An Approach for TV Channel Recognition based on Audio

Lijin Long[1]

[1]Department of Electronic and Information Engineering, Zhejiang University of Media and Communications, Hangzhou, 310018, China

email: zj_education_ljl@163.com

Keywords: Audio Features; TV Channel Recognition; sequence alignment; Tolerance Processing; Pairwise Point Checking

Abstract. In order to amalgamate distributing platform of traditional TV program with mobile Internet on technology and business, it is necessary to study the method of TV channel recognition based on the collaboration between mobile terminals and publishing service of TV program. Due to some advantages based on audio features, such as less processing data volume, lower complexity of signal variation, better realtime performance and non-directional sampling, some processing steps, such as data denoising, data standardization processing, sequence alignment, tolerance processing and pairwise point checking, etc., are studied for TV channel recognition based on audio features. The experimental results show that the overall performance of the proposed approach is better than those based on frequency domain and based on time domain simply because it has some advantages such as reduction of both sampling difference and sampling environment interference for various of handheld mobile terminals, reduction of transmitting interference for content distributing server-side, excellent overall performance on accuracy and efficiency, etc. The proposed approach can also be applied into data pushing, user interactive discussion, realtime vote, etc.

Introduction

With the development of intelligent terminals, media storage and network application technology, by means of taking advantage of latest technology in network storage and data transmission, media content shows a tendency of explosive growth. Its consumption also experienced significant technical renovation from traditional media to new media and omnimedia. Various kinds of media distributing platforms, such as digital TV, IPTV, smart TV boxes developed by different great companies, have provided users on media consumption with rich media content based on network. Under the background of tri- networks integration, three - screen – syncretization , three – screen – interactivity , the generation and distribution of media content have entered into the new phase of diversity and diversified development[1][2].

For traditional TV distributing service platform, the current situation is the number of APP for traditional TV programs is less than other fields such as shopping, taxi, navigation, and so on. The limited interactive services restrain the media content consumption for TV distributing service platform. Compared with PC and other multimedia terminals, some digital TV stations have accomplish two-way transformation on network, however, the development of value-added services, such as video-on-demand, TV content playback(time-shift), premium channels, payment platform, etc. has fallen behind in network technology because of related supporting technology integrated with Internet and mobile Internet, which also limits function extension of digital TV system on operability, user interaction habits, content screening, etc. In this situation, it can’t satisfy user demands on efficient screening of both media and related content. By means of TV channel recognition technology, TV program distributing center can push some content related to distributing TV program to users, which can solve the problem of the interactivity between users and related content of program so as to implement the integration between TV program content distributing platform and convenient interactivity advantage existing in mobile terminals. Based on the technology, TV program distributing center can also implement some other activities for users, such as media content discussion, hot topic guidance, sharing information of TV program, etc. TV

program is composed of video and audio. Compared to video, audio has some advantages for TV channel recognition, such as less processing data volume, lower complexity of signal variation, better realtime performance and non-directional sampling[3][4]. To solve the challenge of TV channel recognition, the main difficulties existing in environment noises, sequence alignment, pairwise point search, realtime processing performance, and so on.

This article studies some methods about audio data, such as audio data sampling, audio data denoising, data standardization processing, sequence alignment, tolerance processing and pairwise point checking, etc., for the purpose of solving challenges in TV channel recognition algorithm. By comparison with other algorithms, the proposed approach has some superiority on robust, accuracy and realtime based on experimental results. With the development of amalgamation between TV program distributing platform and mobile Internet, TV channel recognition algorithm can be further applied into some value-added services, such as data pushing, interactive discussion, realtime vote, and so forth.

TV Channel Recognition Based on Audio

Procedure of TV Channel Recognition

TV audio signals are sampled by APP installed in mobile device with microphone. After denoised by RBF neutral network model, the sampled audio data(also named as audio row vectors or vectors in brief) are transferred to TV program distributing platform in realtime audio streaming way. After that, the proposed TV channel recognition algorithm process the transferred audio data so as to recognize which TV channel the transferred audio data belongs to. The processing procedure of TV channel recognition is shown in Fig.1. The goal of denoising is to minimize the influences on sequence alignment and pairwise point checking, such as volume variation, differences in various of sampling devices, interferences on sampling and transferring, etc. Before recognized by the proposed algorithm, the pre-processing should be used for sampling timestamp extraction, data standardization, parameter settings, and so forth.

According to extracted sampling timestamps and TV listings in database, some audio vectors can be extracted from TV program library, which is distributing for users, and generate data vector-set. In order to eliminate dimensional influences on sequence alignment and pairwise point checking, the data vector-set should be standardized by Z-score method. Based on data features in frequency domain, data vector-set can further reduce dimension quickly so as to improve on processing efficiency. For the reduced-dimension data vector-set by means of sequence alignment based on FFT, the data vector-set only remains the top k matched vectors(see Fig.2), which will be further used to pairwise point checking after they have been updated based on tolerance processing. After an iterative process has finished, the matching collection is updated and the proposed algorithm does not terminated until the data vector-set has all been processed or correlation coefficient between one vector and sampling data vector is greater than set threshold value in terms of iterative conditions. The algorithm flow chart is shown in Fig.2.

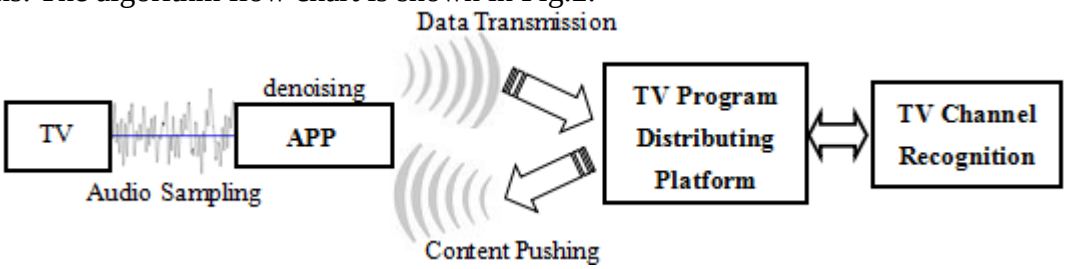

Fig.1. The procedure of TV channel recognition

As shown in Fig.1, sampling noises are denoised by APP and transmitting noises are denoised by TV program distributing platform. The former results from the differences in various of sampling devices and the influences by sampling environment. The latter results from transmitting interference. Both noises have an influence on the accuracy of sequence alignment and pairwise point checking. In order to minimize influences resulted from noises, the proposed approach creates a noise processing model and gets its parameters dynamically by training so as to optimize the

fluctuation in application environment self-adaptively. As shown in Fig.2, the pre-processing for audio data mainly includes sampling timestamp extraction, data standardization, parameter setting, etc. The extracted timestamps is a key factor for extracting data vectors from TV program audio library. Based on timestamps, TV program audio data can be cut according to audio sampling length(see 2.3). The constraint condition for this processing step is the length to cut should be greater than sampled audio data length(see Eq.8). In order to eliminate the impact on sampled audio vector and extracted audio vector, the proposed approach suggests to standardize both audio data vectors by means of Z-score method. The steps of parameter set in algorithm mainly includes length of audio data vector cut by algorithm, the initial shift in sequence alignment based on FFT, the sudden disturbance factor t_d (see 2.4), the threshold value of correlation coefficient, and so on. Briefly, the processing steps can be summarized to three steps. First, to extract audio data vectors from TV program library based on sampled timestamps and TV listings in database. Second, to select the top k data vectors from audio data vector-set and record correspondent offset based on FFT. Third, after tolerance processing, to check pairwise points based on correlation coefficient.

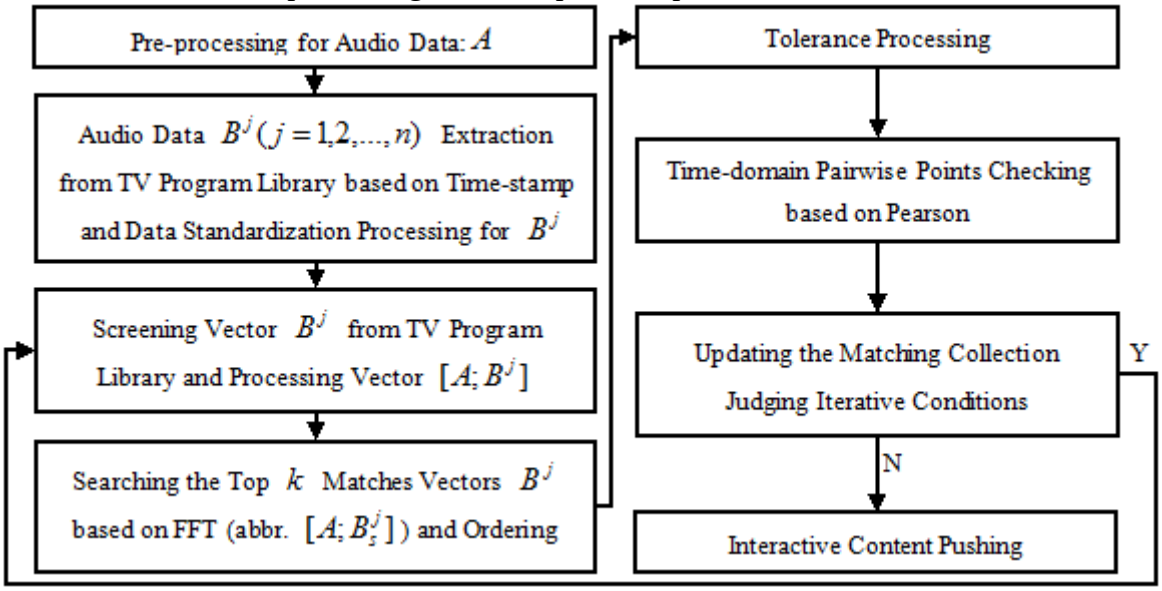

Fig.2. The work-flow of TV channel recognition algorithm

Denoising & Pre-processing

The sequences of audio signal sampling are time-ordered, therefore, their noises can be eliminated by creating noise model, whose parameters can be trained by some examples. The noise eliminating model is simplified as follows:

$$x(t) = s(t) + n(t) \quad (1)$$

In Eq.1, variables $x(t), s(t), n(t)$ represent sampling audio data, ground truth and noise data respectively. Since it is easy to prepare data for $x(t), s(t)$ directly, Eq.1 can be transformed into:

$$n(t) = x(t) - s(t) \quad (2)$$

Eq.2 illustrates weights of $n(t)$ can be trained by $x(t), s(t)$ so that it can be applied into eliminating of data sampling noises and data transmitting noises. In order to solve matrix $n(t)$, there are some main methods such as LMS, spectral subtraction, natural gradient RLS, wavelet packet transform denoising, partial differential, fuzzy rules, neural network model, etc. In view of realtime processing and algorithm robust, RBF neural network model can approximate to any nonlinear function, it has good generalization ability, moreover, it also can converge to optimal solution by training quickly[5], consequently, the proposed approach suggests that the noise eliminating model can be trained by RBF feed-forward neural network. At time t, the output is :

$$n(t) = \sum_{i=1}^{h} w(t)_i h_i \quad t = 1,2,...,n \quad (3)$$

$$h_i = \exp(-\frac{1}{2\sigma_i^2}\left\|x_p - c_i\right\|^2) \quad (4)$$

As shown in Eq.3 and Eq.4, RBF uses Gauss kernel function. In the process of solving $n(t)$, the learning method of RBF uses center method of self organizing screening, that is, during the course of training, to select h samples as centers of clustering c_i from sample library randomly. These h samples can be classified into different groups in terms of distance between sample x_p and center c_i, after that, center of clustering c_i can be recalculated. The process does not terminated until $c_i < v_t$ (variable v_t represents threshold value of variation range from center c_i). The variance of basis function is defined by:

$$\sigma_i = \frac{c_{\max}}{\sqrt{2h}}, \quad i = 1,2,...,h \quad (5)$$

Here supposing output of $N(t)$ is ideal, for weight calculation, its iterative equation is:

$$w(t)_i = w(t-1)_i + \eta(N(t) - n(t))h_i + a(w(t-1)_i - w(t-2)_i)$$
$$\eta + \alpha = 1 \quad (6)$$

In Eq.6, η, α represents weight adjustment factors and are set to 0.2~0.3 and 0.7~0.8 respectively.

Due to sampled audio data is asynchronous, i.e., time-shift is inevitable between sampled audio data and extracted audio data from TV program library, it is necessary to reduce scale of extracted audio data. In consideration of realtime processing, the proposed approach uses FFT to screen the top k extracted audio data and order them, which can also be viewed as k vectors. In order to eliminate influences of different linear ratio between sampling audio data and extracted audio data, the proposed approach uses Z-score method to standardize both sampled audio data and extracted audio data. Formally defined by:

$$x^* = \frac{x - u}{\sigma} \quad (7)$$

where:

u is the mean of the population;

σ is the standard deviation of the population.

Vectors Extracting and Screening

As mentioned above, the sampled audio data can be viewed as one or two row vectors. In order for simplified calculation, the symbol vector A is used as sampled audio data, and vector-set $B^j (j = 1,2,...,n)$ is used as extracted audio data for the reason double track audio data are almost same.

The symbol for sampling timestamps is t_s and length of vector A is t_l, length of every vector B^j is calculated as follows:

$$t_c = t_b - t_a > t_l (t_a > t_s, t_l > 0) \quad (8)$$

In Eq.8, t_a, t_b represent start position and end position of data vectors gotten from TV program library respectively, t_c is length of extracted audio data, and constraint condition $t_a > t_s$ is used to solve time-shift between A and B^j. Because of $t_c > t_l$, in order to improve computation efficiency, the proposed approach screens the top k vectors based on frequency domain analysis such as FFT, STFT，WT, etc. On one hand, the screening is to align pairwise points in data every section on the whole. On the other hand, audio signal is a kind of non-stationary signal, i.e. excitation and resonant characteristics of vocal tract changes over time, it is suitable for using WT, however, for sequence alignment challenges as stated before, vector A has the characteristics of diversity and complexity in sampling and processing, which leads up to uncertain for screening of WT function, therefore the proposed approach suggests to analyze A and B^j based on FFT[6],

which are defined as:

$$X(k) = \sum_{n=1}^{\mathrm{N}} x(n)\omega_N^{(k-1)(n-1)}$$

$$x(n) = (1/N)\sum_{k=1}^{N} X(k)\omega_N^{-(k-1)(n-1)} \quad (9)$$

$$\omega_N = e^{(-2\pi i)/N}(i = \sqrt{-1})$$

On the basis of Eq.9, the proposed approach uses binary search to screen B^j and sets step-size dynamically, i.e. :

$$m_i^j = fft(A)/\ fft(B^j(a:b))$$

$$m_i = \max_j(m_i^j)$$

$$b - a = t_l, a \geq t_a, b \leq t_b \quad (10)$$

$$a = a + s \text{ or } a \pm \frac{s}{2}$$

By reason of $t_l < t_c$, audio data with length t_l can be extract from B^j. Variable s is defined as step-size. In experiment, it is set to between 100 and 500. By means of iterative solution method, let $a = a + s$, then variable m_i can be gotten by Eq.10. In view of the best pairwise point may be less than variable s, the proposed approach uses the iterative refinement further improves the positioning accuracy with $a = a \pm \frac{S}{2}$, namely, according to the results calculated by FFT to solve optimization problems with bi-directional binary search. The optimized m_i corresponding to B^j can be gotten finally.

The reason the proposed approach can use frequency domain to screen the top k vectors exists in the differences on timing sequence between A and B^j. Although, The problem of time-shift can be solved based on FFT, frequency domain analysis can't solve the interrelationship on timing sequence accurately. Therefore, after steps of screening for B^j, pairwise points should be checked in time domain so as to improve recognition accuracy based on high-level semantic[7]. In experiment, k is set to 3~8.

Tolerance Processing

Before pairwise point checking, in order to eliminate interferences of sudden noises on time domain correlation analysis, the proposed approach suggests to eliminate those points, whose values fluctuate violently and are much greater than mean value u so as to improve performance on pairwise point checking and enhance the robustness of the proposed algorithm. The processing model of tolerance is defined as:

$$A_i = \begin{cases} B_i & \frac{1}{N}\sum_{i=1}^{N}\left|(A_i - \overline{A})\right| > T_A \\ A_i & other \end{cases} \quad (11)$$

$$T_A = \frac{1}{t_d \times N}\sum_{i=1}^{N}\left|A_i - \overline{A}\right|$$

$$A_i = \begin{cases} B_i & |A_i - B_i| > T_{AB} \\ A_i & other \end{cases} \quad (12)$$

$$T_{AB} = \frac{1}{t_d \times N}\sum_{i=1}^{N}|A_i - B_i|$$

t_d is used to adjust the deviation from the mean difference, the greater its value is, the more points will be eliminated and the greater the tolerance range of algorithm is correspondently. T_A is used to eliminated points far away from the mean value. T_{AB} is the threshold value used for tolerance of sudden interferences. In experiments, t_d is set to 0.2~0.5. In practical application, the threshold value can be set to 20%~40% of the greatest deviation from mean value automatically by

algorithm.

Pairwise Point Checking

In view of the fact, as stated before, the previous processing steps can't process the differences on time domain, for this reason, after vector screening and tolerance processing, pairwise point checking should be used to further meet the accurate matching requirement based on time domain. The proposed approach uses Pearson correlation coefficient as performance estimation for pairwise vectors and related shifts[8].

$$r = r_{AB} = \frac{\sum_{i=1}^{N}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{N}(B_i - \overline{B})^2}} \tag{13}$$

In Eq.13, $N = t_l$ represents length of vector matching on time domain. Since $n < t_c$, the proposed approach uses step search method on time domain along forward direction so as to get the best matching $r_{\max} = \max_j(r_{AB}^j)$ based on pairwise vectors and related shifts.

Experiments and Analyses

In experiments, we choose 12 different kinds of TV programs from TV program library and distributing them with steaming media way. According to different period, we use the proposed approach to test its performance and accuracy.

In order to easily show the experimental results, vector A and B^j have been compressed on time domain with regular intervals. In Fig.3, it shows the process of pairwise point search for vectors B^j with a binary search based on dynamic step adjustment .

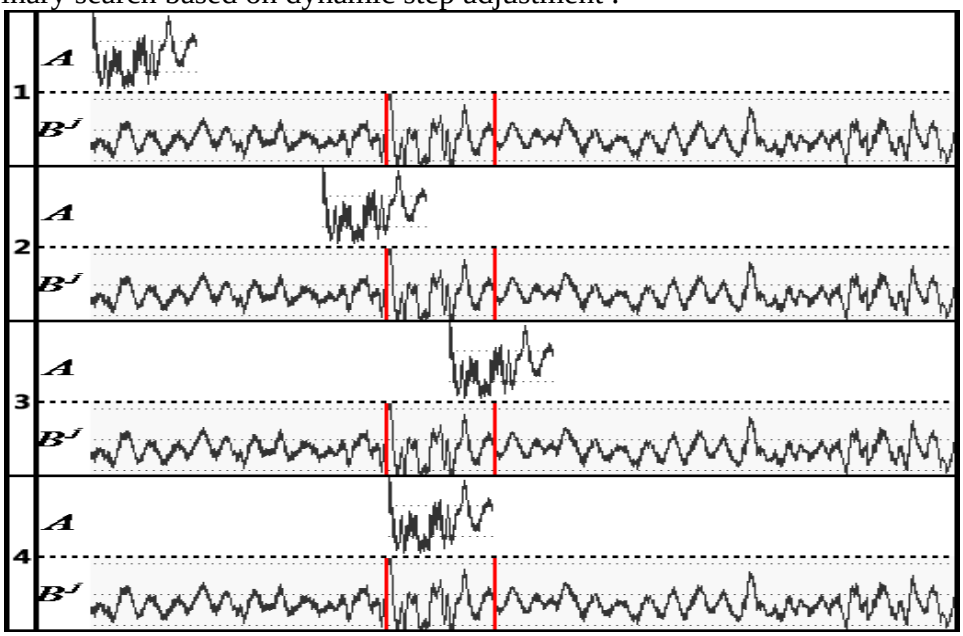

Fig.3. The realtime match point search for B^j

Both the methods based on time domain and the methods based on frequency domain have their own advantages and disadvantages. The advantages of the methods based on time domain are simple, high accuracy degree, good robustness, and the disadvantages of them are low efficiency and weak realtime performance. On the contrary, both the advantages and disadvantages of the methods based on frequency domain is the opposite. The advantages of the proposed approach include difference eliminating for different sampling devices by data standardization, effective screening process based on FFT, accuracy improving by pairwise point checking on time domain. We selected 4 sampled audio data A_1, A_2, A_3, A_4 randomly, in view of the sampled timestamps, 12 audio sections $j = 1,2,...,12$ are cut from distributing TV programs. The experimental results are

shown in Table 1. All of experiments work on the environment with processor i7 5500U, 4GB memory, solid harddisk, and on the operating system with win8.

Table 1 Experimental Results

Sampled Audio Vector	A_1	A_2	A_3	A_4
Matched Vector	B^{j3}	B^{j9}	B^{j4}	B^{j11}
Sampling Length	24KB	24KB	40KB	40KB
Buffer Length	200KB	40KB	64KB	80KB
Matched Point Position	80472	14377	17298	21361
Cost of sequence alignment	3.7s	1.3s	1.8s	2.3s
k	5	3	3	4
Cost of Pairwise Point Checking	0.8s	0.2s	0.2s	0.5s
Correlation	0.93	0.97	0.84	0.91

The experimental results, as shown in Fig.4, compare the proposed approach with the method based on time domain and the method based on frequency domain respectively on accuracy and efficiency.

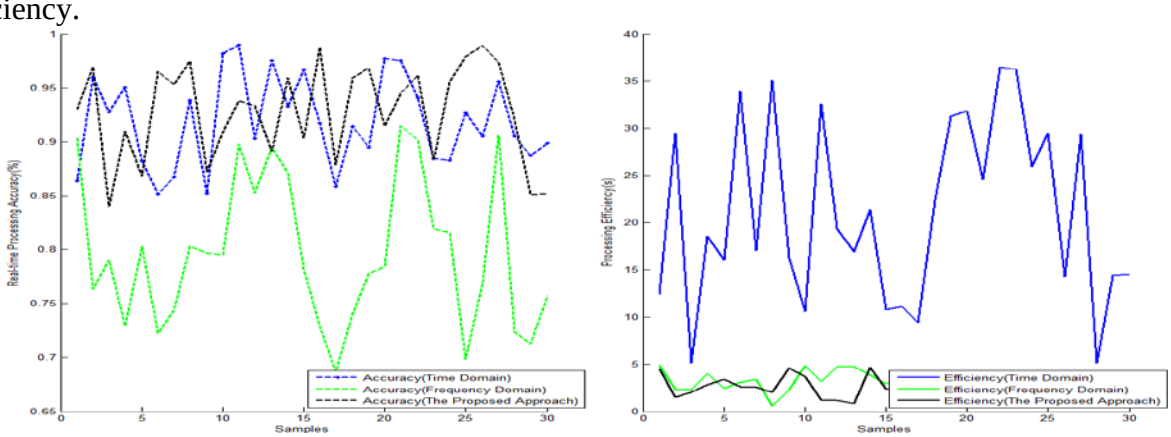

Fig.4. Compared with other methods for test data

In Fig.4, the method based on time domain has good accuracy, however, it has bad efficiency with many times correspondent to the proposed approach and the method based on frequency domain. In contrast, the method based on frequency domain has good efficiency due to the use of statistical method for screening processing. However, it ignores specific difference between A and B^j on time domain so that it can't check the diversity existing in data vector-set. The proposed approach executes sequence alignment by screening the top k vectors based on FFT and further checks pairwise points based on dynamic step-size in time domain. Therefore it combined their advantages so that it improves overall performance.

Conclusion

The TV program distributing platform has the superiority of media content, in contrast, the mobile Internet has the superiority of better interactive functions than the digital TV platform. This article proposes an approach for TV channel recognition based on audio data. The audio data sampled by APP mobile terminals also referred to as client-side and the are transmitted to TV program distributing platform also referred to as server-side. The server-side process the received audio data and recognize TV program channels which they belongs to based on sampling timestamps and TV listings in database so as to push some content related to distributing TV program, discuss with users interactively, vote in realtime, data statistics for program ratings, etc.

The next work of the proposed approach is to further optimize sequence alignment according audio features such as Harris, data trend feature,SIFT, etc. In the process of vector screening,

comparing these features with FFT by experiments and analyzing them on accuracy and efficiency.

Acknowledgement

In this paper, the research was sponsored by the Science and Technology Department of Zhejiang Provincial for Public Project(No.2014C33092).

References

[1] Lim J S, Ri S Y, Egan B D, et al. The cross-platform synergies of digital video advertising: Implications for cross-media campaigns in television, Internet and mobile TV[J]. Computers in Human Behavior, 2015, 48: 463-472.

[2] Meeker M. Internet Trends 2014, Code Conference[J]. Retrieved May, 2014, 28: 2014.

[3] Holmes M E, Josephson S, Carney R E. Visual attention to television programs with a second-screen application[C]//Proceedings of the Symposium on Eye Tracking Research and Applications. ACM, 2012: 397-400.

[4] Shi L, Obregon E, Sung K W, et al. CellTV—On the Benefit of TV Distribution Over Cellular Networks: A Case Study[J]. Broadcasting, IEEE Transactions on, 2014, 60(1): 73-84.

[5] Rajpal N. Comparative analysis of feed forward and radial basis function neural networks for the reconstruction of noisy curves[C]//Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on. IEEE, 2014: 353-358.

[6] Cooley J W, Tukey J W. An Algorithm for the Machine Computation of Complex Fourier Series, vol. 19[J]. Mathematics of Computation, 1965.

[7] Horwege S, Lindner S, Boden M, et al. Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches[J]. Nucleic acids research, 2014: gku398.

[8] Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data[J]. Quaestiones Geographicae, 2011, 30(2): 87-93.