AUDITORY ATTENTION BASED MOBILE AUDIO QUALITY ASSESSMENT

Yang Yuhong[1,2], *Yu Hongjiang*[1], *Hu Ruimin*[1], *Gao Li*[1], *Wang Song*[1], *Zhai Qing*[1], *Xie Songbo*[1]

National Engineering Research Center for Multimedia Software, School of Computer, Wuhan Univ., China[1]
Research Institute of Wuhan University in Shenzhen, China[2]

ABSTRACT

Mobile audio services are growing with rising popularity of smart mobile devices using WiFi or cellular networks. A major issue facing mobile audio quality assessment is occasional background noises due to the prospect of sound recording at anytime and anywhere with smart mobile devices. Psychological study reveals that people pay selective attention to their interested sound in complex auditory input. In this paper, we model the mobile audio objective quality assessment based on auditory attention mechanism, with attention based horizontal azimuth parameters and timbre distortion parameters as additional Model Output Variables (MOVs). The results show that the prediction accuracy can be obtained by using such a method.

Index Terms— quality assessment, auditory attention, mobile audio

1. INTRODUCTION

In recent years, mobile audio services like mobile audio on demand, audio messages and music on cell-phones are confronting rapid growth and increasing number of users. People can easily record and transmit audio signals almost everywhere using smart mobile devices. With the increase customer expectation to mobile audio services and applications, classification of mobile audio becomes more complex and diversified. The issue of how to trustfully assess the mobile audio quality is urgent with the growing markets.

Subjective listening tests are generally regarded as the most reliable way of assessing audio quality. But they are also time consuming, expensive and labor intensive. In 2011, Ribeiro,F. proposed a cost-effective measure called crowdMOS[1], which can outsource subjective listening tests to listeners from an internet crowd. However, it was beneficial to substitute listening tests with objective assessment methods.

Previously developed objective audio assessment algorithms were based on engineering principles such as Total Harmonic Distortion (THD) and Signal to Noise Ratio (SNR), i.e. they did not attempt to model the psychoacoustic features of the human auditory system. These algorithms do not give accurate results for the objective quality assessment of audio signals. In 1979, Schroeder[2] developed the first objective model by using auditory masking properties to improve performance. In 1987, Karjalainen[3] was one of the first to introduce acoustic characteristics into objective model to assess the quality of sound. His model was based on a noise loudness parameter. In the same year, Brandenburg[4, 5] developed a Noise to Mask Ratio (NMR) model. It evaluated the level difference between the noise signal and the masked threshold which is widely used in speech and audio quality assessment models. In 1996 Sporer[6] examined the mean opinion scale for audio quality assessment and completed further work in this area[7]. These early developments ultimately led to the development and standardization of the objective audio assessment.

Based on years of the related research, International Telecommunication Union (ITU) published an audio quality objective assessment standard (ITU C R BS.1387), which is named Perceptual Evaluation of Audio Quality (PEAQ)[8]. PEAQ was an algorithm that models the psychoacoustic principles of the human auditory system. PEAQ is designed for mono audio evaluation. In 2007, Choietal[9] put forward a multichannel audio objective evaluating model aim at surround sound based on PEAQ. In 2009, Sunish George[10] extended the audio quality assessment to the space level, by adding the front audio quality and surround audio quality to objective model. The above mentioned PEAQ advanced models was developed to include sound-field evaluation.

However, it also should be noted that PEAQ has only been designed to grade signals with extremely small impairments. In 2005, Vanam et al[11] show that including an Energy Equalization (EE) parameter as one of the model output variables (MOVs) of PEAQ improves its performance significantly and the performance of this modified version can be used to evaluate low bitrate scalable audio codecs. In 2012, Yuhong et al[12] improve PEAQ model by adding both EE and Jitter Distortion Measure (JDM) parameters to evaluate mobile audio.

Auditory attention can be described as recognition process that listener cast main attention on the specific sound source in the whole complex sound field while neglecting others. That means the interested sound could be hold fast and accurately from the sound scene. To the best of our knowledge, the recent developed objective audio assessment algo-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

rithms did not take account in the auditory attention mechanism of human in complex sound filed with background noise. In this paper, an intrusive model with psychological selective mechanism of auditory attention is proposed. EE parameter and JDM parameter are also incorporated in the PEAQ basic version for mobile audio quality assessment. To validate our model, we choose four typical scenes and conduct experiments to demonstrate the predictive accuracy of the proposed method.

2. AUDITORY ATTENTION BASED APPROACH

The auditory attention based approach is shown in Fig1. A perceptual model is used to compare the reference signal with the test signal. The perceptual model reproduces the key properties of hearing into PEAQ MOVs, JDM, EE parameter and horizontal azimuth MOVs for attention audio. Then a cognitive model uses these measures to estimate Objective Difference Grade (ODG).

Acoustic localization system of human ear mainly relies on the binaural cues and monaural cue, and is more accurate for frontal sound image than for the rear and bilateral one. In this paper, in order to simplify the experimental process, we only consider the binaural horizontal localization of audio source. Interaural time difference (ITD) and interaural level difference (ILD) are used to discriminate attention audio from background noise.

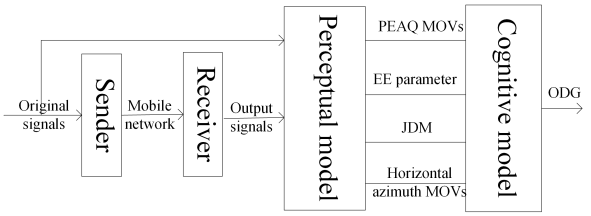

Fig. 1. Overview of auditory attention based approach.

2.1. MOVs for attention audio

When recording audio in complex sound-field, people are tending to fix the attention audio just in front of the audio recording units. Here we assumed that the audio source in azimuth $\theta \leq 6^{\circ}$ namely within the scope of $ILD \leq 0.6dB$, $ITD \leq 55\mu s$ is attention audio source.

ITD is an important cue for sound source localization, especially for a low-frequency sound. ITD can be computed from the following time window-based normalized cross-correlation function (NCF), where$X_{L,k,n}[l]$ and $X_{L,k,n}[l]$ are peripheral ear model outputs of the left ear and the right ear, respectively. k and n are the frequency band and time frame indices. The cross correlation is calculated over 7/8 overlapping rectangular time windows with the length approximately equivalent to 20ms.

$$NCF_{k,n}[d_0] = \frac{\sum_l X_{T,k,n}[l] X_{R,k,n}[l+d_0]}{\sqrt{\sum_l X^2_{T,k,n}[l] X^2_{R,k,n}[l]}} \quad (1)$$

Where d_0 is range from 0 to 55 μs. ITD_Atten is the value of d_0 giving this maximum of NCF.

$$ITD\_Atten[k,n] = \arg\max_{d_0} |NCF_{k,n}[d_0]|_{d_0=0}^{d_0=55} \quad (2)$$

ITD_Atten is measured in both the test and reference signals, and is denoted as ITD_Atten_{test} and ITD_Atten_{ref} in the next computation stage. The perceptual change of the source direction can be appropriately calculated as the Euclidian distance between two positions on a unit circle. The perceptual distance between two source directions due to the difference can be modeled as Formula4.

$$\Delta T = ITD\_Atten_{test}[k,n] - ITD\_Atten_{ref}[k,n] \quad (3)$$

$$ITDDist\_Atten[k,n] = \sqrt{2 - 2\cos\pi \cdot \frac{f_s}{N_{\max}} \cdot \Delta T} \quad (4)$$

Where f_s is the sampling rate and N_{max} is the maximum ITD_Atten represented in sample numbers.

ILD is an important cue for perception of sound direction of high-frequency sounds. ILD is calculated as the logarithm of the intensity ratio between the left ear input X_L and right ear input X_R from the time-frequency segments in the k_{th} frequency band of the n_{th} time frame.

$$ILD[k,n] = 10\log_{10}\left(\frac{\sum_l X^2_{T,k,n}[l]}{\sum_l X^2_{R,k,n}[l]}\right) \quad (5)$$

We can get ILD_Atten_{ref} and ILD_Atten_{test} when $ILD \leq 0.6dB$. The ILD_Atten distortion is calculated as:

$$\Delta L = ILD\_Atten_{test}[k,n] - ILD\_Atten_{ref}[k,n] \quad (6)$$

$$ILDDist\_Atten[k,n] = w_2[k] \times \log_{10}(\sum_l X^2_{T,k,n}[l] \times |\Delta L|) \quad (7)$$

Where ILD_Atten_{test} and ILD_Atten_{ref} are the ILD of the test and reference attention signals. $w_2[k]$ is a non-linear weighting factor, which mirrors the relative importance of the ILD_Atten distortion in each frequency band. By averaging over frequency bands and time frames, we get $ILDDist\_Atten$, which is a measure of perceptual distance between the ILD-based source direction of the test and reference signals.

The horizontal azimuth MOVs of attention audio is based on $ITDDist\_Atten$ and $ILDDist\_Atten$. According to $ILD \leq 0.6dB$, $ITD \leq 55\mu s$, we discriminate corresponding attention audio bark bands from background bark

bands $1 \leq Z \leq 109$. We choose $WinModDiff1\_Atten$, $TotalNMR\_Atten$, $AvgModDiff1\_Atten$ and $AvgMod-Diff2\_Atten$ for attention audio timbre quality. So we can extract 6 MOVs for attention audio.

2.2. MOVs for general audio

The MOVs parameters for general audio are classified as: general PEAQ MOVs, JDM parameter, EE parameter and horizontal azimuth MOVs. The horizontal azimuth MOVs contain: $ITDDist\_Total$, $ILDDist\_Total$ and $IACC\_Total$[9]. 16 MOVs are extracted for general audio, including 11 MOVs from traditional PEAQ.

3. PERFORMANCE COMPARISON

In this section, we compare the performance of the PEAQ metric with and without auditory attention related MOVs. Both methods add EE and JDM parameters for better assessment of mobile audio at mid to low bit-rates.

3.1. Experimental design

The experiments contain subjective tests and objective tests. Comparison is made based on the correlation coefficients that is obtained from subjective and objective test data in predicting the audio quality. The performance comparison is made between our previous developed JDM based approach [11] and the proposed auditory attention based one. Both approaches are trained and verified for the same set of test sequences. In the process of producing audio test sequences, we assumed that the audio source in azimuth $\theta \leq 6^{\circ}$, namely within the scope of $ILD \leq 0.6dB$, $ITD \leq 55\mu s$ are attention audio source. Here we use four typical scenes and select appropriate attention audio scripts for fusion in each scene. Four typical scenes are shown as Table1.

Table 1. Test sequences in the listening test

No.	Scenes	Attention audio
01	office	human voice
02	street	alarm or explosive sound
03	court	human voice
04	concert	music

The synthesized audio sequences are generated by mixing four specific background noises with corresponding attention audio sequences. There are seven different synthesized sequences for each typical scene. Fours are used for training and threes for validation.

The synthesized sequences are encoded by AMR-WB+ codec with 8 different bitrates[12]. In order to simplify the experiment, the error rates are not considered here. Hence we get 4*4*8=128 test sequences for training and 4*3*8 = 96 test sequences for verification. The sequences are 48 kHz sampling rate, and the average duration is 20 seconds.

3.2. Subjective tests

Our tests are based on ITU-R BS.1534 standard recommended "MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)" test platform, and get Subjective Difference Grade (SDG) of the attention audio.

Each test materials consist of 11 audio sequences (reference + 8 impaired + 1 hidden reference + 2 hidden anchors). Two more degraded hidden Anchor conditions, typically low-pass filtered the reference sequences at 3.5 kHz and low-pass filtered at 7.0 kHz.

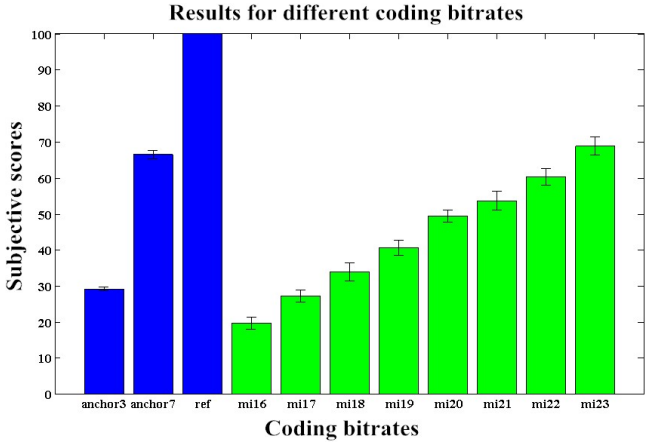

Fig. 2. SDG for different coding bitrates

The subjective test enrolled in 20 listeners. All testers are keen hearing and also well trained. Testers should take timber quality and horizontal position sense into account. Each tester should give corresponding scores to degraded audio sequences. Only testers correctly score hidden reference sequences with 100 points and score two hidden anchors within a limited range, the test results are regard as valid data. The SDG of hidden anchors, hidden references and test sequences are shown as Fig2. Mi16 to mi23 refer to the corresponding 8 different bitrates of AMR-WB+.

3.3. MOVs for attention audio

3.3.1. Objective model training

128 audio sequences are used in the Objective model training phase. We put 11 PEAQ MOVs, EE parameter, JDM parameter and 3 horizontal azimuth MOVs into neural network to match with SDG. Then the JDM based model is trained and built. The training Artificial Neural Network Model is the same as PEAQ. After that we add 6 MOVs for attention audio into neural network to build the auditory attention based model.

3.3.2. *Objective model verification*

96 audio sequences were fed into the two trained objective models. The scatter diagram between objective and subjective scores of the JDM based model is shown in Fig3. And the auditory attention based model is shown in Fig4.

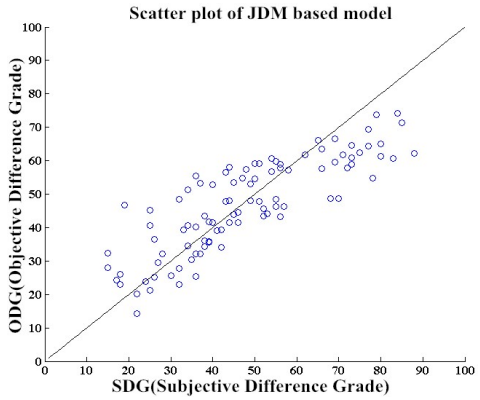

Fig. 3. Scatter plot of JDM based model

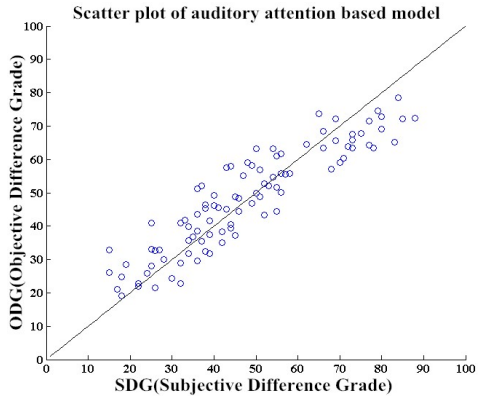

Fig. 4. Scatter plot of auditory attention based model

The scatter figures show that the auditory attention based model works much better. Table2 shows the differences of the correlation between objective and subjective scores, and the correlation is characterized by Pearson's correlation coefficient. The results show that the auditory attention based model outputs has higher correlation(r = 0.912) with the subjective scores.

Table 2. Correlations of different MOVs as input

Objective models	Correlation coefficients
JDM based models	0.831
Auditory attention based models	0.912

Fig5 shows the perceptual importance of each MOV when the neural network use these MOVs to predict ODG in the objective tests. MOVs for attention audio such as $AvgModDiff1B\_Atten$ and $AvgModDiff2B\_Atten$ rank 2nd and 4th in perceptual quality prediction. It reflects that adding auditory attention base MOVs to our model is reasonable and useful. Hence we can conclude that the performance of the auditory attention based model is more accurate, and more suitable for mobile audio codec evaluations with complex background noises.

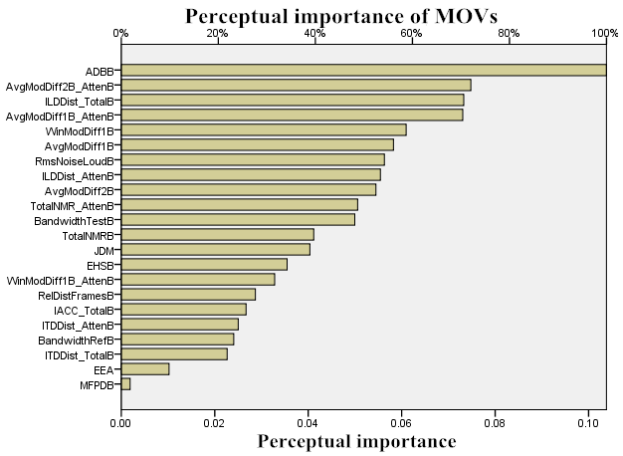

Fig. 5. perceptual importance of MOVs

4. CONCLUSION

We have developed a new objective model to evaluate audio quality by adding attention based horizontal azimuth parameters and timbre distortion parameters as additional Model Output Variables (MOVs) with our previously developed JDM based model. To both design and validate the proposed model, we collected human subjective test data using the MUSHRA method in four typical scenes. The predicted audio qualities show good correlation with subjective quality ratings for the applied test sequences. The performance of our proposed model is shown with high predictive accuracy of 91.2% with the subjective test results.

5. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (No.61102127, 61231015, 61272278, 61201340, 61201169), Major national science and technology special projects (2010ZX03004-003-03).

6. REFERENCES

[1] F. Ribeiro, D. Florencio, C. Zhang "crowdMOS: An approach for crowdsourcing mean opinion score studies", Proc. IEEE ICASSP, 2011, pp.2416 -2419

[2] B. S. A. A. M. R. Schroeder., ”Optimizing digital speech coders by exploiting masking properties of the human ear,” J. Acoust. Soc. Am.,1979, pp. 1647-1652.

[3] M. Karjalainen, ”A new auditory model for the evaluation of sound quality of audio-systems,” Proc. IEEE ICASSP,Vol. 10, 1985, pp. 608-611.

[4] K. Brandenburg, ”OCF-a new coding algorithm for high quality sound signals,” in Proc. IEEE ICASSP, 1987, pp. 141-144.

[5] K. Brandenburg and T. Sporer, ”NMR and Masking Flag: Evaluation of quality using perceptual criteria,” in Proc. of the 11th international AES Conference on audio test and measurement, 1992, pp. 169-179.

[6] U. G. T. Sporer, J. Herre and R. Kapust, ”Evaluating a measurement system,” in 95th AES-Convention, 1996, pp. 3704.

[7] T. Sporer, ”Evaluating small impairments with the mean opinion scale C reliable or just a guess,” in 101st AES Convention, 1996, pp. 4396.

[8] T. Sporer, ”Objective audio signal evaluation -. applied psychoacoustics for modeling the perceived quality of digital audio,” in 103rd AES Convention, 1997, pp. 4512.

[9] ITU ITU-R BS.1387-1, ”Method for objective measurements of perceived audio quality,” Geneva, Switzerland1998-2001 1998.

[10] Choisel S and Wickelmaier F., ”Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference,” J. Acoust. Soc. Am. 121, Issue 1,2007,pp. 388-400

[11] Dermot Campbell, Edward Jones and Martin Glavin, ”Audio Quality Assessment Techniques C A Review, and Recent Developments, Signal Processing,” Vol.89, Issue 8, 2009

[12] Yang Yuhong, Yu Hongjiang and Hu Ruimin, ”A new mobile audio quality assessment using Jitter Distortion Measure approach,” , Fifth International Workshop on IEEE QoMEX, 2013, pp. 182-187.