

Bernhard Rettenbacher, Maria Fellner

INTRODUCTION

Speech Music Discrimination (SMD) can highly improve Automatic Speech Recognition (ASR) and Music Information Retrieval (MIR) results of broadcast or cinema audio signals. Especially in infotainment productions or commercials, speech is often embedded in background music. Whilst earlier research work on SMD focus on the discrimination of pure classes (e.g. [1]), current publications like [2] and [3] take mixture segments into account. These approaches use either HMM segment modelling or the use of a differentiated modelling approach using independent detectors for each class. We present an audio segmentation system for television commercials, which is part of a media monitoring system developed in the EC project "MediaCampaign" [4]. The system is optimized for the handling of speech-music mixture classes using independent detectors for 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.

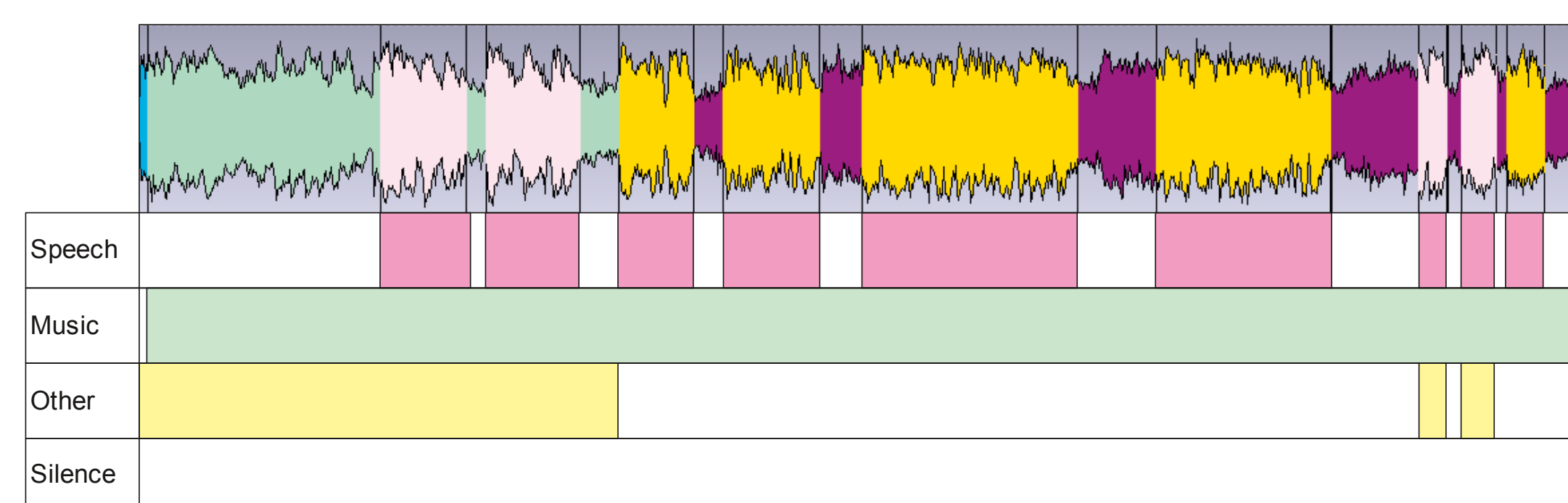


Figure 1: Audio segmentation of a 21 seconds television commercial spot

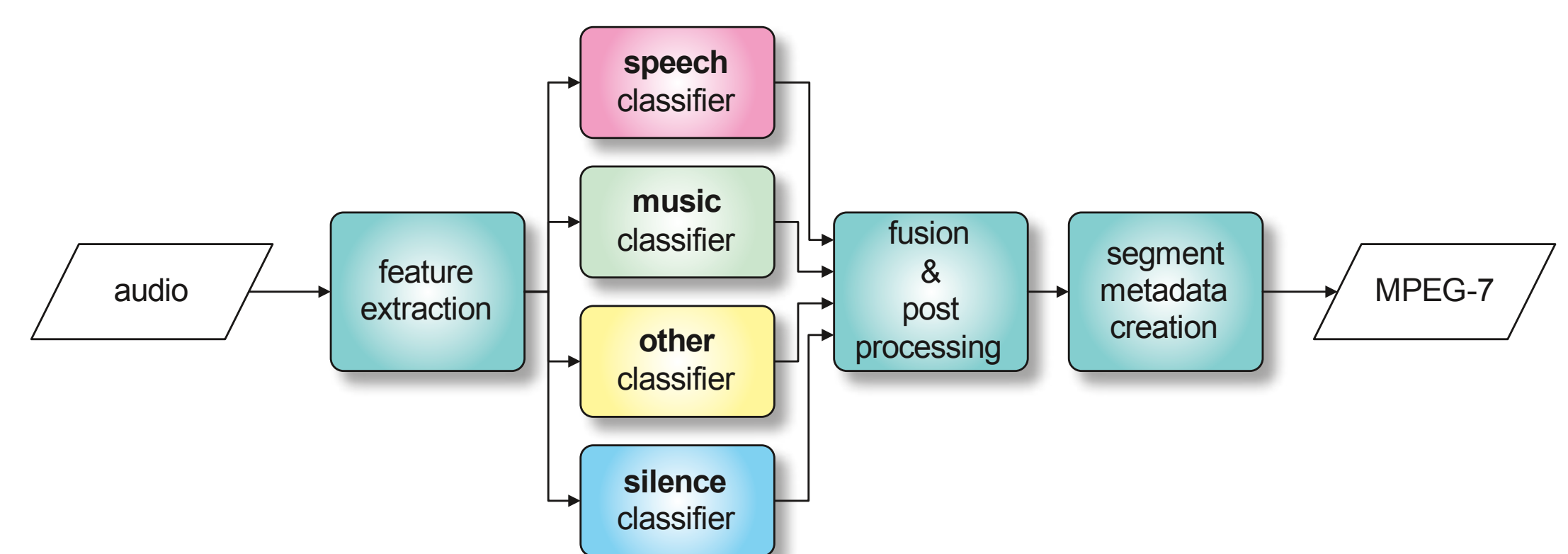


Figure 2: System architecture

CONTENT EVALUATION

Most television commercial spots are audiovisual clips with a typical duration of 20 to 30 seconds in length. Figure 1 shows the segmentation of a typical spot. The dense mixture of different audio classes makes classification and segmentation as well as ASR or 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.

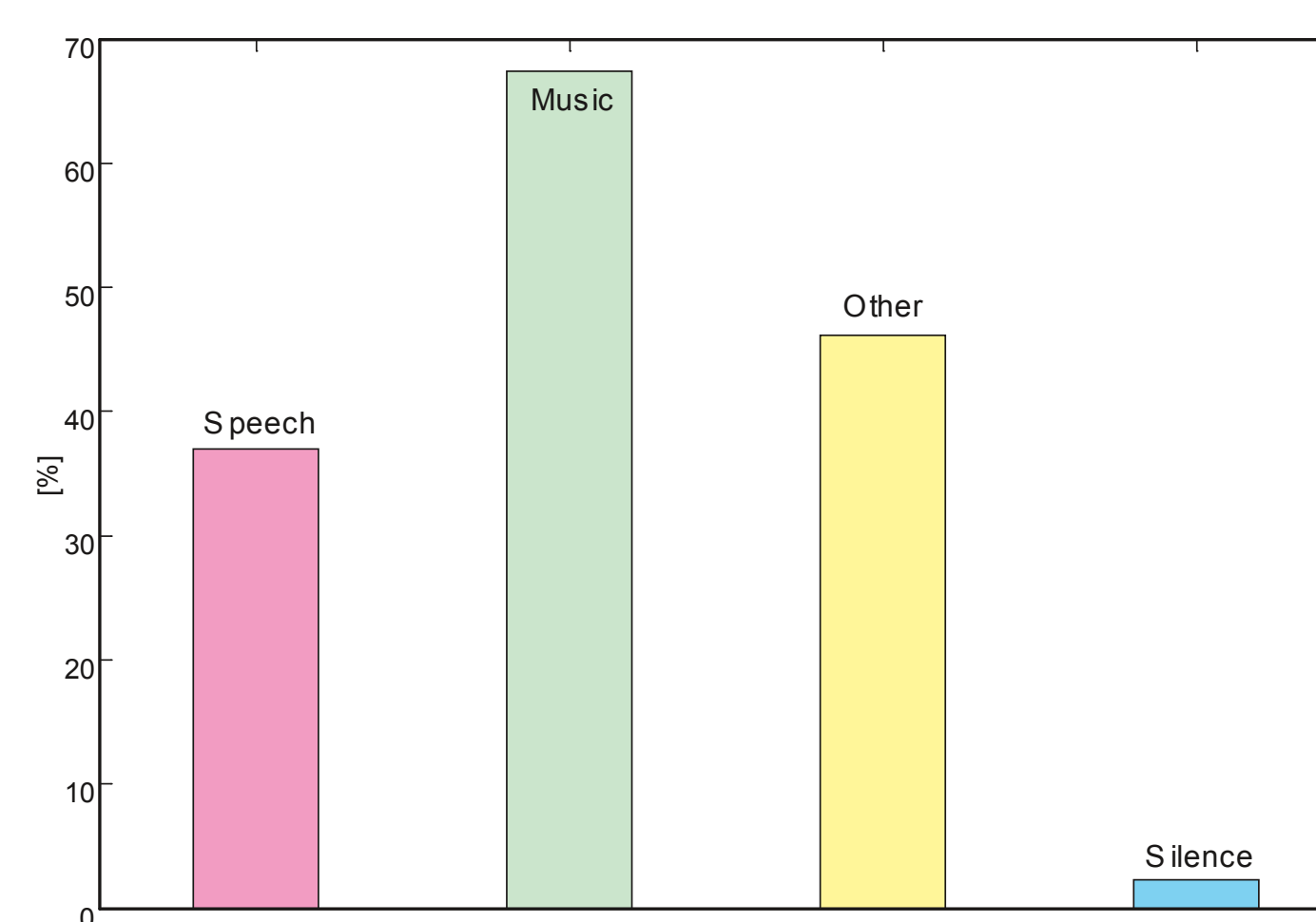


Figure 4: Distribution of overlapping pure classes

308 television commercial spots have been recorded and manually annotated for evaluation (140 spots) and training (286 spots). 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).

In addition to the overall evaluation, the evaluation test set has been divided into a high quality dataset "HQ" (34 spots) and a low quality dataset "LQ" (106 spots, bandwidth < 10 kHz). 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.

Figure 3 and Figure 4 show the class distribution in the data sets. These figures show that pure speech appears very seldom. In most cases, speech is mixed with music or other sounds. In total, television commercials contain more music (63 %) than speech (38 %) overlapping 24 %.

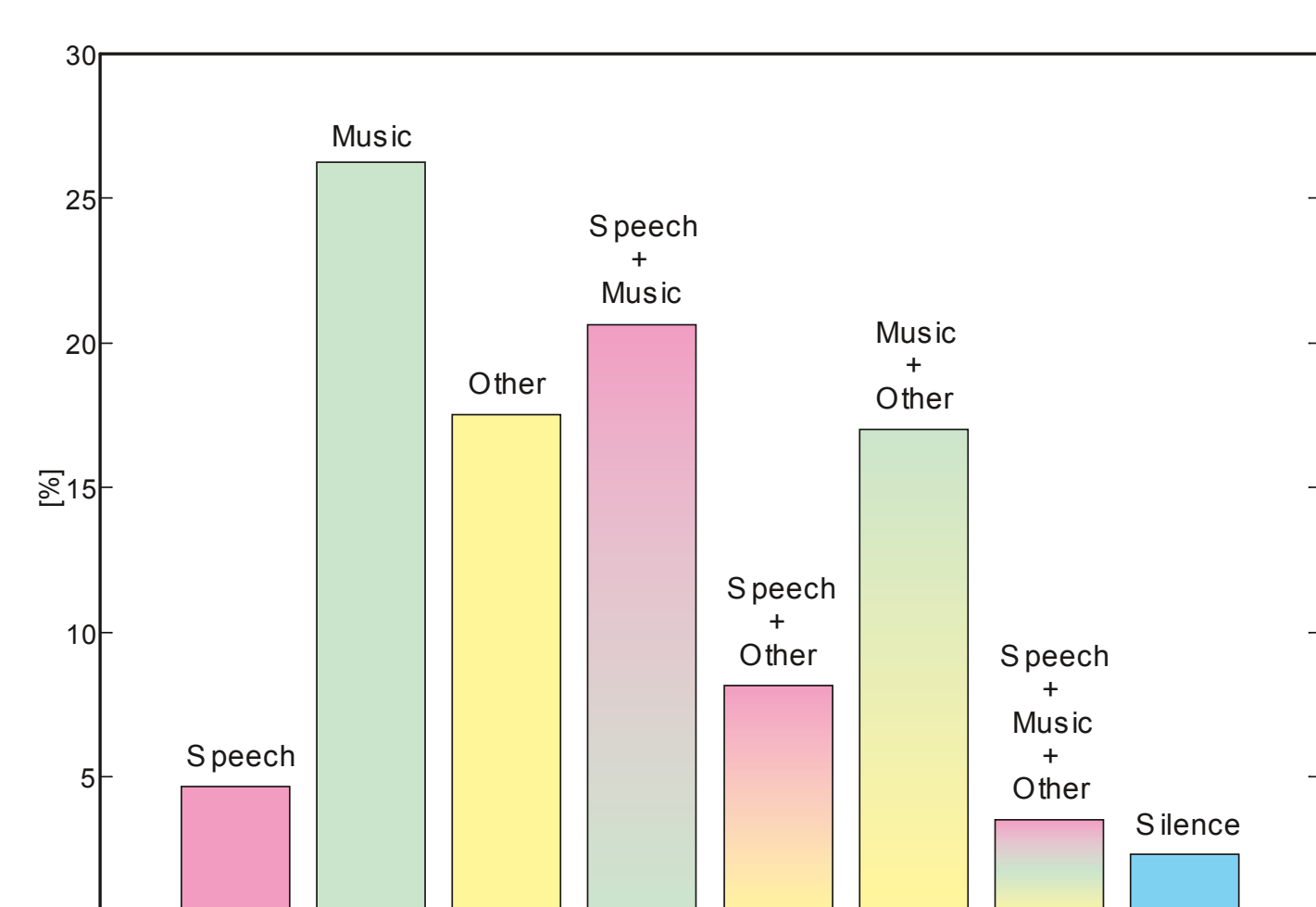


Figure 3: Distribution of classes in the evaluation data sets

SYSTEM DESCRIPTION

The audio segmentation system follows a detector approach for each class and a final fusion stage (Figure 2). The detectors use feature extraction front ends with an individually optimized selection of features for classifying the presence or absence of the respective audio class. Silence is detected by adaptive thresholding the short time psychoacoustic loudness, all other classes are detected by Gaussian Mixture Model classifiers. After a first segment modelling stage, the results from the individual classifiers are combined in the fusion stage.

The system has been trained on 100 minutes audio data from German, Dutch and English commercial spots. The training corpus has been selected through mapping of the detailed semantic annotation to the labels of the destination classes.

SYSTEM EVALUATION

The audio segmentation system has been tested against audio data, which has been annotated manually as non-overlapping segments of 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. System evaluation shows, that the individual classifiers perform quite well, whilst the overall performance has very low rates for individual classes. The poor performance of the speech class comes from confusion between speech and speech+music. As pure speech appears very seldom, this classification error shows up only in the recall rate of the speech class. To improve the classification performance of pure speech, the precision of the music classifier has to be improved.

	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 %

Table 1: Evaluation results for complete system

	Class	Recall	Precision
Complete 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

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 for 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. 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.

JOANNEUM RESEARCH
Forschungsgesellschaft mbH

Institute of
Applied Systems Technology
Intelligent Acoustic Solutions

Steyrergasse 17
8010 Graz, Austria

Phone +43 316 876-1634
Fax +43 316 8769-1634

bernhard.rettbacher@joanneum.at
maria.fellner@joanneum.at
www.joanneum.at/ias

www.joanneum.at/ias

REFERENCES

- [1] Scheirer, E.; Slaney M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, ICASSP, 1997.
- [2] Ajmera, J.; McCowan, I.; Bourlard, 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, www.media-campaign.eu



www.media-campaign.eu

ACKNOWLEDGEMENT

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).



Information Society
Technologies