STRUCTURAL SIMILARITY ANALYSIS OF MODULATION FOR AUDIO QUALITY ASSESSMENT

Mengyao Zhu[1], *Jia Zheng*[1], *Craig Jin*[2] *and Wanggen Wan*[1]

[1]School of Communication & Information Engineering,
Shanghai University, Shanghai, China
[2]School of Electrical and Information Engineering,
University of Sydney, Australia

ABSTRACT

This paper proposes an improved structural similarity method for audio quality assessment, which is Structural Similarity Analysis of Modulation (SSAM). Different from original structural similarity index measure, we introduce the analysis of structural similarity of modulation together with Computational Auditory Signal-processing and Perception (CASP) model. Audio features from CASP are extended to three dimensions: time, frequency and modulation spectrum. The combined architecture of ITU-R BS.1387-1 and proposed SSAM is given in this paper. Our proposed estimation system not only shows highly correlated with subjective results but also overcomes the shortage of ITU-R BS.1387-1 that only suitable for small impaired audio.

Index Terms—Audio quality, SSAM, modulation domain

1. INTRODUCTION

The issue of audio quality assessment was investigated quite intensively in recent years since its importance for multimedia application. The ITU-R BS.1387-1 [1], known as Perceptual Evaluation of Audio Quality (PEAQ), is a representative standard of slightly impaired audio quality evaluation. To overcome the limitation of PEAQ, PEMO-Q [2] was presented by introducing a computational auditory model and a novel Perceptual Similarity Measure (PSM), which predict the structural similarity of audio signals by calculating the linear cross correlation coefficient of the internal representations. Structural similarity measurement is introduced to represent the difference of auditory model output. Actually, the concept of structural similarity measurement is originally from image quality assessment [3]. Structural Similarity Index Measure (SSIM) is statistically based, and it has been applied to audio quality assessment in recent research [4, 5]. In previous work, we considered a new way to implement SSIM in frequency domain for audio quality assessment [6].

The auditory model for audio quality assessment was also improved in PEMO-Q, and one of the latest version of auditory model is CASP model [7]. The most important improvement in CASP is modulation filter bank, which mimic the cortex of human ear. The binaural CASP model is applied in spatial audio quality assessment [8], while the modulation filter bank is neglected due to technical difficulties (not enough RAM for the computation). The efficient and effective method for solving the problem of computational complexity in modulation filter bank has yet to be developed. This paper proposes a structural similarity calculation of 3 dimensional internal representations of modulation filter bank, which is the product of CASP model. Our proposed SSAM method can reduce the data complexity.

In this paper, the combined architecture of PEAQ and SSAM is given. In SSAM algorithm, the SSIM is applied in a different way for audio quality feature extraction. Other than calculating the three structural similarity indexes, we used the indexes as Model Output Variables (MOVs) for analysis of the modulation features implied in the internal representations. Together with another 2 MOVs, we employed the minmax-optimized MOV selection cognitive algorithm [9] for audio quality evaluation. The correlation between the Subjective Difference Grade (SDG) and Objective Difference Grade (ODG) of our proposed approach increased sharply. At the same time, the improved system can assess the quality of highly impaired audio. In this regard, it is a great option of PEAQ to adopt our method for evaluating audio quality more efficiently.

The paper is organized as follows. Section II presents the background and related works. In the section III, we develop an improved objective audio quality evaluation method based on SSAM algorithm while the analysis of the performance of this metric is given in section IV. Finally, conclusions are presented in section V.

2. BACKGROUND AND RELATED WORKS

Recently, structural similarity measurement has been applied to the problem of audio quality assessment and

shows to be a good predictor of perceptual audio quality. It is originated from the idea that a measure of change in structural information is a good approximation to perceive quality change [4]. A weighted mixture of correlation of the means, normalized variances and normalized cross correlations are made up of the structure similarity indexes [5]. For traditional SSIM, three structural similarity indexes: luminance, contrast and structure comparison can be obtained as follows.

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

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

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

where μ_x and μ_y is the means of signal x and y, while σ_x and σ_y are the standard deviations and σ_{xy} is the covariance. C_1, C_2 and C_3 are very small positive constants. Finally, the three components are combined to yield an overall similarity measure,

$$SSIM(x,y)=f(l(x,y),c(x,y),s(x,y)) \tag{4}$$

For the whole sequence, SSIM is often applied to each segments and the final mean SSIM is calculated. Suppose the input sequence is divided into m segments, the MSSIM can be calculated as follows.

$$MSSIM(x,y)=\frac{1}{m}\sum_{i=1}^{m}SSIM(x,y) \tag{5}$$

The research on SSIM is divided into three aspects: Temporal SSIM (T-MSSIM) [4], Frequency SSIM (F-MSSIM) [6] and Time-Frequency SSIM (TF-MSSIM) [4]. The T-MSSIM applied SSIM to each fixed time-domain frames of audio sequence while the F-MSSIM used the SSIM to fixed frequency-domain frames. The TF-MSSIM adopted a time-frequency transform to audio sequence and applied SSIM to the time-frequency representation. It is similar to apply SSIM to a two-dimensional image.

3. PROPOSED SSAM FOR AUDIO QUALITY ASSESSMENT

3.1. Proposed SSAM algorithm

The early researches and related works considered the time domain and frequency domain when applied SSIM. In proposed SSAM approach, we extend the analysis of structural similarity to 3 dimensions: time, frequency and modulation frequency domain.

The modulation frequency domain is produced by the CASP model, which represents time samples, frequency bands and modulation channels. Those data is multi-dimension, and the measurement of difference in modulation frequency domain is very difficult. Inspired by the SSIM, the modulation difference between reference and test signals can be represented by using structural similarity method. In this paper, we propose a structure similarity analysis of modulation method for audio quality assessment. A combined architecture of SSAM and PEAQ for implementation of objective quality evaluation is also given in next section.

3.1.1. Index L

Assuming the reference signal and test signal with N samples and K frames are respectively $x_{ij}(n, k)$ and $y_{ij}(n, k)$, i and j are corresponding of audio with i bands and j modulation channels. The mean value of per-frame audio signal is:

$$\mu_x(k)=\frac{1}{MNq}\sum_{i=1}^{M}\sum_{j=1}^{q}\sum_{k=1}^{N}x_{ij}(n,k) \tag{6}$$

$$\mu_y(k)=\frac{1}{MNq}\sum_{i=1}^{M}\sum_{j=1}^{q}\sum_{k=1}^{N}y_{ij}(n,k) \tag{7}$$

we can get the Index L for per-frame signal as follows:

$$l(X,Y)=\frac{2\mu_x(k)\mu_y(k)+C_1}{\mu^2{}_x(k)+\mu^2{}_y(k)+C_1} \tag{8}$$

where $C_1=(K_1L)^2$ and $K_1<<1$. The variable L is the dynamic range of the elements of x and y.

3.1.2. Index C

The standard deviations ($\sigma_{x_{ij}}(k)$, $\sigma_{y_{ij}}(k)$) of i band and j modulation channel of per-frame audio signal is:

$$\sigma_{x_{ij}}(k)=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{ij}(n,k)-\mu_x(k))^2} \tag{9}$$

$$\sigma_{y_{ij}}(k)=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_{ij}(n,k)-\mu_y(k))^2} \tag{10}$$

Then the average standard deviations of per frame signal are:

$$\sigma_x(k)=\frac{1}{Mq}\sum_{i=1}^{M}\sum_{j=1}^{q}\sigma_{x_{ij}}(k) \tag{11}$$

$$\sigma_y(k)=\frac{1}{Mq}\sum_{i=1}^{M}\sum_{j=1}^{q}\sigma_{y_{ij}}(k) \tag{12}$$

The Index C of the per-frame signal is:

$$c(X,Y)=\frac{2\sigma_x(k)\sigma_y(k)+C_2}{\sigma_x^2(k)+\sigma_y^2(k)+C_2} \tag{13}$$

where $C_2=(K_2L)^2$ and $K_2<<1$.

3.1.3. Index S

The Index S is

$$s(X,Y)=\frac{\sigma_{xy}(k)+C_3}{\sigma_x(k)\sigma_y(k)+C_3} \tag{14}$$

where

$$\sigma_{xy}(k)=\frac{1}{Mq(N-1)}\sum_{i=1}^{M}\sum_{j=1}^{q}\sum_{k=1}^{N}\left(x_{ij}(n,k)-\mu_x(k)\right)\left(y_{ij}(n,k)-\mu_y(k)\right)$$

and $C_3=(K_3L)^2$ where $K_3<<1$.

3.2. Audio quality assessment applied SSAM

PEAQ applies two psychoacoustic models[1]: a low complexity Basic Version and an accurate Advanced Version. In PEAQ advanced version, the FFT-based ear model and filter bank based ear model absorb some ideas of the audio perceptual phenomenon. The FFT-based ear model is responsible for measure the difference of signal and noise energy while the filter bank ear model measure the modulation difference of audio signals. The auditory models of PEAQ do not include the transformation of human ear cortex. Therefore we used the advanced CASP model with cortex transformation instead of the filter bank ear model. Based on the structure of PEAQ advanced version, we combined the FFT-based ear model and CASP model to implement the integrated psychoacoustic processing stage. As for the feature extraction, we consider the application of SSAM algorithm to the problem of modulation difference calculation. Fig.1 shows the framework of our proposed audio quality evaluation system.

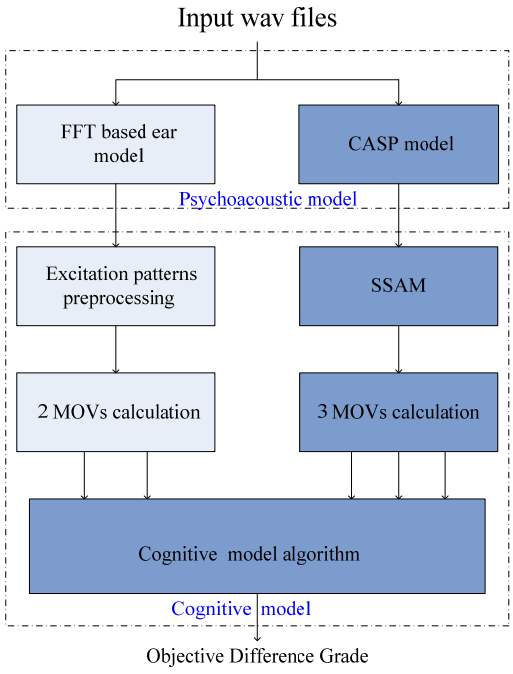


Fig.1. Proposed audio quality evaluation system

From the Fig.1, we can see the FFT based ear model generates 2 Model Output Variables (MOV), including Segmental NMR_B (Noise-to-Mask Ratio) and EHS_B (Harmonic Structure of the Error)[1]. For CASP model, another 3 MOVs: Index L, Index C and Index S can be obtained by the structural similarity analysis of modulation. Supposing K is the number of frames for an audio sequence, the mean L, C and S for an audio sequence are as follows.

$$l_k(X,Y)=\frac{1}{K}\sum_{i=1}^{K}l(X,Y) \quad (15)$$

$$c_k(X,Y)=\frac{1}{K}\sum_{i=1}^{K}c(X,Y) \quad (16)$$

$$s_k(X,Y)=\frac{1}{K}\sum_{i=1}^{K}s(X,Y) \quad (17)$$

SSAM can be regarded as a feature extraction method, and it produces MOV for cognitive part. Artificial network is a cognitive algorithm used in PEAQ. Usually, it is difficult to obtain consistent result by ANN methods. A linear minmax-optimized MOV selection algorithm is widely used in recent research [5, 6, 9] and shows good consistence and performance for audio quality assessment. In our proposed framework, 5 MOVs are mapped to ODG according to minmax-optimized cognitive, which can be referred in more detail at [5].

4. IMPLEMENTATION AND RESULTS

As to measure the effective of our proposed approach, we implemented audio quality assessment on two kinds of audio databases. Database I contains 72 audio files with subjective scores obtained by MUSHRA [11]. Audio files from Database I are formed by audio packet loss concealment and can be regard as high distortion audio. Database II contains 32 audio files and another 4 kinds of 5.1 channels (5-to-2) surround sounds [12]. They are processed by four different methods (time-domain filtering, frequency-domain filtering, MDCT dynamic neighborhood filtering and MDCT fixed neighborhood filtering). The 4 methods above produce small difference of result, so we regard those audio files as slightly impaired audio.

We firstly implemented the PEAQ advanced version. The correlation and Mean Square Error (MSE) between the SDG and ODG of the PEAQ was calculated as benchmark. In the proposed evaluation system, we applied the CASP model and SSAM algorithm to measure the modulation difference of audio signals. By calculating three structural indexes, the modulation difference between the reference and test signals was obtained. As audio has a wider frequency range, the frequency range of the CASP model is fitted to the 80-18000Hz, referring to PEAQ. According the minmax-optimized MOV selection cognitive algorithm, we calculated the final ODG and got the correlation and MSE between the SDG and ODG.

The experiment results are shown in the tables below. In both tables, the PEAQ and PEAQ2 respectively presents the results of original PEAQ advanced version and PEAQ advanced version with minmax-optimized MOV selection algorithm. The proposed method is CASP with SSAM

evaluation approach and used the minmax-optimized MOV selection algorithm.

TABLE 1
RESULTS OF DIFFERENT EVALUATION SYSTEMS ABOUT DATABASE I

Methods / Results	PEAQ	PEAQ2	Proposed
Correlation	0.244129	0.101446	**0.629711**
MSE	0.596015	0.707361	**0.796705**

TABLE 2
RESULTS OF DIFFERENT EVALUATION SYSTEMS ABOUT DATABASE II

Methods / Results	PEAQ	PEAQ2	Proposed
Correlation	0.132164	0.114625	**0.243009**
MSE	1.3272	0.072565	**0.033289**

From the Table 1, the correlation of PEAQ2 between SDG and ODG using minmax-optimal MOV selection algorithm is decreased. Comparing proposed method with PEAQ2, the correlation is increased sharply, and also the highest. In Table 2, the correlation of proposed method is also improved, although it is not as significant as the result in Table 1. Please note that the MSE of the proposed evaluation system based on CASP and SSIM reduced a lot in Table 2. It is proved that the audio quality assessment using CASP model is effective and reasonable.

In order to better describe the good performance of our proposed system, the curve of ODG vs. SDG is given in Fig. 2. Figure (a) describes the fitting curve of PEAQ advanced version and figure (b) is the ODG vs. SDG curve of proposed audio quality assessment system based on CASP and SSAM modulation analysis. For the figures, the *x* axis means the mean SDG while the *y* axis indicates the mean ODG. The blue circles in the picture are all the data spots of the test audio files with coordinates (SDG, ODG). We plot the curves with 95% prediction bounds. The red solid lines are the fitting curves of the SDG vs. ODG and the blue dash lines are the standard lines (ideal results). Compared with the curve of PEAQ advanced version, the proposed approach is more close to the standard line. The circles in figure (b) are more compact than figure (a). Hence, our proposed audio quality assessment metric is more accurate than the traditional PEAQ.

5. CONCLUSION

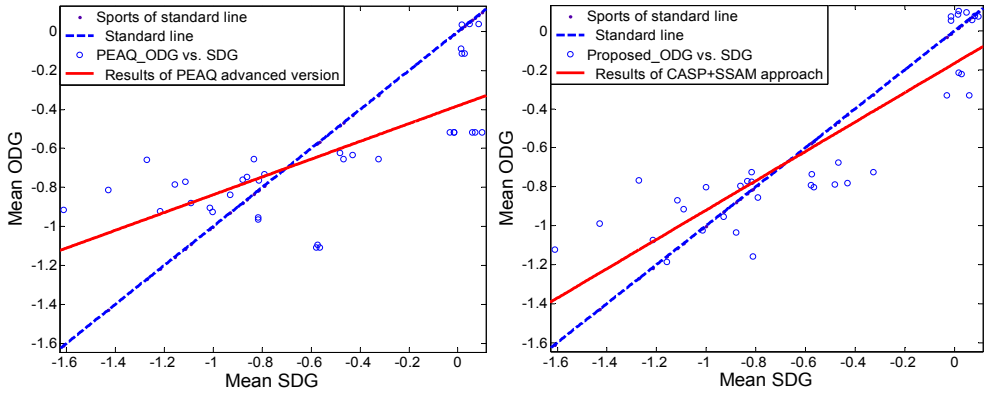
We developed a novel approach for audio quality evaluation based on CASP and SSAM. By the analysis of modulation difference and introduced 3 MOVs in new audio quality assessment system, the accuracy of audio quality is highly improved. Furthermore, our proposed objective evaluation system is compatible with PEAQ advanced version and can predict the quality of highly impaired audio, the PEAQ can be further refined by using the computational auditory model.

ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. 61001161, Innovation Program of Shanghai Municipal Education Commission No.12YZ024, and Leading Academic Discipline Project of Shanghai Municipal Education Commission under Grant No. J50104.

RELATION TO PRIOR WORK

The work presented here has focused on Structural similarity analysis of modulation, which is different from original Structural Similarity Index Measure [4, 5]. The original SSIM is related to time and frequency domain. We combined the analysis of structural similarity of modulation with Computational Auditory Signal-Processing and Perception model. In turn we extend structural similarity to modulation domain, which was not considered in previous works.



(a) Fitting curves of SDG&ODG of PEAQ advanced version (b) Fitting curves of SDG&ODG of Proposed system

Fig.2.Fitting curves of SDG&ODG of Proposed system and PEAQ advanced version

REFERENCES

[1]*Method for Objective Measurements of Perceived Audio Quality*, Rec.ITU-R BS.1387-1, 1998-2001

[2]R. Huber and B. Kollmeier, "PEMO-Q: A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio, Speech*, *Lang. Process.*, vol. 14, no.6, pp. 1902-1911, Nov.2006.

[3]Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: from Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no.4, pp. 600-612, Apr.2004.

[4]S. Kandadai, J. Hardin, and C. D. Creusere, "Audio Quality Assessment using the Mean Structural Similarity Measure," in *IEEE Int. Conf. Acoust.*, *Speech*, *Signal Process. ICASSP'08*, *2008.* pp. 221-224.

[5]C. D. Creusere and J. C. Hardin, "Assessing the Quality of Audio Containing Temporally Varying Distortions," *IEEE Trans. Audio, Speech*, *Lang. Process.*, vol. 19, pp. 711-720, 2011.

[6]C. Gan, X. Y. Wang, M. Y. Zhu, and X. Q. Yu, "Audio Quality Evaluation using Frequency Structural Similarity Measure," in *IET Int. Conf. Wireless*, *Mobile*, *Comput. CCWMC'2011*, *2011*, pp. 299-303.

[7]L. Jepsen Morten, D. Ewert Stephan, and Dau Torsten, "A Computational Model of Human Auditory Signal Processing and Perception," *The Journal of the Acoustical Society of America*, vol. 124, no.1, pp. 422-438, 2008.

[8]Z. Podwińska, "Binaural Auditory Model for Audio Quality Assessment," *Master's Degree Thesis of Science in Engineering Acoustics and Audio Technology, Aalborg University*, *Denmark*, 2012.

[9]C. D. Creusere, K. D. Kallakuri, and R. Vanam, "An Objective Metric of Human Subjective Audio Quality Optimized for a Wide Range of Audio Fidelities," *IEEE Trans. Audio, Speech*, *Lang. Process.*, vol. 16, no.1, pp. 129-136, Jan. 2008.

[10] Z. Y. Guo, "Objective Audio Quality Assessment Based on Spectro-Temporal Modulation Analysis," *Master's Degree Thesis of Stockholm, KTH Signals Sensors and Systems, Sweden*, 2011.

[11] M. Y. Zhu, M. Zhang, X. Q. Yu, and W. G. Wan, "Streaming Audio Packet Loss Concealment Based on Sinusoidal Frequency Estimation in MDCT Domain," *IEEE Trans. Consumer Electronics*, vol. 56, pp. 811-819, 2010.

[12] Meng-Yao Zhu, Ning Chen, Xiao-Qing Yu and Wang-Gen Wan, "Fast convolution for binaural rendering based on HRTF spectrum," in Audio Language and Image Processing (ICALIP), 2010 International Conference on. 2010. p. 353-356.