PERCEPTUAL EVALUATION OF AUDIO WATERMARKING USING OBJECTIVE QUALITY MEASURES

Yiqing Lin[1] and *Waleed H. Abdulla*[2]

Department of Electrical and Computer Engineering,
The University of Auckland, New Zealand
[1]y.lin@ece.auckland.ac.nz [2]w.abdulla@auckland.ac.nz

ABSTRACT

This paper proposes objective quality measures adopted in speech processing for perceptual quality evaluation of audio watermarking. Different from using an auditory perception model that mimics human auditory system, objective quality measures are introduced as alternative approach to perceive the dissimilarities caused by audio watermarking. After embedding the watermark into a variety of audio signals from EBU database, we calculate the distance between the watermarked and host signals in terms of several well-developed quality measures. For correlation analysis, subjective listening tests and a commercial evaluation tool PEMO-Q are also used to grade the differences. Pearson correlation coefficients reveal that the investigated quality measures, especially Weighted Spectral Slope (WSS) measure, correspond well with reference ratings. Moreover, quality measures run much faster than PEMO-Q. The results indicate that objective quality measures can be the perceptual quality predictors for audio watermarking.

Index Terms — Audio watermarking, auditory quality perception, objective quality measures, correlation analysis

1. INTRODUCTION

Along with the advancement of audio watermarking techniques, the necessity for evaluating various algorithms effectively and comprehensively becomes imperative [1, 2]. Imperceptibility, robustness and security are the key criteria in designing any audio watermarking scheme. In particular, imperceptibility is a prerequisite to putting watermarked audio tracks into reality [3]. Hence, the perceptual quality assessment on audio watermarking system is worthy of more attention.

Similar to evaluating the quality of perceptual codecs in the audio, image and video fields [3], perceptual quality assessment on the watermarked audio files is usually classified into two categories: subjective listening tests and objective evaluation tests. Since perceptual quality is defined by human opinion [4], subjective listening tests on audience from different background are pressingly required. However, such audibility tests are not only costly and time-consuming, but also heavily depend on the subjects and the surrounding conditions [5]. Therefore, the industry claims more and more attention on the implementation of objective measurements, such as Perceptual Evaluation of Audio Quality (PEAQ) [6], Evaluation of Audio Quality (EAQUAL) [7] and Perceptual Model-Quality Assessment (PEMO-Q) [8]. Basically, these methods establish an auditory perception model to imitate the listening behavior of human being, so that the test signal is graded relatively to the reference signal. In the context of audio watermarking, the reference and test signal are the host (or cover) and the watermarked signal, respectively. However, a large set of relevant test signals are required to train and characterize such models [5], otherwise the accuracy of evaluation cannot be guaranteed.

Objective quality measures, such as Segmental Signal-to-Noise Ratio (SSNR) measure, Log-Likelihood Ratio (LLR) measure, Itakura-Saito (IS) distortion measure, Log-Area Ratio (LAR) measure and Weighted Spectral Slope (WSS) measure [9], have been widely used in quality evaluation for speech enhancement [10], speech intelligibility estimation [4, 11], speech recognition in blind source separation [12]. Motivated by such findings, we are interested in exploiting the application of those quality measures to objective assessment on perceptual quality of audio watermarking.

The paper is structured as follows. Firstly, the perceptual quality assessments in audio watermarking including subjective listening tests and objective evaluation tests are reviewed in section 2. Section 3 is focused on five objective quality measures under test. In section 4, we set up several experiments to explore the relations between the objective quality measures and the subjective listening grades. Finally, section 5 presents the conclusions and future work.

2. PERCEPTUAL QUALITY ASSESSMENT IN AUDIO WATERMARKING

Audio watermarking intends to embed an unperceivable, robust and secure watermark into host signals. As viewed from communications theory, the watermark is inserted into a cover signal like a kind of noise. Obviously, it is essential that the watermarking should be perceptually transparent, which implies that perceptual quality of the watermarked audio is evaluated related to the host audio.

Generally, there are two approaches to perform the perceptual quality assessment: (1) Subjective listening tests by human's acoustic perception. (2) Objective evaluation tests by perception model or quality measures.

2.1. Subjective listening tests

In the subjective listening tests, the subjects are asked to discern the watermarked and host audio clips. Two popular modes are ABX [13] and MUSHRA [14], described in ITU-R BS.1116-1 and BS.1534 respectively. ABX listing test is interpreted as "Double

bind, triple stimulus, with hidden reference", basically for the assessment of small deterioration. MUSHRA stands for "MUlti Stimulus test with Hidden Reference and Anchors", more reliable than ABX test in the presence of larger distortions.

Moreover, the watermarked signal is graded with respect to the host signal according to a five-grade impairment scale (see Figure 1) defined in ITU-R BS.562 [3]. It is called Subjective Difference Grade (SDG), which equals to the subtraction between the subjective ratings given separately to the watermarked signal and the host signal.

Difference Grade	Description of Impairments
0	Imperceptible
-1	Perceptible but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

Figure 1. Subjective difference grade (SDG)

Subjective listening tests are indispensable to perceptual quality assessment, since the ultimate judgment is made by human perception. However, conducting such listening tests is quite complicated and also not adequate for manufacturing. Therefore, machine-based objective evaluations are aspired to provide a convenient, consistent and fair assessment.

2.2. Objective evaluation tests

Objective evaluation tests are intended to facilitate the implementation of subjective listening tests. To achieve its goal, results of objective evaluation should correlate well with SDGs.

Currently, the commonly used objective evaluation is to assess the perceptual quality of audio data via a stimulant ear, such as PEAQ, EAQUAL and PEMO-Q. The whole process is depicted in Figure 2 [3, 5]. After the watermark is embedded, the host and watermarked signal are separately passed to perceptual model and get their internal representations. Through comparison, the audible differences are calculated and scaled by cognitive model. The final output is called Objective Difference Grade (ODG), whose specifications conform to the SDG discussed above.

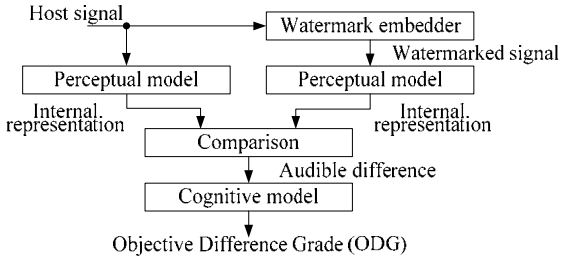

Figure 2. Objective evaluation via perception modeling

Among the implemented models, PEMO-Q is the latest and most advanced predictor of audio quality. It is reported in [8] that PEMO-Q has a higher ability to be applicable to unknown distortions and performs better than the other techniques. In our experiments, a detailed test will be taken to examine all the assessment tools to ascertain their performances.

Besides perception modeling, the extent of dissimilarity between the watermarked and host signal can be quantitated by objective quality measures. A simple example is signal-to-noise ratio (SNR), which reflects the quantity of distortion that a watermark imposes on the host signal [13]. However, SNR actually averages the distortions on the entire signal, so it is not a reliable indicator of perceptual quality [10, 11]. To accurately estimate the dissimilarity, we propose more effectual measures as investigated in the next section.

3. OBJECTIVE QUALITY MEASURES

Objective quality measures have been widely used in the quality evaluation of speech signals [9]. This kind of measurement makes use of sound source information and calculates the "distance" or "distortion" of the test signal with respect to original signal [12], which corresponds to the concept of perceptual assessment in audio watermarking.

Based on the results in the existing literature, five quality measures are selected to evaluate the distance between the host $s_h(n)$ and watermarked $s_w(n)$ signals[1]. Considering that the impact of noise on the signal quality is non-uniform, all the measures are frame-based [10]. So the measure is computed over short segments of the tested signal [9], then its mean is calculated after removing outliers greater than 3σ [10, 12].

- Segmental Signal-to-Noise Ratio (SSNR) measure

SSNR is a variation of SNR, which is formed by averaging frame level SNR as follows [9, 10].

$$d_{SSNR}(\vec{s}_h,\vec{s}_w)=\frac{1}{M}\sum_{m=0}^{M-1}\left\{10\cdot\log\frac{\sum_{n=Nm}^{Nm+N-1}\vec{s}_h^{\,2}(n)}{\sum_{n=Nm}^{Nm+N-1}[\vec{s}_w(n)-\vec{s}_h(n)]^2}\right\} \quad (1)$$

where N is the frame length and M is the number of frames.

Note that due to the definition of SSNR, abnormal SSNR will occur in some situations [9]. To avoid such useless values, the lower and upper limits of SSNR (i.e. –10dB and 35dB) need to be specified in advance [10].

- Log-Likelihood Ratio (LLR) measure

LLR or Itakura distance measure is based on linear prediction (LP) analysis. Given both LP coefficient vectors $\vec{a}_h$ and $\vec{a}_w$, LLR measure is calculated by [9, 10, 12]

$$d_{LLR}(\vec{a}_h,\vec{a}_w)=\log\left(\frac{\vec{a}_w R_h \vec{a}_w^T}{\vec{a}_h R_h \vec{a}_h^T}\right) \quad (2)$$

where R_h is autocorrelation matrix and $(\cdot)^T$ refers to the transpose.

- Itakura-Saito (IS) distortion measure

IS distortion measure is a slightly different form of LLR measure and performs well for signals with additive noise [12], which is defined as [9, 10, 12]

$$d_{IS}(\vec{a}_h,\vec{a}_w)=\left(\frac{\sigma_h^2}{\sigma_w^2}\right)\cdot\left(\frac{\vec{a}_w R_h \vec{a}_w^T}{\vec{a}_h R_h \vec{a}_h^T}\right)+\log\left(\frac{\sigma_w^2}{\sigma_h^2}\right)-1 \quad (3)$$

where σ_h^2 and σ_w^2 represent the all-pole gains.

- Log-Area Ratio (LAR) measure

LAR measure is also involved with LP analysis, but depends on LP reflection coefficients [10, 11, 12].

$$d_{LAR}(\vec{r}_h,\vec{r}_w)=\left|\frac{1}{M}\sum_{m=1}^{M}\left[\log\frac{1+\vec{r}_h(m)}{1-\vec{r}_h(m)}-\log\frac{1+\vec{r}_w(m)}{1-\vec{r}_w(m)}\right]^2\right|^{1/2} \quad (4)$$

[1] Hereunder, the subscript h and w denote the component related to the host and watermarked signals respectively.

Here, $\vec{r}_h$ and $\vec{r}_w$ are LP reflection coefficient vectors, which are defined as $\vec{r}_h = \frac{1+\vec{a}_h}{1-\vec{a}_h}$ and $\vec{r}_w = \frac{1+\vec{a}_w}{1-\vec{a}_w}$.

Since the reflection coefficients are directly related to the power spectra, LAR measure could estimate the differences between the logarithms of the spectra of the host and watermarked signals efficiently [11]. In [10, 11, 12], it has been reported that LAR might be the best measure in some cases.

- Weighted Spectral Slope (WSS) measure

WSS measure is based on an auditory model in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time spectra [10]. Then, it calculates a weighted difference between the spectral slopes in each band [4], and each weight w_a depends on formant locations [10]. The per-frame WSS measure is formulated as [9, 10, 12]

$$d_{WSS} = K_{spl}(K_h - K_w) + \sum_{n=1}^{36}\left\{w_a(n)\cdot[S_h(n) - S_w(n)]^2\right\} \qquad (5)$$

where K_h and K_w are related to the overall sound pressure level, and K_{spl} is a parameter which can be varied to increase overall performance. For WSS measure employs the auditory model, it usually outperforms other measures, as reported in [9, 12].

In the next section, these objective quality measures will be evaluated to gauge their capabilities in predicting the perceptual quality of watermarked audio.

4. EXPERIMENTS AND DISCUSSION

The experiments comprise of four parts: audio watermarking scheme, subjective listening tests, objective evaluation tests and correlation analysis. All the simulations are carried on a computer with 2.0GHz CPU and 384MB RAM.

4.1. Audio watermarking scheme

In these experiments, nineteen pieces of audio signals are involved. All of them are in WAVE format (44.1 kHz, 16 bit, mono) with a length of 4s, and most are excerpts of EBU SQAM disc tracks [15]. For ease of expression hereafter, the host audio signals are marked with a subscript number in ascending order, i.e. (i) Music: Bach_1, Pop_2, Rock_3, Jazz_4; (ii) Percussive instruments: Hihat_5, Castanets_6, Glockenspiel1_7, Glockenspiel2_8; (iii) Tonal instruments: Harpsichord_9, Violoncello_{10}, Horn_{11}, Pipes_{12}, Trumpt_{13}, Electronic tune_{14}; (iv) Vocal: Sopranor_{15}, Bass_{16}, Quartet_{17}; (v) Speech: Female speech_{18}, Male speech_{19}.

Then, a well-developed robust audio watermarking scheme in [16] is used to implement the process of watermarking. During the watermarking, the extent of robustness is controlled by a factor called watermark strength, which is related to the magnitude of the embedded watermark A_w. Here, each host signal is watermarked 20 times with increasing watermark strength. Thus, we use $s_w(i,j)$ (i=1, 2,…, 19; j = 0, 1,…, 20) to represent a certain watermarked signal, where i and j denote the type of signal and the watermark strength respectively[2]. For instance, $s_w(1,10)$ represents a watermarked Bach signal with watermark strength = 10.

4.2. Subjective listening tests

[2] In this case, $A_w = j * A_u$, where A_u is the unitage of watermark magnitude, about 0.3% of the magnitude of host signal.

Subjective listening tests are performed in an isolated chamber, where ten trained listeners are participated. All the stimuli are presented through a high-fidelity headphone.

In the tests, the participants are asked to grade the quality of the watermarked signal compared with its host signal, and then provide an absolute SDG. In view of the difficulties in the real audibility tests, the human subjects are not required to undergo all the 20 watermark strengths, but at an interval of four, i.e. (j' = 0, 4, 8, 12, 16, 20). It means that every listener needs to do 19 tests, each of which consists of 6 watermarked signals with different strengths. Then, by averaging the SDGs over all the listeners, each host signal gets 6 SDG scores based on the watermark strengths, denoted as $SDG(i, j')$. For instance, four SDG sets of Bach_1, Jazz_4, Castanets_6 and Soprano_{15} are plotted in Figure 3, i.e. $SDG(1, j')$, $SDG(4, j')$, $SDG(6, j')$ and $SDG(15, j')$.

4.3. Objective evaluation tests

Before evaluating the performance of objective quality measures, the accuracy of three assessment tools using perception model are investigated, i.e. PEMO-Q [8], EAQUAL [7] and PEAQ [6]. The aim is to find the most effective quasi-subjective predictor of audio quality, which best conforms to SDG. Then, its ODG will be regarded as quasi-SDG for correlation analysis in the next sub-section. In correlation analysis, we prefer quasi-SDG over the SDG sets, because it would be hard to get the best fit due to the insufficient amount of SDG scores in each set.

Using these tools, all the watermarked signals $\{s_w(i,j)\}$ are evaluated completely. Then, for each host signal $s_h(i)$, it has three sets of ODGs against watermark strengths, i.e. $\{ODG_{PEMO\text{-}Q}(i)\}$, $\{ODG_{EAQUAL}(i)\}$ and $\{ODG_{PEAQ}(i)\}$. For illustration, ODG sets of the previous four examples are plotted in Figure 3 as well.

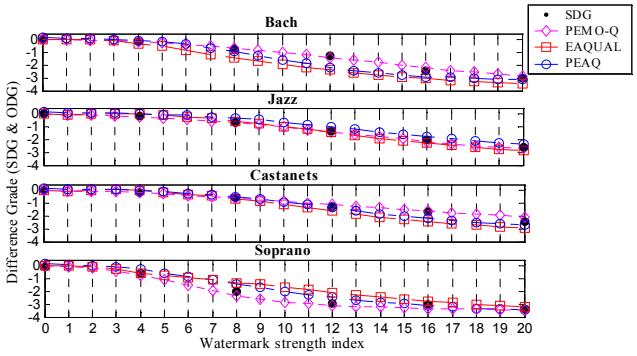

Figure 3. Evaluation of PEMO-Q, EAQUAL and PEAQ

Note that a few ODGs in Figure 3 are slightly positive, like some values with small watermark strengths. According to its definition, ODG should normally be in the range [-4, 0]. However, if the distortion caused by watermarking is very low, then the cognitive model calculates positive values. In such cases, it is interpreted that the distortion is mostly inaudible for human [2].

After comparing all the SDG and ODG sets, it is observed that PEMO-Q indeed provides a better correspondence between SDG and ODG. Thus, $\{ODG_{PEMO\text{-}Q}(i)\}$ or simplified as $\{G(i)\}$ will be adopted for correlation analysis subsequently.

Then, five objective quality measures discussed in section 3 are computed between the host and watermarked signals with different watermark strengths. Hence, each host signal gets five sets of measures in terms of SSNR, LLR, IS, LAR and WSS, which are

denoted by $\{Q_k(i)\}$ (k = 1, 2, 3, 4, 5) respectively[3].

It is worth notice that PEMO_Q takes around 70 seconds to complete each evaluation with the default settings in [8]. For each quality measure, the computation time is mostly less than 6 seconds, which is much faster than PEMO-Q. For example, Table 1 lists the processing time of quality measures on $s_w(4,10)$.

Table 1. Comparison of the total processing time (sec)

PEMO_Q	SSNR	LLR	IS	LAR	WSS
71.0	4.47	5.66	5.58	4.99	6.44

4.4. Correlation analysis

Performances of objective quality measures are evaluated in terms of Pearson correlation coefficient $\rho(i, k)$ (absolute value). Note that $\rho(i, k)$ refers to the correlation coefficient between the k^{th} quality measure and the quasi-SDG of the i^{th} host signal, defined as [4,12]

$$\rho(i,k) = \left| \frac{\sum_j \left\{ \left[Q_k(i,j) - \overline{Q}_k(i) \right] + \left[G(i,j) - \overline{G}(i) \right] \right\}}{\left\{ \sum_j \left[Q_k(i,j) - \overline{Q}_k(i) \right]^2 \sum_j \left[G(i,j) - \overline{G}(i) \right]^2 \right\}^{1/2}} \right| \qquad (6)$$

where $\overline{Q}_k(i)$ and $\overline{G}(i)$ are the means of $Q_k(i)$ and $G(i)$ respectively.

All the audio signals were taken into consideration in the experiments. But because of space limitation, only the results of the previous four examples are presented in Table 2.

Table 2. Pearson correlation coefficient $|\rho(i,k)|$

	SSNR	LLR	IS	LAR	WSS
$Bach_1$	0.5176	0.4249	0.5923	0.5552	0.6671
$Jazz_4$	0.5903	0.2668	0.4859	0.4546	0.7514
$Castanets_6$	0.5740	0.3176	0.5964	0.6042	0.7432
$Soprano_{15}$	0.2574	0.0365	0.3052	0.2767	0.3362

In most cases, WSS measure exhibits a higher correlation with the subjective quality grades, then IS and LAR measures come next, while LLR measure performs worst. Apart from one case, the correlation is high enough to be accepted.

5. CONCLUSIONS

Motivated by their wide application in speech quality evaluation, five objective quality measures are assessed for their capabilities in the new field of audio watermarking. Compared to traditional perception model, quality measures provide a faster and more efficient method of evaluating the watermarked audio with reference to the host audio. Extensive experiments have shown that WSS, IS and LAR measures correlate well with SDG and quasi-SDG from PEMO_Q, although their performances vary with different host audios. The results indicate these measures can be used reliably to estimate the perceptual quality of watermarked audios. The correlation of quality measures with bit error rate (BER) in watermarking detection as well as using different audio watermarking techniques will be studied in future work.

6. REFERENCES

[1] F.A.P. Petitcolas, "Watermarking Schemes Evaluation," *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 58-64, 2000.

[2] A. Lang, J. Dittmann, "Transparency and Complexity Benchmarking of Audio Watermarking Algorithms Issues," *MM &Sec Workshop*, pp. 190-201, 2006.

[3] M. Arnold, "Subjective and Objective Quality Evaluation of Watermarked Audio Tracks," *Intl. Conf. on WEDELMUSIC*, pp. 161-167, 2002.

[4] W.M. Liu, K.A. Jellyman, J.S.D. Mason and N.W.D. Evans, "Assessment of Objective Quality Measures for Speech Intelligibility Estimation," *ICASSP*, vol. 1, pp.1225-1228, 2006.

[5] J.G. Beerends, "Audio Quality Determination Based on Perceptual Measurement Techniques", *Applications of Digital Signal Processing to Audio and Acoustics* (Edited by M. Kahrs and K. Brandenburg), Kluwer Academic Publishers, Boston, 1998.

[6] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality", *TSP Lab Technical Report*, McGill University, 2003.
http://www- mmsp.ece.mcgill.ca/Documents/Reports/index.html

[7] A. Lerch, software: EAQUAL - Evaluation of Audio Quality, V.0.1.3alpha, zplane.development, 2002.
http://www.rarewares.org/others.php → Eaqual

[8] R. Huber, B. Kollmeier, "PEMO-Q — A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, No. 6, pp. 1902-1911, 2006.
http://www.hoertech.de → products → downloads

[9] S.R. Quackenbush, T.P. Barnwell, M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, New Jersey, 1988.

[10] J.H.L. Hansen, B.L. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms," *INTERSPEECH*, vol. 7, pp. 2819-2822, 1998.

[11] F. Mustière, M. Bouchard, and Bolić M., "Quality Assessment of Speech Enhanced Using Particle Filter," *ICASSP*, vol. 3, pp. 1197-1200, 2007.

[12] L.D. Persia, M. Yanagida, H.L. Rufiner, D. Milone, "Objective Quality Evaluation in Blind Source Separation for Speech Recognition in A Real Room," *Signal Processing*, vol. 87, no. 8, pp. 1951-1965, 2007.

[13] A. Garay Acevedo, "Audio Watermarking Quality Evaluation," *e-Business and Telecommunication Networks* (Edited by J. Ascenso et al.), Springer, Netherlands, 2006.

[14] G. Stoll, F. Kozamernik, "EBU Listening Tests on Internet Audio Codecs," *EBU Technical Review*, 2000.

[15] EBU, "SQAM - Sound Quality Assessment Material",
http://sound.media.mit.edu/mpeg4/audio/sqam/

[16] Y.Q. Lin, W.H. Abdulla, "Multiple Scrambling and Adaptive Synchronization for Audio Watermarking", *IWDW*, LNCS 3304, Springer-Verlag, pp. 456-469, 2007.

[3] Specifically, indices i, j, k are for denoting host signal, watermark strength, quality measure, respectively.