AUDIO-VISUAL ANALYSIS OF MULTIMEDIA DOCUMENTS FOR AUTOMATIC TOPIC IDENTIFICATION

Uri Iurgel, Steffen Werner, Andreas Kosmala, Frank Wallhoff
Gerhard-Mercator-University of Duisburg
Department of Computer Science
{iurgel,werner,kosmala,wallhoff}@fb9-ti.uni-duisburg.de

Gerhard Rigoll
Munich University of Technology
Institute for Human-Machine Communication
rigoll@ei.tum.de

ABSTRACT

This paper presents a system that shall automatically scan multimedia data like TV or radio broadcasts for the presence of specific topics and, whenever topics of users' interests are detected, alert the related user. Our current work on the three main modules of the system will be shown.

(1) The speech recognition system (with 18.7 % WER) is already among the most advanced German broadcast speech recognition systems. (2) The new and innovative topic identification approach, which is especially designed to work on the output of a speech recognizer, is compared to a standard text based approach. (3) The topic segmentation module has a good performance detecting real topic boundaries, not just scene cuts or speaker turns.

KEY WORDS

Audio and Video, Multimedia, Speech Processing, Topic Segmentation, Topic Identification

1 Introduction & Motivation

Newspapers, magazines, radio, television, world wide web - information is of strategic importance for business and governmental agencies as well as for citizens. The exponential evolution of multimedia makes it difficult to overview the opulence of information, and that's why important pieces of information have to be filtered and processed automatically.

Nowadays, information is mainly obtained by manually analyzing (reading, listening and watching) large audio and video databases and current broadcast multimedia sources (such as broadcast TV, radio or Internet streams). After having assigned topics to the incoming news and stories, only the items of interest or items regarding a specific request will be selected and further processed. The use of automatic methods for selective dissemination of information would enable such monitoring companies to cover a much larger variety of media sources by working more cost efficiently and providing 24 hours availability.

The objective of the presented work is to develop an intelligent software system that shall automatically scan multimedia data like TV or radio broadcasts for the presence of specific topics. Whenever topics of users' interests are detected, the system shall alert the related user by email or other means. Figure 1 illustrates on a functional level how the multimedia documents will be processed by such a system.

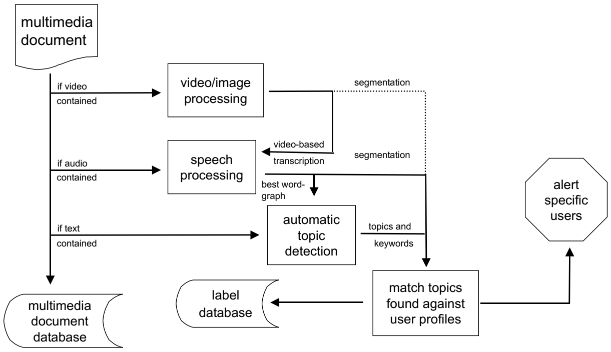

Figure 1. *Labeling of multimedia data and alert generation.*

As it can be seen in this figure, the architecture is capable of processing multimedia documents containing audio, video and text data. If the data contains video, video and audio processing techniques are used to segment the data into scenes, such as Newscaster or Report and to detect story boundaries as described in Section 3 [1]. The audio track is transcribed using the speech recognition methods described in Section 2. It is also possible to handle pure text data (e.g. acquired from the Internet) which is directly passed to the topic detection module. Otherwise, all available information will be used for topic detection, which is presented in Section 4.

2 Automatic Transcription of Multimedia Documents

The current state-of-the-art in broadcast speech recognition is characterized by the existence of a few experimental systems world-wide. Recognition error rates for these systems have been improved from about 50 % a few years ago down to around 20 % for the currently best systems, which often work on specific databases for American English. Compared to error rates for recognition of read speech, which are often below 10 %, improvements for broadcast speech

From: SPPRA 2002, ACTA Press, Calgary

recognition are required in order to develop systems useful e.g. for retrieval tasks in real life scenarios.

The processing of multimedia documents, especially broadcast news, requires partitioning into homogeneous segments and the identification and exclusion of non-speech segments (such as music or jingles). This was done partly manually (in this case overlapping speech was excluded, too) and partly by means of the developed audio segmentation described in [2] which is based on the Bayesian Information Criterion (BIC). This audio segmentation method is briefly described in Section 2.1. The first results obtained with the developed Duisburg broadcast recognition system in German called "DuBREC" are presented in Section 2.2.

2.1 Audio Segmentation

Several methods for audio segmentation have been proposed, like Akaike's Information Criterion (AIC) , the Bayesian Information Criterion (BIC) , the Consistent AIC (CAIC) and the Minimum Description Length (MDL). These and other methods have been compared in [3] and it has been shown that with optimal parameters, almost all algorithms perform comparably well.

The audio segmentation algorithm deployed in this work uses the BIC, which was among the best performing methods. The BIC follows the method of Tritschler and Gopinath [4] which will be described briefly. The algorithm takes a window of n audio features $x_1, \ldots, x_n$ and arbitrarily places a boundary at position i, resulting in two segments. It then decides whether it is more likely that one single model θ_1 has produced the output $x_1, \ldots, x_n$, or that two different models θ_{21} and θ_{22} have generated the two segments' output $x_1 \ldots x_i$ and $x_{i+1} \ldots x_n$ respectively. The decision rule to check if there is a boundary at point i is

$$\Delta BIC_i \overset{!}{<} 0 \qquad \text{with} \tag{1}$$

$$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_w| + \frac{i}{2}\log|\Sigma_f| + \frac{n-i}{2}\log|\Sigma_s| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n. \tag{2}$$

Σ_w denotes the covariance matrix of all window feature vectors $x_1, \ldots, x_n$, Σ_f and Σ_s are the covariance matrices of the features of the first and second segment respectively. d is the feature vector dimension. According to theory, the penalty weight λ should equal 1, but practical applications show better results with $\lambda \neq 1$.

If for a point i $\Delta BIC_i < 0$, then also for some points j surrounding i there will be $\Delta BIC_j < 0$. The algorithm decides that the boundary is at the point with the lowest ΔBIC value.

To detect all audio segments of a news show, the window is shifted over all feature vectors with varying lengths n and varying i. See [4] for details.

After implementing the above described algorithm, it was noticed that sometimes segment boundaries are set too early, roughly one or two syllables before the speaker finishes. Instead of considering the point i at which the minimum of ΔBIC occurs as a boundary, the middle of those two points where the ΔBIC value crosses the 0 line was chosen. This modification improves the segmentation accuracy and reduces the number of boundaries appearing too early.

As feature vectors we use 39 - dimensional Mel - Cepstral vectors without mean subtraction. The penalty weight has been set to $\lambda = 3.0$.

2.2 The Duisburg Broadcast Recognition System in German: DuBREC

This section presents the developed broadcast recognition system (called "DuBREC") and describes the underlying decoder (called "DuCoder"), the training material and the results of the recognition experiments.

2.2.1 The baseline system - LVCSR DuCoder

DuCoder is the LVCSR decoder developed at Duisburg University [5]. It performs the Viterbi search for the most probable hypothesis on word level using Hidden Markov Models (HMMs). In principle, the decoding procedure is similar to the one of the stack decoders [6, 7], which set up and initialize stacks $S_1, ..., S_T$ at each time-step t. A stack S_t contains a sorted list of word hypotheses H_t at time t. After choosing a stack, all the stored stack's hypotheses get expanded simultaneously by performing a single word recognition starting at t, which results in new hypotheses $H_{t+\tau}$ that get pushed to the specific stacks $S_{t+\tau}$.

Operating in its standard mode, the DuCoder uses stacks of fixed size for each time-step, in which the best hypotheses ending at that time are stored. The different implemented stack selection and exclusion strategies are outlined in detail in [5, 8]. The single word expansion is organized as a time-synchronous search (which follows the principles of Token Passing) through a recognition network. The reduction of the number of nodes which have to be expanded at each time step results in a tree structure of the search space.

A large dictionary is required in order to achieve reasonable out of vocabulary (OOV) rates, specially for the German language with its frequent use of compound words. However, such large dictionaries in combination with 3- or 4-gram language models would require a huge memory space for the decoding process. On the other hand, regarding a simple time step during decoding, large parts of the language model (LM) are deactivated and can be neglected. Thus, the DuCoder is designed to process specially formatted cache based language models, where only the relevant parts of a language model are buffered and, if needed, additional parts can be reloaded during decoding. The removal of irrelevant buffer entries is carried out according to the simple least recently principle. This enables the processing of dictionaries with more than 65k entries in combination with complex language models even on standard PCs with memory requirements between 100 and 150 MByte.

2.2.2 Training material

The training of large vocabulary speech recognition systems, including broadcast systems, requires large amounts of transcribed audio data. Especially the HMM training performs an alignment between the audio signal and the phone models, which is usually derived from a nearly perfect orthographic transcription of the speech data and a good phonetic lexicon.

Within the ALERT project [9], minimum requirements have been defined for the multilingual corpus (consisting of French, Portuguese and German) that will serve as the basic corpus for the training and the evaluation of speech recognition components. This definition is 50 hours for training, 3 hours for development and 3 hours for evaluation.

The DuBREC system was trained with about 100 hours of manually transcribed broadcast news data. During the development of the DuBREC system a 30 - minute preliminary test set was used, comprising three different radio and three different TV stations. The final evaluation of the system is planned within the ALERT project using a test set of about 3 hours.

2.2.3 Experimental results

Starting from a baseline system [8], which was trained with a mixture of spontaneous speech and read sentences (such as Verbmobil and Phondat), the broadcast system was trained on the manually transcribed radio and TV data. Therefore, a general 95k dictionary and a trigram language model (LM) were derived from newspaper texts. The recognition results with the preliminary test set are displayed in Table 1 as monophone DuBREC system 1 and triphone DuBREC system 5. The monophone system consists of 50 different phone and 17 non-speech models (e.g.: pause, silence, filler, breath, cough, ...). The triphone system consists of 9307 models (with 96417 mixtures) which were derived with a tree-based clustering method.

When checking that general dictionary it was realized that a lot of phonemization errors are present. There were some systematical (such as ”ch” separation into phone ”x” or ”C” and the removal of double phonemes) and some manual corrections necessary. The manual word checking was done in accordance to the likelihood of the word occurrence. The impact of the correction was only measurable if the corresponding changes occur within the most likely 5000 words. The recognition results are shown as System 2 in Table 1.

After the dictionary correction an extension of the general dictionary size was done by including the most likely 1000 words derived from the transcriptions, which had not already occurred in that one. It was recognized that the best results (System 3) were obtained with a dictionary size of 98k. Because the general LM was derived mainly from newspaper texts, a new trigram LM was derived from the transcripts. That new language model was interpolated into the general one. Using that interpolated LM the word error rate could be reduced another time (System 4 in Table 1).

Table 1. *Recognition results for the preliminary test set and various systems.*

System	DuBREC system with:	WER
1	general dict. and LM (Monophone)	39.2 %
2	corrected general dict. and general LM (Monophone)	37.5 %
3	specific dict. and general LM (Monophone)	36.5 %
4	specific dict. and LM (Monophone)	33.7 %
5	general dict. and LM (Triphone)	22.5 %
6	specific dict. and LM (Triphone)	19.2 %
7	gender dependent Models (Triphone)	18.7 %

By using the extended specific dictionary and the interpolated LM for the triphone DuBREC system the recognition performance has been improved from 22.5 % (WER) to 19.2 % (Sytem 6). In addition to the standard ML training a gender dependent training was conducted too. In that case only the state transitions and the means were updated (System 7).

All recognition results presented in Table 1 were achieved with respect to a real time factor (RTF) of about 10. As it can be seen in Table 1, System 7 has already a very good performance and can be considered as one of the most advanced German broadcast speech recognition systems.

3 Automatic Audio-Visual Topic Segmentation of Multimedia Documents

The extraction of topic information from a multimedia document (such as broadcast news) requires a correct detection of topic boundaries. The well-known methods for audio segmentation are mainly capable of detecting significant changes in the audio signal, thus indicating speaker turns or transitions from speech to non-speech segments. One approach to topic segmentation is to consider audio boundaries as topic boundaries. However, this will lead to an over-segmentation of the document, as there will be many more audio boundaries than topic boundaries. Our observations have shown that 76.7 % of the existing topic boundaries are detected, but 81.2 % of the boundaries are mistakenly inserted by such an audio segmentation algorithm (see Table 2).

If a multimedia document contains video information, this additional information might be helpful to detect correct topic boundaries. Eickeler and Müller [10] presented a novel approach to scene classification based on Hidden Markov Models which was extended in order to extract real *topic* boundaries.

Our approach describes the topics of a news show using an HMM-based topic model that makes use of video as well as of audio features (Section 3.2). Each show is thus modeled as a stochastic sequence of topics.

3.1 Video Segmentation

Eickeler and Müller [10] used only the visual track of a news show. They segmented it into content classes (Begin, End, Newscaster, Report, Interview, Weather Forecast) and into edit effects (Cut, Dissolve, Window Change, Wipe). Each of these classes is modeled by a Hidden Markov Model (HMM). The models are combined to a flexible news show structure.

A feature vector consisting of 12 video features represents each image [10]. Among these features are the center, the velocity and variance of motion, intensity of motion, difference histogram and a feature which improves the detection of dissolve edit effects. All these features are based on luminance only. Three more values are added to the vector, giving the average value of the luminance (Y) and the two chrominance (U,V) components. The scene segmentation and classification of a show are the result of calculating the sequence of HMMs that most probably has generated the observed feature vector.

As the described approach does not allow to detect topic boundaries, the following extensions have been introduced.

- Video and audio features are combined into the HMM structure.
- An adapted video model is used which represents the topic structures within a news show. (see Figure 2).

3.2 Audio-Visual Segmentation

To combine the audio and video information, we first segment the audio track using the modified BIC algorithm described in Section 2.1. The resulting boundary positions are rounded to the nearest video frame. This method outputs a 1-dimensional audio feature stream which is added to the 12 video features described in Section 3.1. One new edit effect has been introduced, which we call AVcut (audio and video cut). It describes a hard video cut with an audio cut nearby. Consequently, the Cut edit effect is defined as a hard video cut without any audio cut nearby.

Besides combining audio and video features, a novel news model has been introduced that reflects typical *topic* structures instead of the structure of a whole news show. Figure 2 depicts such a model with topic beginnings marked by gray circles. The model on the top represents a news show with embedded topic structures (N_topic and R_topic) and edit effects (ed_eff) that are defined below. The classes Newscaster and Report are abbreviated by N and R respectively. Square brackets denote optional elements.

algorithm	audio	audio-visual
precision	18.8 %	64.8 %
recall	76.7 %	91.5 %

Table 2. *Topic segmentation performance.*

3.3 Experiments and results

As a test and training set recordings of the most important German TV news show were used. All shows last 15 minutes. Nine shows, for a total time of 2:15 hours, have been used for training and testing each algorithm. For the experiments, we used the hold-out method, i.e. we tested each show with a system trained on the other eight.

For all shows a topic boundary list was manually created as a reference. These reference lists were compared to the results of our algorithm in terms of precision and recall rate. The topic segmentation rates are shown in Table 2.

The results we achieved for the combined audio and video approach show its great ability to detect almost all topic boundaries that are present in the news show. However, there are quite a high number of boundaries that are mistakenly inserted. This leads to over-segmentation, which could be compensated for by a subsequent topic identification step that clusters the segmented parts with the same topics. With its high recall rate, this algorithm is very well suited for cases that need to detect all boundaries.

4 Automatic Topic Identification

Our topic identification module implements two different algorithms for topic detection. The first one works on a word basis by identifying each word in the vocabulary with a unique number. As this approach shows limitations when applied to topic identification on erroneous transcriptions generated by automatic speech recognition, we are currently investigating a novel hybrid approach to topic identification that is based on a sub-word feature extraction. It does not depend on a pre-defined vocabulary, and it is to some extent robust to speech recognition errors.

4.1 Standard approach on a word feature basis

The standard approach assigns to each word in the text an index number of the vocabulary list. Each topic is modeled with a discrete single state HMM using the word index numbers as observations. The vocabulary is extracted from the test and training set and from the vocabulary of the speech recognizer in such a way that all words can be assigned an index. This system is motivated by [11].

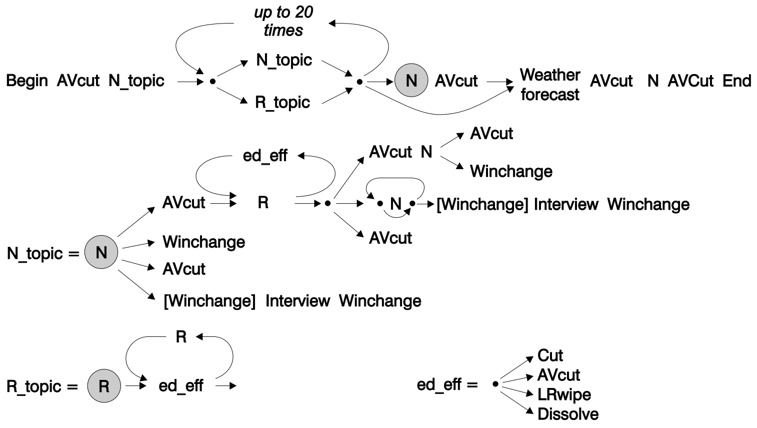

Figure 2. *Model of a news show with embedded topic structures.*

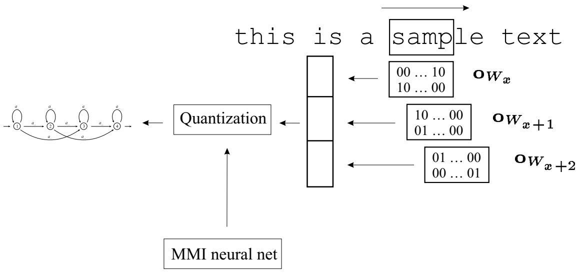

Figure 3. *Overview of the architecture of the character-based topic identification system.*

4.2 Character-based MMI approach

An alternative method has been implemented that extracts features on a character basis and quantizes the feature vectors using the Maximum Mutual Information (MMI) criterion. It has been developed with respect to erroneous speech recognition, especially with compound words being mistakenly split or combined.

A sliding window W of size w (typically $w = 1..7$) characters scans the text. From each character C in the window, we extract a 32-dimensional binary feature vector $\mathbf{o}_C$. Exactly one component of $\mathbf{o}_C$ gets a value of +1, the others are assigned a value of 0. The vector representing an a has a +1 value at its first component, a b gets a +1 at the second component, and so on. The feature vector of each text window $\mathbf{o}_W$ thus has a size of $w * 32$ with w components being +1 (see Figure 3). Additionally, center characters are implicitly duplicated by combining adjacent feature vectors. This leads to improved recognition results.

In German, words tend to change their stem vowel or umlaut when changing their grammatical function; this change may not be detected by the speech recognizer. The idea behind using these big feature vectors is that if characters are wrongly recognized, the distance between the vector of the correct spelling and the vector with one wrong character is the same whatever the wrong character may be. Scanning with a window, provided its size is properly chosen, emphasizes the morphemes of the text, and thus the semantic information carriers.

The feature vectors are then quantized using a quantization codebook that has been created by maximizing the mutual information between the prototype vectors and the topics. Each topic is modeled with a discrete HMM by using the indices of the prototype vectors as observations.

4.3 Experiments and Results

As a text source, we use pre-segmented, manually created summaries of TV news in German and the output of our speech recognition system. For our experiments we have defined the following training and test sets:

- **A**: summaries, no stop word removal, no stemming, erroneous texts. **A1**: 22 topics **A2**: 173 topics
- **B**: summaries, deletion of 150 stop words, optimal text. 22 topics
- **C**: training set: summaries from 898 topics; test set: 48 pre-segmented, automatically transcribed stories.

In the summaries of test set A some words are separated into two single words. Besides, there are some special abbreviations. This means the text basis is not optimal, thus simulating in a more or less rough way errors which are made by automatic speech recognition. In test set B, which is made up of different texts, there are no such errors. We give our recognition rates as the ratio of the number of correctly classified summaries to the total number of tested summaries. There is no extra model for out-of-topic summaries, as all tests topics were restricted to the trained topics.

The results of the new and the standard approach are listed in Table 3. The standard approach works significantly better on test set B (optimal text), whereas it is only slightly better than the proposed approach on test sets A1 and A2 (erroneous text). However, it will have problems with speech recognition errors when transcribed words do not appear in the vocabulary. When we removed all spaces from the test and training set, the character-based approach showed only a slight decrease in performance, whereas the word-based standard approach will fail to work on this test set. Removing spaces simulates the effect of a speech recognizer that mistakenly combines or splits compound words.

First results were obtained on the output of the automatic speech recognition system described in Section 2 (test set C). Much more training topics (898) were usesd than in the other sets. The recognition rate significantly decreases for both systems; as can be seen, the new system does not perform as good as the standard system for a high number of potential topics.

new system	standard system	test set
49.5 %	50.4 %	A1
45.3 %	—	A1, no spaces
31.3 %	35.3 %	A2
66.6 %	78.0 %	B
22.9 %	35.4 %	C

Table 3. *Comparison of the best rates of the presented system to a standard approach.*

5 Conclusion

We have presented a system that automatically scans multimedia data like TV or radio broadcasts for the presence of specific topics. Each of the three main modules (speech recognition, topic segmentation and topic identification) show a good performance.

The speech recognition module achieved a performance of 18.7 % word error rate using a gender dependent triphone system. This module can be considered as one of the most advanced German broadcast speech recognition systems.

The well-performing audio-visual topic segmentation module is, unlike many other approaches, able to detect real topic boundaries instead of just audio or video cuts.

For topic identification, two algorithms were presented. The new innovative approach is more robust against transcription errors, e.g. mistakenly combined or split compound words, than the standard one, but recognition rates should be improved for a high number of potential topics. The standard approach shows better overall recognition rates.

6 Acknowledgments

The content and themes discussed in this paper largely benefited from the ALERT project. This project (IST-1999-10354) is sponsored by the European Commission within the 5^{th} Framework Programme in the area of Human Language Technologies (HLT).

References

[1] Uri Iurgel, Ralf Meermeier, Stefan Eickeler, and Gerhard Rigoll, “New Approaches to Audio-Visual Segmentation of TV News for Automatic Topic Retrieval,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City , Utah, May 2001.

[2] Ralf Meermeier, “Automatische Transkribierung von Radionachrichten,” M.S. thesis, Faculty of Electrical Engineering - Computer Science, Gerhard-Mercator-University Duisburg, June 2000, in German.

[3] Mauro Cettolo and Marcello Federico, “Model selection criteria for acoustic segmentation,” in *Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, Paris, France, 2000, pp. 221–227.

[4] Alain Tritschler and Ramesh Gopinath, “Improved speaker segmentation and segments clustering using the bayesian information criterion,” in *Proc. EUROSPEECH*, 1999, vol. 2, pp. 679–682.

[5] Daniel Willett, Christoph Neukirchen, and Gerhard Rigoll, “Ducoder - the Duisburg University LVCSR Stackdecoder,” in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, pp. 1555–1558.

[6] S. Renals and M. Hochberg, “Start-Synchronous Search for LVCSR,” in *IEEE Trans. on Speech and Audio Processing (SAP)*, Sept. 1999, vol. 7, No. 5, pp. 542–553.

[7] M. Schuster, “Nozomi - A fast, Memory-Efficient Stack Decoder,” in *5th International Conference on Spoken Language Processsing (ICSLP)*, Sydney, Dec. 1998, pp. 1835–1838.

[8] Daniel Willett, *Beitraege zur statistischen Modellierung und effizienten Dekodierung in der automatischen Spracherkennung*, Ph.D. thesis, Faculty of Electrical Engineering, Gerhard-Mercator-University Duisburg, Nov. 2000.

[9] “ALERT homepage,” http://alert.uni-duisburg.de.

[10] Stefan Eickeler and Stefan Müller, “Content-based video indexing of tv broadcast news using hidden markov models,” in *Proc. IEEE ICASSP*, 1999, pp. 2997–3000.

[11] Richard Schwartz, Toru Imai, Francis Kubala, Long Nguyen, and John Makhoul, “A maximum likelihood model for topic classification of broadcast news,” in *Proceedings of Eurospeech ’97*, Rhodes, Greece, Sept. 1997, pp. 1455–1458.