From: AAAI Technical Report SS-03-08. Compilation copyright © 2003, AAAI (www.aaai.org). All rights reserved.

Π-Avida - A Personalized Interactive Audio and Video Portal

Gerhard Paass[†], **Edda Leopold**[†], **Martha Larson**[‡], **Jörg Kindermann**[†], **Stefan Eickeler**[‡]

[†] Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
[‡] Fraunhofer Institute for Media Communication (IMK)
53754 Sankt Augustin, Germany

Abstract

We describe a system for enregistering, storing and distributing multimedia data streams. For each modality – audio, speech, video – characteristic features are extracted and used to classify the content into a range of topic categories. Using data mining techniques classifier models are determined from training data. These models are able to assign existing and new multimedia documents to one or several topic categories. We describe the features used as inputs for these classifiers. We demonstrate that the classification of audio material may be improved by using phonemes and syllables instead of words. Finally we show that the categorization performance mainly depends on the quality of speech recognition and that the simple video features we tested are of only marginal utility.

Introduction

During the last years new methods for the detection of information relevant to user needs in text, audio and video have been developed. At the same time new channels for information delivery of multimedia, like high-speed Internet and UMTS mobile communications, have been introduced. In 2001 we started the project Π-Avida in order to pursue research and development integrating both of these aspects. The aim of the project is the development of segmentation, classification, and retrieval techniques for multimedia data streams that are to be used in web portal applications.

Text, audio, video and voice streams will be prepared and analyzed in an almost entirely automated manner, enabling the distribution of information over different media and communication channels. At the core of our multimedia stream enregistration research is the development of a sophisticated classification methodology. These classification techniques will be integrated in the design and development of a production and distribution environment for a personalized multimedia information portal during the project runtime.

To achieve this goal, development of not only of media specific, but also of integrated text, audio, video and voice classification techniques is necessary. Combination of features from different media should inherently lead to improved classification performance. Starting from information acquisition via information processing to information distribution, our content preparation technique will be embedded in the complete chain of media production. Ultimately it will allow the expansion of today's mostly exclusively text based personalized online databases to encompass additional media like audio, video and voice.

By using automated data detection, segmentation and classification techniques, the system developed in the Π-Avida project aims to provide support for multimedia editing as well as meet the needs of individual end users of the information portal. An intriguing consequence is that for each individual user, user-specific classifiers may be trained and exploited for end-device-independent personalized information distribution.

The portal access for users is customized to their preferences and will be possible using a variety of end devices, such as a PDAs, PCs or mobile phones. Interactions within the streaming portal will be conducted via voice entry or through standard protocols (like WML, HTTP or UMTS based protocols).

The integration of Π-Avida in already existing production environments (editorial systems) will provide the possibility to realize a multitude of application scenarios in areas of advertising, TV-analysis, digital information archiving, trend and news analysis.

The overall structure of the final project system is shown in figure 1. First, the audio/video stream is time stamped and decoded into MPEG-7 multimedia format. Then the different modalities text, speech, music and audio as well as video are analyzed by different modules, which extract meta-information. The music and audio module analyzes the audio information and partitions the data from this channel into speech, music and noise/silence segments. Speech segments are forwarded to speech recognition. Special features, discussed in more detail in the following section, are extracted for music and noise. The speech recognition system generates phonemes, syllables and words. The video subsystem segments the video stream into shots and characterizes each shot by keyframes as well as image statistics. Other types of features extraction such as extracting text from images, segmenting images into coherent regions,

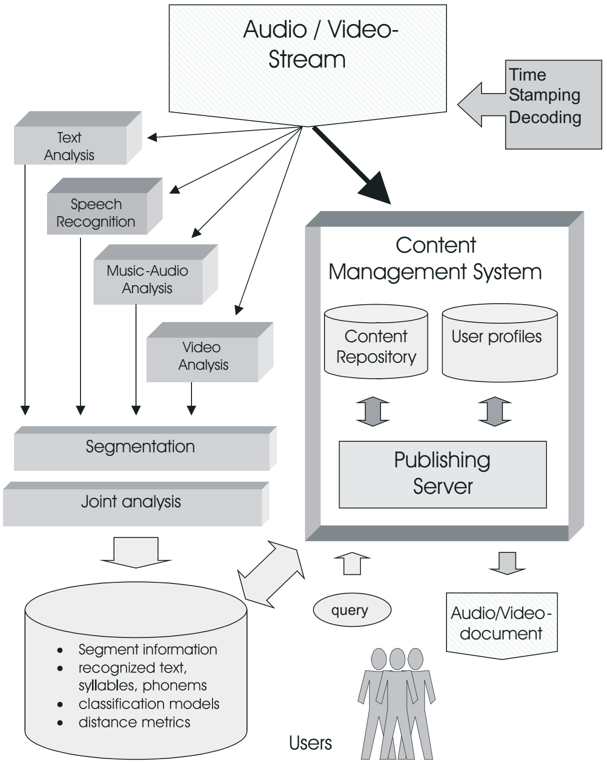

Figure 1: The overall structure of the Π-Avida system.

and detection of faces are slated for future investigation.

Together all features are coded as meta-information in MPEG-7 and are forwarded to the joint analysis system. The joint analysis system combines features from all modalities to perform integrated classification. This system basically generates different classifiers: starting with example classifications in training data it classifies the multimedia segments into different meaningful categories. These may be general categories, e.g. IPTC press categories, or user-specific categories, which are formed according to the personal preferences of different users. The aim is to allow each user of the system to form his own set of categories. The corresponding classification models are stored and are later used to classify new multimedia documents from the video stream.

In this paper we describe some aspects of this ongoing project. First we summarize the features we extract from the audio/video streams to be used for classification. Second we report on the results of preliminary topic classification experiments based on features generated. We carry out two experiments, classification of radio documentaries recorded from Deutsche Welle and classification of television news recorded from two television channels, n- tv and N24.

Features for Multimedia Classification

Audio Features

Processing of the audio data[1] is one of the first tasks that must be carried out in order to distinguish between silence, noise, music and speech. Also the genre of music – like pop/rock, classical music – may be determined. For this classification the PCM audio stream is partitioned into segments of constant length.

For each of 16 frequency bands the energy is measured each 30 msec. In addition the average intensity in each frequency band and its variance are determined for every second. The audio classification basically compares these features with the corresponding values of prototypes and determines the best matching class, e.g. classical music. In a subsequent post-processing step, segments with similar characteristics are merged. In this way the results may be used for short-time classification as well as segmentation.

This approach can be extended to enable the recognition of known pieces of music. In the same way as described above MPEG-7 conformant characteristics (AudioSignatureType) are extracted. As before, similar features are collected to a ”fingerprint” covering one second. This fingerprint is stored together with other metadata (title of the piece, musician, etc). For retrieval the classifier compares the fingerprint of the new piece of music with the stored fingerprints. Currently it is able to select a piece of music out of a database of 80000 pieces with very high reliability.

Speech Features

A source of speech recognition error relevant to the document classification task is error due to the language model used in the speech recognition system. The worst case is when the word pronounced in the audio is completely missing from the language model inventory. After the vocabulary reaches a certain threshold, the new words are mostly proper names, or, especially in the case of German, compound nouns. The inventory of words is in principle infinite in any natural language. The inventory of linguistic forms present in the language model is, however, finite. In order to reduce the size of the language model vocabulary and increase its ability to cover the word forms appearing in incoming spoken audio, we use syllables as the base unit in the language model for the speech recognizer.

In fact for language processing in general words are a conventional and convenient basic feature, but they are not justified on independent grounds. Sub-word alternatives to the orthographic word include phonemes, syllables and morphemes. Phonemes are short and thus acoustically less distinct, and inflectional morphology is often nothing more than a single phoneme, making syllables the most promising alternative.

Although phonemes are a closed class, surprising neither morphemes nor syllables are. The syllable-based system will constantly encounter new syllables in the incoming input, now not due to compound words, but due to

[1] The work on audio stream processing is performed by Fraunhofer IIS in Erlangen (Allamanche et al. 2001).

Table 1: Agreement of human annotators in classifying the Kalenderblatt Documents

DW Kalenderblatt data from year	top choice of both annotators identical	complete disagreement between annotators
1999	67 %	11 %
2000	74 %	9 %
2001	70 %	20 %

proper names, coinages, and loan words. The syllable inventory of the statistical language model represents a delicate balance between coverage and confusion. There must be enough syllables to take care of a substantial portion of out-of-vocabulary error, but not so many as to tempt the recognizer into mis-recognitions.

The speech recognition system that we use is a typical Hidden-Markov-Model-based system in which the basic acoustic models are phoneme models and consist of three states connected by forward and self-transitions. At each state the probability that that state emitted the given feature vector is modeled by a probability density function composed of a Gaussian or a mixture of Gaussians. In order to minimize acoustic error, acoustic models can be adapted to background conditions or to the particular voice of the person speaking.

Word models, or in our case actually syllable models, are built up by stringing phoneme models together according to the prescription of a pronunciation lexicon. The class conditional probability $P(O|W)$ is the probability of observing the acoustic input vectors, 25 msec frames of speech transformed into spectral features (36 Mel-frequency Cepstral Coefficient (MFCC) and three energy terms), given the syllable model.

The prior probability is delivered by an n-gram syllable language model. The syllable inventory is determined when the language model is trained. We smooth the language model using Katz-style backoff and Good-Turing smoothing (Jurafsky and Martin 2000). Language models are trained on text, and not on spoken language, since text is readily available in sufficient quantities. In fact, the separation of acoustic and language models is necessary exactly in order that text can be used to model word occurrences, since although a given phoneme event occurs relatively often in spoken audio, many word events are too rare to supply sufficient cases in word audio.

The Viterbi algorithm is used to find the best state sequence through the model. The enormous search space is limited by integrating the prior probability, a word or a syllable probability supplied by the language model, into the search lattice as soon in the search as is feasibly possible and by judiciously pruning unpromising search paths. The output of the recognizer consists of a string of syllables deemed most likely to have produced the acoustic observation. The output is evaluated by aligning it with reference texts and counting insertions, deletions or substitutions.

The automatic speech recognition system (ASR) used in Π-Avida was built using the ISIP (Institute for Signal and Information Processing) public domain speech recognition toolkit from Mississippi State University. For our experiments we use two different implementations of the system, a ”generic” system, which is not trained for any particular domain, and an ”improved” system, which uses improved acoustic and language models and is adapted to the news domain.

The generic ASR system used monophone acoustic models that we trained on a set of 30k sentences from the Siemens/Phondat 100 database available from Bavarian Archive for Speech Signals [2] The syllable language model was a bigram model and was trained on 11 million words transcribed in the German Parliament[3]. To train the syllable model, the training text must be transformed into syllables. We use the transcription module of the Bonn Open Source Synthesis System (BOSSII) developed by the Institute Communication Research and Phonetics of Bonn University (Klabbers et al. 2001). This module applies a combination of linguistic rules and statistical correspondences to transform written German words into syllable decompositions. Syllables with the same pronunciations are merged. We restricted the syllable inventory to the 5000 most frequently occurring syllables in this corpus, all of which were well represented enough to be useful to the model. For processing by the speech recognizer, the audio stream was divided into segments of 20 second length.

In addition to this initial ”generic” speech recognition system described above we developed a second ”improved” system that used improved acoustic and language models and was adapted to the news domain. The improved ASR system used cross-word tri-phone models trained on 7 hours of data from radio documentaries. A 3-gram syllable model was trained on 64 million words from the German dpa newswire. The training set for the improved language model was not only larger than for the generic language model, but also originated from a domain close to the target domain. Each television news document was segmented using the BIC algorithm. Then we located breaks in the speech flow with a silence detector and cut the segments at these points as was necessary in order to insure that no segment be longer than 20 seconds. The improvement to the system allowed us to achieve a syllable accuracy rate improvement of about 30% absolute.

It is important to note that although the news documents were segmented, for these experiments, no speech detection was used. For this reason the output of the speech recognizer also consists of nonsense syllable sequences generated by the recognizer during music and other non-speech audio. As features we used not only syllables from the recognizer but also additional derived features. These were overlapping syllable n-grams and phoneme n-grams and were formed by renormalizing and concatenating the recognizer output.

[2] see http://www.bas.uni-muenchen.de/Bas/BasPD1eng.html.

[3] see http://dip.bundestag.de/parfors/parfors.htm

Video Features

The video[4] to be classified is segmented into shots using standard cut detection and a keyframe is automatically identified to represent each shot. This keyframe is then mapped to a triple of three prototypical video frames, which we call buoys, each originating from a different codebook. Our codebooks of buoys are derived from a partitioning of a color-based feature space created by clustering of frames from a development data set. The buoys are intended to act analogously to the linguistic units we use as features representing spoken language. This triple is the basic video feature and is extracted for each shot in the television news segment. For classification we also use n-grams of basic features in order to have features that span several shots, and thus capture temporal dependencies.

The three codebooks are created by partitioning frames from the development set using a k-medians clustering, according to the algorithm proposed by (Vollmer 2002). Each codebook is a partition created on a different feature space. The first is based on a histogram of 29 colors, the second is based on a correlogram calculated on the basis of the 9 dominant colors, the third is based on first and second moments of the 9 major colors.

We encode basic temporal patterns contained in the video by stringing buoys together to form an n-gram. In this way, basis patterns of shots can be represented. The color-based features are intended to make differentiations between, for example, studio and field shots. By including n-grams of video features, we retain important temporal information from, for example, a news segment which goes from the newsroom, to the soccer field, back to the newsroom, a development characteristic for a sports report. The classifier can exploit the progression from newsroom to soccer field and back without it ever having been necessary for us to explicitly identify either locus in the feature extraction process.

Previous work demonstrated that codebooks containing 400 buoys were suited to video classification, since smaller and larger codebooks led to diminished performance. Also in previous work we confirmed that efficacy of the codebook is independent of the time at which the development data used to generate it is recorded. Experiments with video classification using codebooks defined on the test set did not lead to markedly improved classification results.

Classification with Support Vector Machines

A support vector machine (SVM) is a supervised learning algorithm. In its simplest form, an SVM is a hyperplane in the input space that separates the set of positive examples from the set of negative examples (see figure 2). It can be shown that the classification performance is optimal, if the smallest distance of the hyperplane to the data points – the margin – is large (Vapnik 1998). Maximizing the margin can be expressed as an optimization problem:

$$\text{minimize}\frac{1}{2}\|w\|^2 \quad \text{subject to } y_i(w \cdot x_i + b) \geq 1 \quad \text{for all } i$$

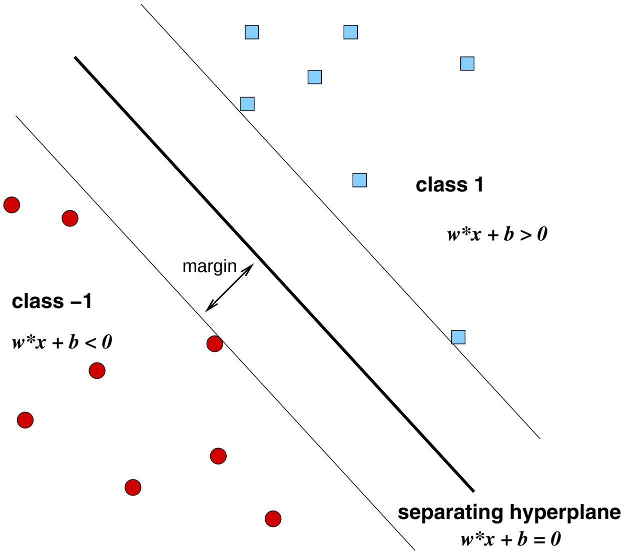

Figure 2: The principle of support vector machines. In the training phase the algorithm searches for the hyperplane that separates the different classes with maximal margin. This hyperplane is defined by its support vectors.

where x_i is the i-th training example and $y_i \in \{-1, +1\}$ is the correct output of the SVM for the i-th training example. Note that the hyperplane is only determined by the training instances x_i on the margin, the *support vectors*. Since not all problems are linearly separable (Vapnik 1998) proposed a modification — SVM with soft margins — to the optimization formulation that allows, but penalizes, examples that fall on the wrong side of the decision boundary.

The SVM may also be applied if the input vectors are mapped into a high-dimensional 'feature space', e.g. the different products of all input variables. The decision hyperplane then is determined in this feature space. Because the margin of the decision boundary is maximum, the Vapnik-Chernovenkis dimension, a measure for classifier complexity, can be controlled independently of the dimensionality of the input space (Vapnik 1998). This leads to an excellent classification capability of the SVM for new data.

Instead of restricting the number of features, support vector machines use a refined structure, which does not necessarily depend on the dimensionality of the input space. By constructing new non-linear feature maps, so-called 'kernels', SVMs may be adapted to many domains in a flexible way. For our experiments we used the SVM^{light} package developed by (Joachims 1998).

Support Vector Machines (SVM) have proven to be fast and effective classifiers for text documents (Joachims 1998, Dumais et al. 1998). Since SVMs also have the advantage of being able to effectively exploit otherwise indiscernible regularities in high dimensional data, they represent an obvious candidate for spoken document classification, offering the potential to effectively circumvent the error-prone speech-to-text conversion.

In the *bag-of-words-representation* the number of occur-

[4] Video processing is performed by Fraunhofer IGD, Darmstadt.

rences in a document is recorded for each word. A typical text corpus can contain more than 100,000 different words with each text document covering only a small fraction. Previous experiments (Leopold and Kindermann 2002) have demonstrated that the choice of kernel for text document classification has a minimal effect on classifier performance, and that choosing the appropriate input text features is essential.

To facilitate classification with sub-word units one can generate n-grams, i.e. sequences of n units. They may take the role of words in conventional SVM text classification described above (Leopold and Kindermann 2002). We use subsequences of linguistic units — phonemes, syllables or words — occurring in the text as inputs to a standard SVM.

For the classification of video, we generated one video feature per shot. These features were considered to be analogous to words on the speech domain. We concatenated video features to build video n-grams parallel to speech n-grams. For classification with the joint-features system, video features were added directly to the bag-of-words representation in the type-frequency vector.

During preprocessing all n-grams up to a certain degree, e.g. $n = 3$, were generated, yielding several million features. To reduce this number we used a statistical test to eliminate unimportant ones. We employed a Bayesian version of the likelihood ratio test for independence. The procedure is discussed in (Gelman et al. 1995). A higher degree n-gram was selected only if it had a additional predictive power with respect to the class not contained in its lower degree parts. In the experiments different threshold values for the test statistic are evaluated. For a high threshold more terms are omitted than for a lower threshold.

Recently a new family of kernel functions — the so called string kernels — has emerged in the SVM literature (Watkins 98, Haussler 99). In contrast to usual kernel functions these kernels characterize sequences of symbols by pairs, triples, etc. of words. String kernels have been applied successfully to problems in the field of bio-informatics (Leslie, Eskin, and Noble 02) as well as to the classification of written text (Lodhi et al. 01). Our approach is equivalent to a special case of the string kernel. We go beyond the original string kernel approach insofar as we investigate building strings from different basic units. We employ the word “n-gram” to refer to sequences of linguistic units and reserve the expression “kernel” for traditional SVM-kernels. The kernels that we use in the subsequent experiments are the linear kernel, the polynomial kernel, sigmoid kernel and the Gaussian RBF-kernel (Cristianini and Shawe-Taylor 2000).

We present the results of experiments which applied SVMs to the classification of radio documentaries and to the classification of television video news from the N24 and n-tv channels. The video classification task is similar to the TREC-9 Spoken Document Retrieval track, which evaluated the retrieval of broadcast news excerpts using a combination of automatic speech recognition and information retrieval technologies[5].

Experiments

Our classification consists of deciding for each previously unseen document in test collection, whether or not it belongs to a given category. Let c_{targ} and e_{targ} respectively denote the number of correctly and incorrectly classified documents of the target category $y = 1$ and let e_{alt} and let c_{alt} be the same figures for the alternative class $y = -1$. We use the *precision* $prec = c_{targ}/\left(c_{targ} + e_{alt}\right)$ and the *recall* $rec = c_{targ}/\left(c_{targ} + e_{targ}\right)$ to describe the result of an experiment. In a specific situation a decision maker has to define a loss function and quantify the cost of misclassifying a target document as well as a document of the alternative class. The F-measure is a compromise between both cases (Mannning and Schütze 1999)

$$F_{val} = \frac{2}{\frac{1}{prec} + \frac{1}{rec}}, \qquad (1)$$

If recall is equal to precision then F_{val} is also equal to precision and recall. As the F-value seems to be more stable because it is not affected by the tradeoff between recall and precision we use it as our main comparison figure. We perform classification experiments on two different multimedia collections. First, the *Kalenderblatt* corpus of German-language radio documentaries and second, the German television news corpus, containing video news recorded from two different television stations.

Experiments with the *Kalenderblatt* Corpus

In order to evaluate a system for spoken document classification, a large audio document collection annotated with classes is required. It is also helpful to have a parallel text document collection consisting of literal transcriptions of all the audio documents. Classification of this text document collection provides a baseline for the spoken document system.

The Deutsche Welle *Kalenderblatt* corpus meets these requirements. The data set consists of 952 radio programs and the parallel transcriptions from the Deutsche Welle *Kalenderblatt* web-page http://www.kalenderblatt.de aired from January 1999 to August 2001. Each program is about 5 minutes long and contains 600 running words. The programs were written by about 200 different authors and are read by about 10 different radio reporters and are liberally interspersed with the voices of people interviewed. This diversity makes the Deutsche Welle *Kalenderblatt* an appealing resource since it represents a real world task. The challenge of processing these documents is further compounded by the fact that they are interspersed with interviews, music and other background sound effects.

Although the text transcriptions of the Deutsche Welle *Kalenderblatt* data set are not perfect, they are accurate enough to provide a text classification baseline for spoken document classification experiments. The transcriptions were decomposed into syllables for the syllable experiments and phonemes for the phoneme based experiments using the transcription module of the BOSSII system (Klabbers et al. 2001).

In order to train and to evaluate the classifier we needed topic class annotations for all of the documents in the data

[5] see http://www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.html

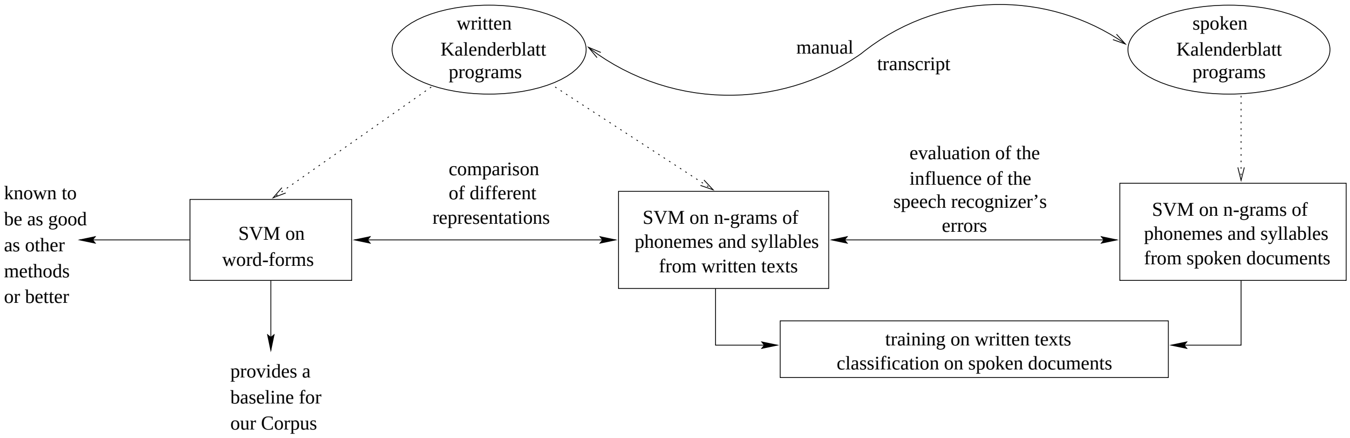
written Kalenderblatt programs
manual
transcript
spoken Kalenderblatt programs
comparison of different representations
evaluation of the influence of the speech recognizer's errors
known to be as good as other methods or better
SVM on word-forms
SVM on n-grams of phonemes and syllables from written texts
SVM on n-grams of phonemes and syllables from spoken documents
training on written texts classification on spoken documents
provides a baseline for our Corpus

Figure 3: The available data and the structure of our experiments.

set. We chose as our list of topics the International Press Telecommunications Council (IPTC) subject reference system (http://www.iptc.org). Annotating the data set with topic classes was not straightforward, since which topic class a given document belongs to is a matter of human opinion. The agreement of the human annotators about the class of documents represents an upper bound for the performance of the SVM classification system (table 1). Our corpus poses a quite difficult categorization task. This can be seen from figure 1 where the discrepancies of the different human annotators are shown. If we assume that one annotator provides the "correct" classification then precision and recall of the other annotator is about 70%. As the final classification was defined by a human this is some upper limit of the possible accuracy that can be achieved.

We performed experiments with respect to three topic categories: 'politics' of about 230 documents and 'science' with about 100 documents and 'sports' of about 40 documents. This should give an impression of the possible range of results. Experiments with smaller categories led to unsatisfactory results. In preliminary experiments RBF-kernels turned out to be unstable with fluctuating results. We therefore concentrated on linear kernels.

In the standard bag-of-words approach texts are represented by their type-frequency-vectors: word-forms of the written text, syllables derived from written text using BOSSII, phonemes derived from written text using BOSSII, syllables obtained from spoken documents by ASR, and phonemes obtained from spoken documents by ASR

In our experiments we compare the properties of three different representational aspects and their effect on classification performance: (1) Representation of a document by words. (2) Representation by simple terms or n-grams of terms, where 'non-significant' n-grams are eliminated. (3) Terms generated from the written representation or terms produced by ASR. As all representations are available for the same documents we can compare the relative merits of the representations. The setup is shown in figure 3.

We used ten-fold cross-validation to get large enough training sets. These experiments were repeated four times with different random seeds to reduce the variance of results due to different training sets. In the tables shown below the average values over all experiments with different seeds are shown.

First we performed experiments on the text transcriptions of the Deutsche Welle *Kalenderblatt* dataset in order to achieve a baseline. Second we performed the same experiments on output from the generic implementation of the speech recognition system.

Experiments with Written Material

As can be seen in the upper part of table 2 topic classification using simple words starts with an F-value of 68.9% for 'politics', 63.5% for 'science', and 86.0% for sports.

For all classes the syllables yield better results than words. For 'politics' syllables reach an F-value of 71.3% which is 2.4% better than the best word figure. There is a gain by using n-grams instead of single syllables which nevertheless reach an F-value of 70.4%. Longer n-grams reduce accuracy. This can be explained by their increasingly lower frequency of occurrence.

For the smaller category 'science' there is a dramatic performance increase to an $F_{val} = 73.2\%$ compared to an optimal 63.5% for words. Here n-grams perform at least worse than simple terms, perhaps as they are more affected by the relatively large noise in estimating syllable counts. The good performance of syllables again may be explained by more stable estimates of their frequencies in each class. It is interesting that in the larger 'politics' class n-grams work better in contrast to the smaller 'science' class.

For sports the results are rather stable across different inputs. Here it is remarkable that a low test threshold, which leads to a higher number of input features, usually has inferior performance.

Also phoneme n-grams yield lower errors than words. For 'politics' there is no significant difference F-values compared to syllables, whereas for the small category 'science' there is again a marked increase to an F-value of 75.2% which is 1.9% larger than for syllables. The average length of German syllables is 4 to 5 phonemes, so phoneme tri-

Table 2: Classification results on spoken and written material. Linear kernels and ten-fold cross-validation are applied.

linguistic		n-gram		**politics**			**science**			**sport**		
units	source	degree	thresh.	prec.	recall	F_{val}	prec.	recall	F_{val}	prec.	recall	F_{val}
words	written	1	0.1	76.8	60.8	67.8	54.5	76.20	63.53	100.0	54.17	70.3
		1	4.0	66.6	71.5	**68.9**	72.7	56.49	**63.55**	91.4	81.25	**86.0**
		2	0.1	71.2	63.9	67.3	63.8	62.26	63.03	98.6	45.83	62.5
		2	4.0	71.2	66.1	68.6	82.3	45.67	58.73	93.7	72.22	81.5
		3	0.1	71.5	64.0	67.5	65.8	50.48	57.14	98.6	46.53	63.2
		3	4.0	71.4	66.1	68.6	83.9	43.75	57.51	93.5	69.44	79.6
syllables	written	1	0.1	73.6	68.0	70.6	70.7	75.96	**73.23**	100.0	72.92	84.3
		1	4.0	67.6	73.4	70.4	66.3	79.33	72.22	100.0	77.78	**87.5**
		2	0.1	69.2	72.7	70.9	68.4	74.52	71.35	100.0	51.39	67.9
		2	4.0	72.0	69.9	70.9	72.3	64.66	68.26	94.2	79.17	86.0
		3	0.1	70.8	71.1	71.0	73.8	65.62	69.46	100.0	52.78	69.1
		3	4.0	69.5	70.9	70.2	77.7	57.69	66.18	93.3	76.39	84.0
		4	0.1	71.2	71.4	**71.3**	78.0	61.54	68.81	100.0	52.08	68.5
		4	4.0	71.0	68.5	69.7	77.3	55.53	64.58	91.1	77.78	83.9
		5	4.0	71.4	67.2	69.2	76.3	52.64	62.25	91.0	77.08	83.4
		6	4.0	70.0	68.3	69.1	76.7	52.88	62.54	89.4	76.39	82.4
phonemes	written	2	0.1	55.5	85.9	67.5	58.4	91.83	71.35	94.9	77.08	85.0
		2	4.0	57.4	85.7	68.8	58.4	88.70	70.42	90.8	75.00	82.1
		3	0.1	70.9	69.9	70.4	74.6	75.72	**75.17**	100.0	68.75	81.4
		3	4.0	60.7	81.6	69.6	71.8	77.64	74.60	100.0	74.31	85.2
		4	0.1	65.3	77.0	**70.7**	80.1	65.62	72.13	82.3	73.61	77.7
		4	4.0	63.1	78.5	70.0	77.8	66.35	71.60	94.8	75.69	84.1
		5	4.0	72.0	68.8	70.3	79.7	56.73	66.29	95.1	80.56	**87.2**
		6	4.0	67.6	74.0	70.6	73.2	55.77	63.31	93.5	80.56	86.6
syllables	spoken	1	0.1	58.1	75.8	**65.8**	30.4	67.31	41.87	33.7	47.22	39.3
		1	4.0	56.9	74.7	64.6	40.1	45.67	**42.68**	77.0	27.08	**40.0**
		2	0.1	73.1	51.0	60.1	45.6	28.37	34.95	52.5	6.25	11.2
		2	4.0	54.2	64.9	59.1	37.7	21.63	27.46	0.0	0.00	0.0
		3	4.0	64.1	41.1	50.1	50.8	3.37	6.28	50.0	1.39	2.7
		4	4.0	68.7	35.5	46.7	100.0	1.92	3.77	75.0	4.17	7.9
		5	4.0	69.0	34.3	45.8	100.0	1.92	3.77	75.0	4.17	7.9
		6	4.0	68.6	33.6	45.1	100.0	1.92	3.77	75.0	4.17	7.9
phonemes	spoken	2	0.1	55.9	66.8	60.9	44.4	37.02	40.34	51.6	41.67	**46.1**
		2	4.0	47.8	79.3	59.7	39.7	45.43	**42.38**	41.0	45.83	43.2
		3	4.0	57.8	70.6	**63.6**	50.6	30.53	38.07	13.9	61.11	22.6
		4	4.0	64.1	61.6	62.8	20.7	60.34	30.83	27.9	35.42	31.2
		5	4.0	66.8	50.1	57.3	28.0	25.24	26.53	50.0	1.39	2.7
		6	4.0	57.7	57.3	57.5	30.4	8.17	12.88	0.0	0.00	0.0

grams in average are shorter and consequently more frequent than syllables. This explains the high F-value of phoneme trigram in the small category. Note that for both categories we get about the same accuracy which seems to be close to the possible upper limit as discussed above. For sports we get similar results as for words and syllables.

The effect of the significance threshold for n-gram selection can be demonstrated for bigrams, where the levels of 0.1 and 4 were used. The selection of features according to their significance is able to support the SVMs capability to control model complexity independently of input dimension.

Experiments with Spoken Documents

As discussed above the language model of the generic speech recognizer was trained on a text corpus which is different from the spoken documents to be recognized. Only 35% of the syllables produced by ASR were correct when counting insertions, omissions and wrong assignments as error. With the experiments we can investigate if there are enough regularities left in the output of the ASR such that a classification by the SVM is possible. This also depends on the extent of systematic errors introduced by the ASR.

Again we performed experiments with respect to the two topic categories ‘politics’ and ‘science’. In the next section we evaluate the results for spoken documents and compare

Table 3: SVM classification of spoken documents when trained on written material. Only the topic category 'politics' is considered. Linear kernels are applied and ten-fold cross-validation is performed.

linguistic	n-gram		results for **politics**		
units	degree	thresh.	prec.	recall	F_{val}
syllables	1	0.1	57.5	57.7	57.6
	1	4.0	55.6	54.1	54.8
	2	0.1	72.5	33.6	46.0
	2	4.0	64.5	53.6	**58.6**
	3	0.1	79.1	30.9	44.4
	3	4.0	71.2	42.7	53.4
	4	0.1	79.3	31.4	45.0
	4	4.0	74.3	38.2	50.5
	5	4.0	74.5	34.5	47.2
	6	4.0	74.5	33.2	45.9
phonemes	2	0.1	48.8	82.3	61.3
	2	4.0	55.5	69.1	61.5
	3	0.1	57.6	77.7	**66.2**
	3	4.0	59.5	74.1	66.0
	4	0.1	65.4	60.9	63.1
	4	4.0	59.8	69.5	64.3
	5	4.0	60.8	70.5	65.3
	6	4.0	62.2	67.3	64.6

them to the results for written material. As before the SVM was trained on the output of the ASR and used to classify the documents in the test set. The results are shown in the lower part of table 2.

For 'politics' simple syllables have an F-value of 65.8%. This is only 5.5% worse than for the written material. The effect of errors introduced by the ASR is relatively low. There is a sharp performance drop for higher order n-grams with $n > 3$. A possible explanation is the fact that the language model of the ASR is based on bigrams of syllables.

For 'science' classifiers using syllables for spoken documents yield only an F-value of 42.7% and perform far worse than for written documents (73.2%). Probably the errors introduced by ASR together with the small size of the class lead to this result. The same decrease can be observed for 'sports', where we get an F-value of 40.0% compared to 87.5% for written material.

Surprisingly phonemes yield for the topic category 'politics' on spoken documents an F-value of 63.6% which is nearly as good as the results for syllables. This result is achieved for 3-grams. For the small category 'science' phonemes yield 42.4% which is nearly identical to the result for syllables. For 'sports' the phonemes with an F-value of 46.1% perform much better than syllables on spoken documents.

Classification of Spoken Documents with Models Trained on Written Material

To get insight into the regularities of the errors of the speech recognizer we trained the SVM on synthetic syllables and phonemes generated for the written documents by BOSSII and applied these models to the output of the ASR.

Table 4: Optimal F-values for the Kalenderblatt experiments.

topic	data used for		optimal F-values		
category	training	test	word	syll.	phon.
'politics'	written	written	68.9	71.3	70.7
'politics'	spoken	spoken	—	65.8	63.6
'politics'	written	spoken	—	58.6	66.2
'science'	written	written	63.5	73.2	75.2
'science'	spoken	spoken	—	42.7	42.4
'sports'	written	written	86.0	87.5	87.2
'sports'	spoken	spoken	—	40.0	46.1

The results for this experiment are shown in table 3. Whereas models trained on syllables arrive at an F-value of 58.6% the phonemes get up to 66.2%. This is nearly as much as the maximum F-value of 66.0% resulting from a model directly trained on the output of a speech recognition system. This means that — at least in this setting — topic classification models may be trained without loss on synthetically generated syllables instead of genuine syllables obtained from a ASR.

Experiments with the German News Corpus

We test our speech and video features on a task involving classification of television news into basic topic categories. Data was collected from two German language television news stations, N24 and n-tv. The corpus was segmented by hand into 693 audio-visual documents, each constituting a homogenous segment from the TV-news. The reports are between 30 seconds and 3 minutes long. The data from N24, 353 documents total, were collected in May and June 2002. Sports reports during this period were often dedicated to coverage of World Cup soccer. The data from n-tv, 340 documents total, was collected during April 2002. Dominating news at that time was coverage of the high school shootings in Erfurt, Germany.

The Topic Classes

The news were again annotated by hand, this time by a team of two human annotators. Due to the nature of the domain, there was not a large discrepancy between the opinions of the two annotators. In this case there is no upper limit on the F-measure that the classifiers should be able to attain. As for the radio documentary corpus, we used category labels from the IPTC system. We added two categories that were appropriate for the television news domain, advertisement and jingles. Each news segment was assigned at least one class, but could be assigned to multiple classes. We used the following thematic categories: politics (200), justice (120), advertisement (119), sports (91), conflict (85), economy (68), labor (49), human interest (40), disaster (38), jingles (22), culture (22), health (19), environmental issues (17), leisure (15), science (13), education (10), weather (8), social issues (6), religion (4). In parentheses are the number of documents assigned to each category. We performed tests on the largest 7 of these categories, which ensured that each

Table 5: Classification using best speech features from generic speech recognizer.

	F-measure for class . . .						
	just.	econ.	labor	polit	sports	confl.	ads
rnd.	17.3	9.8	7.0	28.9	13.0	12.3	17.1
syll.	54.9	50.0	**65.6**	**64.9**	50.2	42.2	90.5
phon.	53.1	43.5	61.8	64.3	51.0	43.4	89.0
video	49.8	24.5	33.3	43.9	47.7	35.0	84.6
audio	37.6	15.1	13.3	42.0	28.6	21.4	90.6
joint	**59.0**	**54.2**	37.0	59.1	**68.0**	**48.6**	**95.8**

Table 6: Classification results using speech features from the domain adapted speech recognition system.

	F-measure for class . . .						
	just.	econ.	labor	polit	sports	confl.	ads
rnd.	17.3	9.8	7.0	28.9	13.0	12.3	17.1
syll.	65.0	59.3	**85.3**	**74.7**	80.3	**73.5**	85.0
phon.	**67.3**	**60.2**	83.6	**74.7**	84.8	72.2	88.9
video	49.8	24.5	33.3	43.9	47.7	35.0	84.6
audio	37.6	15.1	13.3	42.0	28.6	21.4	90.6
joint	62.7	59.2	78.0	65.3	**86.6**	71.0	**93.4**

category had at least 40 positive examples. The remaining documents were used as additional negative examples.

We experimented with two different sets of speech features. First, with speech features from the generic, unadapted automatic speech recognition (ASR) system and then speech features extracted with the adapted system. In each case we tested the effect of combining speech features with video features. There are many parameters involved optimizing classification with support vector machines. In the tables we report results for each experiment with the best settings and in the text we comment in cases in which parameter settings yielded interesting performance differences.

Speech Features from Generic ASR

These experiments were performed with the speech recognizer before it was adapted to the domain. Table 5 reports the F-values of the best performing type of speech features. The second row contains the F-measure that would result, the documents were assigned randomly. Syllable n-grams were tested up to order 3 and phoneme n-grams were tested up to order 5. For sports and conflicts syllable 1-grams achieved the best speech results, while for the other classes phoneme 2- or 3-grams were optimal. Optimal video features were 1-grams for justice, labor and politics, while for the other classes video 2-grams were best.

Classifier performance of the category of advertisement is already markedly better than on other categories. Advertisements are characterized by large amounts of non-speech audio, and the classifier seems capable of generalizing over the nonsense syllables which the recognizer has generated for non-speech segments.

With the addition of video features to speech features ("joint"), classification results improved in 5 of the 7 categories, we experimented with. Classification for politics is worse and also the results for labor decreased drastically. In the fifth line of Table 5 it can be seen that the (non-speech) audio features did not yield good classification performance, except in the case of advertisements.

Speech Features from Improved ASR

We performed a second round of experiments using the output from the improved speech recognizer, which had been adapted to the news domain. As shown in table 6 classification results dramatically improved in 6 of 7 categories. When video features are added to speech features ("joint"), only two classification results were improved, sports and advertisement. The improvement reflects the relative importance of the images for conveying content in these two classes. Leaving out advertisement the error rate on the whole was reduced by 13% (for economy) up to 58% (for sports) with a mean of about 37% compared to the generic ASR.

Exploratory investigation tested four different SVM kernels, rbf, sigmoid, linear and polynomial and demonstrated that different kernels yielded better performance for different classes and different modalities. The difference between the best performing kernels and the second best performing kernels remained below 1% absolute. In general the best performing kernel was the sigmoid kernel.

Discussion and Conclusions

This paper reports on our efforts to extract effective "speech features" from spoken audio for use classifying German language radio documentaries and broadcast news into basic topic categories. We are also working towards the identification of optimal video features that can be used to supplement the speech features to improve classifier performance. Our experiments confirmed that simple, low level features are appropriate and useful for our classification task. From the audio stream we extract syllables, which are used to build non-orthographic speech features. Dependent on the category, phoneme 2- or 3-grams or syllable 1- or 2-grams yield maximal classification rates. We show that these speech features can be satisfactorily supplemented with simple color-based video features.

The main results of the classification experiments on the *Kalenderblatt* radio documentary data are summarized in table 4. From the experiments on the radio documentary data we were able to draw several important conclusion. First, n-grams of sub-word units like syllables and phonemes are better features than words for the classification of text documents. The improvement of classification performance may be dramatic for some document classes. Second, if the output of an automatic speech recognition (ASR) system is used for training and testing there is a drop of performance with respect to performance on text transcriptions. This drop is relatively small for larger classes and substantial for small classes. On the basis of syllable n-grams the SVM can compensate errors of a low-performance speech recognizer. Third, in our setup it is possible to train syllable classifiers on written material and apply them to spoken documents. This technique is important since written material is far easier to obtain in larger quantities than annotated spoken documents.

The classification experiments on the corpus of television news also yielded interesting results. On this corpus syllables and phonemes also proved to be effective for categorization. An enormous improvement resulted from improving the speech recognizer and adapting it to the news domain. An interesting result of our investigation is that combining video features with speech features does not consistently yield improved classifier performance. Instead, video features should only be used in cases where speech recognition rates are low, such as with generic speech recognition. Video features also proved helpful for classifying the categories of sports and advertisement, possibly due to the fact that these categories carry relatively more visual information than other categories. contain little speech and much background noise.

Future work will concern continued improvement of the syllable-based statistical language model. We have been able to improve syllable recognition rates without compromising coverage by adding a certain portion of words to the syllable-based language model, training what we call a mixed-unit language model. We intend to continue investigation on how the basic inventories of the statistical language model for speech recognition can be optimized to improve classification performance.

We will also concern ourselves with in depth exploration of the features that should be used for video classification. The three color-based feature spaces experimented with here represent only a fraction of video features potentially interesting for our task.

Acknowledgements

This study is part of the project “Personalisierte Interaktive Audio-, Voice- und Video-Informationen für Portal- und Auswertungsanwendungen” (Pi-AVIda) which is funded by the German Ministry for Research and Technology (BMFT, proj. nr. 107). We thank the Institute for Communication and Phonetics of the University of Bonn for contributing the Bonn Open Synthesis System (BOSSII) which we used for the syllabification of the written material of the Kalenderblatt radio programs and we thank Thorsten Joachims (Department of Computer Science at Cornell University) who made us available the SVM-implementation SVM^{light}. We also thank Deutsche Welle for allowing us to use their broadcastings.

References

Allamanche, E., Herre, J., Hellmuth, O., Frba, B., Cremer, M. (2001): AudioID: Towards Content-Based Identification of Audio Material. 110th Convention of the Audio Engineering Society, Amsterdam, The Netherlands.

Cristianini, Nello and Shawe-Taylor, John 2000. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Dumais, S., Platt, J., Heckerman, D., Sahami, M. 1998: Inductive learning algorithms and representations for text categorization. In: 7th Int. Conf. on Information and Knowledge Management, 1998.

Drucker, H, Wu, D., Vapnik, V. 1999. Support vector machines for spam categorization. IEEE Trans. on Neural Networks, 10 (5): 1048-1054.

Gelman, A., Carlin J.B., Stern, H.S., Rubin, D.B. 1995: *Bayesian Data Analysis*. Chapman, Hall, London.

Glavitsch, U., Schäuble, P. 1992: A System for Retrieving Speech Documents, SIGIR 1992.

Haussler, David 1999: Convolution Kernels on Discrete Structures, Technical Report, UCSL-CRL-99-10.

Joachims, Thorsten 1998. Text categorization with support vector machines: learning with many relevant features. Proc. ECML ’98, (pp. 137–142).

Jurafsky, Daniel and Martin, James. H. 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.

Klabbers, E., Stöber, K. , Veldhuis, R. Wagner, P., and Breuer, S. (2001): Speech synthesis development made easy: The Bonn Open Synthesis System. EUROSPEECH 2001.

Larson, M. 2001. Sub-word-based language models for speech recognition: implications for spoken document retrieval. Proc. Workshop on Language Modeling and IR. Pittsburgh.

Leopold, E. and Kindermann, J. 2002: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning* 46: 423–444.

Leslie, Christa; Eskin, Eleazar, Noble, and William Stafford 2002: The Spectrum Kernel: A String Kernel SVM Protein Classification. To appear: Pacific Symposium on Biocomputing.

Lodhi, Huma, Shawe-Taylor, John, Cristianini, Nello & Watkins, Chris 2001. *Text classification using kernels*. NIPS 2001, pp. 563-569. MIT Press.

Manning, Christopher D. and Schütze, Hinrich 2000: *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Paass, G., Leopold, E., Larson, M., Kindermann, J., Eickeler, S. 2002: SVM Classification Using Sequences of Phonemes and Syllables. Proc. ECML.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. John Wiley and Sons, New York.

Volmer, S. 2002. Fast Approximate Nearest-Neighbor Queries in Metric Feature Spaces by Buoy indexing. Proc. Conf. Visual Information Systems, Hsinchu, Taiwan, p.36-49.

Watkins, Chris 1998. Dynamic alignment Kernels. Technical Report, Royal Holloway, University of London. CSD-TR-98-11.