Evaluation of Auditory Saliency Model Based on Saliency Map

Xiansong Xiong
School of Information and Communication Engineering
Communication University of China
Beijing, China
xxs0820@163.com

Zhijun Zhao
School of Information and Communication Engineering
Communication University of China
Beijing, China
zhaozhijun@cuc.edu.cn

Lingyun Xie
State Key Laboratory of Media Convergence and Communication
Communication University of China
Beijing, China
xiely@cuc.edu.cn

Abstract—For the bottom-up auditory attention process, many auditory attention models have been proposed, including the earliest four auditory saliency models developed from visual saliency models, namely Kayser model, Kalinli model, Duangudom model and Kaya model. In order to compare the correlation between the output results of the four models and subjective perception, firstly the four models were evaluated by carrying out a subjective saliency evaluation experiment in this paper. In the subjective evaluation experiment, 20 kinds of sound scene materials were scored with relative saliency and absolute saliency, and two rankings were obtained. Secondly in the saliency model, the saliency scores were calculated for the same 20 kinds of sounds, and the saliency of the sounds were scored by extracting the mean, peak, variance and dynamic characteristics of the saliency score of each sound, and then correlations were calculated between model saliency scores and two subjective scores. The conclusion was that Kalinli model had the best effect among the four models and had the highest correlation with subjective perception; among the four features of the saliency score, the variance had the highest correlation with subjective perception. The main reason for the better results of Kalinli model was that the method of extracting auditory spectrograms and features was more consistent with the auditory characteristics of human ear and the extracted features were more comprehensive. By analyzing the structure and perceptual features of the models with high correlation between model output and subjective perception, we can improve the models in the future based on the conclusions drawn, so as to enhance their performance and make them more consistent with the auditory characteristics of the human ear.

Keywords—Auditory attention, Subjective evaluation, Auditory saliency model

I. INTRODUCTION

In daily life, sounds rarely appear in isolation. A wide variety of sounds are constantly entering our ears, yet we can identify the most interesting sound while ignoring irrelevant background and distracting sounds. For example, at a cocktail party, we can hear a specific person speaking even though the environment is noisy. To accomplish this process, the listener engages in a process of auditory attention, by focusing our limited cognitive resources and the brain's limited computational resources on the sounds that interest us most.

Auditory attention is not a single process, but is regulated by both "top-down" goal-driven and "bottom-up" stimulus-driven modalities [1]. "Top-down" auditory attention is a conscious process regulated by specific goals, expectations, or habits. "Bottom-up" auditory attention is an unconscious process that directs our attention to the most salient sound in the scene through sensory cues.

Auditory saliency can be defined by bottom-up auditory attention, defining salient sounds as those that do not require focused attention to be noticed. Auditory saliency can help us to classify information in complex auditory scenes, and since computational resources are limited, we must identify the most important sounds in the scene for processing.

Based on the bottom-up auditory attention pattern, a bottom-up auditory saliency model can be developed. In vision field, bottom-up visual attention was understood through saliency maps [2]. Based on the behavior and neural structure of the early primate visual system, a visual attention system was proposed. Firstly, different features of the image were extracted and multi-scale feature maps were generated, such as color, luminance, orientation and other features. Next, the different feature maps were combined together to obtain the final visual saliency map. In the auditory field, a similar approach had been proposed, which considers the time-frequency map of audio signal as "auditory image", ran a visual-like saliency mechanism on it, and finally obtained the auditory saliency map.

Kayser proposed the first auditory saliency model by using feature maps to obtain auditory saliency maps [3]. On this basis, Kalinli [4], Duangudom [5], and Kaya [6] proposed different auditory saliency models respectively. In order to compare the effects of these models, i.e., to compare the correlation between the results obtained by different models and the subjective saliency perception, the results obtained by subjective experiments were compared with those obtained by the models, and the better performing models are obtained and analyzed in this paper.

The structure of this paper is as follows: the first part introduces the auditory saliency models; the second part introduces the basic principles of the four auditory saliency models; the third part introduces the subjective experiments and

978-1-6654-6248-8/22/$31.00 ©2022 IEEE
DOI 10.1109/CoST57098.2022.00061

the experimental results; the fourth part compares the subjective experimental results and the model results, draws conclusions, and analyzes the reasons; and the fifth part concludes.

II. Introduction to Auditory Saliency Model

A. Introduction to Kayser Model

Kayser model describes the saliency of a sound segment through a saliency map. Firstly, multi-scale auditory feature maps are extracted from the auditory spectrogram based on the processing stages of the central auditory system, which are intensity features, temporal contrast features, and frequency contrast features. In the second step, based on the feature maps, the center-surround difference processing is performed to highlight the salient parts, and then the feature maps are normalized and the feature maps of different scales are summed across scales. Finally, the feature maps corresponding to each feature are linearly summed to obtain the final saliency map, and the framework of the model is shown in Fig. 1 [3].

By validating the model in experiments with natural acoustic scenarios, it was demonstrated that it had reproduced human judgments of auditory saliency and predicted salient sounds embedded in noisy backgrounds. In addition, the model uses the structure of the visual attention model for the first time in a computational model of auditory attention, and subsequent computational models of auditory attention begin to evolve based on this model.

B. Introduction to Kalinli Model

Based on Kayser model, Kalinli proposed an auditory saliency model. The basic framework of Kalinli model is similar to that of Kayser model. Firstly, the feature map is extracted from the auditory spectrogram, followed by center-surround difference processing, normalization, cross-scale summation, and linear combination. However, the differences include the method of calculating the auditory spectrogram, the type of extracted features, and the normalization method.

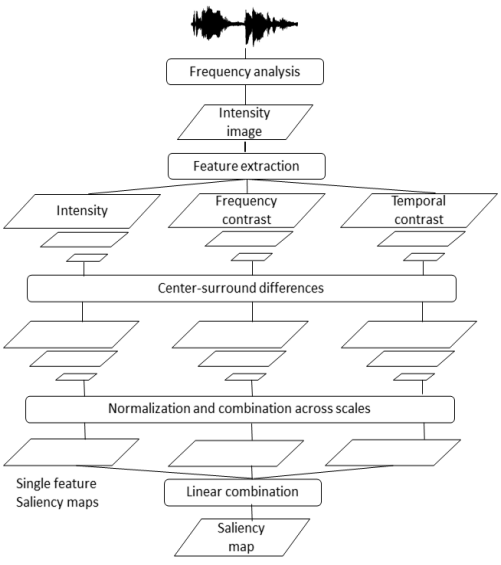

Fig. 1. Kayser Auditory Saliency Model Framework

Kalinli model addresses the problems of Kayser model in which the method of calculating auditory spectrograms does not sufficiently match the human auditory system, the lack of comprehensiveness of the extracted features, and the defects of the normalization method. A novel and biologically plausible auditory saliency map is proposed to simulate saliency-based auditory attention. Kalinli model uses an early auditory system to calculate auditory spectrograms, which includes cochlear filters, inner hair cells (IHC) and a lateral inhibition phase that simulates the process of the human auditory system from the basilar membrane to the cochlear nucleus. The extracted features add orientation features, as well as pitch features, to more fully reflect human perception of saliency. Iterative nonlinear normalization is used in the normalization process, which can better highlight the salient locations.

C. Introduction to Duangudom Model

Compared with Kayser model, Duangudom model has the same basic framework of extracting feature maps from the auditory spectrogram, normalizing them and then linearly combining them to obtain the saliency map, with the differences including the method of extracting features and the absence of center-surround difference processing. And the method of calculating the auditory spectrogram is the same as that of Kalinli model.

Duangudom model incorporates a more biologically meaningful mechanism that simulates the processing of the peripheral central auditory system. Feature mappings generated by spectral-temporal receptive field (STRF) simulate the neural responses of the mammalian primary auditory cortex (A1), effectively replacing the parallel feature mappings of the previous model [7]. A1 can be considered as a set of modulation filters [8], with selectivity as well as variability for time response, responding fast to some fast-changing spectra and slower to slow-changing ones. Using this property, the spectral-temporal modulation rate of the auditory spectrogram can be analyzed, thus reflecting the saliency of the sound in different dimensions. Using different spectral-temporal modulation filters, the extracted features include global energy features, temporal modulation features, spectral modulation features, and high temporal-spectral modulation features. These features are then inhibited and linearly combined to obtain the final saliency map.

D. Introduction to Kaya Model

The basic framework of Kaya model is compared with Kayser model, both of which extract features first, followed by center-surround difference, normalization, cross-scale summation and linear combination, but the differences are the type and method of extracting features and the form of the final auditory saliency map obtained. And the method of calculating the auditory spectrogram and the normalization method are the same as Kalinli model.

The multi-scale features extracted by Kaya model include one-dimensional pitch features extracted directly from the audio signal as well as envelope features, two-dimensional temporal modulation features and spectral modulation features extracted from the auditory spectrogram using STRF, and spectral features. Next, the extracted multiscale features are normalized, summed across scales, and linearly combined. In the process of

linear combination, since Kaya model contains two one-dimensional features, the two-dimensional feature saliency map needs to be averaged in the second dimension to obtain a one-dimensional curve over time. Then the one-dimensional feature saliency maps are combined to obtain the final auditory saliency map.

In order to evaluate the effectiveness of the four saliency models, i.e., to compare the correlation between the results of the four models in different types of sounds and the results of subjective human perception, a subsequent subjective evaluation experiment of auditory saliency is required.

III. Subjective Evaluation Experiment of Auditory Saliency

In order to evaluate the sound saliency more comprehensively, the experiment was divided into two parts: the first part calculated the relative saliency of the sound; the second part calculated the absolute saliency of the sound, and the two types of saliency corresponded to the two experimental methods.

In the first part of the experiment investigating auditory relative saliency, the pairwise comparison method was chosen, combining each material in pairs, comparing the saliency differences between them [9], and scoring the saliency differences. In the second part of the experiment investigating auditory absolute saliency, the series category method was chosen to score the absolute saliency of each of the 20 sound materials. The scoring scale was from 1 to 9, with higher scores indicating greater salience.

A. Experimental Signals

20 types of scene sound materials were randomly selected from the material library containing different types of sounds, including 5 types of animal sounds, 4 types of mechanical sounds, 8 types of natural sounds and 3 types of human sounds. It was represented by serial numbers S1 to S20.

Each scene sound was selected for 1.5 seconds, and in order to reduce the energy differences due to the different types of scene signals, each scene signal was superimposed in a background sound of 3 seconds. The background sound contained the average signal of 20 scene signals and Gaussian white noise [10], which allowed subjects to judge more salient sounds in the same environment.

In the first part of the experiment to study the relative salience of auditory, 20 experimental materials were grouped in pairs. Because there is a psychological sense of "predominance" in a group of signals, there is a distinction between the front and back order, and the structure of each group of 380 signals is shown in Fig. 2. In the second part of the experiment to study the absolute salience of auditory, 20 experimental signals were presented directly to the subjects.

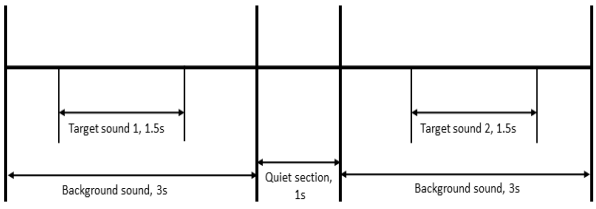

Fig. 2. Test Signal Structure in Pairs

B. Principles of Subject Selection

Since the hearing of the subjects is related to their age, e.g., as age increases, the ability of the auditory cells in the cochlea to acquire low-frequency as well as high-frequency signals decreases, and if they are exposed to a noisy environment for a long time, the decay of the auditory cells is also accelerated in a state of prolonged excitement. Therefore, the subjects should not be very old. The final subjects selected for this experiment were 30, 15 males and 15 females, all between the ages of 18 and 34, all with normal and healthy hearing, who gave informed consent to participate in the experiment. Before starting, each subject was asked if he or she had any auditory or musical training (and if so, how many years of training), and the subjects were informed of the requirements of the experiment and of the basic concepts of auditory saliency scores by means of experimental instructions prepared in advance by the experimenter.

C. Experimental Procedure

This experiment could only be done online because of the impact of the COVID-19, and the influence of environmental differences on the experimental results should be minimized, and the influence of visual stimuli should be reduced [11]. Subjects received sound stimuli through headphones on a PC, and the experiment was performed with monitor headphones with the best possible sound quality, while the experimental data were recorded in a table on the PC.

In the first part of the experiment, the 380 pairs of signals were divided into 20 groups of 19 pairs each, and subjects listened to the sound clips one by one after having read the instructions and listened to the examples in advance. Subjects selected the relatively salient signal in a pair of scenes by subjective perception, and then rated the relative salience of the difference between two scenes in each pair. A score of 1 indicates a very small difference, 2 indicates a small difference, 3 indicates an average difference, 4 indicates a large difference, and 5 indicates a very large difference.

In the second part of the experiment, the selected 20 sound materials were presented to the subjects independently, and subjects listened to the sound clips one by one after having read the instructions and listened to the examples in advance. The subjects first selected the most salient (9 points) and the least salient (1 point) clips. For the remaining 18 clips, the subjects rated the salience of each sound clip on a scale of 2 for very not salient, 3 for relatively not salient, 4 for somewhat not salient, 5 for neither salient nor not salient, 6 for somewhat salient, 7 for salient, and 8 for very salient.

D. Saliency Score Results and Reliability Analysis

In order to determine whether the subjective experiment was reliable, the data obtained from this experiment need to be tested for reliability. The reliability test used in this experiment was the internal consistency reliability, which means that the scores of different subjects were used to test the reliability of the overall experimental data.

The reliability coefficient of this experiment was calculated by using the variance or covariance of the subjects' responses to

the items to calculate the Cronbach's alpha coefficient [12]. In general, the value of alpha coefficient should be between 0 and 1, and the larger the value represents the more reliable the data in general, and in the practical research an alpha coefficient of 0.6 was required.

By calculating the Cronbach coefficients for each of the two parts of the experiment, the first part yielded an overall reliability coefficient of 0.992 for the 30 subjects, indicating that the reliability was good and the results could be retained in their entirety. In the second part, the overall reliability coefficient of the 30 subjects was only 0.560. In order to make the experimental data meet the due standard, the reliability coefficient was recalculated by screening the data and finally reached 0.615 by removing the data of the first three subjects, which met the standard.

Next, the saliency results were calculated, and in the process of calculating the first part of relative saliency, the data that were more affected by the chronological order were removed to make the results more reliable and then calculated, and normalized. The normalized relative saliency scores were obtained as shown in Table I. After normalizing the data for the second part of the absolute saliency scores, the results obtained were shown in Table II. Based on the obtained relative saliency scores and the absolute saliency scores, they can be used in the next step, to evaluate the auditory saliency model.

IV. Auditory Saliency Model Evaluation

The same 20 scenes of sound material from the above experiment were used as input to the four auditory saliency models for processing. First, the sound signals were divided into frames with a frame length of 20ms and a frame shift of 10ms, so that each frame was calculated as a corresponding auditory saliency map, and the saliency map of each frame was combined to obtain the auditory saliency map corresponding to this sound. By calculating the mean value in the second dimension of the auditory saliency map, a one-dimensional time-dependent saliency score was obtained for each sound. Since this score was a series of values that cannot be directly compared with the results obtained above, four features were extracted for each sound's saliency score, including mean value, peak value, variance, and dynamics. The dynamics referred to the maximum value of the saliency score minus the minimum value. In this way each sound was given four feature values under each model, and these four feature values were used for subsequent calculations.

TABLE I. Normalized Score of Relative Saliency

Experiment signals	***S10***	***S14***	***S11***	***S1***	***S3***
Saliency	1	0.965	0.947	0.863	0.776
Experiment signals	***S7***	***S6***	***S4***	***S2***	***S12***
Saliency	0.579	0.579	0.568	0.535	0.496
Experiment signals	***S17***	***S20***	***S5***	***S13***	***S18***
Saliency	0.455	0.446	0.414	0.268	0.169
Experiment signals	***S16***	***S9***	***S15***	***S8***	***S19***
Saliency	0.12	0.114	0.105	0.076	0

TABLE II. Normalized Score of Absolute Saliency

Experiment signals	***S10***	***S11***	***S1***	***S14***	***S3***
Saliency	0.82	0.81	0.76	0.71	0.63
Experiment signals	***S7***	***S20***	***S2***	***S6***	***S17***
Saliency	0.61	0.54	0.53	0.53	0.51
Experiment signals	***S4***	***S13***	***S5***	***S12***	***S18***
Saliency	0.5	0.49	0.45	0.45	0.42
Experiment signals	***S9***	***S15***	***S16***	***S8***	***S19***
Saliency	0.41	0.32	0.24	0.22	0.13

A. Correlation of Subjective Relative Saliency Scores with Model Saliency Scores

The four features extracted from each saliency score were compared with the subjective relative saliency scores obtained from Table I to calculate the Spearman's correlation coefficient as well as the Pearson's correlation coefficient, and the p-value was calculated to determine whether the results were reliable. Compared with the Pearson correlation coefficient, the Spearman correlation coefficient has no restriction on the data distribution and is also applicable to nonlinear data.

The calculated results were screened using the P-value, and the results with P-values less than 0.05 corresponding to the features were retained. Of the two correlation coefficients, if one P-value was less than 0.05, both results were retained for later analysis, and the results obtained were shown in Table III.

B. Correlation of subjective absolute saliency scores with model saliency scores

Using the same method as above, the four features extracted from each saliency score were compared with the subjective absolute saliency scores obtained in Table II to calculate the Spearman correlation coefficient as well as the Pearson correlation coefficient, and the P-values were calculated to determine whether the results were reliable. The results with P-values less than 0.05 were retained, and the results obtained by screening were shown in Table III.

C. Analysis of results

By calculating the correlation between the saliency scores of the models and the saliency scores of the two subjective in Table III, it can be seen that the models that satisfied the p-value screening condition were Kalinli model as well as Duangudom model. The correlation coefficient showed that there was a negative correlation between the saliency scores of Duangudom model and both subjective saliency scores, which showed that it was not effective. The variance of Kalinli model saliency scores had a Spearman correlation coefficient of 0.6 (P=0.006) and a Pearson correlation coefficient of 0.604 (P=0.005) with the subjective relative saliency scores, and a Spearman correlation coefficient of 0.633 (P=0.003) and a Pearson correlation coefficient of 0.63 (P=0.003) with the subjective absolute saliency. It can be seen that the variance feature of the saliency scores of Kalinli model was positively correlated with both saliency scores, and the small P-value indicated that the results were significant. Therefore, it can be concluded that Kalinli model had the best effect among the four models and the variance feature had the best effect among the four features.

By analyzing the computational process of Kalinli model, the following conclusions were obtained.

TABLE III. CORRELATION OF SELECTED MODEL SALIENCY SCORE WITH TWO SUBJECTIVE SALIENCY SCORE

	Correlation Between the Selected Model Saliency Score and the Subjective Relative Saliency Score					**Correlation Between the Selected Model Saliency Score and the Subjective Absolute Saliency Score**				
	Kalinli model	***Duangudom model***				***Kalinli model***	***Duangudom model***			
	Variance	***Mean value***	***Peak value***	***Variance***	***Dynamics***	***Variance***	***Mean value***	***Peak value***	***Variance***	***Dynamics***
Spielman correlation coefficient	0.600	-0.555	-0.690	-0.517	-0.690	0.633	-0.493	-0.564	-0.389	-0.564
P-values	0.006	0.012	0.001	0.021	0.001	0.003	0.027	0.010	0.090	0.010
Pearson correlation coefficient	0.604	-0.457	-0.395	-0.235	-0.395	0.630	-0.314	-0.354	-0.207	-0.354
P-values	0.005	0.043	0.084	0.319	0.084	0.003	0.178	0.125	-0.382	0.125

1) Compared with Kayser model: Firstly, Kalinli model includes cochlear filtering, inner hair cell processing and lateral inhibition stages in the computation of auditory spectrogram, which is more consistent with the human auditory system. Secondly, more features are extracted, including pitch features and orientation features. Pitch is a basic perception of sound, which is essential for people to appreciate the rhythm of words and the melody of music, as well as to organize the sound environment into different sources; the orientation filter that extracts orientation features simulates the dynamic response of auditory neurons to moving ripples, which can better simulate the response process of neurons.

2) Compared with Duangudom model: Duangudom model uses STRF to extract features, oversaturates the output at different STRF, does not represent the complete representation of the brain for sound [6], and does not have a center-surround difference process to better highlight salient locations.

3) Compared with Kaya model: The methods of extracting features in Kaya model include extracting one-dimensional features directly from waveform, and extracting features by using similar STRF in Duangudom model. In contrast, Kalinli model uses the Gabor filter, which is a filter formed by applying different stimuli to the primary auditory cortex of the cat and observing its spectral temporal receptive fields corresponding to different activation or inhibition under different stimuli, and by simulating this activation or inhibition, which is more consistent with physiological processes.

Among the four features extracted from Kalinli model, the variance feature worked best. Firstly, the variance feature can reflect the variation range and the distribution of the model saliency scores more comprehensively. Because the subjective saliency score of a sound was one value, and the model saliency scores was a series of values, the larger the variance of the series of values, the larger the range of fluctuation of the saliency score, i.e., there was a larger value, indicating that there was a more salient part of the sound, and thus the variance had a stronger correlation with the subjective saliency score. Secondly, the mean value feature, the dynamic feature and the peak value feature were affected by other factors such as noise in the sound material, so the variance feature results were more reliable and effective.

V. CONCLUSIONS

In this paper, four auditory saliency models based on visual development, namely Kayser model, Kalinli model, Duangudom model, and Kaya model, were used to evaluate the effectiveness of these four models by conducting subjective evaluation experiments. In the subjective experiments, relative saliency scores and absolute saliency scores were calculated for the 20 sound scene materials, and correlations were calculated using the two subjective saliency scores with the scores calculated by the saliency models. It was concluded that Kalinli model worked best among the four models. Through the conclusions obtained in the experiments and analysis, the model can be improved in the future work. In the process of feature extraction, the features with high correlation with subjective auditory saliency are selected as the features to be extracted; in the process of feature fusion, the weights when the features are summed are changed so that the features with high correlation with subjective auditory saliency have more weight. Through the above improvements, the model can be more consistent with the auditory characteristics of the human ear, thus enhancing the performance of the model and making it work better in practical applications. In this paper, only four auditory saliency models based on saliency maps were compared, and there are other types of auditory attention models, such as probabilistic statistical models and linear discriminant models. The application scenarios of auditory saliency models include environmental sound detection, sensitive sound detection, sound source localization, etc. Future work can be done by comparing auditory attention models based on different algorithms and analyzing the advantages of each and where they work better, so that the models can be better applied in different scenarios.

REFERENCES

[1] J. B. Fritz, M. Elhilali and S. V. David, et al, "Auditory attention - Focusing the searchlight on sound," Current Opinion in Neurobiology., vol. 17, no. 4, pp. 437-455, 2007.

[2] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, 1998.

[3] C. Kayser, C.I. Petkov and M. Lippert, et al, "Mechanisms for allocating auditory attention: an auditory saliency map," Current Biology, vol. 15, no. 21, pp. 1943-1947, 2005.

[4] O. Kalinli and S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," INTERSPEECH 2007, Antwerp, Belgium, 2007, pp. 1941-1944.

[5] V. Duangudom and D. V. Anderson, “Using auditory saliency to understand complex auditory scenes,” 2007 15th European Signal Processing Conference(EUSIPCO), Poznań, Poland, 2007, pp. 1206-1210.

[6] E. M. Kaya and M. Elhilali, “A temporal saliency map for modeling auditory attention,” 2012 46th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 2012, pp. 1-6.

[7] E. M. Kaya and M. Elhilali, “Modelling auditory attention,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 372, no. 1714, pp. 20160101, 2017.

[8] T. Chi, Y. Gao and M. C. Guyton, et al, “Spectro-temporal modulation transfer functions and speech intelligibility,” The Journal of the Acoustical Society of America, vol. 106, no. 5, pp. 2719-2732, 1999.

[9] Z. H. Meng, An experimental psychological approach to the subjective evaluation of sound quality. National Defense Industry Press., 2008.

[10] X. Zhang, X. Y. Xia and X. J. Wang, et al, “A saliency map extraction model for auditory attention,” Journal of Sichuan University (Natural Science Edition), vol. 29, no. 9, pp. 1142-1147, 2013.

[11] P. Yang and Z. H. Meng, “Visual-auditory mode: Relativity between Auditory Masking Effect and Attention of Visual-Auditory,” Elementary Electroacoustics, vol. 35, no. 2, pp. 42-44+48, 2011.

[12] Z. G. Xi and S. Y. Wang. “On negative Cronbach alpha coefficient and split-half reliability coefficient,” Journal of Chongqing University of Posts and Telecommunications(Natural Science), vol. 2007, no. 6, pp. 785-787, 2007