

Two Level Discriminative Training for Audio Events Recognition in Sport Broadcasts

Konstantin Biatov

Fraunhofer Institute Intelligent Analysis and Information Systems

Abstract

In this paper, two level discriminative learning for audio events recognition in sport broadcasts archive is described. The audio events recognition is based on the idea that audio events are composed of basic units. Basic units are some elementary events. Audio events used for semantic interpretation (mid-level concepts) are presented as a combination of the basic units. Models for the basic units are GMM models. Each 5 frames of audio data are recognized using models of the basic units. Each mid-level concept is described by the distribution of the basic units. The distribution of the basic units in each class of segment corresponding to mid-level concepts is considered as a macro model of this class. For events recognition the tree based framework is used. In each level of the tree two macro models are compared. The two level discriminative learning for macro models is applied. First discriminative training level is on the level of basic units, second is on the level of macro models. The suggested approach is compared with maximum likelihood decision and SVM with polynomial kernel. The results of experiments indicate significant improvement in comparison with the conventional approaches in the task of acoustically closely audio events recognition.

1. Introduction

In multimedia content analysis audio events recognition is an important source for semantic content interpretation. Some research has been done on audio events detection and recognition in sport broadcasts [1], [2], [3] and [4]. Mostly semantic analysis of the sport recording is based on video features. However audio is also an important source for sport events interpretation and understanding. In [1] an extraction of audio highlights from three sports, such as, baseball, golf and soccer is presented. In the experiments the following highlights are recognized: applause, ball-hit, cheering, music, speech and speech with music. In [2] a framework called audio keywords is presented. In soccer, for example, the following audio keywords could be used: long-whistling, double-whistling, excited commentator speech, excited crowd sounds, etc. Each keyword can correspond to one or other semantic events. For example, long whistle in soccer can correspond to the start of a free kick, penalty kick, game start or end, etc.

Some interesting results were obtained in the area of speaker verification, which could be applied to audio events detection. In [5] experiments to characterize a speaker using long term speaker characteristics are described. Long term characteristics are the speaker entropy calculated from N-grams of phonemes. In [6] the technique for speaker verification based on short events is suggested. These events don't use a priori information about the types of basic acoustic events such as phonemes.

In the presented paper the approach that uses micro events distribution for semantic mid-level descriptors is exploited. The event recognition from light athletics disciplines presented in sport broadcasts archive, in particular audio events recognition in disciplines such as high jump, pole vault, long jump and running are considered. In the list of such events could be included applause, support, the immediate crowd reaction after jumps, starting signal, stadium background noises, commercials, speech and non-speech. Usually in sport audio recording in the stadium some parallel events are presented including commentator speech, radio advertisement in the stadium, music, etc. Audio events should be detected from the mixture of the sounds in a highly noisy environment. Since the audio events appear mixed with other sounds we suggest to describe these as a distribution of the basic micro units. The basic units are relatively stable short sounds that can be

used to describe more complex real audio segments corresponding to particular mid-level concepts such as applause, support, etc. We also suggest to use two step discriminative learning both for basic units and macro models to improve accuracy of audio events recognition.

The next section describes overall approach for audio events recognition. In the section 3 the discriminative training approach for basic units and macro models is presented. In the section 4 overview of the system and in the section 5 experiments are described. Finally the conclusion is presented.

2. Algorithm Overview

In this paper the recognition of such type of audio events as applause, support, stadium background noises, advertisements, immediate reaction of the stadium to the sport discipline completion is described. Part of these events such as applause, support, reaction corresponds to mid-level concepts used for further semantic sport scene interpretation. Often sport broadcast recordings, in particular, light athletics competitions include some parallel events and the commentator periodically switches from one event to another. Usually audio segments labeled by hand, for example as support event is not really homogeneous events and could includes applause, stadium background, music and etc. We suggest to describe audio segments by using a combination of basic units presented in broadcast audio recordings. The basic units are considered as short relatively pure audio events that could appear as a mixture of sounds that corresponds to mid-level concepts. In the list of basic events speech, music, silence, applause, reaction, sounds of crowd, noise, stadium background noises are included. The model for each basic unit is presented as Gaussian Mixture Models (GMM). Each model is trained using Expectation Maximization (EM) and each basic unit model has 64 GMM. Each frame of the audio is recognized using GMM models of basic units and then the results of recognition are averaged within each 5 frames. Then each 5 frames are labeled by one of the basic units having the highest score. In the training phase the basic unit distribution for each set of segments corresponding to the mid-level concepts is evaluated. Let us denote $e_1, e_2, e_3, \dots, e_n$ as the basic audio units. The value of $\text{count}(e_i / X_j)$ refers to the number of basic unit e_i in segment X_j . Relative frequency of the event e_i in segment X_j is calculated as:

$$p(e_i / X_j) = \frac{\text{count}(e_i / X_j)}{\sum_{k=1}^T \text{count}(e_k / X_j)} \quad (1)$$

Each segment is described by the sequence of relative frequencies of all basic units. The vector describing the segment is presented as:

$$(p(e_1 / X_j), p(e_2 / X_j), p(e_3 / X_j), \dots, p(e_n / X_j)), \quad (2)$$

where n is the number of different basic units. In the training phase for each class we calculate such vector using training audio data corresponding to this class. Each audio class is described by the vector of the basic units distribution.

For audio events the ideas described in [5] are exploited. Let us denote test segments as $Y_1, Y_2, Y_3, \dots, Y_k$, where k – number of segments.

Each test segment Y_i under condition of the event class j is described as:

$$P(Y_i / j) = \sum_{k=1}^T p(e_k / Y_i) \log p^j(e_k / X_j), \quad (3)$$

where T is the number of events, $p^j(e_k / X_j)$ - distribution of basic unit for event class j . For two class j_1 and j_2 comparison the decision rule for acceptance j_1 is $P(Y_i / j_1) - P(Y_i / j_2) \geq \Delta$

and for rejection is: $P(Y_i/j_1) - P(Y_i/j_2) < \Delta$, where Δ is the predefined threshold.

We adopt hierarchical top-down approach to model events using distribution of basic units. For recognition improvement for each level of the tree discriminative learning is applied.

3. Discriminative Training

Previous works in the speech recognition and speaker recognition have shown that combining discriminative and generative training can substantially improve performance of the recognition [8]. The discriminative training is applied to update models previously obtained.

In this paper two step discriminative training is described. In the first step means of the Gaussian models for basic units are updated. In the second step on macro level the relative frequency of basic units that are considered as the description of the class are updated.

Misclassification measure for two classes i and j and for training data X_i and X_j where X_i are the positive examples for class i and X_j are the negative examples for class i is defined as:

$$d(X, i, j) = -(\log(P(X_i/i)) - \log(P(X_i/j)) - \Delta) + (\log(P(X_j/j)) - \log(P(X_j/i)) + \Delta) \quad (4)$$

The loss function with respect to the input data is defined in the term of misclassification measure:

$$F(X, i, j) = \frac{1}{1 + e^{-d(X, i, j)}} \quad (5)$$

Discriminative training is performed by obtaining the derivative of $d(X, i, j)$ with respect to models parameters and then applying gradient descent method to all training data for both classes to mutually update parameters of the concurrent models.

In the first step of the discriminative learning we update the means of Gaussian probabilities of the basic units. Let denote μ_{kl}^t the mean of the l component of the mixture k of basic unit t . We used derivative of the loss function with respect to means of mixtures of basic functions. The means are updated using following formula:

$$\widetilde{\mu}_{kl}^t = \mu_{kl}^t - \varepsilon (F(X, i, j)(F(X, i, j) - 1)) \frac{\partial d(X, i, j)}{\partial \mu_{kl}^t} \quad (6)$$

$$\frac{\partial d(X, i, j)}{\partial \mu_{kl}^t} = \frac{1}{N(i) + N(j)} \left(\sum_i \frac{(x_{1l}^i - \mu_{kl}^t)}{\sigma_{kl}^2} - \sum_j \frac{(x_{1l}^j - \mu_{kl}^t)}{\sigma_{kl}^2} \right) \quad (7)$$

where x_{1l}^i is l component of the positive examples of X_i , x_{1l}^j is l component of the negative examples of X_j , $N(i)$ number positive examples for GMM model corresponding to the unit i and $N(j)$ number of negative examples corresponding to the unit i .

In the second level we update the frequencies of basic units distributions obtained in the training phase. These frequencies can be considered as the weights.

$$P(Y_i/j) = \sum_{k=1}^T p(e_k/Y_i) \log w_k^j \quad (8)$$

The misclassification measure is defined as:

$$d(X, i, j) = -((P(X_i / i) - P(X_i / j) - \Delta) + (P(X_j / j) - P(X_j / i) + \Delta)) \quad (9)$$

$$\tilde{w} = w - \varepsilon \frac{\partial F(X, i, j)}{\partial w} \quad (10)$$

$$\tilde{w} = w - \varepsilon (F(X, i, j)(F(X, i, j) - 1)) \frac{\partial d(X, i, j)}{\partial w} \quad (11)$$

$$\frac{\partial d(X, i, j)}{\partial w_k} = \frac{1}{w_k} \left(\frac{1}{N(i)} \sum p(e_k / Y_i) - \frac{1}{N(j)} \sum p(e_k / Y_j) \right), \quad (12)$$

where $N(i)$ number positive examples for macro model i and $N(j)$ number of negative examples for macro model i .

4. System Description

For primary features 15 mel-cepstral coefficients plus energy are used. The audio analysis is conducted using 30 msec analysis window with the 10 msec step. Before audio events recognition audio segmentation is carried out. For segmentation the Bayesian Information Criterion (BIC) is used. Segmentation via BIC was initially proposed in [7]. In general BIC is defined as

$$BIC(M) = \log L(X, M) - \lambda \frac{\#(M)}{2} \log(N), \quad (13)$$

where $\log L(X, M)$ denotes segment X likelihood given by model M , N is the number of data points, $\#(M)$ is the number of free parameters of the model and λ is a tuning parameter.

The models corresponding to the basic units are trained via EM algorithm using 2 hours of labeled data from sport broadcasts. The models of basic units are GMM. Then macro models for each mid-level concept are trained using the same data via the algorithm described in section 2. In the training step relative frequencies of the basic units distributions are evaluated.

The event recognition is carried out in a series of steps: feature extraction, audio segmentation, segments classification.

The algorithm described in section 2 is compared with the conventional maximum likelihood decisions and SVM. Each frame is labeled using maximum likelihood decision rules. Then using the voting rule the results on the frame level are averaged within each segment to get segment label.

5. Experiments

In experiments Eurosport broadcasts audio data that includes European Light Athletics Championship in Goteborg in 2006 is used. The two hours of data are used for training and 1.5 hours data of are used for testing. For training and testing purposes all data are segmented and labeled using mid-level concepts such as applause, support, crowd reaction, commercials and stadium background noise.

The most complex for recognition pair of events is support and applause. There are very similar events. Support is periodic applause that is used by crowd to support sportsmen before start. Applauses usually appear after completion of sport discipline. Applauses are not so emotional and have longer duration in comparison with the immediate reaction of the crowd after completion of sport discipline. Immediate reaction could be positive as exult and negative as disappointment. For evaluation of the training approaches for pair of events (applause, support) is supposed that the boundaries between segments are recognized correctly. The total duration of applause test data is

700 sec. The total duration of support test data is 1056 sec. The training data for these two events have approximately the same duration.

In the first experiment 64 mixture GMM are trained for applause and support. These are used for applause and support separation. Then these two models are updated using discriminative training for GMM. Only means are updated. The results are presented in Table 1. The experiments show that generative models after discriminative training give improvement for considered audio events.

In the next experiment SVM with polynomial kernel using the same as for GMM training support and applause training data is trained [9]. The results are presented in the Table 2. The results of recognition using SVM are better than the results obtained using generative models without discriminative training. Discriminative training of GMM performs better than SVM with polynomial kernel.

Table 1. Two events recognition on the level of basic units without and with the discriminative learning

	GMM models without discriminative training	GMM with discriminative learning
Applause segments	58% correct	61% correct
Support segments	84% correct	92% correct

Table 2. Comparison with SVM classifier

	SVM with polynomial kernel
Applause segments	63.7 %
Support segments	81.6%

Finally macro models are trained. Then macro models are updated in the following way. Basic units GMM models corresponding to macro models of applause and support are updated using discriminative training for GMM. Then macro models using discriminative training for macro models are also updated. The results of experiments using macro models without and with the discriminative training are presented in Table 3.

Table 3. Two events recognition using macro models without and with the discriminative learning

	Using macro models without discriminative learning	Macro models with two step discriminative learning
Applause segments	62% correct	75% correct
Support segments	86% correct	86% correct

The experiments show that macro models with the two level discriminative training performs better than other tested models such as SVM with polynomial kernel and GMM models updated by discriminative training.

6. Conclusion

A macro model technique that exploits two levels discriminative training is presented. It employs distributions of basic units, short elementary audio units, for sport events (mid-level concepts) description. A comparative evaluation of macro model based recognition with several classifiers and also comparative evaluation of two step discriminative learning with discriminative learning for GMM were performed on the task of audio events recognition in sport broadcasts archive. The macro model approach in combination with two step discriminative learning demonstrated the highest accuracy. The audio events (mid-level concepts) recognition is intended as a part on multimedia information extraction system in sport domain.

7. Acknowledgments

This work was done in context of BOEMIE project “Bootstrapping Ontology Evolution with Multimedia Information Extraction”.

References

1. Bowman, B., Debray, S. K., and Peterson, L. L. Audio events detection based highlights extraction form baseball, golf and soccer games in unified framework. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, China, April 6-10, 2003.
2. Min Xu, Namunu C. Maddage, Changsheng Xu, Mohon Kankanhalli and Qi Tian, Creating audio keywords for events detection in soccer video. In Proceedings of the International Conference on Multimedia and Expo (ICME), 2003.
3. Lu L., Cai R. and Hanjalic A. Toward a unified framework for content-based audio analysis. In Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP), Philadelphia, March 18-23, 2005.
4. Zhang D. and Ellis D. Detecting sound events in basketball video archive. Dept. Electronic Eng., Columbia Univ., New York, 2001.
5. Doddington G. Speaker recognition based on ideolectical difference between speakers. In Proceedings of Eurospeech, Aalborg, Denmark, 2001.
6. Scheffer N. and Bonastre J-F. Speaker detection using acoustic events sequences. 9th European Conference on Speech Communication and Technology (Interspeech'2005-Eurospeech), Lisbon, September, 4-8, 2005.
7. Chen S., and Gopalakrishnan P. Clustering via the Bayesian Information Criterion with the applications in speech recognition. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, Washington, USA, May 12-15, 1998.
8. Juang B.-H. and Katagiri S. Discriminative learning for minimum error classification”, IEEE Trans. Signal Processing, 40:3043-3054, 1992.
9. Joachims T. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. B. Schölkopf, C. Burges and A. Smola (ed), MIT-Press, 1999.