Audio Scene Semantic Similarity Computing Approach

Wei Wei[1] , Ye Bin[2], Chen Bang-sheng[3]
Department of Computer Science and Technology
Chengdu University of Information Technology, Chengdu, China
weiwei@cuit.edu.cn, ybing@21cn.com, bshchen@cuit.edu.cn

***Abstract*—Audio in the video carries abundant semantic message. An audio scene is temporal audio segments which represented by a few basic audio effects. The semantic similarity of pair audio scenes is very useful for high-level audio semantic understanding. A computing approach for audio scene semantic similarity is proposed in this paper. Firstly, audio track is pre-segmented to audio scenes. Then, basic audio effects dominating each audio scene are recognized. Finally, the similarity of two audio scenes is calculated based on a model consist with information theoretic similarity principles and Tversky's set-theoretic similarity. The results of experiments indicate the audio scene semantic similarity computing approach could count quantitative semantic similarity of two scenes.**

Keywords-audio scene; semantic similarity; audio effects; HMMs; semantic affinity

I. INTRODUCTION

An audio scene is a semantically consistent sound segment that is characterized by a few dominant sources of sound [1]. In this paper, the audio scene is defined as a semantically consistent chunk of audio data, which consists of several key audio effects with semantic affinity relationship. The quantitative semantic similarity between audio clips is very important for multimedia semantic understanding. Audio scene semantic similarity is necessary for high-level semantic inference from audio data.

There has been many works done dealing with the problem of audio scene segmentation [1, 2]. Audio scene segmentation method is not the key issue discussed in this paper. Thus, the input audio stream is first segmented into audio scene, which is a prerequisite work of the approach in this paper. As shown in Figure 1, basic audio semantic effect could be detected and extracted. The proposed framework is consists of two parts: Basic audio effect recognition and Semantic similarity computing. For basic audio effect recognition, the fixed interval audio spectrum feature is extracted. Then, an HMM model is trained for adjusting the model parameters of each basic audio semantic. Through semantic similarity computing process, audio scene semantic similarity could be computed finally. Key audio effects are used as middle-level representations in the auditory scene high-level semantic understanding.

In this paper, we present a novel framework for quantitative measuring semantic similarity of pair audio effects. The remainder of this paper is organized as follows. In Section 2, we briefly describe semantic space of audio scene. In Section 3, basic audio effect recognition is given in detail. In Section 4, audio scene semantic similarity computing is introduced. All experimental results are given in Section 5. Finally, Section 6 concludes this paper.

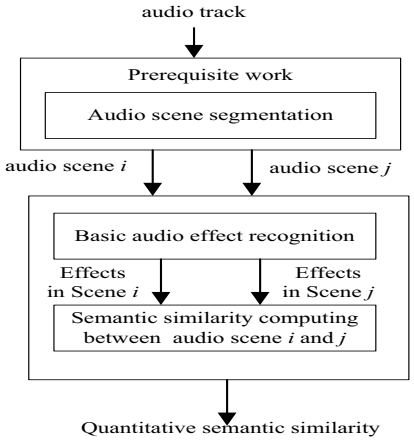

Figure 1. Framework of method in this paper

II. SEMANTIC SPACE OF AUDIO SCENE

An illustration of audio space is given in Figure 2. Left part of the figure is multi-level semantic space. SB_i（i =L, M, H） is the sub-semantic space. Where, SB_L is bottom sub-semantic space, which is spaned by low-level acoustic (audio) feature. The level of video semantic concept is increases with *i*. Mid level audio semantic could represented by basic audio effects. Basic audio effects can be extracted by the method described in Section 3. In audio semantic analysis, key audio effects recognition plays an important role in process of bridging the gap between low-level features and high-level semantics. The *semantic affinity* of several key audio effects is a key clue to infer high-level semantics.

Audio effects semantic affinity in audio scene directly related to the high-level semantics. Different composing of effects means distinct high-level semantic. For instance, *explosion*, *gun-shot* and *burning-snap* often indicates *violence scene*. Otherwise, *applause*, *cheer*, *singer* and *laughter* are associated with *joyance scene*.

978-1-4244-5265-1/10/$26.00 ©2010 IEEE

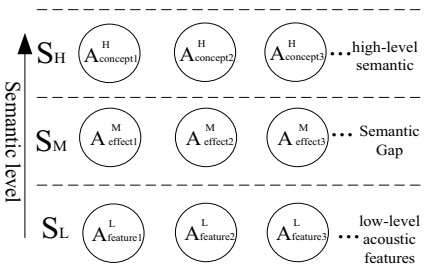

Figure 2. Audio semantic space

Combination of different effects plays critical roles in high-level audio scene understanding. In general, the cognizing among resemble audio scenes with similar audio effects is easy for human brain. However, the computing for quantitative semantic similarity is not very easy for computer.

III. BASIC AUDIO EFFECT RECOGNITION

A. Model for Basic Audio effect recognition

Hidden Markov models (HMMs) have become the method of choice for modeling stochastic processes and sequences in applications such as speech and handwriting recognition. We apply HMMs to extracting basic audio effect like the method in reference[3]. In real world, voice is continuous. At the first step of audio processing, voice signal is discrete by the sample values of some time interval. Then, audio spectrum features are extracted.

The temporal context information in audio is one of significant cues for content understanding. The purpose of basic audio semantic analysis is to discover the hidden states or semantics behind video signals. In this view, audio signals could be looked as the observations of basic audio semantics. Denoted audio information in video as a statistical process, therefore in terms of probabilities, the basic approach in semantic analysis may be formulized as follows:

$$\lambda = (\pi, A, B) \tag{1}$$

Five important parameters of the basic audio semantic model are:

1) N is the number of audio semantic states in the model. Although the semantic states are hidden, for audio spectrum there is some physical significance attached to the semantic states. The individual states are denoted as $S = \{S_1, S_2, \cdots, S_N\}$, the state at time t is q_t.

2) M is the number of distinct observation symbols for per basic audio state. The observation symbols correspond to the spectrum of the audio signal. The individual spectrum symbols are denoted as $V=\{v_1, v_2, v_3, \ldots, v_M\}$. The observation symbols at time t is T O_t .

3) The basic audio state transition probability distribution matrix is $A = (a_{ij})_{N \times N}$. Where, $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, $1 \le i, j \le N$.

4) The semantic observation symbol probability distribution in basic audio semantic state j is $b_j(k) = P((V_k)_t \mid q_t = S_j)$. Where, $1 \le j \le N, 1 \le k \le M$.

5) The initial state distribution is $\pi = \{\pi_i\}$ Where, $\pi_i = P(q_1 = S_i)$, $1 \le i \le N$ .Given appropriate values of N, M, A, B and π , the basic audio model can be used as a generator to give an audio spectrum observation sequence $\boldsymbol{O} = O_1 O_2 \cdots O_T$. Where, each observation OB_{tB} is one of the symbols from V, and T is the number of observations in the sequence.

B. Spectrum feature extraction

The spectrum feature extraction process includes several descriptors of MPEG-7, such as Audio Spectrum Basis, Audio Spectrum Projection and Audio Spectrum Envelope. As described in Figure 3, it mainly consists of normalized audio spectrum envelope[4, 5]. Audio Spectrum Envelope Descriptor, which is a logarithmic-frequency spectrum, spaced by a power-of-two divisor or multiple of an octave. This Audio Spectrum Envelope is a vector that describes the short-term power spectrum of an audio signal. To represent compactly the independent subspaces of a spectrogram, a principal component analysis and independent component analysis are used.

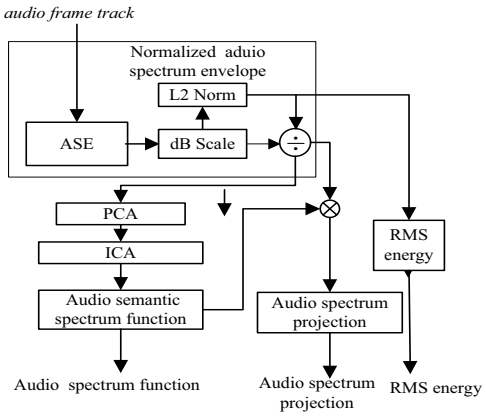

Figure 3. Spectrum feature extraction

For the temporal sampling signal, the observed audio signal (discrete audio sampling signal) is divided into overlapping frames by sliding window, namely a hamming window. Thus, the temporal sampling signal id analyzed using the short-time Fourier transform (STFT).

$$S(l,k) = \sum_{n=0}^{N-1} s(n + lM) w(n) e^{-j\frac{2nk}{N}} \tag{2}$$

Where, *n* is the size of the STFT, *k* is the frequency bin index, *l* is the temporal frame index of audio frame, *w* is hamming analysis window with size *lw*. *M* is the hop size.

In order to preserve power:

$$P(k,l) = \frac{|S(k,l)|^2}{nf \cdot N} \tag{3}$$

Where, *nf* is defined:

$$nf = \sum_{n=0}^{lw-1} w^2(n) \tag{4}$$

To extract reduced-rank spectral features, the spectral coefficients are grouped in logarithmic sub-bands. Frequency channels are logarithmically spaced in non-overlapping 1/8-octave bands spanning between 62.5 Hz ("low edge"), and 8 kHz ("high edge"). The output of the logarithmic frequency range is the sum of the power spectrum in each logarithmic sub-band.

The resulting log-frequency power spectrum is converted to the decibel scale:

$$D(f,l) = 10\log_{10}(ASE(f,l)) \tag{5}$$

After that, each decibel-scale spectral vector is normalized with the root mean square (RMS) energy envelope, thus yielding a normalized log-power version of the ASE (NASE):

$$R_l = \sqrt{\sum_{f=1}^{F}[10\log_{10}\{ASE(f,l)\}]^2} \tag{6}$$

$$X(f,l) = \frac{10\log_{10}\{ASE(f,l)\}}{R_l}, 1 \le l \le L \tag{7}$$

Where, F is the number of ASE spectral coefficients, and L is the total number of frames. The basis function is non-normative. Then, the basis follows two steps to get *L2-norm* information in *k* dimensions. First, the columns should be centered by subtracting the mean. Second, the rows should be standardized by removing any dc offset and normalizing the variance.

Basis functions may be orthogonal, as given by PCA extraction, or non-orthogonal as given by ICA extraction. Basis projection and reconstruction are described by $\mathbf{Y} = \mathbf{XV}$ and $\mathbf{X} = \mathbf{YV}^+$.Where, $\mathbf{Y}$ is the observation matrix with *m* x *k*, $\mathbf{X}$ is spectral matrix with *m* x *n*. $\mathbf{V}$ is a $n \times k$ matrix of basis functions, which is arranged in the columns. $\mathbf{V}^+$ is the pseudo inverse of $\mathbf{V}$ for the non-orthogonal case.

C. The training of basic audio model

How to adjust the model parameters $\lambda = (\pi, A, B)$ to maximize $P(\mathbf{O}/\lambda)$ is the training process of model, as showed in Figure 4. Namely, we attempt to optimize the model parameters so as to best describe how a given audio spectrum observation sequence comes about. The observation sequence used to adjust the model parameters is called a training audio clip. In practice, we can choose the parameters of $\lambda = (\pi, A, B)$ so that $P(\mathbf{O}/\lambda)$ is locally maximized with Baum-Welch method. Model training is a re-estimation process denotes as[6]:

$$\bar{\pi} = \gamma_1(i) \tag{8}$$

$$\bar{a}_{ij} = \sum_{t=1}^{T-1}\xi_t(i,j) \Big/ \sum_{t=1}^{T-1}\gamma_t(i) \tag{9}$$

$$\bar{b}_j = \sum_{\substack{t=1 \\ s.t.O_t=V_K}}^{T}\gamma_t(j) \Big/ \sum_{t=1}^{T}\gamma_t(j) \tag{10}$$

The final result of this re-estimation procedure is called a maximum likelihood estimate of the HMM. Eventually, the likelihood function converges to a critical point. In the experiments of this paper, the number of audio semantic hidden states is 15.

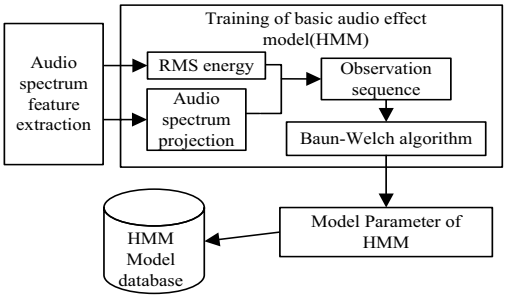

Figure 4. Training process of basic audio effect

D. Elimination of un-predefined event

Elimination of un-predefined events is a Bayesian decision process. To calculate the *i*th basic audio event possibility of *m*th window audio frame is a typical Bayesian decision. Let α_1 denotes the decision behavior of w_1, and α_2 denotes the decision behavior of w_2. Where, w_1 denotes that the window represents *i*th BE. For the window, w_2 means no possibility of *i*th BE (elimination). $\lambda_{u,l}=\lambda(\alpha_k|w_l)$ is a loss function, for the event in the state w_l and the decision w_u made. Condition loss is defined as[6]:

$$R(\alpha_1 \mid x) = \lambda_{11}P_i(w_1 \mid x) + \lambda_{12}P_i(w_2 \mid x) \tag{11}$$

$$R(\alpha_2 \mid x) = \lambda_{21}P_i(w_1 \mid x) + \lambda_{22}P_i(w_2 \mid x) \tag{12}$$

If $R(\alpha_1|x) < R(\alpha_2|x)$, the decision is w_1. For the decision rule, $P(w|x)$ is posteriori probability, which can be denotes by prior probability $P(w)$ and conditional probability $p(x|w)$. Thus, if the likelihood ratio is bigger than k_i the decision is w_1. if it is smaller than k_i, the decision is w_2. Where, k_i is a threshold depended on *x*.

$$(\lambda_{21} - \lambda_{11})p_i(x \mid w_1)P(w_1) > (\lambda_{12} - \lambda_{22})p_i(x \mid w_2)P(w_2) \tag{13}$$

$$\frac{p_i(x \mid w_1)}{p_i(x \mid w_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}\frac{P_i(w_2)}{P_i(w_1)} = k_i \tag{14}$$

Where, $p_i(x|w_l)$ in Equation (25) denotes the probability distribution of log-likelihood that *i*th basic event under *i*th HMM model. Furthermore, $p_i(x|w_2)$ denotes the probability distribution of log-likelihood that all basic events except *i*th basic event under *i*th HMM model. Generally, both the distribution probability density functions of $p_i(x|w_1)$ and $p_i(x|w_2)$ are negative Gamma distribution[7]. $P(w_1)$,$P(w_2)$, $p_i(x|w_1)$ and $p_i(x|w_2)$ can be estimated by the priori information in train sets. In practice, it is demonstrated that

the risk of false decision is bigger than correct decision. Therefore, $\lambda_{2,1} > \lambda_{1,1}$ and $\lambda_{1,2} > \lambda_{2,2}$ (the values in experimentation are $\lambda_{2,1}$=1, $\lambda_{1,1}$=0, $\lambda_{1,2}$=6, and $\lambda_{2,2}$=0). For variable *i*(from 1 to *n*), $c(0 \leqq c \leqq n)$ decisions of w_1 can be got by Equation (14).

E. Bayesian classification

If c=0, the event corresponding to mth analysis window is not predefined. If c>0, the post probability could be calculated by[8]:

$$P_{r,l}(\mathrm{BE}_{r,l} \mid x) = \frac{p_{r,l}(x \mid w_1) P_{r,l}(w_1)}{\sum_{j=1}^{r} p_{j,l}(x \mid w_1) P_{j,l}(w_1)} \tag{15}$$

For $\mathrm{BE}_{r,l}$, r denotes the basic event order in c, and l denotes the class order of BE in Section 2.3. The class semantic label of maximum post probabilities is final class label of analysis window.

IV. AUDIO SCENE SEMANTIC SIMILARITY COMPUTING

Similarity could be defined in diverse formats based diversified theory. In paper[9], information theoretic similarity definition for general application is proposed. This section will briefly expatiate audio scene semantic similarity approach, any detailed content referencing paper[9].

Intuition 1: The semantic similarity between audio scene A and audio scene B is related to their commonality of basic audio effects. The more commonality audio effects they share, the more semantic similar they are.

Intuition 2: The similarity between audio scene A and audio scene B is related to the difference audio effects between them. The more effects difference they have, the less semantic similar they are.

Intuition 3: The maximum similarity between audio scene A and audio scene B is reached when A and B are composed of same audio effects, no matter how much commonality they share.

The common of audio scene A and audio scene B is counted as:

$$I(common(A,B)) \tag{16}$$

Where, common(A,B) is the same basic audio effects.

The difference of audio scene A and audio scene B could be measured by:

$$I(description(A,B)) - I(common(A,B)) \tag{17}$$

Where, description(A,B) denotes whole key audio effects that describes the audio scene A and audio scene B.

Audio effects semantic affinity in $sim(A,B)$, the similarity of audio scene A and audio scene B, is a function of common and description for A and B.

$$sim(A,B) = f(I(common(A,B)), I(description(A,B))) \tag{18}$$

While audio scene A is same as audio scene B, the similarity of the pair audio scenes is 1.

Based on information theoretic similarity principles[5]and Tversky's set-theoretic similarity models[10], the the similarity of audio scene A and audio scene B is:

$$sim(S_a, S_b) = \frac{|A \cap B|}{|A \cap B| + \delta(S_a, S_b)|A/B| + (1-\delta(S_a, S_b))|B/A|} \tag{19}$$

Where, $\delta(S_a, S_b) = \frac{\|A \cap B\|}{|A|}$.

A is the set of basic audio effects (middle-level audio semantic) that describe audio scenc S_a. S_b is the same signification as S_a. the sign || denotes radix audio effects. δ is weight of difference feature.

According the study of cognition, the similarity of son to father is bigger than the similarity of father to son[11]. therefore, the audio scene similarity is unsymmetrical. Namely, if $|A/B| < |B/A|$, then $sim(S_a, S_b) > sim(S_b, S_a)$. In Equation (19), δ reflects the unsymmetrical similarity of audio scene[12].

V. EXPERIMENTAL

A. Basic audio semantic extraction

TABLE I. BASIC AUDIO SEMANTIC CLASSIFICATION AND RECOGNITION

Semantic concept	Bayesian Decision method in this paper		
	Recall	Precision	F-Score
explosion	0.925	0.872	0.921
speech	0.950	0.974	0.962
car engine	0.900	0.878	0.889
missile in trajectory	0.950	0.975	0.962
water flow	0.925	0.881	0.902
Average	0.930	0.916	0.927

This experiment uses quality evaluation metrics of TREC (precision, recall and F-Score). The length of each audio sample is about 3-8s, and the sampling rate is 22.05 kHz. The semantic category of each auditory clip is manually labeled for all of the audio tracks in the database. Based on HMM model, seven basic audio semantic categories are further detected: *explosion*, *speech*, *car engine*, *flight missile in trajectory*, *water flow*. For each class, there are 40 training clips and 40 test clips. To compare the performance of directly classification using maximum-likelihood scores and the Bayesian decision method in this paper, the recognition results of basic audio semantic are showed in Table I.

B. *Quantitative computing of audio scene semantic similarity*

This experiment will compute two audio scenes semantic similarity. Two pre-segmented audio scenes called Scene A and Scene B. Scene A and Scene B are extracted feature by the step describing in section 3. Then, the audio feature will be input into basic audio effect recognition module. Using the basic audio effect recognition technique proposed in section 3, *applause*, *laughter*, *singing*, *cry*, *music*, *speech*, *laughter* are recognized basic audio effects for Scene A, and *applause*, *laughter*, *singing*, *cry*, *laughter*, *cheer* are recognized basic audio effects for Scene A. Supposing. Therefore, $|A \cap B|$ is 4, $|A/B|$ is 3, $|B/A|$ is 2. Then, δ is 4/7. According to equation (19), the semantic similarity of audio A and audio B is 0.609.

VI. CONCLUSIONS

Audio information plays a very important role in the understanding of video semantic content. To compute semantic similarity of two audio scenes is a matter of great urgency. In this paper, a computing approach for audio scene semantic similarity is proposed. However, the similarity measure approach proposed in this work is based on the analysis of audio effects. Namely, audio effects are the middle-level semantic representation, and audio effects semantic affinity in audio scene directly related to the high-level semantics. Therefore, audio scene semantic similarity measures based on more general middle-level audio semantic representation will be studied in the future.

ACKNOWLEDGMENT

This paper is supported by the Scientific Research Foundation of CUIT under Grant No. KYTZ200916.

REFERENCES

[1] Sundaram H., Chang S.-F, "Audio Scene Segmentation Using Multiple Feature", Models and Time Scales. In: IEEE International Conference on Acoustic, Speech and Signal Processing, Istanbul, Turkey, pp. 2441-2444 ,2000.

[2] 2 Cai R., Lu L., Cai L.-H, "Unsupervised auditory scene categorization via key audio effects and information-theoretic co-clustering", In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. ii/1073- ii/1076, 2005.

[3] WEI WeiP, ZOU Shu-rong, LIU Feng-yu, " Study of Basic Audio Semantic Analysis and Extraction Techniques for Video Data", Journal of Chinese Computer Systems, vol, 1728, No. 1719pp. 1715-1719 ,2007.

[4] JTC1/SC29/WG11 I I. Coding of Moving Pictures and Audio, "Overview of the MPEG-7 Standard" Int'l Organization for Standariation, Oct. 2000.

[5] Kim H-G, Berdahl E., Moreau N. et al, "Speaker Recognition Using MPEG-7 Descriptors", Proceeding of EUROSPEECH, Geneva, Switzerland, pp. 1-4 September, 2003.

[6] R. O.D., P. E.H., DavidG. Stork., Pattern classification second edition, China Machine press, Beijing, 2003.

[7] Cai R., et al, "Highlight sound effects detection in audio stream", In: . ICME '03. Proceedings. 2003 International Conference on Multimedia and Expo, Maryland, USA, 2003, pp. III - 37-40

[8] Bian Zhao-qi., Zhang Xue-gong et al, Pattern Recognization, TsinghuaUniversity Press, Beijing, 2000.

[9] Lin D, "An Information-Theoretic Definition of Similarity. In: Proceedings of the 15th International Conference on Machine Learning Madison, Wisconsin USA, 296-304, 1998.

[10] A. T. Features of similarity. Psychological review. 1977, vol, 84, pp. 327-352.

[11] Rosch E, "Cognitive Representations of Semantic Categories", Journal of. Experimental Psychology, vol, 104(3) , pp. 192-233,1975.

[12] Krumhansl. C.L, "Concerning the applicability of geometric models to similarity data: The interrelationship between similarityand stimulus density", Psychological review, vol, 85, No. 5, pp. 445-463,1978.