

SPEECH MUSIC DISCRIMINATION IN MIXED AUDIO CONTENT

Bernhard Rettenbacher, Maria Fellner

JOANNEUM RESEARCH Forschungsgesellschaft mbH
Email: {bernhard.rettentbacher, maria.fellner}@joanneum.at

Abstract: *Speech Music Discrimination (SMD) can highly improve Automatic Speech Recognition (ASR) and Music Analysis results of broadcast or cinema audio signals. Especially in infotainment productions or commercials, speech is often embedded in background music. We present an audio segmentation system for television commercials, optimized for the handling of speech-music mixture classes using independent detectors for the classes “speech”, “music”, “other” (sounds) and “silence”. The system is optimized to process high and low quality audio in terms of bandwidth limitation and distortion. We evaluated the system on a representative set of commercial spots in three different languages with different audio qualities.*

Key words: speech music discrimination, segmentation, gaussian mixture models, evaluation, television commercials

1. INTRODUCTION

In automatic content analysis of multimedia data, Automatic Speech Recognition (ASR) and Music Information Retrieval (MIR) are important tools for extracting semantic information from audio data. Especially ASR systems rely on a correct adaption of their acoustic models, when background noises or music is present. To assist an ASR or MIR system, a prior classification and segmentation of the audio content into speech, music and mixed content can highly improve the performance of such systems. Classification and segmentation into speech and music is often called “Speech Music Discrimination” (SMD) in literature. Many publications on SMD address only pure speech and pure music signals. When such systems have to classify mixed content, they tend to assign the content either to the more dominant class or oscillate between the classes. Class oscillation may be used to classify class mixtures, but often leads to an over-segmentation of the audio signal.

These Speech/music mixtures appear quite often in radio and television programmes. Movies, infotainment productions and commercials contain speech, music, sound effects and background sounds. Especially in commercials these signal classes appear often in a mixed and fast changing manner.

We present a system for automatic temporal segmentation of audio data in television commercial spots. It temporally divides the audio data into homogeneous segments of the

signal classes „speech“, „music“, „silence“ und „other sounds“ and into segments containing mixtures of these signal classes. Our approach is based on a differentiated modelling approach, where each class is modelled by an individual class/no-class classifier. The results from the classifiers are combined in a preceding fusion stage, which also creates the segments.

This system is part of the EC IST project “MediaCampaign”. MediaCampaign deals with detecting media campaigns in three different media, namely press, television and internet. This is achieved by interrelating the results of a multimodal content analysis of commercial spots. In this system, the audio segmentation system processes the television audio data stream as a pre-processing stage for ASR, Word Spotting and Jingle Recognition.

1.1. Related Work

For Automatic Speech Recognition (ASR), the importance of a prior classification and segmentation into speech, music and other sounds has been stated in many scientific publications. Relevant research work on this “coarse” level classification and temporal audio segmentation can be found especially for “speech/non-speech” discrimination or “speech/music” discrimination (SMD). Earlier approaches, e.g. the quite often cited paper from E. Scheirer und M. Slaney [1], focus on feature extraction and the classification of presegmented audio data. These publications concentrate more on the extraction of charac-

teristic features for classification than on the correct temporal segmentation of the audio data. Further, these methods have been evaluated only for pure speech or music segments but not on mixed content.

Later works introduce segment modelling and regard segments containing class mixtures. In [2], the output of a speech recognition system is used to extract “entropy” and “dynamism” features. The segments are modelled by Hidden Markov Models whereas the emission probabilities are classified using Gaussian Mixture Models or Multilayer Perceptrons.

In [3] a differentiated modelling approach is presented, where each class is detected using a class/non-class classifier (e.g. “speech/non-speech”, “music/non-music”). The segments are created for each detector and combined afterwards. The motivation, for using independent classifiers is to fusion the approaches from the ASR community, trying to discard non-speech audio, and the Music Information Retrieval (MIR) community, which of course is more interested in extracting musical segments. This approach shows good results for mixed audio content from a television movie and seems to be applicable to television commercial demands.

2. CONTENT EVALUATION

Most television commercial spots are audiovisual clips with a typical duration of 20 to 30 seconds in length. In this short time period a lot of information has to be passed directly or indirectly to the consumer. For audio, this can be messages from a professional speaker, interviewed “real-world” persons, background speech, music, natural and artificial sound effects or soundscapes. These elements are often mixed to intensify and condense the information the consumer should receive. The dense mixture of different audio classes makes classification and segmentation as well as Automatic Speech Recognition (ASR) or Music Information Retrieval (MIR) challenging tasks. To gain knowledge about the distribution of the investigated audio classes, we carried out a detailed analysis of television commercial audio data.

308 television commercial spots have been recorded and manually annotated for evaluation and training of the audio segmentation system.

The signal quality of the used material ranges from television quality recordings (15-16 kHz bandwidth, 44.1 kHz sample rate, 16 bit, stereo) to extreme low quality data (5.5 kHz bandwidth, 12 kHz sample rate, 8 bit, mono).

The evaluation set contains 140 spots with a total length of 59 minutes. In addition to the overall evaluation, the evaluation test set has been divided into a high quality dataset “HQ” and a low quality dataset “LQ”. The HQ dataset contains 34 spots (14 minutes) with TV-quality audio data sampled at 32-44.1 kHz, 16 bit, stereo. The signal bandwidth is around 15 kHz, resulting from the bandwidth of the TV transmission channel.

The LQ dataset contains 106 spots (45 minutes) with a signal bandwidth of the audio data at less than 10 kHz.

Most of the audio data is available only at a sample resolution of 12 kHz, 8 bit, mono. The average signal bandwidth is 5.928 kHz. For some spots the signal-to-noise ratio is very low because of the low amplification. Figure 1 shows the distribution of classes in the data sets.

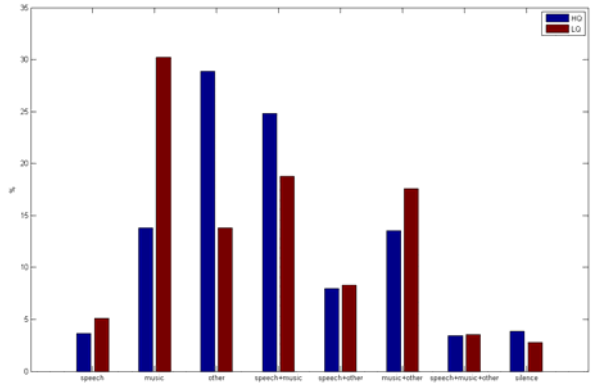


Fig. 1. Distribution of classes in the evaluation data sets

The annotation can also be represented as overlapping segments of “speech”, “music”, “other” and “silence”. Figure 2 shows the class distribution of the overlapping segments. The statistics are gained by summing up all segments containing “speech”, “music”, “other sounds” and “silence” (e.g.: “speech” + “speech”/”music” + “speech”/”other” + “speech”/”music”/”other”). The values are normalized to the total length of the specific test set. The sum of all durations is greater than 100%, showing the degree of segment overlap.

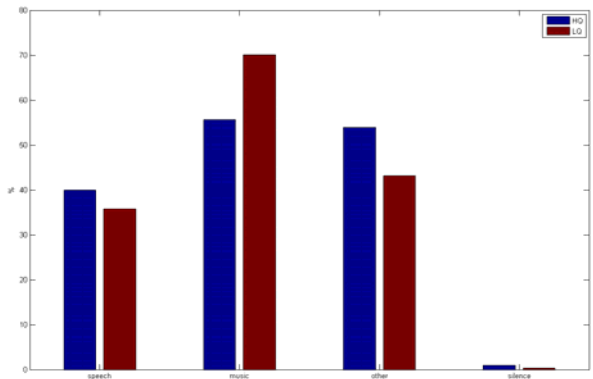


Fig. 2. Distribution of overlapping pure classes

The training set contains 268 spots, which are 99 minutes of audio data. The training set contains only high quality audio data, which has been pre-processed for the use with the high and low quality evaluation sets. The class distribution of the training corpus is shown in figure 3.

When we look at the class appearances in the evaluation set (which we gained from looking at the overlapped representation), we can see that “music” is dominant (avg. 63%) in TV commercials, followed by “other” (avg. 49%), “speech” appears only in avg. 38% of the data and silence is almost not present (avg. 0.6%).

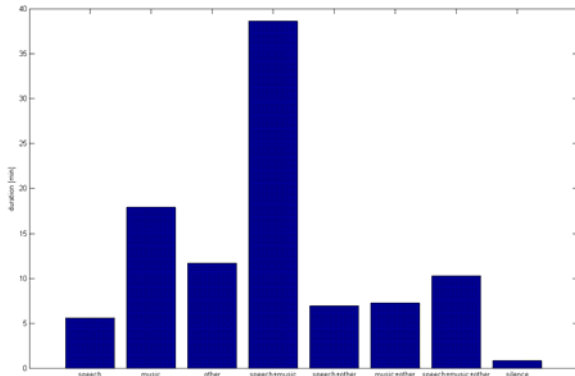


Fig. 3. Class distribution in training data

When looking at the non-overlapping class distribution, the classes “music”, “other” and “speech+music” are dominant (each class around 21%), followed by “music+other” (avg. 16%). Pure “speech” (avg. 4%) appears very seldom. Speech and music overlap in 24% of the test data. In 65% of the “speech” data, we have “speech” mixed with “music”.

3. SYSTEM DESCRIPTION

The presented audio segmentation system is based on state of the art window based feature extraction and probabilistic classifiers for detecting each signal class. The classification results are combined and optimized in a subsequent fusion stage. The system uses a data stream approach, so it can be used both for file processing and real-time audio stream processing.

For feature extraction, the audio stream is divided into frames with a duration of 25 ms and 60 % overlap. Out of these frames, temporal and spectral features are extracted. The resulting features are combined to 60 % overlapping texture windows with a window length of 250 ms, to extract central moments of the feature vectors and their first order derivatives.

The extracted features form feature vectors, which are presented to classifiers. Each classifier gets an individual feature subset. For the classification of “speech”, “music”, “other”, Gaussian Mixture Models (GMM) are used. Each classifier maps the frames containing the pure class (e.g. “speech”) and also mixtures with other classes (e.g. “speech” AND “music”) to a “positive” class and all other frames to a “negative” class (e.g. non-“speech”). “Silence” detection is based on a calculation of the psychoacoustic loudness. A dynamic threshold decision is taken to distinguish between “silence” and all other classes.

The classifier outputs are combined and segments are modelled by a fusion stage. First, the posterior probabilities output of the Gaussian Mixture Model classifiers are low-pass-filtered to remove single outliers. The corrected results from the individual classifiers are combined and a second single outlier removal step is applied on the combined classification results. Afterwards, the combined

classification results are searched for class changes, which build up labelled segments. These segments are written into an MPEG-7 document for output.

For training and evaluation, German, Dutch and English television commercial spots have been recorded. In total, 308 spots have been annotated manually. The segments have been annotated in more detail than needed for training to allow an automatic selection of segments for the training corpus. With the definition of mappings from the detailed semantic annotation to the labels of the classifiers, different corpora can be designed. Specialised corpora for e.g. pure speech or speech mixtures as well as corpora with different class compositions, e.g. 50% male speech, 50% female speech, can be created.

4. SYSTEM EVALUATION

The audio segmentation system has been tested against audio data, which has been annotated manually as non-overlapping segments of the classes “music”, “speech”, “other sounds” and “silence”. Mixture segments have been annotated with terms for the appearing classes.

The evaluation results of the segmentation system are shown in table 1 for each classifier and in table 2 for the complete system in terms of recall and precision.

	Class/Mixture	Recall	Precision
Complete test set	Speech	6%	35%
	Music	62%	49%
	Other	24%	79%
	Speech+Music	85%	61%
	Speech+Other	1%	5%
	Music+Other	29%	25%
	Speech+Music+Other	1%	5%
	Silence	80%	47%
LQ test set	Speech	5%	19%
	Music	64%	60%
	Other	9%	68%
	Speech+Music	88%	48%
	Speech+Other	1%	6%
	Music+Other	22%	28%
	Speech+Music+Other	0%	0%
	Silence	83%	45%
HQ test set	Speech	8%	52%
	Music	59%	37%
	Other	38%	91%
	Speech+Music	82%	74%
	Speech+Other	1%	5%
	Music+Other	37%	22%
	Speech+Music+Other	2%	9%
	Silence	78%	49%

Table 2. Evaluation results for complete system

	Class	Recall	Precision
Com- plete test set	Speech	83%	80%
	Music	97%	73%
	Other	43%	75%
	Silence	78%	47%
LQ test set	Speech	88%	70%
	Music	97%	81%
	Other	28%	68%
	Silence	81%	46%
HQ test set	Speech	78%	90%
	Music	97%	66%
	Other	59%	83%
	Silence	76%	49%

Table 2. Evaluation results for the individual classifiers

5. DISCUSSION

A differentiated modelling approach for audio classes like speech or music has some advantages compared to approaches, where all classes are modelled by one classifier. The main advantage is that feature selection and classifier parameters can be adapted to the individual signal characteristics. Further, in contrast to the single classifier approach, the classifiers have to handle fewer class changes. In the “speech/music/other/silence” task, segment modelling has to deal with 8 different classes in the single classifier case, whilst the segment modelling in the differentiated approach has to deal only with two classes, namely the presence or absence of the class, the classifier has to handle.

The evaluation of the presented audio segmentation system shows results for the individual classifiers, which are comparable to results from other publications. But it also shows the difficulties with highly mixed audio content after fusion. Especially pure “speech” has very low rates. This results from a mismatch between the “speech” class and the “speech+music” mixture class, which is a dominant class in the class distribution. This mismatch is caused by the “music” classifier. Also the “other” classifier has very low rates. This is caused by the annotation of the segments assigned to the “other” class, because the handled sound effects and background noise sounds do not appear throughout the whole segment. Therefore, a more fine grained annotation would be needed.

The results for “silence” have to be interpreted carefully, because an objective definition of silence is not possible, as real silence does not appear in the audio data (sample values at zero). We defined silence as audio signals with a psychoacoustic loudness (in Sone) of less than 25% of the maximum loudness. Nevertheless, “silence” detection has a relatively high agreement with the perception of the manual annotators.

Finally, the comparison between the HQ and LQ evaluation set show the robustness of the system.

6. CONCLUSION

The presented audio segmentation system can deal with highly mixed audio content and therefore is appropriate for processing audio data in television commercial spots. The used differentiated modelling approach has several advantages to systems, which use a single classifier for all classes. The most important advantage is the easier and more robust segment modelling. Experimental results show, that even a system trained on an incomplete corpus, the segmentation of the audio data, which is covered by the classifier, shows only little under- or over-segmentation. Figure 4 shows an example segmentation, where the first 6 seconds contain audio data, where no similar data is contained in the training corpus.

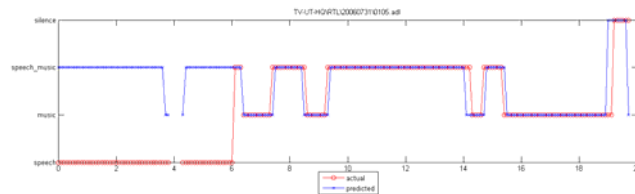


Fig. 4. Examples of segmented TV commercial spots.

The insight got from the evaluation of the appearing audio content facilitates the improvement of the system. For improvement, the training dataset has to be enlarged. Especially pure speech is underrepresented, whilst “speech+music” is dominant. Further, it is important to optimize the annotation especially for “other” sounds. Other optimizations include the extension of the potential feature set and feature subset selection.

7. ACKNOWLEDGEMENTS

The R&D work carried out for the MediaCampaign project is partially funded under the 6th Framework Programme of the European Union within the strategic objective “Semantic-based knowledge and content systems” of the IST Work programme 2004 (IST FP6-027413).

REFERENCES

- [1] Scheirer, E.; Slaney M.: **Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator**, ICASSP, 1997.
- [2] Ajmera, J; McCowan, I; Bourslard, H: **Speech/Music Segmentation Using Entropy and Dynamism Features in a HMM Classification Framework**, *Speech Communication* Vol. 40, no. 3, pp. 351-363. May 2003
- [3] Pinquier, J.; Senac, C.; André-Obrecht, R.: **Speech and Music Classification in Audio Documents**, ICASSP, 2002.
- [4] **MediaCampaign** homepage, URL: <http://www.media-campaign.eu>