AUDIO QUALITY ASSESSMENT USING THE MEAN STRUCTURAL SIMILARITY MEASURE

Srivatsan Kandadai, Joseph Hardin and Charles D. Creusere

Klipsch School of Electrical and Computer Engineering
New Mexico State University, Las Cruces, NM 88003-8001

ABSTRACT

Efficient transmission and storage of digital audio signals can be accomplished using a wide variety of compression algorithms. To compare and optimize the performance of such algorithms, objective metrics are often used to measure the quality of such compressed audio signals since subjective testing is extremely time consuming. In this paper, we consider the application of the structural similarity measure, originally developed for image quality assessment, to the problem of audio quality assessment. Specifically, we study two different implementations of the structural similarity index: the first applies it to short and fixed time-domain frames of an audio sequence while the second decomposes the audio signals into a non-redundant, time-frequency map and then compares the structural similarity in the resulting 2-dimensional domain. We compare the accuracies of the two structural similarity measures with those of other accepted objective audio quality metrics relative to MUSHRA-based subjective audio evaluations.

Index Terms— Structural Similarity, SSIM, MSSIM, Audio quality analysis, PEAQ, objective metric, Quality measurement.

1. INTRODUCTION

Audio compression is widely used to efficiently store and transmit audio data, and many different algorithms have been developed. Algorithms that are particularly well-know include MPEG-1 audio layer 3 (MP3), MPEG-2/4 Advanced Audio Coder (AAC) and Bit Slice Arithmetic Coding (BSAC), MPEG-4 Transform Weighted Interleaved Vector Quantization (TWIN-VQ), Microsoft Windows Media Audio (WMA), and Dolby Digital (originally called AC-3). These algorithms significantly reduce the number of bits needed to represent audio signals while maintaining acceptable perceptual quality. For compression or source coding, objective metrics like mean square error and segmental signal-to-noise ratios have often been used to characterize the quality of the reconstructed signal, but such metrics are generally not popular for evaluating modern audio codecs (encoder/decoders) because such codecs allocate bits using detailed psychoauchoustic models of human hearing.

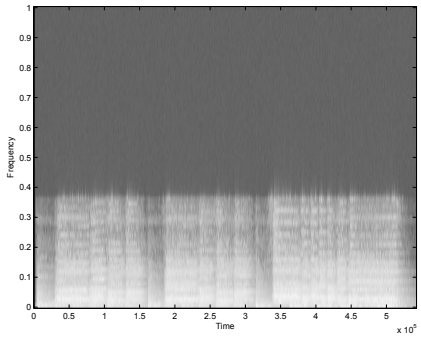
Since the advent of the modern perceptual audio codec in the late 1980s, many objective quality metrics have been developed that try to measure human subjective audio quality [1][2][3][4]. The International Telecommunications Union (ITU) has put forth recommendation BS.1387, also known as Perceptual Evaluation of Audio Quality (PEAQ) as an approach for doing exactly this [5]. PEAQ combines many of the best available quality metrics available in the early 1990s, attempting to merge their various strengths. PEAQ however has been shown to be a poor indicator of perceptual quality for highly impaired audio [6]. More recently, a new method called Energy Equalization Truncation (EET) has been introduced, and has been shown to be especially effective as a measure of the quality in highly impaired audio [7]. EET has also been combined with the some of the model output variables (MOVs) from PEAQ to form an audio quality metric that is more robust over a wide range of audio fidelities [6].

In this paper we study the application of the Mean Structural Similarity (MSSIM) measure [8], developed to estimate the reconstruction quality of compressed images, to the problem of audio quality evaluation. MSSIM is a statistical method that compares corresponding segments of a given degraded image with the same segments of the original image, and it has been shown to give results that closely match subjective test results. This paper is organized as follows. In Section 2, the MSSIM metric is discussed while in Section 3, we develop two different ways of applying MSSIM to audio sequences. Section 4 details our experimental results, and conclusions and future research directions are presented in Section 5.

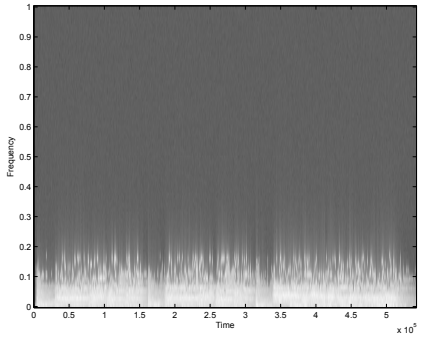
2. STRUCTURAL SIMILARITY INDEX (SSIM)

The structural similarity index is based on the idea that a measure of change in structural information is a good approximation to perceived quality change. For example, the audio sequences corresponding to the spectrograms in Figure 1 (a) and (b) have the same mean square errors but very different mean opinion scores (MOS) and structural similarity index scores. The sequence in Figure 1 (a) has frequencies truncated above

Research supported by NSF Grant CCR-0133115.



(a)



(b)

Fig. 1. (a) Spectrogram of frequency (with positive frequencies normalized to a range of 0 and 1) truncated audio sequence with additive white Gaussian noise with MSE 0.04 and MSSIM = 0.38 and (b) audio sequence quantized using BSAC algorithm at 16 kbps with MSE 0.048 and MSSIM = 0.1042.

8 kHz w.r.t. the original (reference) signal, and has a constant hiss generated by an additive Gaussian noise. On the other hand, the sequence in Figure 1 (b) is generated by the compressing the audio sequence with the AAC-BSAC algorithm at 16 kbps. This sequence contains many clicks and clacks which are much more discernable than the constant audible hiss of sequence (a).

The structural similarity measure considers three different measured differences between the original and reconstructed signals: luminosity, contrast and structure. The luminosity is a comparison of the mean values of the signals. If $\mathbf{x}$ and $\mathbf{y}$ are corresponding segments of audio with N samples each, the luminosity comparison is given by

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{1}$$

where $\mu_x = 1/N\sum_{i=1}^{N} x_i$, $\mu_y = 1/N\sum_{i=1}^{N} y_i$ and $C_1 = (K_1L)^2$ where $K_1 << 1$. The dynamic range of the elements of $\mathbf{x}$ and $\mathbf{y}$ is denoted by the variable L.

Ignoring the C_1 term, the form of (1) is very similar to that of the correlation coefficient except with respect to the two means. The luminosity comparison is not particularly useful for audio since the mean values do not change much even with large degradation in the audio sequences (all audio sequences are essentially zero mean over long segments). Not surprisingly, we find in Section 5 that when the relative weight of the luminosity is optimized for the subjective audio test data, it is small.

The contrast or variance comparison is defined similar to the luminance comparison given by (1) but with respect to the relative standard deviations of the two segments: i.e.,

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2}$$

where $\sigma_x = \left(1/(N-1)\sum_{i=1}^{N}(x_i-\mu_x)^2\right)^{\frac{1}{2}}$, $\sigma_y = \left(1/(N-1)\sum_{i=1}^{N}(y_i-\mu_y)^2\right)^{\frac{1}{2}}$ and $C_2 = (K_2L)^2$ where $K_2 << 1$.

Structure comparison is done after the local mean subtraction and local variance normalization. Structure comparison measures the similarity between the two N-dimensional unit-norm vectors, $\overline{\mathbf{x}} = (\mathbf{x}-\mu_x)/\sigma_x$ and $\overline{\mathbf{y}} = (\mathbf{y}-\mu_y)/\sigma_y$, and it is simply the dot product between the two unit vectors $\overline{\mathbf{x}}\cdot\overline{\mathbf{y}} = \overline{\mathbf{x}}^t\overline{\mathbf{y}}$ which is an effective way to quantify the structural similarity between them. This is equivalent to the correlation coefficient between the original $\mathbf{x}$ and $\mathbf{y}$. The structural measure in terms of the original $\mathbf{x}$ and $\mathbf{y}$ vectors is given by

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3}$$

where $\sigma_{xy} = 1/(N-1)\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$ and $C_3 = (K_3L)^2$ where $K_3 << 1$.

The similarity comparisons of (1), (2) and (3) also satisfy the following properties (where 'S' below can be either 'l', 'c', or 's'):
1. Symmetry: $S(\mathbf{x},\mathbf{y}) = S(\mathbf{y},\mathbf{x})$,
2. Boundedness: $S(\mathbf{x},\mathbf{y}) \leq 1$,
3. Unique maximum: $S(\mathbf{x},\mathbf{y}) = 1$ if and only if each element of $\mathbf{x}$ is equal the corresponding element of $\mathbf{y}$.

Finally, the three components are combined to yield an overall similarity measure for the segment,

$$\begin{aligned} S(\mathbf{x},\mathbf{y})) &= f(l(\mathbf{x},\mathbf{y})c(\mathbf{x},\mathbf{y})s(\mathbf{x},\mathbf{y})) \\ &= l(\mathbf{x},\mathbf{y})^{\alpha}c(\mathbf{x},\mathbf{y})^{\beta}s(\mathbf{x},\mathbf{y})^{\gamma} \end{aligned} \tag{4}$$

where, $\alpha > 0$,$\beta > 0$ and $\gamma > 0$ are used to adjust the relative importance of the three components. The function in (4) also satisfies the three properties listed above.

The SSIM measure described above is clearly statistically based. While it is possible to apply SSIM to an entire audio sequence to extract a single quality number for that sequence, information about local structure would be lost and the complexity would be quite high. Instead, we calculate the mean of the SSIM values taken over segments of the original data. This is advantageous both from a complexity standpoint and also because it allows for the possibility of an unequal weighting of the segmental SSIM values (e.g., values towards the

end of the sequence might be weighted more heavily to take into account that the perceived quality of the most recently heard audio influences the listener's opinions of its quality most heavily). This global measure is referred to as mean SSIM or MSSIM.

3. APPLYING MSSIM TO AUDIO

Structure in audio can be viewed in two ways. In the first case, we assume that the structure depends on each time sample and its position with respect to a small temporal neighborhood of samples around it. To apply the MSSIM from this point of view, we split the sequences into temporal frames of length 128 with 50% overlap and then apply the SSIM to each frame separately. The mean SSIM is calculated by averaging the individual SSIM values for each frame. Doing this compares only the temporal structure of the audio sequences. We refer to this method as the temporal MSSIM (T-MSSIM).

In the second approach, we apply a time-frequency transform to the audio sequences. Specifically, we use a 256-point Modified Discrete Cosine Transform (MDCT) with a 50% overlapping window. This represents the audio data as a time-frequency decomposition. Clearly, this representation of a 1-dimensional audio sequences is similar to an image as shown in Figure 1. By applying SSIM to 2-dimensional blocks of the time-frequency representation, we can evaluate structural similarities in both the time and frequency domains simultaneously. We will refer to this method as the time-frequency MSSIM (TF-MSSIM).

MDCT used here can be viewed as a critically-subsampled, quadrature mirror filter bank (QMF). In our case, we implement the MDCT using a raised sine-shaped window followed by a cosine transform. The input frame has a 50% overlap with each adjacent temporal frame.

4. EXPERIMENTS AND RESULTS

We use seven different monoaural audio sequences sampled at 44.1 kHz for the following experiments. To generate the different test sets we modify these audio sequences by adding different types of noise, band-limiting the audio sequences and applying different audio compression algorithms. The subjective test results used to compare the MSSIM results with those of other metrics are obtained from 15 test subjects using the MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) protocol [9].

For the T-MSSIM, the audio sequences are split into time frames of 128 samples each. In the initial testing, the values of α, β and γ are simply set at 1, implying that each of the three components of the metric are equally weighted. Figure 2, shows the scatter plot of the temporal MSSIM with respect to subjective test results for 17 different cases. The solid line represents the regression line between the MSSIM and subjective results. The correlation coefficient obtained is 0.98, which shows that the MSSIM and the subjective tests are highly correlated. However, from the regression line we see that there is a bias of 0.21 in the MSSIM data.

Table 1. Exponential weights obtained for the different MSSIM components

Parameters	T-MSSIM	TF-MSSIM
α	0.2	0
β	0.5	0.8
γ	0.7	0.2

Similarly, Figure 3 shows the scatter plot for TF-MSSIM with respect to subjective test results. The time-frequency representation is obtained by performing a 256 point MDCT with an overlap of 128 samples. The individual block size over which SSIM is performed is of size 8×8 samples. The correlation coefficient obtained by comparing the time-frequency MSSIM is 0.976, which implies that there is good correlation between the MSSIM and subjective tests.

The T-MSSIM and TF-MSSIM appear to perform equally well based on the correlation coefficient with respect to subjective data. From (4) we know that parameters α, β and γ control the relative importance of the mean, variance and structure components of the SSIM. To obtain the values of these parameters so that the MSSIM matches the subjective tests results, we perform a least squares approximation as follows. We assume that the MSSIM is approximately equal to the weighted product of the mean of the luminosity, contrast and structure components, i.e.

$$S_m(\mathbf{x},\mathbf{y}) \approx l_m(\mathbf{x},\mathbf{y})^{\alpha} c_m(\mathbf{x},\mathbf{y})^{\beta} s_m(\mathbf{x},\mathbf{y})^{\gamma} \qquad (5)$$

The subscript m in (5) implies the mean of the three components. Taking the natural logarithm on both sides of (5), yields the linear equation,

$$\ln S_m(\mathbf{x},\mathbf{y}) = \alpha \ln l_m(\mathbf{x},\mathbf{y}) + \beta \ln c_m(\mathbf{x},\mathbf{y}) + \gamma \ln s_m(\mathbf{x},\mathbf{y}) \qquad (6)$$

from which a least square approximation of the exponential parameters can be found, while constraining the parameters to be positive. Table 1 lists the values for the parameters for the temporal and time-frequency cases. From this table we see that the mean or the luminance component is weighted very low for both the temporal and time-frequency cases. In T-MSSIM the structural component is emphasized more than the variance or contrast component while TF-MSSIM is exactly the opposite. Figures 4 and 5 show the scatter plots of the subjective test results with respect to the weight-optimized temporal and time-frequency MSSIMs, respectively. The test set represented in these plots are different from the ones used to estimate the weights. The correlation coefficients for the weighted temporal and time-frequency MSSIMs are 0.998 and 0.988 respectively. From the plots we see that the constant bias is reduced and that the correlation coefficients have not improved significantly.

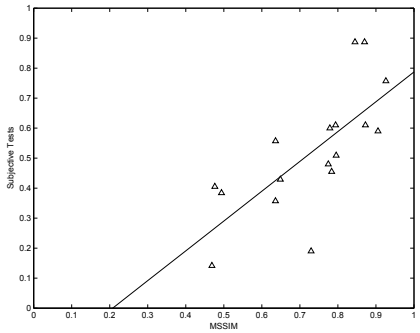


Fig. 2. Scatter plot of subjective test results vs. T-MSSIM

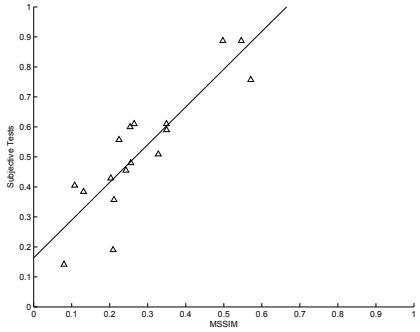


Fig. 3. Scatter plot of subjective test results vs. TF-MSSIM

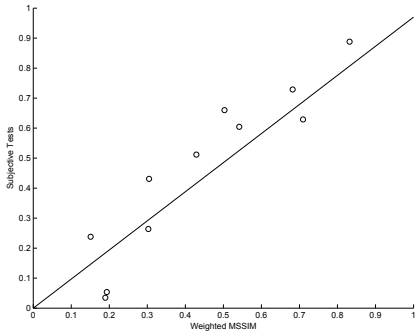


Fig. 4. Scatter plot of subjective test results Vs. T-MSSIM after optimization

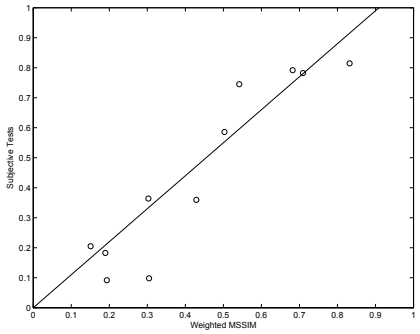


Fig. 5. Scatter plot of subjective test results Vs. TF-MSSIM after optimization

5. CONCLUSIONS AND FUTURE WORK

In this paper we describe the use of MSSIM in the context of assessing audio quality. We have presented two different ways of applying MSSIM to audio data. Experimental results show that both these methods are equally effective in finding the audio quality. The optimal weights used to combine the three components of SSIM are, however, different for each case. Both these techniques have good correlation to subjective data even with equal weights.

In the future we would like to validate this technique with more analysis. Since MSSIM is differentiable, we can maximize or minimize over it while keeping the MSE constant to create pairs of audio sequences which can then be subjectively assessed: if MSSIM is a good perceptual metric, such min/max-optimized perceptual comparisons should make it very apparent. Furthermore, if MSSIM can be shown to truly be a good predictor of perceptual audio quality, then using it in place of frequency-masking models in audio codecs could significantly simplify their implementations.

6. REFERENCES

[1] T. Thiede and E. Kabot, "New perceptual quality measure for the bitrate reduced audio," *Preprint 4280, Copenhagen, Denmark*, 1996.

[2] J.G. Beerends and J.A. Stemerdink, "A perceptual audio quality measure based on psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 40, Dec. 1992.

[3] S. Morissette, B. Paillard, P. Mabilleau, and J. Soumagne, "Perceval: Perceptual evaluation of the quality of audio signals," *Journal of the Audio Engineering Society*, vol. 40, pp. 21–31, January/February 1992.

[4] "Method for objective measurements of perceived audio quality," in *Recommendation ITU-R BS.1387-1, Geneva, Switzerland*, 1998-2001.

[5] "Methods for objective measurement of perceived audio quality, recommendation itu-r bs.1387-1," 1998-2001.

[6] C. D. Creusere, R. Vanam, and K. Kallakuri, "An objective metric of human subjective audio quality optimized for a wide range of audio fidelities," *IEEE Trans. Audio, Speech, and Language Process.*, 2006.

[7] C.D. Creusere, "Understanding perceptual distortion in mpeg scalable audio coding," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 422–431, May 2005.

[8] Z. Wang, A. C. Bovik, H. R. Sheik, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, 2004.

[9] "Method for subjective assesment of intermediate quality levels of coding systems," *Recommendation IRU-R BS.1534-1,(Question ITU-R 220/10)*.