MOBILE VIDEO CAPTURE TARGETED NARROWBAND AUDIO CONTENT CLASSIFICATION

Satu-Marja Mäkelä, Johannes Peltola, Mikko Myllyniemi

VTT Electronics, P.O.Box 1100, FI-90571 Oulu, Finland
phone: +35885512111, fax: +35885512320,
email: Firstname.Surname@vtt.fi

ABSTRACT

Audio content analysis and automatic content management research field is facing new challenges from personal home video material created with mobile camera phones. We developed an audio content analysis and segmentation system that is robust to low sampling rate used in mobile phone camcorder tools and operates also well for AMR compressed data. Audio signal is segmented in five different classes, which can be used to classify videos for mobile video content management applications. The classification was done with Bayesian Networks using topology of four-stage binary tree network that works in a hierarchical manner. Results were further smoothed within three second window using weighted sum of Bayesian networks class probabilities. The results show an average recognition accuracy of 88.6% for the high quality audio, 88.6% for the narrowband audio and accuracy of 82.9% for an AMR compressed audio. The effect of sampling rate to the recognition results is not significant and the effect of the AMR compression is also relatively low.

1. INTRODUCTION

Many of the recent research on audio content analysis and segmentation concentrate on material from news archives, digital libraries and TV programs/movies [1], [2], [3]. In above applications audio material is professionally created and edited, it typically also contains certain application specific prior information which may guide the audio analysis tools. So far we have not been able to find a publication approaching the audio content analysis from the mobile video point of view.

The content management of a home made personal video material is also a less researched area and it brings new challenges for audio analysis tools. The personal home videos are created with camcorders or camera phones which do not have as good recording properties as professional equipments. This will degrade the quality of video material. Also the amateur users will typically not create a clear structure for their videos; the recording runs are long with no shot partition. For audio analysis the use of camera phone will additionally increase the challenge. The video clips are very short and the audio quality is typically worse than in camcorder videos. Currently used sampling rate in mobile video capture tools is as low as 8 kHz and the microphone is optimised for speech. Also the audio is processed with speech specific gain control and quality enhancement and finally the audio is AMR (Advanced Multi Rate) compressed to be included into the .3gp video clip. An AMR codec is ACELP (Algebraic Code Excited Linear Prediction) based speech codec specifically designed for narrowband telephony communications, but it is also used as an audio codec for current mobile phone camcorder tools.

The content of the video material might also vary between the camera users according to personal interest. In [4] they studied the use of camera phones where users were asked to take photos or video clips. They found out that the most general content type was people (51%) and it was followed by a category of specific things (32 %). Also Zhao et al. came into conclusion that personal video material would contain lot of human related material and thus human made sounds [5]. This makes them to believe that audio analysis can provide important metadata which cannot be revealed by visual features. Based on this assumption we have developed audio classification tools that are targeted for mobile video content management.

Typically audio content is classified first into a basic set of main audio categories. Generally these classes contain speech, music, silence and additionally different noise classes or mixed classes, i.e. speech with music. Two typical approaches are hierarchical rule-based classification and statistical classification. An example of the rule-based classification is in [6] where they classified seven basic audio types achieving classification accuracy over 90% for audio from movies and TV programs and movies. A kNN classifier was used together with hierarchical classification system in [1]. They achieved an accuracy of 96.51% for three classes for news material.

We used Bayesian Networks, which have earlier used successfully in the context aware audio classification [7]. Bayesian Networks is also used in the audio analysis scheme by Lu et al. where they used it for the classification of audio effects to semantic contexts, but not to classify low-level features to audio classes [8].

In this paper a development and comparison of performance of the audio content analysis classifier for high quality audio, narrowband 8 kHz sampled audio and AMR compressed audio

is presented. The purpose of the classifier is to provide tools that may operate simultaneously for the high quality camcorder home video material and the camera phone recorded mobile video material. The audio is classified into five different basic classes; speech, music, constant noise (CN), environmental sound (ES) and silence. The feature set contains temporal and spectral features which are classified using Bayesian Networks. The classification is performed in one second windows thus audio track is also segmented into these 5 audio classes. The classification results are further smoothed in three second window using weighted sum averaging of the probabilistic outputs from the network.

2. FEATURE SET

This section describes the selected features. The selection criteria for the features was that they should provide a good discrimination performance for audio with low and high sampling rates and also offer robustness against AMR compression. The energy features - power variance, low energy ratio and energy onset - are simple to calculate but can separate some of the classes efficiently. The spectral features and harmonic ratio have been selected from the MPEG-7 standard [9].

Power variance is the variance of average frame log power values from the past one-second interval.

Low energy ratio (LER) is the ratio of frames with average power less than a pre-defined threshold (here 20%) of the mean of the frames in the past one-second interval.

Energy onset is the number of valid peaks in audio signal energy contour during two second window. Algorithm calculates first the derivative signal from the low frequency energy contour. Energy peaks or onsets are detected if the energy in the audio signal between two zero crossing in the derivative signal grows more than predefined threshold and the duration of the onset is more than 96 ms (3 frames).

Harmonic ratio is loosely defined as the ratio of harmonic power to total power within the signal. The feature is calculated as the maximum of normalized cross-correlation of the signal segment and its lagged counterpart. This feature is implemented as in the MPEG-7 standard [9].

The **lag** of the maximized cross correlation is also used as a feature. The lag can be considered as an estimation of fundamental period.

Spectral centroid is a measure of the "brightness" of the signal. This measure is obtained by calculating the center of the gravity of the logarithmic power spectrum in octave bands around 1 kHz. This feature is implemented as in the MPEG-7 standard [9].

2.1 Distribution of features

A distribution analysis for the features in different audio classes can be used to evaluate the discrimination power of the features in heuristic way. We used this method to select the presented features from a larger set of features. Also distributions were used for designing the architecture of the classifier. Fig 1 presents an example of the spectrum centroid histograms for music, constant noise and environment sound for high quality

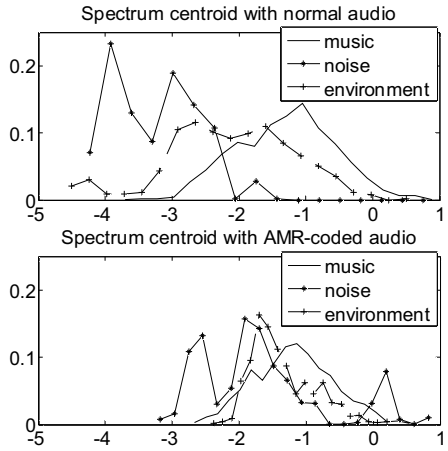

Fig.1. Spectrum Centroid distributions for music, noise and environmental noise for high quality above and AMR coded audio below.

(above) and AMR coded (below) audio. For the AMR coded audio feature distributions are typically quite similar, although the distribution of constant noise and environmental sound are more overlapping in the AMR compressed case.

Based on the distribution analysis we selected four feature sets for recognizing the audio classes. The feature sets are presented in Table 1.

Table 1. Feature selection for different audio classes

Discriminative classes	**Feature set**
Silence/ Other classes	Power variance
Speech/ Music, CN, ES	Low energy ratio
Music/ CN, ES	Power variance, lag, harmonic ratio, energy onset, spectral centroid
CS (ConstantNoise)/ ES (Env. Sound)	Harmonic ratio, spectral centroid

3. BAYESIAN NETWORKS

We used Bayesian Networks to classify the selected audio classes. Bayesian networks are directed graphical models that can model joint probability distributions. Networks consist of nodes that represent random variables and arcs that model the conditional joint probability between the variables. The topology of our classifier was a four-stage binary tree network. First hierarchy level calculates the probability for *silence,* second calculates the probability for *speech*, third the probability for *music* and the last node evaluates the probability for *constant noise* and *environmental noise*. Each hierarchy level takes input from the audio class specific set of features and also from the previous node, if there is one. Fig. 2. presents this network topology.

By designing our Bayesian Networks classifier in a hierarchical way the classifier resembles a rule-based hierarchical classifier that has been used in previous works. The rule-based hierarchical classifiers are powerful since hierarchy levels can be used to exclude or confirm the presence of specific

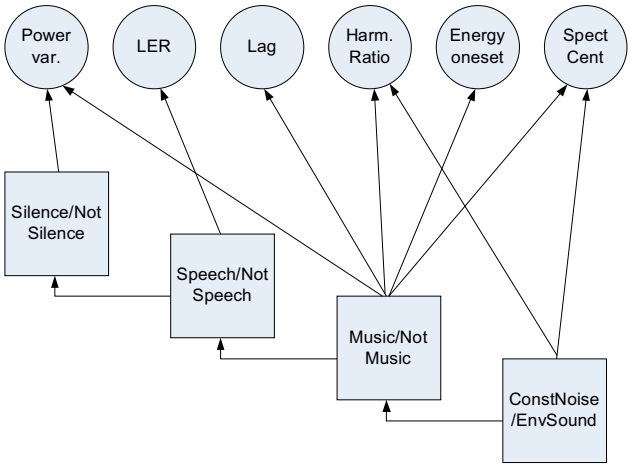

Fig. 2. Topology of the implemented Bayesian Networks

audio classes in the audio track. Going down in hierarchy one by one it's possible to test if audio track consists of any specific class until a predefined acceptance threshold is exceeded and the winner class is selected. Each hierarchy level is a classifier on its own, thus it is possible the select only the most discriminative set of features for each hierarchy level. The negative point with hard decision rule-based hierarchical classifiers is that the classification errors in higher levels prevent the recognition of the audio class located in lower layers.

Our Bayesian Networks classifier has an advantage over rule-based hierarchical classifiers since the classification errors in upper layers do not prevent inference in lower layers. Errors affect only to the class probabilities in lower layers, because there is a conditional dependence between the nodes. Each layer utilizes only those audio features that have the strongest discrimination capability for the particular audio class that the given node represents. We have estimated the discrimination capabilities of the features by analyzing the distribution of each feature for each audio class. The features that had the least overlapping distribution for each node specific audio classes were selected as an input information for given node. The feature sets are presented in Table 1.

4. DATABASE AND EXPERIMENT

4.1 Database

The database contained high quality audio recordings from various sources. High quality database has sampling frequency of 16 kHz. The database was selected so that it contained 5 different audio classes, totalling over 166 minutes of sound. Music samples contain classical, various pop and rock genres as well as instrumental and vocal pieces. Speech samples include female and male speech in Finnish, Swedish, English and German. Constant Noise (CN) samples include "stationary" machine noises from car, train, computer cooler etc. Environment Sound samples include sounds from amusement park, airport check-in, swimming pool, restaurant, shop etc. Constant noise and environment samples were taken from the NTT-AT Ambient Noise Database for Telephonometry 1996 [10]. Silence audio samples were recorded at quiet office environment containing some occasional sounds.

The partition between constant noise and environmental sound classes is somewhat difficult and the sound clips categories contain partly similar material.

The whole database was down sampled to 8 kHz in order to get a narrowband low sampling rate version of the database. The narrowband audio was also AMR compressed with the 3GPP TS26.073 fixed-point version 5.1.0 with a bit rate of 10.2 kbps for testing the effect of AMR coding for the audio analysis. For feature calculation the 8 kHz and AMR coded data was upsampled back to 16 kHz.

4.2 Experiment

The classification algorithms were implemented in Matlab environment. The Bayes Networks implementation was based on the Bayes Net Toolbox [11]. All features described in section 2 were calculated in 32 ms frames for the analysis (512 samples) and averaged in 1 sec. window for the classifier input.

All model parameters were trained from the training data that was selected to be about 50 % of the available audio material. Rest of the material was used for testing the system.

The output of the Bayesian Networks results were averaged in 3 sec. windows with probability weight for each class to smooth the temporal fragmentation. For these results a hard decision was made to achieve the final decision for each 1 s segment

5. RESULTS

Experiments were performed to evaluate the proposed method. The classification results were obtained for high quality, narrowband and AMR compressed audio data. The accuracy of correct classification results for 1 second segments are presented in a confusion matrix in Table 2 for the high quality audio, Table 3 for the narrow band audio and Table 4 for the AMR coded audio. The accuracy is given in percentages. The audio classes in the table are: music, speech, constant noise (CN), environmental sound (ES) and silence.

Table 2. Accuracy for high quality audio in percentages

	Music	Speech	CN	ES	Silence
Music	91,3	4,0	0,4	3,9	0,4
Speech	3,1	96,9	0	0	0
CN	1,8	0	96,1	2,1	0
ES	3,3	3,2	32,6	60,9	0
Silence	0	0,4	1,6	0	98,0

Table 3. Accuracy for narrowband audio in percentages

	Music	Speech	CN	ES	Silence
Music	90,7	4,5	0,2	4,3	0,3
Speech	2,9	97,1	0	0	0
CN	1,9	0	96,2	1,9	0
ES	3,2	3,2	32,9	60,7	0
Silence	0,1	0,1	1,5	0	98,2

Table 4. Accuracy for AMR coded audio in percentages

	Music	Speech	CN	ES	Silence
Music	87,1	3,8	8,8	0,3	0
Speech	3,5	96,5	0	0	0
CN	0,3	0	90,6	9,1	0
ES	3,7	3,8	49,8	42,7	0
Silence	0	1,1	1,3	0	97,6

The average performance for the high quality audio is 88.6%, 88.6% for the narrowband audio and for the AMR coded audio 82.9%. The effect of band limiting the audio files is insignificant. There is even a small change to a better direction. This might be caused by some occasional high frequency components that confuse the classification with the higher sampling rate case. As seen from the Table 2. the most difficult case has been environmental sound which has been confused with constant noise. By fusing the two noise types we achieve an accuracy of over 95% for both non-compressed databases.

AMR coding has had some degrading effect on results, but the accuracy is still reasonable. The smallest change has been on speech class, which is the type of audio the AMR compression has been designed for. The confusion between CN and ES is even greater in this case, and by listening the audio samples, it was evitable the certain type of ES samples had turned more like CN. The other class that has been suffering from AMR coding is music. This is understandable since the effect of AMR coding to the music is also easily audible.

6. CONCLUSIONS

We have implemented a Bayesian Network based audio classification and segmentation with weighted average smoothing for 5 audio classes. The designed network topology mimics traditional rule-base hierarchical structure by implementing conditional dependencies between decision nodes. Our classifier is targeted to a mobile phone based video content management, thus we developed a robust feature set for narrowband audio and also investigated the effect of the AMR coding to the classification performance. The average recognition accuracy for each 1 second audio segments for the high quality audio is 88.6%, 88.6% for the narrowband audio and for the AMR compressed audio 82.9%. There is no significant difference in results between 16 kHz and 8 kHz sampled audio. Also the effect of AMR coding is acceptable to the recognition accuracy for most content classification applications. The effect of real mobile phone captured content on the recognition results should be evaluated in the next phase.

ACKNOWLEDGEMENT

This work is partly funded by Nokia Research Center, Finland and partly by National Technology Agency of Finland (TEKES) in the framework of EUREKA ITEA/CANDELA project. Special thanks to Paavo Pietarila for technical support.

7. REFERENCES

[1]. L. Lu, H.-J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no 7, pp 504 – 516, Oct. 2002.

[2]. C.-H. Wu, C.-H. Hsieh, "Multiple Change-Point Audio Segmentation and Classification Using an MDL-Based Gaussian Model", *IEEE Transactions on Speech and Audio Processing,* Accepted for future publication, vol PP, is 99, pp.1 – 11, 2005.

[3]. C.-H. Lin, S. –H. Chen, T.-K. Truong, Y. Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," *IEEE Transactions on Speech and Audio Processing,* vol. 13, is. 5, Part 1, pp. :644 – 651, Sept. 2005

[4]. T. Kindberg, M. Spasojevic, R. Fleck, A. Sellen, "The ubiquitous camera: an in-depth study of camera phone use," *IEEE Pervasive Computing*, vol. 4, is. 2, pp. 42 -50, Jan.-March 2005.

[5]. M. Zhao, J. Bu, C. Chen, "Audio and video combined for home video abstraction", *Proc. ICASSP, IEEE 2003*, vol. 5, pp V - 620-3.

[6]. T. Zhang and C.-C. J Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation And Classification", *IEEE Trans. on Speech and Audio Processing,* vol.9. no.4, pp 441 – 457, May 2001.

[7]. P.Korpipää, M, Koskinen, J. Peltola, Johannes, S.-M. Mäkelä, T. Seppänen, "Bayesian approach to sensor-based context awareness, " *Personal and Ubiquitous Computing,* vol. 7, no 2, pp 113 - 124 , July 2003.

[8]. L. Lu, R. Cai, A., Hanjalic, "Towards a unified framework for content-based audio analysis," *Proc. of ICASSP,* IEEE, 2005, vol. 2, pp. 18-23.

[9]. MPEG7 standard, ISO-IEC 15938-4, Information Technology - Multimedia Content Description Interface - Part 4: Audio.

[10]. http://www.ntt-at.com/products_e/noise-DB/index.html.

[11]. K. Murphy, "The Bayes net toolbox for Matlab," *Computing Science and Statistics*, vol.33, 2001.