A HIERARCHICAL APPROACH FOR AUDIO STREAM SEGMENTATION AND CLASSIFICATION

Wei Liang
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
wliang@hitic.ia.ac.cn

Shuwu Zhang
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
swzhang@hitic.ia.ac.cn

Bo Xu
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
xubo@hitic.ia.ac.cn

ABSTRACT

This paper describes a hierarchical approach for fast audio stream segmentation and classification. With this approach, the audio stream is firstly segmented into audio clips by MBCR (Multiple sub-Bands spectrum Centroid relative Ratio) based histogram modeling. Then a MGM (Modified Gaussian modeling) based hierarchical classifier is adopted to put the segmented audio clips into six pre-defined categories in terms of discriminative background sounds, which is pure speech, pure music, song, speech with music, speech with noise and silence. The experiments on real TV program recordings showed that this approach has higher accuracy and recall rate for audio classification with a fast speed under noise environments.

Keywords: Audio classification, MBCR, MGM, histogram

1 INTRODUCTION

In real world, media streams, such as TV programs, broadcasts, etc., are a rich source of multimedia information, containing audio, speech, text, image, motion, and so on. How to build an efficient mechanism for AV (Audio/Video) indexing and retrieval is becoming extremely important, which require automatic understanding of semantic contents. As a counterpart of visual information in video sequence, audio stream got more attention recently for its semantic content discrimination capability. As the first step of audio content processing, audio stream segmentation and classification are required.

As for audio classification, most studies are focused on speech/music/silence/others separation [1,2]. Scheirer and Slaney [1] proposed to use thirteen features in time, frequency, and cepstrum domains and model-based (MAP, GMM, KNN, etc.) classifier, which achieved an accuracy rate over 90% on real-time discrimination between speech and music. As in general, speech and music have quite different spectral distribution and temporal changing patterns, it is not very difficult to reach a relatively high level of discrimination accuracy. Further classification of audio data may take other sounds into consideration besides speech and music. Srinivasan, et al [3] proposed an approach to detect and classify audio that consists of mixed classes such as combinations of speech and music together with environment sounds. The accuracy of classification is more than 80%. An acoustic segmentation approach was also proposed by Kimber and Wilcox[4], where audio recordings were segmented into speech , silence, laughter and non-speech sounds. They used cepstral coefficients as features and the hidden Markov model (HMM) as the classifier.

In this paper, we propose a MGM-based (Modified Gaussian Modeling) hierarchical classifier for audio stream classification. Compared to traditional classifiers, MGM can automatically optimize the weights of different kinds of features based on training data. It can raise the discriminative capability of audio classes with lower computing cost. The experiments on real TV program recordings show that the average precision of classification can reach above 89%.

The remainder of the paper is organized as follows: Section 2 is an overview of proposed segmentation and classification. Section 3 and section 4 explain the details of the segmentation and classification algorithms respectively. The experiments is describes in section 5; and final conclusion is given in section 6.

2 OVERVIEW

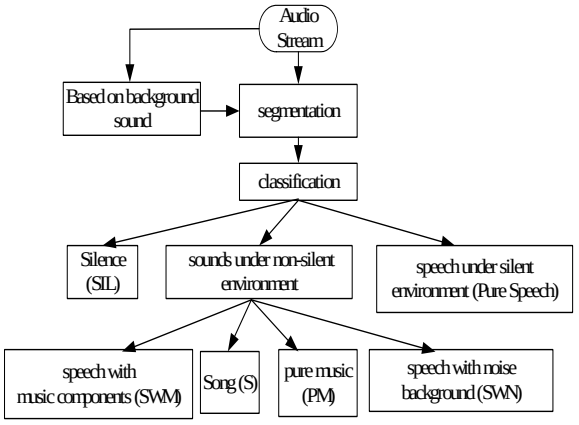

Figure1: The flowchart of segmentation and classification algorithm

Figure 1 shows the flowchart of proposed audio segmentation and classification algorithm. It is a hierarchical structure. In the first level, a long audio stream can be segmented into some audio clips according to the change of background sound by MBCR based histogram modeling. Then a two level MGM (Modified

Gaussian modeling) classifier is adopted to hierarchically put the segmented audio clips into six pre-defined categories in terms of discriminative background sounds, which is pure speech (PS), pure music (PM), song (S), speech with music (SWM), speech with noise (SWN) and silence (SIL).

3 SEGMENTATION ALGORITHM

Since background sounds always change with the change of scenes, the acoustic skip point of an audio stream may be checked by background sounds. As shown in Figure 2, the MBCR feature vectors are firstly extracted from the audio stream. We set a sliding window which consists of two sub-windows with equal time length. The window on input signal is shifted with a range of overlapping. Then two histograms are created from each sliding sub-windows. The similarity between two sub-windows can be measured by histogram matching. The skip point can thus be detected by searching the local lowest similarity below a threshold.

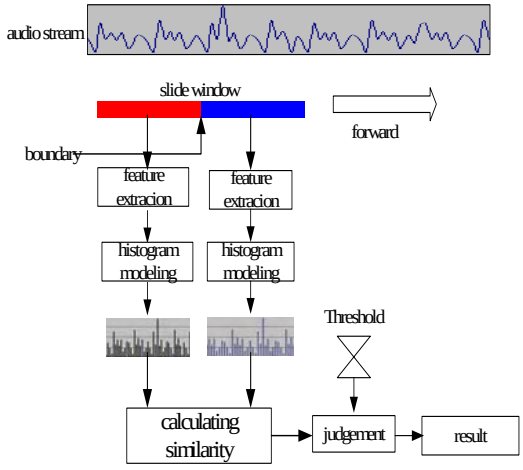

Figure2: Block diagram of segmentation algorithm

3.1 Feature extraction

Considering the lower frequency spectrum are too sensitive to even a bit of changes of the scenes and speakers, it could cause segmented clips too small. It will have effects on succeeding audio classification. We, thus, use Multiple sub-Bands spectrum Centroid relative Ratio (MBCR) [5] over 800Hz as basic feature. This feature may depict centroid movement trend in a time-frequency-intensity space. Its mathematical description can be described as follows.

$$SCR(i,j) = \frac{SC(i,j)}{\max_{j=1:N}(SC(i,j))} \quad (1)$$

$$SC(i,j) = \frac{f(j) * FrmEn(i,j)}{\sum_{k=1}^{N} FrmEn(i,k)} \quad (2)$$

where $SCR(i,j)$ is MBCR of the ith frame and the jth sub-band, $SC(i,j)$ is the frequency centroid of the ith frame and the jth sub-band, and N denotes the number of frequency sub-bands. The element of $f(j)$ is the normalized central frequency.

$$FrmEn(i,j) = \log(\int_{\omega_L(j)}^{\omega_H(j)} |F(i,\omega)| d\omega) \quad (3)$$

where $\omega_L(j)$ and $\omega_H(j)$ are lower and upper bound of sub-band j respectively, $F(i,\omega)$ represent denotes the Fast Fourier Transform (FFT) at the frequency ω and frame i, and $|F(i,\omega)|$ is square root of the power at the frequency ω and frame i.

3.2 Histogram modeling and similarity measurement

After feature extraction, we need to train model for each sub-window. Some modeling approaches, such as GMM, HMM, SVM, have already been employed for audio modeling. However, because of computationally expensive processing, it is hard to meet the speed demand of quick audio segmentation. Histograms can be used as a type of non-parametric signal model. It doesn't need computationally expensive processing and it is relatively stable under adverse environments [5].In order to remove the influence of noises, the feature vector firstly need to be quantized (VQ) before modeling histogram.

The similarity distance between the two sub-window feature vector histogram can be measured by histogram intersection. The histogram intersection for a window is defined as

$$S_n(h_1(n), h_2(n)) = \frac{1}{L}\sum_{j=1}^{L} \min(h_1^j(n), h_2^j(n)) \quad (4)$$

where $h_1(n)$ and $h_2(n)$ is the histogram of two sub-window at frame n; $h_1^j(n)$ and $h_2^j(n)$ represent the value of two sub-window histogram's jth bins at frame n respectively; L denotes the number of bins.

4 CLASSIFICATION ALGORITHM

4.1 Feature Extraction

The classification process consists of two levels. The first level is used to discriminate the pure speech and non-pure speech by setting threshold simply. The second level adopts MGM to classify the non-pure speech based on six kinds of features.

On the first level of classification, we adopt Energy Change Ratio (ECR), Silence Ratio (SR) and Spectrum Entropy(SE) as basic features. Its mathematical description can be described as follows.

$$ECR(k) = \sqrt{\frac{\sum_{i=1}^{N-1} \left(\frac{E_k(i) - E_k(i-1)}{\max(E_k(i), E_k(i-1))}\right)^2}{N-1}} \quad (5)$$

where $ECR(k)$ is average energy change ratio of the kth time clip; $E_k(i)$ denotes the ith frame energy of

the *kth* time clip; And $\max(x, y)$ denotes the maximum of x and y.

$$SR(k) = \frac{nSilenceFrms(k)}{nTotalFrms(k)} \quad (6)$$

where $SR(k)$ is silence ratio of the *kth* clip; $nSilenceFrms(k)$ denotes the number of silence frames in the *kth* clip.

$$SE(n) = \sum_{f=0}^{Fm} P_f(n,f) * \log_2(P_f(TFD(n,f))) \quad (7)$$

where

$$P_f(TFD(n,f)) = \frac{TFD(n,f)}{\sum_{f=0}^{Fm} TFD(n,f)} \quad (8)$$

where, $TFD\ (n, f)$ represent the energy of the signal at time frame n and frequency index f; and F_m refers to the maximum frequency.

On the second level of the classification, the features used in work are modified from those describe in [4-5].

Spectrum flux

$$SF = \frac{1}{(N-1)*(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta)]^2 \quad (9)$$

Spectrum Centroid:

$$f(n) = \frac{\sum_{f=0}^{Fm} fTFD\ (n,f)}{\sum_{f=0}^{Fm} TFD\ (n,f)} \quad (10)$$

Band-Width:

$$B(n) = \sqrt{\frac{\sum_{f=0}^{Fm} abs\ (f - f(n)TFD\ (n,f))}{\sum_{f=0}^{Fm} TFD\ (n,f)}} \quad (11)$$

HZCRR (High Zero-Crossing Rate Ratio):

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} [\mathrm{sgn}(ZCR\ (n) - 1.5 avZCR\ ) + 1] \quad (12)$$

$$avZCR = \frac{1}{N} \sum_{n=0}^{N-1} ZCR\ (n) \quad (13)$$

LSTER (Low Short-Time Energy Ratio):

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} [\mathrm{sgn}(0.5 avSTE - STE\ (n)) + 1] \quad (14)$$

$$avSTE = \frac{1}{N} \sum_{n=0}^{N-1} STE(n) \quad (15)$$

where, STE(n) denotes the short-time energy of nth frame.

Spectrum Entropy [see formula 7-8]

Since the dynamic ranges of these features differ a lot, we normalize them by their standard deviation, which are computed based on the training data.

4.2 MGM classifier

Compared to traditional Gaussian model (GM), MGM has modified the drawback that all features are of the same weight. The MGM classifier consists of two processes: model training and similarity measurement.

Model Training

Firstly GM of each dimension feature is created for each audio class by statistics. Then the model parameters, μ_m^i and σ_m^i, can be calculated. Here μ_m^i and σ_m^i represent the mean and variance of the *mth* dimension feature of audio class i respectively.

Based on μ_m^i and σ_m^i of all audio classes, we can calculate the weight of all features. For a feature weight ψ_m, this can be calculated by the following equation:

$$\psi_m = \frac{w_m}{\sum_{i=1}^{M} w_m} \quad (16)$$

where,

$$w_m = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\mu_m^i - \mu_m^j|}{\sum_{i=1}^{N} \sigma_m^i} \quad (17)$$

$$\mu_m^i = \frac{1}{N} \sum_{j=0}^{N-1} f_m^i(j) \quad (18)$$

$$\sigma_m^i = \sqrt{\frac{\sum_{j=0}^{N-1} (f_m^i(j) - \mu_m^i)^2}{N}} \quad (19)$$

$$m = 1,2,3,....M$$

where $f_m^i(j)$ denotes *jth* feature value of the *ith* dimension feature of audio class m. M is the dimension of the features, N is the number of audio classes.

Similarity measurement

Based on the feature vector $f = [f_1, f_2, ..., f_M]$, the similarity distance between feature f and class i can be calculated by the following equation:

$$i = \max_i(\theta_i) \quad (20)$$

where

$$\theta_i = \sum_{m=1}^{M} \psi_m \eta_m^i \quad (21)$$

$$\eta_m^i = \frac{1}{\sqrt{2\pi}\sigma_m^i} e^{-\frac{(f_m - \mu_m^i)^2}{2\sigma_m^{i\,2}}} \quad (22)$$

where θ_i represent the similarity between the feature vector f and audio class i.

5 EXPERIMENTS

We conducted a series of experiments based on proposed audio segmentation and classification approach. The experimental platform we used is a workstation Pentium4 2.4G CPU, 256M memory. The performance was evaluated on the recordings of real TV program.

The segmentation and classification results were evaluated by the recall rate δ, accuracy rate ξ, and average precision η. These are defined as

$$\delta = \frac{the\ number\ of\ correctly\ objects}{the\ number\ of\ objects\ that\ should\ be\ correct} \quad (23)$$

$$\xi = \frac{the\ number\ of\ correctly\ objects}{the\ number\ of\ all\ get\ objects} \quad (24)$$

$$\eta = \frac{\delta * \xi}{0.5*(\delta + \xi)} \quad (25)$$

We randomly picked out 6 hours TV news program from CCTV1 (China Central Television Station Channel 1) as test set of audio stream. With MBCR-based segmentation, the testing audio stream was split into about 12000 audio clips. The segmentation accuracy was calculated by the alignment of segmenting result and human examination. When the background sound change acutely or there is a silent interval of more than 0.6s, we consider there should be a skip point. As a result, the segmentation accuracy is 92.5%, recall ratio is 93.2%, and the average precision is 92.85%.

We pre-defined six categories as audio classes, which is pure speech (PS), pure music (PM), song (S), speech with music (SWM), speech with noise (SWN) and silence (SIL). Table1 and Tabl2 give the first level and the second level classification results.

Table1 gives the result of first level classifying. It firstly puts the audio clips into three classes, pure speech, silence and others. The others is further classified into four classes: speech with music, song, pure music, and speech with noise, in the second level classifying. The result is shown in Table 2.

Since the performance of classification is partially dependent on the result of segmentation, we also give the results based on simple equal time segmentation in comparison with MBCR-based segmentation. In the mean time, we also compared MGM with traditional GM approach.

The results showed that MGM classifier with MBCR-based segmentation can achieve an average precision of about 89%. It outperforms GM classifier with either equal time segmentation or MBCR-based segmentation.

Table1: The results of first level classification

Algorithm	**Audio Type**	**Accuracy**	**Recall**	**Precision**
Equal Time(2s)	Pure speech (PS)	85.15%	87.52%	86.32%
	Silence (SIL)	97.10%	86.14%	91.29%
	Others	77.95%	95.08%	85.67%
MBCR	Pure speech (PS)	91.33%	93.65%	92.47%
	Silence (SIL)	98.22%	92.97%	95.52%
	Others	85.68%	95.45%	90.3%

Table2: The results of the second level classification

Algorithm	**model**	**type**	**Acc. (%)**	**Rec. (%)**	**Pre. (%)**
Equal Time (2s)	GM	SWN	44.55	72.77	55.2
		SWM	65.12	56.72	60.6
		S	61.05	86.56	71.6
		PM	79.86	61.15	69.2
MBCR	GM	SWN	71.42	74.3	72.8
		SWM	69.43	74.19	71.7
		S	84.88	89.46	87.1
		PM	85.31	82.54	83.9
	MGM	SWN	83.08	87.48	85.2
		SWM	73.28	83.3	77.9
		S	95.44	93.78	94.6
		PM	97.1	86.93	91.7

Here, SWM: speech with music, SWN: speech with noise

6 CONCLUSION

In this paper, we have introduced a MBCR-based segmentation and MGM hierarchical classification approach for audio stream scene processing. The experimental results reported here are meant to show the promise of applying the method in AV stream scene classification. Based on this work, we will further explore more robust features and modeling approaches for finer column classification on TV program.

7 ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (NSFC) under the grant No. 60475014 and National Hi-tech Research Plan under the grant No. 2003AA115520 & 2005AA114130..

REFERENCES

[1] Scheirer E., Slaney M. (1997). "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator", Proc. of ICASSP97, vol II,pp 1331-1334,April 1997.

[2] Williams G. and Ellis D. (1999), "Speech/music discrimination based on posterior probability features", Proc European Conf. on Speech Communication and Technology, Budapest, Hungary, pp.687-690,Sept.1999.

[3] Srinivasan,S., Petkovic,D., Ponceleon, D. (1999), "Towards robust features for classifying audio in the Cue Video System". Proc. of the 7th ACM Intel. Conf. on Multimedia'99, pp.393-400,1999

[4] Kimber D. and Wilcox L. (1997), "Acoustic segmentation for audio browsers". Computing Science and Statistics. Vol.28. Graph-Image-Vision. Proceedings of the 28th Symposium on the Interface. Interface'96, (295-304),1997.

[5] Lu, L., Zhang, H.-J., Li, S. (2003), "Content-based audio classification and segmentation by using support vector machines", Multimedia Systems 8:482-492(2003).