NEW APPROACH TO AUDIO SEARCH BASED ON KEY-EVENT DETECTION PROCEDURE

Igor E. Kheidorov, Peter D. Kukharchik and Yan Jingbin
Belarussian State University, 4, Nezalezhnosti av., Minsk, Belarsu, Belarus

Keywords: Audio search, Keyword spotting.

Abstract: In this paper a new approach to audio search in multimedia data bases is developed based on key-event detection procedure. The main idea of the proposed approach is to present each audio fragment as sequence of context dependent "key-events", specially determined or taken from the data base. Wavelets and support vectors method were used as the basis for the created procedure of key-events selection and classifier training. The experiments show good results and perspectives for the proposed approach.

1 INTRODUCTION

Information flow grows year by year, and as a result multimedia databases in different applied areas are already collected and ready for indexing. Computational capabilities of modern signal processing devices become higher and higher, and new information technologies allow the processing of not only text and image information but also video, speech and audio signals. Managing the huge amount of data leads to some difficulties connected with information storing and retrieving in image databases. In order to solve the data managing problem new instruments for indexation, searching and visualization were developed (Chelba, Silva and Acero, 2007). The first system suitable for indexation and searching in large image databases was developed in the end of 1990th. Modern systems support indexation, retrieval and displaying of different types of multimedia content, such as audio, video, and some types of images, but up to the moment there is no perfect universal system able to index any type of audio with high accuracy (Chen, 2008).

In the paper the new approach for universal indexing system creation is proposed based on "key-event" detection ideology. The section 2 is devoted to the introduction to the "key-event" method, section 3 gives an idea of "key-event" creation procedure, in section 4 there are experimental results for music genre recognition and keyword search tasks.

2 KEY-EVENT IDEOLOGY

Audio data indexing can be considered as a typical recognition task. The main problem is to find the procedure, which minimizes average risk when having a finite set of data, i.e. the procedure which that minimizes empirical risk. The success in indexation system creation is connected with development and implementation of new fundamental mathematical methods, algorithms and schemes which allow the modelling of acoustic signal as a structure with several parallel existing conditions. The most effective are the graphical probability models, hierarchical hidden Markov models (HMM) (Wilpon, Rabiner etc., 1990.), neural networks (NN), Support Vector Method (SVM).

The main problem of modern audio indexing systems is to find the appropriate feature vector for the signal. Characteristics for each type of signal differ each other for music, speech, noise sounds, etc, and this fact essentially complicates the indexing procedure. In order to decrease the calculation score and construct fast search algorithm it can be used the multi-stage procedure which include several one-by-one stages: speech/music/noise/silence discrimination, speech language recognition, determination of speaker changing points, male/female speech distinguishing, etc.

Such approach provides good and practically acceptable results but has the drawbacks:

Kheidorov I., Kukharchik P. and Jingbin Y. (2009).
NEW APPROACH TO AUDIO SEARCH BASED ON KEY-EVENT DETECTION PROCEDURE.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 33-37
DOI: 10.5220/0002185900330037

- The error caused on upper processing stages leads to the total error of the search procedure. For example, if the speech fragment was falsely recognized as the music one the keyword search will never even start.
- In a number of cases it is very hard to distinguish speech from music on the preliminary stage because of their similarity on the most parameters (for example, Chinese speech);
- In the real system we need to pass throw all stages of analysis to find the required audio fragment not depending on its context. In the case of speech indexing different words have different features with different distinction power, and this apriory information has to be used for fast search. For example, it very hard to find very short words like "yes" or "no", but not a problem to find the specific long word "synchrophasotron".

In order to overcome these drawbacks there were developed main approaches of content-sensitive adaptable system, suitable for direct indexing an search of different type of information in audiodocuments.

In order to introduce "key-event" method lets do three assumptions about audio signal. The first one is that wavelet-image of audio contains all significant information about time and spectrum-based signal features (Kukharchik, Kheidorov etc. I.E., 2008). All features like formants, fundamental frequency, cepstrums, formats, short energy parameters, etc. can be calculated using wavelet image of audio, and can be considered as transformations of wavelet image. These means that we suppose that all necessary information (including time dependences) about audio can be obtained based on wavelet.

The second assumption is that any audio fragment can be described as a set of specially determined "events", each event presents acoustically significant property for the certain audio part. The event has to be an important feature which is typical for the specific type of audio objects. In the broad sense such acoustical events (we will call them "key-events") presents the audio fragment semantics, and can be used for the direct description and indexing of audio. Specially selected "key-events" can be transformed into indexes suitable for storing in data bases.

The third assumption is that SVM classifier can be trained to model any selected event in audio signal with different accuracy. In other words, it is supposed the SVM kernel can approximate any feature parameter (transformation) of audio signal.

The developed and proposed "key-event" search method includes two main stages: audio content dependent "key-event" creation for target audio, and target audio search based on its "key-events" presentation. Each target audio has to be found by looking for the key events appearance in a continuous audio stream. The "key-event" search can be done in the order of calculation complexity growth, i.e. the simplest "key-event" has to be found in the first order, then the second one and so on. Each next search can be done within the framework of the previous search results, and the final decision has to be done by specially trained decoder (for example, HMM). The search scheme is presented on fig.1.

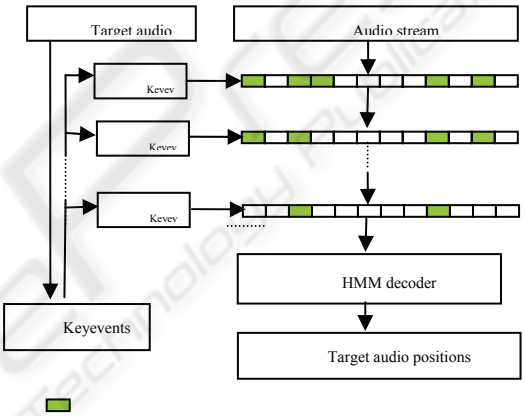

Figure 1: Audio search scheme based on key-event ideology.

The main advantage of the proposed approach is that we try to find the feature vector (which presents the "key-event") mostly suitable and typical for the given content. This activity lets us to describe audio in the terms of features of the specific audio, not by common features like cepstrums and others. Such approach allows usage of a priory audio information as much as possible, and this information is to be presented by "key-events". The number of "key-events" can differ for different audio fragments. Of course, the procedure of "key-event" selection for each audio is very complicated and time consuming, but it can be iteratively. The most successful key-events can be stored in the data base as the basis for new "key-events". Time by time the "key-event" data base will be enlarged and finally will contain the most of possible events for audio, and the new target audio can be easily described in terms of "key-events". Different "key-events" can be joined in an arbitrary order to form the semantic description

of target audio in human comprehensible terms. genre recognition and keyword search tasks.

3 SUPPORT VECTOR MACHINES FOR KEY-EVENTS DETECTION

As it was noticed earlier the idea of "key-events" is very powerful from the viewpoint of direct semantic audio analysis. The "key-event" detection procedure can provide the rough semantic context of audio. The main problem of "key-event" search ideology is to find the appropriate "key-events". In this paper the "key-event" search idea is developed and introduces on two sample tasks: keyword spotting, and music recognition.

For keyword search task at the singular case the phonemes can be taken as "key-events", and the key-word spotting procedure is typical for this task and based on automatic speech recognition (ASR) engine. But the presentation of speech as sequence of phonemes is not an optimal one because of several reasons. First of all, a lot of significant information which is contained on phoneme boundaries can be lost. The second reason is that usage of ASR needs full presentation of speech as a sequence of phonemes, such presentation is abundant for keyword search procedure.

Its main idea is to substitute the term "keyword" by the term "key-event", which gives number of points with the maximum probability of keyword appearance. The "key-event" can include a series of features from wide range of characters with different discrimination power. For example, the "key-event" can be an appearance of a certain pair of - frequencies (in tonal dialing, for example), as well as the form or behavior of speech first formant (quick grow, for example). In this case the keyword search can be done as a sequential search of key events, typical to the given keyword. One key word can be characterized by several key events, which has to be detected in an order reverse to their complexity.

In order to create a set of simple key-events it was proposed to use SVM technique. There are several reasons to use SVM for this task. The principle distinguish of SVM classifiers from common classification methods like HMM or Gaussian models is that SVM directly approximates class boundaries and not describe the probability densities on training set. Within the framework of the proposed approach the SVM will be used in order to determine if the given "key-event" is presented at audio fragment. Two-class SVM (class "yes" and class "no") is suitable for these task.

Let we have a set of training vectors which belong to two different classes, $(x_1, y_1, ..., x_l, y_l)$, where $x_i \in R^n$ and $y_i \in \{-1, ..., 1\}$. The train aim is to find the optimal hyper plane $wx + b = 0$ which divides data and maximizes the distance between this plane and the closest data points for each class. In order to solve this task it is possible to use Lagrange function:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum \alpha_i \{y_i [(wx_i) + b] + 1\}, \quad (1)$$

where α_i - Lagrange multiplier (Krebel U., 1998.). Such approach lets to divide the target task on two sub-tasks which can be easily solved. We can find the solution like:

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha}_i y_i x_i \ , \ \overline{b} = -\frac{1}{2}\overline{w}[x_r + x_s], \quad (2)$$

where x_r x_s - two support vectors, $\overline{\alpha}_r, \overline{\alpha}_s > 0$, $y_r = 1$, $y_s = -1$. In order to solve the non-separable task Cortes и Vapnik involve fictive values $\zeta_i \geq 0$ and penalty function $F(\zeta) = \sum_{i=1}^{l} \zeta_i$, where ζ - measure of false classification. The solution is identical to the case of separable task with the exception of necessity to modify Lagrange multipliers $0 \leq \alpha_i \leq C$, $i = 1, ...l$. The choice of C is not fixed and we will suppose $C = 200$ in all experiments

But in practice data mostly can not be divided by the linear plane, but can be divided in other features space after non-liner transform. For such case SVM with nonlinear discrimination can be used based on kernel functions. There are three main kernel functions of non-linear reflection: polynomial, radial base Gauss function, exponential radial base function.

The "key-event" selection procedure was organized in the following way. Wavelet image of the key-word was analyzed in different scales in order to find the candidates on simple events, typical for the certain word. The selection procedure consists of two stages. The first one is necessary to create a set of candidate features to become a "key-event". The second stage is SVM training and testing for each candidate, in order to select the candidates with the highest discrimination power. The pair "Feature vector-SVM" forms the "key-event". The discrimination power of such "key-event" can be defined as the probability of right recognition with the previously limited probability of false alarm. The right recognition rate is determined for each "key-event" during testing as

the number of event appearances in target audio fragments related to the total number of such fragments. In such manner the probability of each "key-event" for each type of audio is known before the recognition experiment, and forms the observation probability matrix which used by HMM decoder.

In other words, selected "key-events" forms a set of observations $O = \{o_1, o_2, ...o_M\}$, where M - number of "key-events". For target audio for each observation o_m, $1 < m < M$, there is a corresponding probability $P(o_m)$. The main task of SVM classifier is to answer the question: does the "key-event" occur in an audio stream, or not. If the "key-event" o_m was detected it means that the estimated probability of this event is $P(o_m)$. Then these probabilities can be processes by the HMM decoder instead of typical description of observation sequences by Gaussian mixtures.

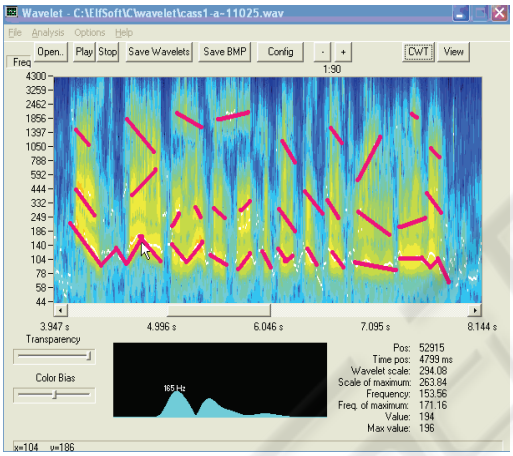

Figure 2: Wavelet image of audio.

One of the great problems within the framework of this approach is to create reasonable candidates to be "key-events". On this stage any apriory information about audio signal structure is welcome. But in general it was proposed the following procedure for creation candidates to be "key-events". Wavelet image of audio (fig.2) is processed in order to form different "time-frequency" cells of different forms and sizes.

In the simplest case the domain divider is a rectangular grid which contains cells with different widths and heights. The relations between energy means inside cells can be considered as feature vector which is able to present not only energy correlations in one moment, but between different time moments.

4 EXPERIMENT

First series of experiments was devoted to genre recognition task. Experiments were conducted in order to prove the idea for four music genres on a small data base. The main efforts were devoted to the search of features and corresponding SVM kernels for different music genres and musicians. Ideally each musician or genre has to be described by his own "features vector-SVM" pair, which reflects its peculiarities. For music signal it was supposed that the most essential information inside the music is the cross-correlations between main spectrum lines (fig.2), this relationships were selected as "key-event" candidates (prototypes). This information together with behavior of spectral lines is the main distinctive feature for music, but the problem is to choose the "key" essential feature for the given music. The experiment shows that the common average error is 2.69 % (table 1). The worst classification results are shown for audio files segmentation which contain classics (accuracy 96.25%) and disco (accuracy 96.53%). Speech segments classification accuracy is 98.3 %. The average accuracy level is 97.31 %, and this fact shows the effectiveness of the proposed segmentation scheme.

Table 1: Music genre recognition result.

	Correct	Critical errors	Non-critical errors
Classics	96,25 %	2,03 %	1,72 %
Pop	97,68 %	1,03 %	1,29 %
Rock	97,6 %	2,17 %	0,024 %
Disco	96,53 %	3,47 %	0 %
Average mean	97,31 %	2,03 %	0,66 %

The second experiment was devoted to the keyword spotting task. In order to make the experiment on the key-event based keyword search the following data base was created. Data base vocabulary contains 520 phonetically balances Russian phrases. This data base was collected at Radiophysics department of Belarussian State University. As a minimal acoustic unit it was chosen a context dependent phoneme-allophone. Full set of allophones describes the whole speech variety. This data base guarantees presence of all types of phoneme and can be used for the adequate estimation of segmentation algorithm. Phonetically balanced database contains 53 speakers, 18 females and 35 males, all phrases are processed by speech –detector and contains speech only. For each phrase the phonetic transcription is known, allophones boundaries are detected using specially developed software and checked manually.

This data base was used to train the phone recognizer based on continuous HMM, each state is connected with the correspondent phoneme. This recognizer was used in order to construct the ASR-prototype and estimate the output phoneme probabilities during the keyword spotting procedure, as well as in order to find the context-dependent key-events mostly suitable for the given key-word.

The key-word detection accuracy of 5, 10 and 20 words for different number of key-events and ordinary phonetic ASR-based searcher are presented in table 2. The used ASR system was based on HMM and trained on the same data base as it was used for key-event search procedure.

Table 2: Key-word search accuracy for different techniques.

	Number of words for search		
Searcher type	5 words	10 words	20 words
Key-events	80,5%	80,2%	79,6%
ASR-based	68,5%	67,2%	66,3%

The False acceptance rate for the same key-word search system is presented in table 3.

Table 3: False acceptance rate for key-word search.

	Number of words for search		
Searcher type	5 words	10 words	20 words
Key-events	3,5%	5,8%	6,2%
ASR-based	26,1%	30,8%	32,2%

5 DISCUSSION

It can be noticed from the experiment results that the proposed context dependent key-events technique provides the higher search accuracy in comparison with ordinary ASR techniques based on HMM. The false acceptance rate is much lower than for another technique because each SVM classifier for the "key-event" was trained in order to minimize FAR.

The main drawback of the method of key-words search based on "key-events" is that it very complicate to chose the suitable "key-events". It is not very reasonable to start the complicated key-event determination procedure for all possible keywords. The idea is that all possible "key-events" suitable for key-word search can be united into common data base which will be used for initial "key-events" estimation in specific cases.

6 CONCLUSIONS

The algorithm of "key-events" creation and usage was proposed for audio sequences indexing and search. The "key-event" is a "feature vector-SVM classifier" pair, and the fixed probability of each event can be determined during SVM testing. This probability was proposed to use as estimation of observation probabilities for HMM-based decoder. The experiments show good results and perspectives for the proposed approach.

ACKNOWLEDGEMENTS

The paper was proposed under the partial support of International Science and Technology Center, project B-1375.

REFERENCES

Chelba, C., Silva J. and Acero, A., 2007. In *Computer Speech and Language, "Soft Indexing of Speech Content for Search in Spoken Documents"*. 21(3), pp. 458-478.

Kukharchik, P.D., Kheidorov, I.E., Bovbel, E.I., Ladeev, D.D., 2008. In *Image and Signal Processing, "Speech signal processing based on Wavelets and SVM for vocal tract pathology detection"*. LNCS 5099, Springer, pp.385-389.

Wilpon, J.G., Rabiner, L.R., Lee, C.H., and. Goldman, E.R, 1990. In *IEEE Transactions on Acoustics, Speech and Signal Processing, "Automatic recognition of keywords in unconstrained speech using Hidden Markov Models"*.

Chen, B., Chen, Y.T., 2008. In *Pattern Recognition Letters, "Extractive Spoken Document Summarization for Information Retrieval"*. 29(4), pp. 426-437.

Krebel U., 1998. *Advances in Kernel Methods: Support Vector Learning.* MIT Press.