, 0.7, and 1.0 according to the four SNR brackets described in [13]. The results are summarized in Figs. 6–10. The curves shown in the figures are second-order polynomial predictors fitted by least-square linear regression to the scatter plots. The numbers shown in the figures have the following meanings. r is the correlation coefficient between the actual and predicted MOSD values. Perfect prediction would yield the value $r = 1$. The variable s is the standard deviation of the prediction error, and in ideal conditions, it would be zero.

4.4. Discussion

It can be concluded from our experimental results that, the proposed E-Itakura measure, which incorporates noise masking effects of the human auditory system into the original Itakura measure, outperforms many other measures in predicting MOSD. The correlation of the original Itakura measure with subjective evaluation ratings of speech quality has been improved from 0.73 to 0.89 by using the enhancement procedure. Also, the E-Itakura distortion measure is more consistent with the subjective assessment of speech quality than the FWISD, CD and BSD measures, as shown in Table 2.

From Figs. 6–10, it can be seen that the plots of the MOSD values in all figures cluster along the estimated MOSD values in general except for a slightly more deviation of the Itakura measure when the values of MOSD is less than 2. Namely, all distortion measures perform reasonably well for high MOS values. On the other hand, when the values of MOSD is greater than 2, the estimated values of the Itakura, FWISD, CD and BSD distortion measures distribute more dispersedly than those of the E-Itakura measure. In other words, for low MOS values, the E-Itakura measure has a higher correlation with subjective assessment of speech quality. This characteristic of the E-Itakura measure is due to its salient feature of having to consider noise masking effect of the human auditory system. Since inaudible noise components masked by

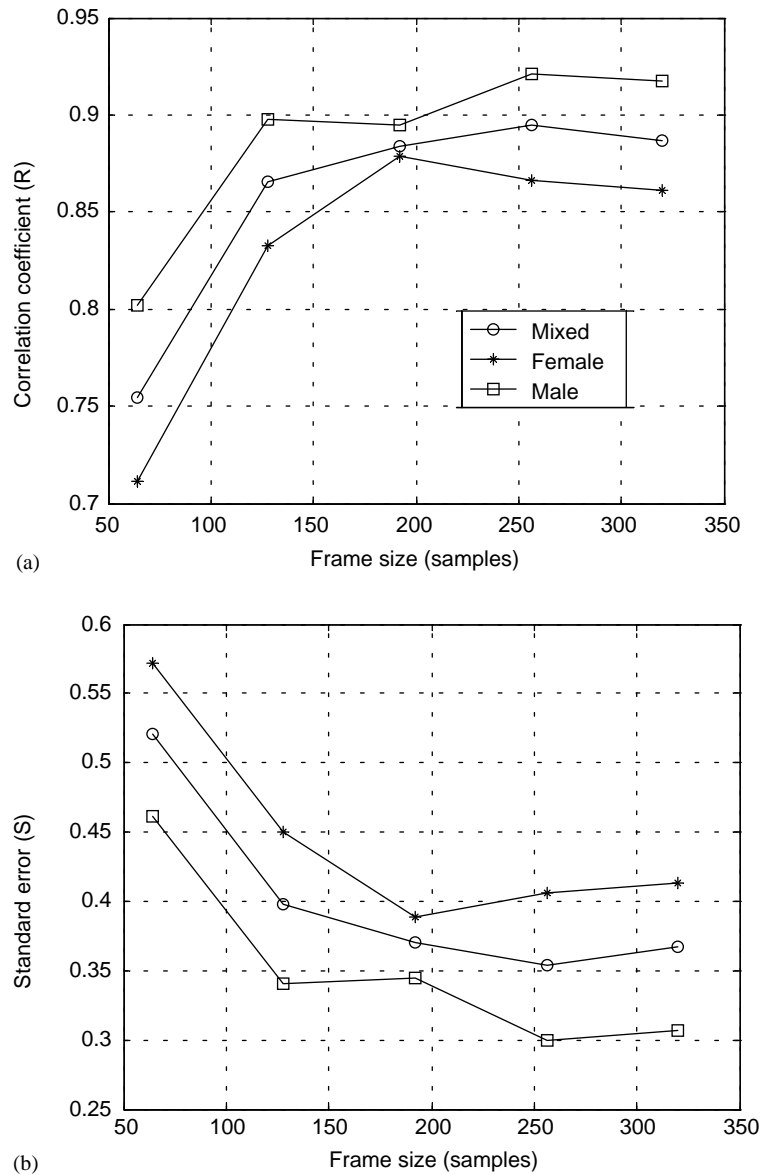


Fig. 5. (a) Analysis window size of E-Itakura versus correlation coefficient and (b) analysis window size of E-Itakura versus stand error.

speech signals are excluded from the distortion calculation, the E-Itakura measure has achieved a preferable performance.

Another advantage of the E-Itakura measure is that it calculates the distortion between two short-time speech spectra by using the formulation of the Itakura measure which is a very good measure of matching

errors between two speech spectra. This is unlike the BSD measure which estimates the overall distortion by using the average Euclidean distance method.

The computational cost of the E-Itakura distortion measure is not much more than that of the original Itakura measure. Computationally, the most expensive operation is the FFT spectral analysis

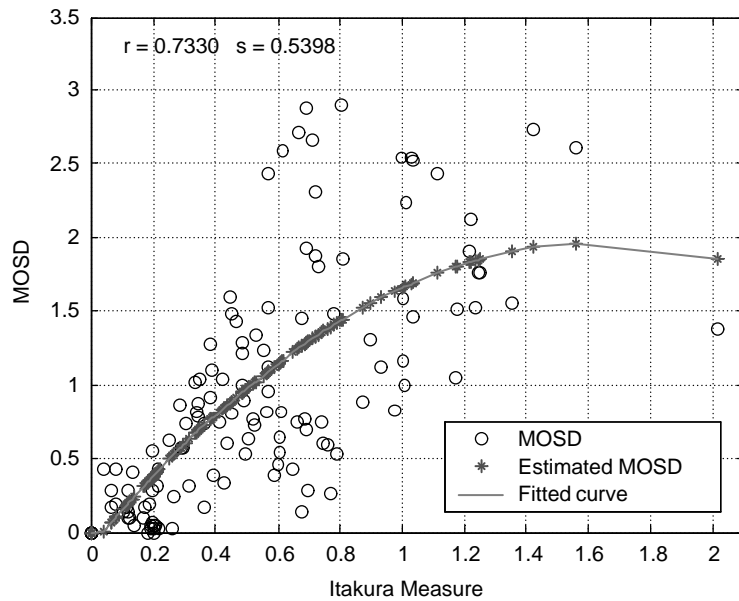


Fig. 6. Relationship between MOSD values and estimated MOSD via the Itakura measure. Mixed speakers.

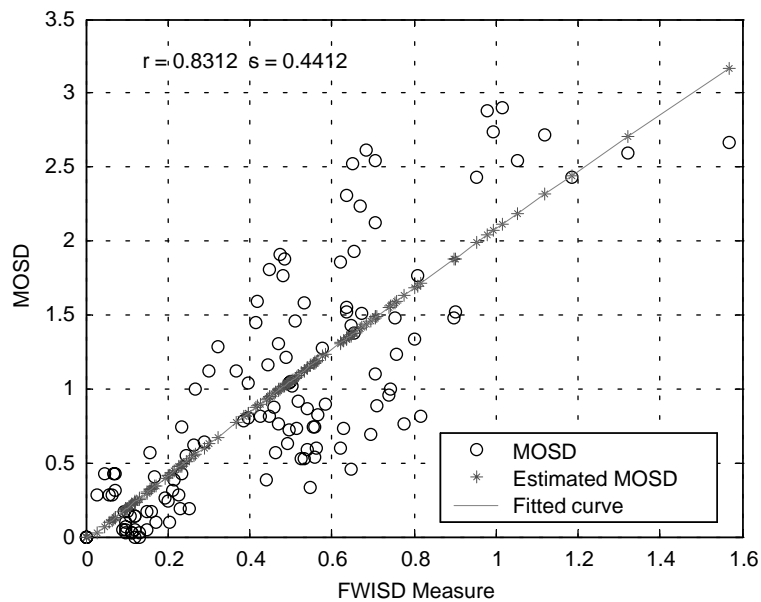


Fig. 7. Relationship between MOSD values and estimated MOSD via the FWISD measure. Mixed speakers.

calculation, followed by the calculation of the noise masking threshold. For a 32 ms speech frame at a 8 kHz sampling frequency, the additional computational load approximately comes from the following

parts: (1) the FFT analysis with $2N \log_2 N$ real multiplications and $2N \log_2 N$ real additions, where N is 256 in this paper, (2) critical band analysis with 128 real additions, (3) masking across critical bands with

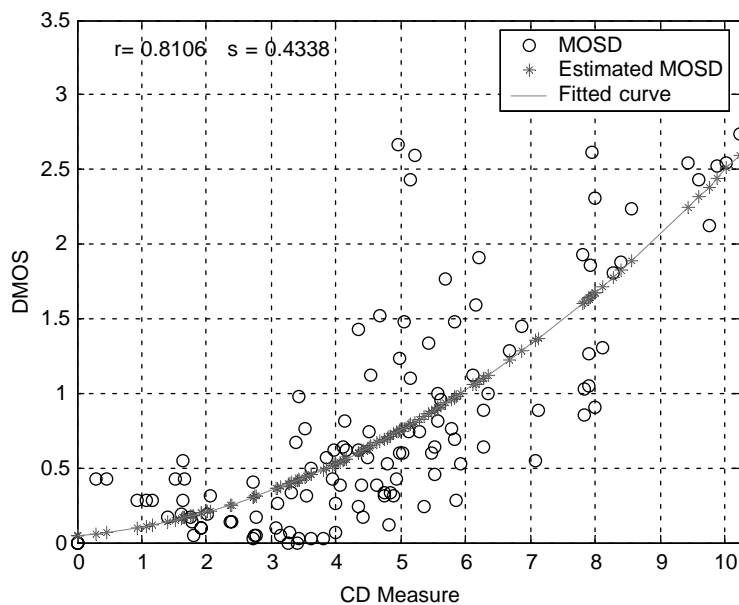


Fig. 8. Relationship between MOSD values and estimated MOSD via the CD measure. Mixed speakers.

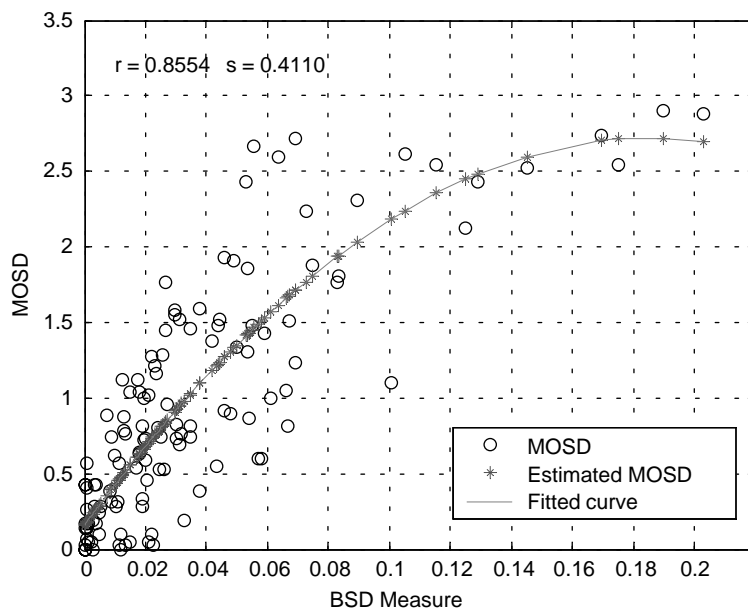


Fig. 9. Relationship between MOSD values and estimated MOSD via the BSD measure. Mixed speakers.

Q^2 real multiplications and $Q(Q - 1)$ real additions, where the total number of critical bands, Q , is 18 in our study, (4) computation of noise masking threshold with Q real subtractions, and (5) renormalization with

Q real divisions and Q real multiplications. It is noted that the E-Itakura measure, in the process of calculating noise masking threshold, warps the spectrum of speech signal along the frequency axis into the Bark

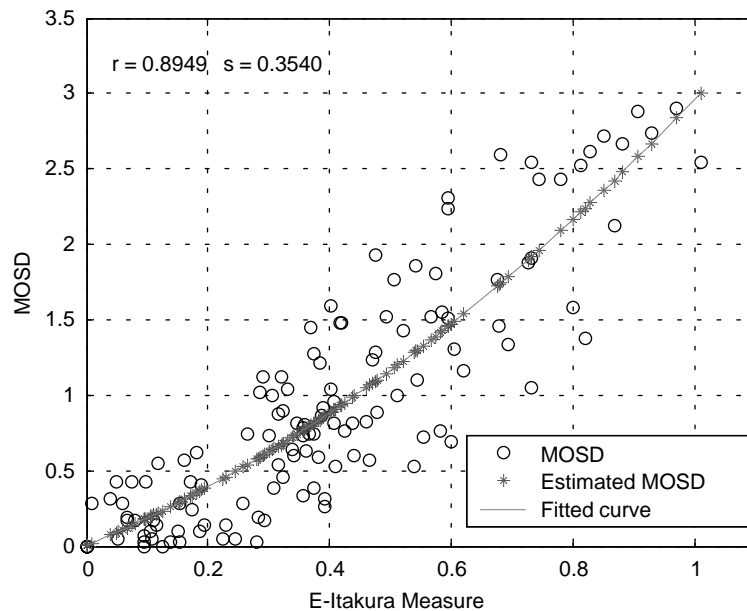


Fig. 10. Relationship between MOSD values and estimated MOSD via the E-Itakura measure. Mixed speakers.

Table 2
Performance of different distortion measures

| Measure | Mixed | | Female | | Male | |
|-----------|--------|--------|--------|--------|--------|--------|
| | r | s | r | s | r | s |
| Itakura | 0.7330 | 0.5398 | 0.7671 | 0.5216 | 0.7122 | 0.5420 |
| CD | 0.8106 | 0.4338 | 0.7853 | 0.5028 | 0.8412 | 0.3494 |
| FWID | 0.8312 | 0.4412 | 0.8425 | 0.4380 | 0.8302 | 0.4305 |
| BSD | 0.8554 | 0.4110 | 0.8552 | 0.4215 | 0.8873 | 0.3561 |
| E-Itakura | 0.8949 | 0.3540 | 0.8664 | 0.4060 | 0.9214 | 0.3010 |

scale axis, and therefore the number of the samples in the Bark domain is reduced to 18 Bark spectral samples. The rest of the computational cost is the same as the original Itakura measure. The overall computational cost of the E-Itakura measure is therefore only slightly higher than that of the original measure.

5. Conclusions

In this paper, a new E-Itakura measure is proposed. It is based on the auditory masking phenomenon which is modeled by the calculation of the noise masking

threshold. In the process of calculating the E-Itakura measure, the noise components below the noise masking threshold are excluded from the distortion distance because they are inaudible, while the intrinsic advantage of the original Itakura measure is retained. The proposed algorithm has been tested and compared with the original Itakura, FWISD, CD and BSD measures. The results show that the correlation of the original Itakura measure with subjective evaluation ratings of speech quality has been improved from 0.73 to 0.89. Also, the E-Itakura measure offers a more consistent indication of the subjective speech quality than other measures.

References

- [1] F. Itakura, Minimum prediction residual principle applied to speech recognition, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-23 (1) (February 1975) 67–72.
- [2] F. Itakura, S. Saito, An analysis-synthesis telephony based on the maximum-likelihood method, *Proceedings of the Sixth International Congress on Acoustics, Japan, 1968*, pp. C17–C20.
- [3] J.D. Johnston, Transform coding of audio signals using perceptual noise criteria, *IEEE J. Sel. Areas Commun.* 6 (2) (February 1988) 314–323.
- [4] N. Kitawaki, H. Nagabuchi, K. Itoh, Objective quality evaluation for low-bit-rate speech coding systems, *IEEE J. Sel. Areas Commun.* 6 (3) (February 1988) 242–248.
- [5] S.N. Koh, G.H. Chua, Application of auditory masking in improved multiband excitation model, *Appl. Acoust.* 63 (2002) 693–698.
- [6] J. Li, A.K. Krishnamurthy, A modified frequency-weighted Itakura spectral distortion measure, *IEEE Trans. Acoust. Speech Signal Process.* 37 (10) (October 1989) 1614–1617.
- [7] N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt, Comparative study of several distortion measures for speech recognition, *Speech Commun.* 4 (December 1985) 317–331.
- [8] S.R. Quackenbush, T.P. Barnwell III, M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [9] M.R. Schroeder, Models of hearing, *Proc. IEEE* 63 (9) (September 1975) 1332–1350.
- [10] M.R. Schroeder, B.S. Atal, J.L. Hall, Optimizing digital speech coders by exploiting masking properties of the human ear, *J. Acoust. Soc. Am.* 66 (16) (December 1979) 1647–1651.
- [11] D. Sen, D.H. Irving, W.H. Holmes, Use of an auditory model to improve speech coders, *IEEE ICASSP 2* (1993) 411–414.
- [12] D. Sinha, A.H. Tewfik, Low bit rate transparent audio compression using adapted wavelets, *IEEE Trans. Signal Process.* 41 (December 1993) 3463–3479.
- [13] F.K. Soong, M.M. Sondhi, A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise, *IEEE Trans. Acoust. Speech Signal Process.* 36 (January 1988) 41–48.
- [14] S. Wang, A. Sekey, A. Gersho, An objective measure for predicting subjective quality of speech coders, *IEEE J. Sel. Areas Commun.* 10 (5) (June 1992) 819–829.