Appl. Math. Inf. Sci. 6 No. 2S pp. 397S-402S (2012)

Applied Mathematics & Information Sciences
An International Journal

Spatial Localization Evaluation Model for Parametric Stereo Audio

Shingchern D. You[1] **and Fan-Yu Cheng**[2]

[1] Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106, Taiwan
[2] American Megatrends, Inc., Taipei 104, Taiwan
[1]*Corresponding author: Email: scyou@ntut.edu.tw*

Received Aug 15, 2011; Revised Dec 16, 2011; Accepted xxxx, 2011
Published online: xxxx 2012

Abstract: Version 2 of high-efficiency (HE) advanced audio coding (AAC) introduced in MPEG-4 audio (called HE AAC v2 or eAAC+) has received wide attention recently. One of the key tools in HE AAC v2 is parametric stereo (PS) tool, whose decoded stereo audio sometimes exhibits rapid shifts of the spatial locations of instruments. This situation could be avoided if the encoder is properly designed. Inspired by the concept of perceptual evaluation of audio quality (PEAQ), we propose a model to quantitatively evaluate the quality of a PS encoder in terms of the spatial accuracy of the reproduced sound field. Inside the model, the input signal is analyzed by 32 subband filters. The inter-channel time difference (ITD) and inter-cannel intensity difference (IID) of low-frequency subbands are calculated. The ITD and IID values from the original signal are compared with those from the decoded signal to calculate the overall scores. The experimental results show that the model is able to detect localization discrepancy, and the provided scores are correlated with the preference of audiences. Though the resulting scores are not proportional to that from human listeners, the model nevertheless provides a useful tool for evaluating the perceived spatial quality of the decoded audio.

Keywords: Spatial Localization, HE-AAC v2, Parametric Stereo, ITD, IID

1 Introduction

After the great success of the MPEG-1 (moving picture experts group) audio [1], commonly known as MP-3, many new schemes for audio coding have been developed with more advanced techniques for higher coded quality and/or lower bitrates. Among these coding schemes, the version 2 (v2) of high efficiency (HE) profile of advanced audio coding (AAC) in MPEG-4 audio [2], commonly known as MPEG-4 HE AAC v2 or eAAC+, has received wide attention due to its higher coded quality in lower bitrates. Currently, many new services have adopted HE AAC v2 as the audio coding standard, such as digital radio mondiale (DRM) [3], upgraded version of digital audio broadcasting (DAB+) [4], and digital video broadcasting–handheld (DVB-H) [5], to name just a few. With the increasing number of applications, HE AAC v2 coding is expected to be more popular in the near future.

The core of the HE AAC v2 encoder is the low-complexity (LC) AAC coder, which is a perceptual audio coder [6] for waveform coding with a built-in psychoacoustic model. Since AAC LC is unable to provide satisfactory coded quality at low bitrates, additional tools are introduced. The first tool is spectral band replication (SBR) [2]. It is used to replicate high-frequency components of the reproduced audio by transposing up the low-frequency components of the same block during decoding. Therefore, instead of encoding high-frequency components, only necessary parameters are encoded, such as the energy of a given frequency band. Consequently, SBR uses around 2 to 4 kbps (kilo-bits per second) to encode the parameters. Combining AAC LC and SBR is called HE AAC or eAAC. To further reduce the bitrate for stereo audio, parametric stereo (PS) tool [7] is introduced. During encoding, PS converts a stereo input into mono, and then uses HE AAC to encode the mono source. To reproduce stereo audio during decoding, inter-channel information is also calculated during encoding. The PS encoding scheme, therefore, works only for music in stereo form.

Obviously, it is not easy to reproduce two (stereo) channels with only one channel and side information. Therefore, the decoded music sometimes exhibits rapid shifts (or jumps) of spatial locations of instruments. For example, assume that the perceived spatial location of an instrument in the original audio is on the front right direction. After coded with PS scheme, the spatial location of the same instrument may suddenly jump to front left for a short period of time, and then jumps back to its original location. This annoying situation is easily noticed by an audience.

To aid the developers of PS encoders to address this problem, we propose a model to evaluate the accuracy of spatial localization in the decoded audio. The model is inspired by the concept of perceptual evaluation of audio quality (PEAQ) [8], which is designed to perform objective evaluation of the quality of coded audio. With the proposed model, the heavy burden of listening tests during the development toward a good PS encoder can be alleviated. In addition, this model may also be used to compare the perceived spatial accuracy among different PS encoders.

2 Spatial Hearing

The PS encoding is based on the knowledge of psychophysics of human sound localization [9], or spatial hearing. Suppose that a sound source is placed on front left of the audience. When the sound wave travels to the audience's ears, due to the distance difference, left ear hears the sound shortly before right ear does. The time difference is known as inter-aural time difference (ITD). In addition, since the sound wave is partially blocked by the head (called head shadowing) before reaching the right ear, the sound intensity to right ear is attenuated. The intensity difference between ears is known as inter-aural intensity difference (IID). For a single-frequency tone, the time difference can be converted to the phase difference. In such a case, ITD may be replaced by inter-aural phase difference (IPD).

In terms of spatial hearing, ITD plays a dominant role at lower frequencies because diffraction of the sound wave occurs at lower frequencies. For higher frequencies, IID is more important due to head shadowing. The border frequency is around 1.5 kHz, which is determined based on the diameter of a human head (around 0.2 meter) and the speed of sound (around 340 m/s).

With the knowledge of spatial hearing, it is possible to synthesize a sound field with the virtual sound source at any spatial location. To do so, we need to know the transfer functions (or impulse responses) relating the sound source and the ears. The transfer functions are called head-related transfer functions (HRTFs) and the impulse responses as head-related impulse responses (HRIRs). Since it is more difficult to derive a theoretical model for HRTFs, they are usually obtained by measurement. By placing microphones in ears of a dummy or a real person, the measured HRTFs implicitly include the factors of IID, ITD, the shape of the head, and so on. The concept has been realized by various research groups, such as MIT's Media Lab [10]. Typically, the measured HRTF dataset contains many zero-only transfer functions for different elevation and azimuth angles.

With the HRTF dataset, it is not difficult to generate binaural signals for headphones from a mono source. Suppose that $h_{\alpha,L}[n]$ and $h_{\alpha,R}[n]$ represent the HRIRs relating left and right ears and a sound source at α° (azimuth). If the mono source $x[n]$ is to be perceived as if it were placed at α° azimuth, we may use the following operations:

$$\begin{aligned} y_L[n] &= x[n] * h_{\alpha,L} \\ y_R[n] &= x[n] * h_{\alpha,R}, \end{aligned} \tag{1}$$

where "*" is the convolution operator and $y_L[n]$ and $y_R[n]$ are signals to drive the left and right transducers of the headphones. In our experiments, we use HRTFs to generate audio pieces with pre-determined spatial (azimuth) angles to test our model.

3 Encoding Flow of Parametric Stereo

Figure 1 is the block diagram of the PS encoding flow. Each channel of input samples is individually decomposed into subband samples using quadratic mirror filters (QMF). The subband samples of both channels are mixed into mono for AAC coder to encode. Samples from the same indexed subband in both channels pass through a parameter-extraction block to calculate inter-channel intensity difference (also abbreviated as IID as it is related to inter-aural intensity difference), inter-channel phase difference (IPD), overall phase difference (OPD), and inter-channel coherence (ICC). These values are encoded and packed in the bitstream. During decoding, the decoder uses all-pass filters to generate another channel from the mono channel. By properly mixing the levels of the mono and the generated channels based on the side information, stereo channels are reproduced.

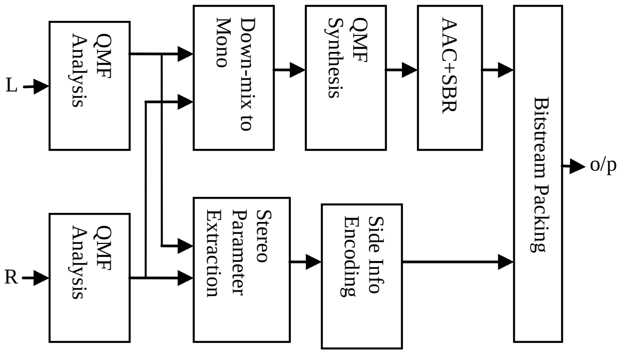

Figure 1: Block diagram of parametric stereo encoding flow.

4 Proposed Model

The proposed model, given in Figure 2, includes subband analysis filters, IID extraction, ITD extraction, and scoring blocks. These blocks are described in the following.

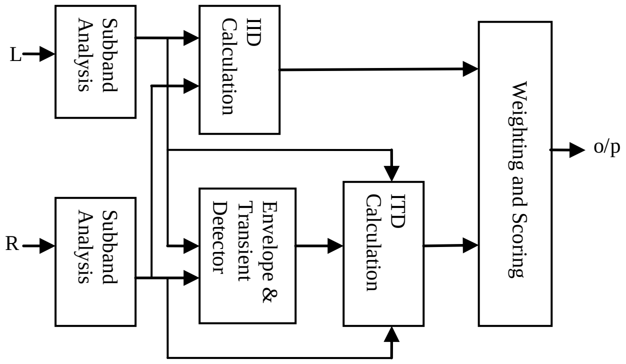

Figure 2: Block diagram of the proposed model.

Subband Analysis Filters The first stage of the proposed model is subband analysis filters to decompose input samples of each channel into 32-subband samples. The impulse responses of the subband filters are

$$h_i[n]=h_p[n]\cdot\cos\left[\frac{(2i+1)(n-16)\pi}{64}\right],\ 0\le i\le 31,\ 0\le n\le 511 \quad (2)$$

where $h_p[n]$ is the prototype low-pass filter used in MPEG-1 audio. Therefore, the analysis filter in (2) is essentially equivalent to that used in MPEG-1 audio except that no decimation is performed after subband analysis. Without decimation, the subband samples are oversampled and, thus, no aliasing distortion. If the sample rate of the input samples is 44.1 ks/s, the effective bandwidth (BW) of each subband is around 690 Hz. In contrast, the PS encoder uses QMF filters with much smaller BW (around 30 Hz) at low frequencies. For PS encoding, ITD is replaced by IPD, therefore higher frequency resolution (or smaller BW) is necessary at low frequencies. In our case, if a transient is present, we shall use the transient to calculate ITD. Therefore, to partially preserve the transient waveform, the BW of low-frequency subbands should not be too small.

Calculation of IID After the subband analysis filter, subband samples are divided into blocks, and ITD and IID are calculated for each subband block. In our model only subbands with indices in $0\le k\le 7$ are used. That is because subband samples with frequencies higher than 4 kHz mainly consist of harmonics, which are not used in the score calculation. In the following, we shall drop subband index k for brevity. In our model, the length of a block is 1152 (subband) samples, the same number of samples as in an MP-3 frame. The IID of the block is calculated as the ratio of the energy between the two channels by

$$\mathrm{IID}[i]=\frac{\sum_{n=0}^{1151} x_L^2[i\cdot 1152+n]}{\sum_{n=0}^{1151} x_R^2[i\cdot 1152+n]}, \quad (3)$$

where $x_L[n]$ and $x_R[n]$ are subband samples from left and right channels, and i is the block index.

Calculation of ITD It is known that, when compared with a stationary sound, human beings can identify the spatial location of a transient sound more easily. Therefore, the proposed model performs transient detection before actually calculating ITD. Depending on whether transients are present, different calculation methods are adopted.

The transient detection consists of envelope detector and transient detector operating on subbands. For envelop detector, its output $e[n]$ is obtained as

$$e[n]=\max(|x[n]|, 0.9995\cdot e[n-1]) \quad (4)$$

where $x[n]$ is the input to the detector and 0.9995 is experimentally determined.

Following envelope detector is the transient detector. Similar to that recommended in [11], the proposed envelope detector consists of the following four steps:

i. Find the standard deviation $\sigma[i]$ for block i of $e[n]$ with $1152\cdot i\le n<1152\cdot(i+1)$.

ii. Calculate a threshold $\gamma[i]$ for the current block (block i) by

$$\gamma[i]=0.66\cdot\gamma[i-1]+0.34\cdot\sigma[i] \quad (5)$$

iii. For j from 1 to 3, find candidates of transients whose values $v[n]$ are determined by the following iteration

$$v[n] \leftarrow v[n] + \frac{\delta_j}{\gamma[i]}, \text{if } \delta_j > \gamma[i], \qquad (6)$$

where $\delta_j = x[n+j] - x[n-j] + \delta_{j-1}, 1 \le j \le 3$ with the initial condition of $\delta_0 = 0$.

iv. If $v[n] < 0.9 \cdot v[n-1]$, the position of transient $p[i]$ is obtained by $p[i] = n$. On the other hand, if $v[n] < 0.9 \cdot v[n-1]$ does not hold for the current block, then no transient is present.

If a transient is present in the current block, we calculate the cross-correlation $c[i,m]$ between left and right channels in a range of 61 samples with $c[i,m]$ calculated as

$$c[i,m] = \sum_{n=-30}^{30} x_L[p[i]+n+m] \cdot x_R[p[i]+n+m], \quad (7)$$

with $-30 \le m \le 30$. If the value of $c[i,0]$ is greater than a threshold of

$$0.3 \cdot [(\sum_{n=-30}^{30} x_L[p[i]+n])^2 \cdot (\sum_{n=-30}^{30} x_R[p[i]+n])^2]^{1/2}, \qquad (8)$$

the ITD of block i is obtained by finding the argument of the maximum of $c[i,m]$ over m, i.e., $\text{ITD}[i] = \arg(\max_m c[i,m])$. On the other hand, if no transient is detected or the value of $c[i,0]$ is too small, we perform the cross-correlation on the entire block, i.e.,

$$d[m] = \frac{\sum_{n=0}^{1151} x_L[n+m] \cdot x_R[n+m]}{\sqrt{\sum_{n=0}^{1151} x_L[n+m]^2 \cdot \sum_{n=0}^{1151} x_R[n+m]^2}}, \qquad (9)$$

In this case, $\text{ITD}[i] = \arg(\max_m d[i,m])$ is the ITD of block i.

Calculation of Overall Score Before summing the individual IID and ITD values, we need to multiply these values with frequency-dependent weighting functions. Since ITD is more important at low frequencies, we use the following weighting function to reflect this fact:

$$w_{ITD}[k] = (32 - 2 \cdot k)^3 / 32^3, 0 \le k \le 7, \qquad (10)$$

where k is the subband index. Similarly, the weighting function for IID is calculated as

$$w_{IID}[k] = (32 - 2 \cdot 7)^3 / (32 - 2 \cdot k)^3, 0 \le k \le 7, \qquad (11)$$

With the weighting functions, the ITD and IID scores are weighted sums of the ITD[i] and IID[i].

To use the proposed model, the reference music (file) should be presented for analysis first. The model records the IID and ITD values of each block during analysis. The PS-decoded file is then presented to analyze the IID and ITD values of the coded file. For every subband block, the model calculates the differences of IID and ITD between the original and the coded files. The final scores are the sums of the absolute differences of IID and ITD. In our model, higher score means larger difference, or relatively poor localization accuracy. As the PEAQ model, the proposed model requires that both the reference and the coded file have the same duration for proper evaluation.

5 Experiments and Results

Three experiments are conducted to evaluate the proposed model. The experimental procedures and results are given below.

Experiment with Entire Music File Steered with an Azimuth Angle The first experiment uses ten mono music files (data1 to data10) as references. The references are altered using (1) for the entire files to get test files. The HRTF dataset is from MIT's Media Lab [10]. The azimuth angles used in the experiments are 15, 30, 45, 315, 330, and 345 degrees. Since a mono reference has an azimuth angle of zero, a larger angle deviation should produce higher scores. The results for ITD and IID are given in Figures 3 and 4, respectively. It is observed that the scores are proportional to the deviated angles, as expected. Note that angles of 45 and 315 degrees have the same deviation, therefore the same scores. Figure 3 also reveals that ITD scores are different for different pieces of music at the same azimuth angle. Therefore, the scores are comparable only for the same piece of music, but not for different pieces of music.

Experiment with a Small Fraction of a Music File Steered with an Azimuth Angle The second experiment alters the spatial angles of the test files for only a small period of time (500 ms). This experiment is to test if the proposed model is able to detect a short-term change of the spatial angle in a file. The ITD and IID results confirm that the proposed model is able to detect a short-term change as expected. The results are given in Figures 5 and 6. Since the scores of the proposed model are calculated based on the accumulation of the IID and ITD differences, the scores calculated in this experiment are very small as only a small fraction of the music is altered with an azimuth angle. Therefore, the proposed model usually requires a test file with sufficient duration to provide reliable scores. This situation is similar to that of the PEAQ model, which also requires a test piece having sufficient duration, such as 100 seconds, for proper scoring [12].

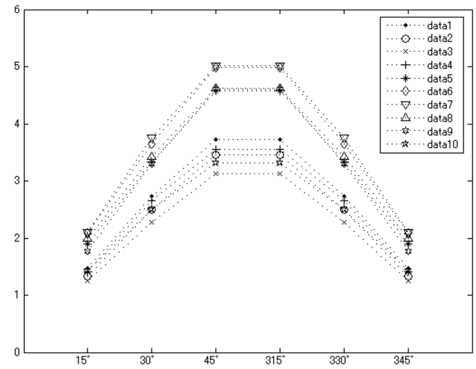

Figure 3: ITD scores for the first experiment.

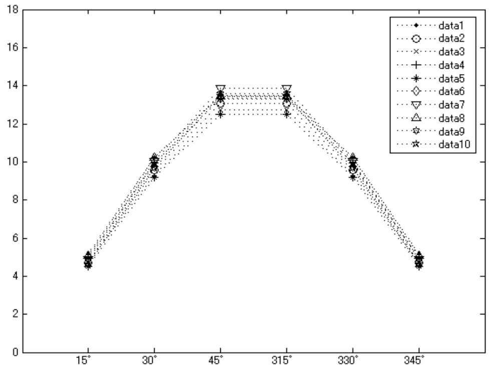

Figure 4: IID scores for the first experiment.

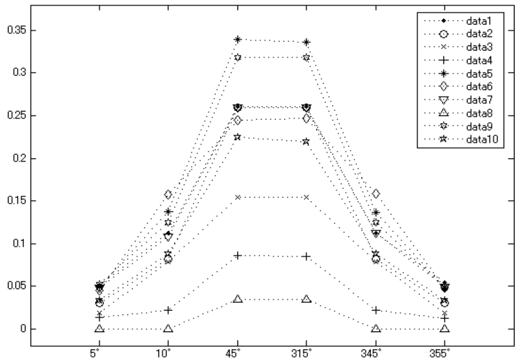

Figure 5: ITD scores for the second experiment.

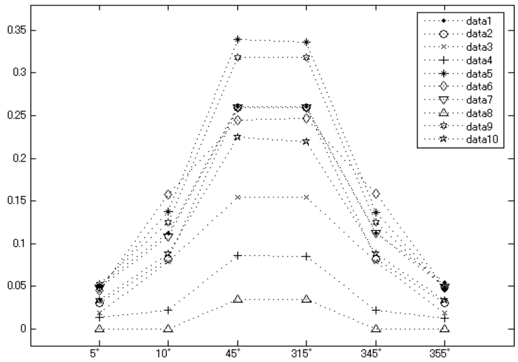

Figure 6: IID scores for the second experiment. Note that Y-axis is in log scale due to relatively small values in the figure.

Experiment with PS-Coded Music The third experiment uses the PS-coded music as the test files. The experiment includes two parts. The first part compares the scores obtained from PS-coded and non-PS-coded files to see if PS-coded files receive higher scores. The second part is a listening test to correlate the human listening preference and the scores from the model.

For the first part of the third experiment, a dataset of 531 stereo soundtracks are individually encoded with 24 kbps using HE AAC v2 profile and 44 kbps using HE AAC (v1) profile. Using different bitrates in different profiles ensures both have similar coding (waveform) distortion. The IID and ITD scores for both types of coded files are obtained with the original music as the reference. For each soundtrack, we calculate the IID and ITD differences between files coded in v2 and v1 profiles. Then, the obtained ITD and IID differences are normalized by the average score differences. The results are shown in Table 1 for eight pieces of music with different genres. The reason that we present differences of IID and ITD in Table 1 is because ITD scores are content dependent, and cannot be directly compared. We hope that the normalized score differences are comparable. In the table, a positive value means that the model judges that a file coded with v1 profile has a better quality in terms of sound localization. The results show that the model judges that files coded with v1 are better.

The listening experiment uses a simplified comparison category rating (CCR) method to obtain comparison mean opinion scores (CMOS) [13]. In this experiment, seventeen grad students are asked to give scores based on spatial localization after listening to music arranged in R/A/B format, where R is the original signal, and A and B are files coded with v1 (44 kbps) and v2 (24 kbps) profiles. The sequence of A and B are random and the audience does not know the sequence. The average scores are also given in Table 1. Again, if a v1 file is better, the score is positive (up to +2). Based on Table 1, we confirm that some PS files (such as sound no. 4 and 6) are so worse that even non-expert listeners notice the difference in spatial accuracy and stability. When comparing the scores from the model and CMOS from listeners, we notice that the listeners give song no. 1 and 7 negative values and the model also gives negative values in either ITD or IID score. Therefore, generally speaking, the model has a reasonable correlation with the preference of human listeners. However, we are unable to obtain a linear relationship between ITD and IID differences and CMOS from listeners. Therefore, a song with higher scores may not be worse than another one with lower scores. To provide reliable "absolute" scores, a better score conversion is required. Nevertheless, the proposed model is a useful tool for measuring relative spatial accuracy of the same source coded with different encoders.

Table 1: Results for experiment three. A non-PS file is better if the score is positive.

Song index	ITD diff.	IID diff.	Average CMOS score
1	-0.062	0.128	-0.12
2	1.100	2.205	0.00
3	0.329	3.025	0.00
4	0.804	1.350	0.65
5	0.072	0.181	0.35
6	0.248	0.144	0.59
7	0.315	-0.017	-0.12
8	0.983	1.038	-0.06

6 Conclusions

We propose a model for evaluating spatial accuracy of PS-encoded audio in this paper. The model uses subband analysis filters to decompose input signal into subbands. Then, the weighted ITD and IID scores are calculated for each subband block. The experimental results show that the proposed model successfully reports scores proportional to the shifts of spatial angles, even for a very short period. As expected, the proposed model also gives higher scores, or equivalently worse quality, for music files coded with PS than ones coded without PS. The IID and ITD scores given by the proposed model currently shows some correlation with CMOS given by listeners, though these scores are not in proportion. Therefore, the scores from different music files cannot be compared. Overall, the proposed model is a useful tool for developing and evaluating PS encoders in terms of the accuracy of the spatial localization.

References

[1] ISO/IEC, Information technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s - Part 3: Audio, IS 11172-3 (1993).

[2] ISO/IEC, Information technology - 14496-3; Coding of Audio-Visual Objects - Part 3: Audio, IS 14496-3 (2009).

[3] EBU, Digital Radio Mondiale (DRM); System Specification, ES 201 980, v3.1.1 (2009).

[4] EBU, Digital Audio Broadcasting (DAB); Transport of Advanced Audio Coding (AAC) Audio, ETSI TS 105 563, v. 1.1.1 (2007).

[5] EBU, Digital Video Broadcasting (DVB); Transmission System for Handheld Terminals (DVB-H), ETSI TS 302 304, v.1.1.1 (2004).

[6] T. Painter and A. Spanias, Perceptual coding of digital audio. Proc. IEEE, Vol. 88, No. 4, (2000), 451 - 513.

[7] J. Breebaart, S van de Par, A. Kohlrausch, and E. Schuijers, Parametric Coding of Stereo Audio. EURASIP J. Applied Signal Process., Vol. 9, (2005), 1305-1322.

[8] ITU-R, Method for Objective Measurements of Perceived Audio Quality, Recommendation ITU-R BS.1387 (1998).

[9] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, Revised Edition, MIT Press, Cambridge, MA (1997).

[10] W. G. Gardner and K. D. Martin, HRTF Measurements of a KERMAR Dummy-Head Microphone, MIT Media Lab (1994).

[11] Measured HRTF dataset is available at http://sound.media.mit.edu/resources/KEMAR.html

[12] 3GPP, General Audio Codec Audio Processing Functions; Enhanced aacPlus General Audio codec; Enhanced aacPlus Encoder Spectral Band Replication (SBR) Part, TS 126 404, v8.0.0 (2008).

[13] S. D. You and W.-K. Chen, Efficient Quantization Algorithm for Real-time MP-3 Encoders. Multimedia Tools and Applications, Vol. 40, No. 3, (2008), 341 – 359.

[14] ITU-T, Method for Subjective Determination of Transmission Quality, Recommendation ITU-T P.800 (1996).

Shingchern D. You received the Ph.D. degree in Electrical Engineering from the University of California, Davis, USA in 1993. Currently he is Associate professor in Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan. Dr. You's research interests include audio signal processing, audio identification, and applications of signal processing.

Fan-Yu Cheng received the M.S. degree in Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan in 2008. Mr. Cheng currently works as an engineer for

American Megatrends, Inc. in Taiwan.