THE ALERT SYSTEM: ADVANCED BROADCAST SPEECH RECOGNITION TECHNOLOGY FOR SELECTIVE DISSEMINATION OF MULTIMEDIA INFORMATION

Gerhard Rigoll

Gerhard-Mercator-University Duisburg
Faculty of Electrical Engineering, Dept. of Computer Science
D-47057 Duisburg, Germany
rigoll@uni-duisburg.de

ABSTRACT

This paper presents a brief description of the ALERT system, which is under development by a consortium working on a research project sponsored by the European Commission. The ALERT system uses advanced speech recognition technology and video processing techniques in order to process large broadcast speech archives and multimedia information resources for the purpose of extracting specific information from such databases and inform selected customers about its contents. It is one of the most ambitious projects currently carried out in the Human Language Technologies (HLT) area (see also **http://alert.uni-duisburg.de**). The paper describes the objectives of the overall system, its basic system architecture and the scientific approach taken in order to realize the specified demonstrators.

1. INTRODUCTION

The ALERT system is currently developed by an international consortium consisting of some of Europe's leading experts in speech recognition and multimedia information processing. The project is sponsored by the European Commission as a result of the first call for R&D proposals in the area of Human Language Technologies (HLT) within the European Commission's 5th Framework Programme. The ALERT project aims to demonstrate, that by associating state-of-the-art speech recognition with audio and video segmentation and automatic topic detection, an automatic media monitoring demonstration system can be developed that detects topics in large amounts of multimedia data and alerts those users about the detection of this information that it is relevant for. The major motivation for this project results from the fact that keeping aware of information is of strategic importance for many businesses and governmental agencies. With the rapid expansion of different media sources (newspapers, newswire, radio, television, internet) for information dissemination, there is a large market for monitoring these sources and an increasing need for automatic processing of the data. Therefore, media monitoring is a crucial activity. For the most part today's methods are manual, with human reading, listening and watching, annotating topics and selecting items of interest for the user. The ALERT system intends to drastically change this situation, by providing media companies a brand new tool, enabling them to carry out these activities almost fully automatically, resulting in a remarkable competitive edge for these institutions. It is an impressive example in order to demonstrate how advanced speech recognition technology and other related methods can be already applied usefully in some of today's most competitive business sectors.

2. OBJECTIVES OF THE ALERT SYSTEM

The objective of the ALERT project is to develop an intelligent software system that automatically scans multimedia data like TV or radio broadcasts for the presence of specific topics and that alerts users whenever topics of their interest are detected. This includes the following functionalities:

1. The system will be capable of identifying specific information in multimedia data consisting of text/audio/video streams, using advanced speech recognition and video processing techniques, as well as automatic topic detection algorithms.
2. It will alert a client about the existence of such information
3. It will send detailed information (on the client's further request), consisting of extracted text, or annotated audio data and video clips, where appropriate.
4. Demonstrators in French, German and Portuguese with the above mentioned capabilities will be provided.
5. The demonstrators will be evaluated mainly by the industrial partners in the consortium.

The overall goal in terms of quality and usability of the demonstrator and its components, like speech recognition and topic detection, is that the industrial partners find it to be useful for their daily business.

3. DESCRIPTION OF THE CONSORTIUM

The consortium consists of three research institutions, two industrial companies, mainly responsible for implementing the developed algorithms, and three industrial partners active in the media monitoring and business, who will act as users of the developed demonstrators.

Duisburg University (Germany), **LIMSI** (France) and **INESC** (Portugal) are research oriented organizations with substantial experience in the field of automatic speech recognition, topic detection in written texts, video segmentation and categorization and multimedia data processing. They provide and improve their experience, know-how and tools in these areas in order to fulfill the scientific objectives targeted with this project. They develop tools to automatically transcribe and partially label the data provided by the users. They also contribute to the demonstrator specification, development and its evaluation**. Duisburg University** furthermore is in charge of **project management**.

Vecsys (France) and **Ergoprocesso** (Portugal) are independent industrial companies that develop software, hardware and integrated systems based on advanced video- and audio-processing technology. The emphasis of the Vecsys company is in audio-processing, while Ergoprocesso's major experience is in video and image processing and indexing. Beside a few contributions with respect to the scientific research objectives, Vecsys and Ergoprocesso will work as system integrators building the targeted ALERT system.

RTP (Portugal), **SECODIP** (France) and **Observer RTV** (Germany) are the users of the target application. RTP is the Portuguese television broadcast public service organisation. They are running there own SDI department. SECODIP is the leader in France for press review. It is a subsidiary of the Sofres/Taylor-Nelson group. Observer RTV is the leading media monitoring group in Germany, and is a subsidiary of the Sifo group with media-monitoring activities in 8 Northern European countries. They are providing SDI services for clients. Besides being the system's users who evaluate the demonstration application RTP, Observer and SECODIP contribute by providing data and by an involvement in the demonstrator specification.

4. SCIENTIFIC APPROACH

The state-of-the-art in automatic multimedia information retrieval is characterized by the existence of retrieval tools that mainly work on text databases and can be considered as intelligent keyword search engines (such as e.g. search engines used for the Internet, for instance Altavista). The ALERT system will advance this state-of-the-art by introducing the following scientific innovations:

- large vocabulary continuous speech recognition with high enough accuracy and fast enough processing to satisfy the needs of media watch (in French, German and Portuguese)
- automatic topic detection from imperfect transcriptions (in French, German and Portuguese)
- automatic topic detection on word graphs and confidence labeled speech recognizer output (in all three languages)
- unsupervised vocabulary and language model adaptation using contemporary textual data in order to maintain a high lexical coverage and consistent recognition performance (in all three languages)
- supervised topic collection and adaptation
- multimedia data segmentation based on audio and video signal
- definition of a language-independent representation for topic detection
- definition of a multimedia data representation that allows to develop automatic indexing techniques for different media types

Figure 1 on the next page illustrates how the multimedia data will be processed by the ALERT system: Multimedia data is segmented and topic indexed with the approaches and tools developed by the research partners. Users can specify certain topics that they are interested in, so that whenever new data concerning one of these interests has been detected, an alert is invoked (by email or other means) that informs the user about this new data and how to retrieve.

The following approach can be used in order to solve this problem: First of all, it is possible to derive a coarse segmentation of the data, based on the video track by detecting cuts or wipes, i.e. obvious image changes. In many cases (e.g. in TV news), this will indicate the change from a newscaster to a report, or within a report, it might indicate the change of the topic to be reported on. Detection of those changes in the video track will not automatically indicate the change of topics, but will yield some valuable information about the structure of the video document. While cuts are good indicators for topic changes, editing effect like wipes or dissolves are often used to merge two separate shots that belong to the same topics. In [4], it is described how a statistical approach can be used for the integrated detection of so-called editing effects (including cuts, wipes, dissolves, etc.) plus the major content classes of TV news (such as "news caster", "report", "interview", etc...). The approach is illustrated in Fig. 2. Here, it can be seen that a hierarchical model is employed that contains in its upper level the above mentioned content classes and their possible transitions for a broadcast news show in a grammar-like style. Each of these high level elements is realized by a Hidden-Markov-Model that takes as observation sequence a vector composed of visual cues for each video frame (e.g. simple motion & color information). Running the Viterbi-algorithm on an observed video sequence will thus lead to a segmentation of the video into the classes shown in Fig. 2. Thus, the video processing step will lead to an indexing of the data concerning the presence of editing effects (which might indicate topic changes or not) and possibly the presence of certain content classes, where e.g. the detection of a new content class is a strong indication for a topic change. Additionally, the data will be segmented into so-called “shots”, where each shot contains a specific visual content and thus is a potential candidate for a topic analysis.

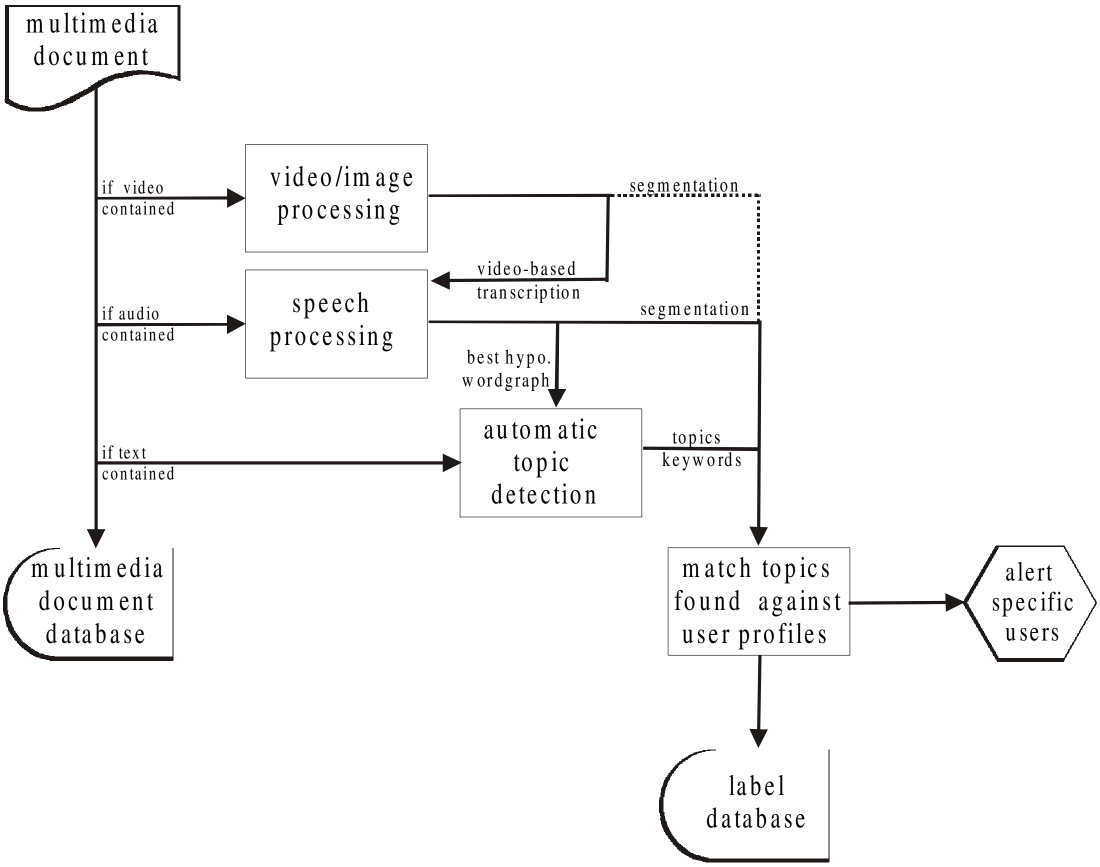

Figure 1: Labeling of multimedia data and alert generation

Once a pre-segmentation has been carried out on the video level, a more detailed analysis of the data can now be performed using the audio track. Of course, if the raw data is not taken from TV news but pure audio data from broadcast news, the above described step will be omitted. The audio portion of radio and television broadcasts contain signal segments of various linguistic and acoustic natures, with abrupt or gradual transitions between segments. Analogously to the video segmentation, the audio data can be partitioned into homogeneous segments, and appropriate labels associated with each segment. While it is evidently possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-forward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities. Prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. By using acoustic models trained on particular acoustic conditions, overall performance can be significantly improved, particularly when cluster-based adaptation is performed. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long) reduces the computation time and simplifies decoding.

Merging the video-based transcription, that contains the detected editing effects and possible topic boundaries, and the audio-based segmentation that is based on the detection of speaker changes and non-speech segments, is a challenging task. In [13], a method is described that detects suitable candidates for acoustic segments using the popular BIC criterion and considers these candidates as “audio cuts” and incorporates these audio cuts as additional data stream into the video model shown in Fig. 2. The result is an audio-visual system capable of segmenting TV news **directly into topics**.

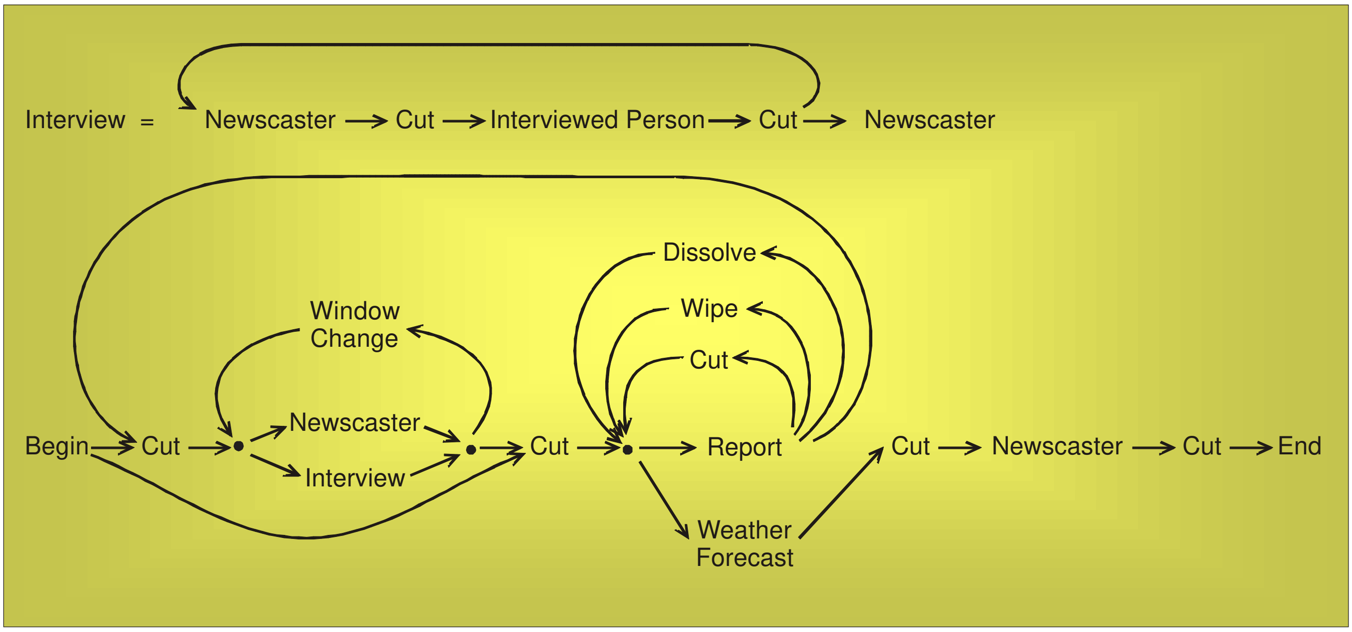

Fig. 2: Stochastic video model used for HMM-based video segmentation of broadcast news

Transcription of broadcast data is very challenging, as the system must be able to deal with changing conditions in the continuous stream of data. As a result, the word accuracy that can be obtained using the best hypothesis of the speech recognizer on radio and television speech is substantially lower than those reported for dictation systems on clean read speech. Hence, the transcription will always contain a fair number of recognition errors. However, it should be kept in mind that, although transcription of the speech might be only 70% correct, it is still possible to derive the topics from such erroneously transcribed sentences, if the sentences contain enough topic-related words to indicate the correct topics. As a consequence, the chances for high quality topic detection are considerably higher.

Therefore, special strategies have to be developed in order to deal with the erroneous transcription. First of all, although a perfect transcription of broadcast data is beyond available technology, it is important to have as precise a transcription as possible. A substantial part of the project efforts are devoted to developing high quality and flexible speech recognition for the three languages. This includes special algorithms for dealing with broadcast speech, for adaptation to varying quality conditions, and for performing unsupervised adaptation (see e.g. [1,5]) to be able to train the system with large speech databases for which no transcriptions exist. In order to deal with the large uncertainty in the transcribed words, the use of word lattices will allow multiple alternative decoding results to be considered, enabling the evaluation of words with lower scores. Methods are developed in order to re-score word lattices according to posterior probabilities, which can be directly interpreted as confidence measures (see [3]). Such confidence-scored lattices will then be used as input to the topic detection modules.

The broadcast speech recognition systems developed at LIMSI (see [7-12]) as well as the recognizers for German (see [6]) and Portuguese at Duisburg University and at INESC respectively, serve as a basis for the automatic transcription of broadcast news in the diverse languages. Approaches for topic indexing developed at LIMSI, the University of Duisburg ([2]) and at RTP serve as a basis for the developments concerning topic annotation of the output of the speech recognizers.

As an example, the evolution of the German broadcast speech recognition system shall be described in more detail: This system, called Ducoder, is the LVCSR decoder developed at Duisburg University (see [6]). It performs the Viterbi search for the most probable hypothesis on word level using HMMs and Backoff N-gram language models. In principle, the decoding procedure is similar to the one of the stackdecoders, which set up stacks at each time-step. A stack contains a sorted list of word hypotheses. After choosing a stack, all the stack's hypotheses get expanded simultaneously by performing a single word recognition, resulting in new hypotheses that get pushed to the specific stacks.
Operating in its standard mode the DUcoder uses stacks of fixed size for each time-step, in which the best hypothesis ending at that time are stored. There are several stack selection and exclusion strategies implemented, which are outlined in detail in [6]. The single word expansion is organized as a time-synchronous search (which follows the principles of Token Passing) through a recognition network. To reduce the number of nodes to be expanded at each time step the dictionary is organized in a tree structure. 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. With the general dictionary consisting of 100k entries an OOV rate of 8.5% has been computed on news texts. This OOV rate can be reduced to 5.2% if specific

dictionary is extracted from large news-text corpora. 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 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 PC's with memory requirements between 100 and 150 MByte. The LVCSR system described above was trained with about 60 hours of speech data. For this training a mixture of spontaneous speech and read sentences with almost no acoustic background noise was used. This system was used as the baseline for the development of a broadcast speech recognizer, using a 100k vocabulary. The following Table 1 shows the development of the recognition rate, starting from the baseline system up to a context-dependent system especially trained for broadcast news:

system	phone models	# mixtures	WER
1. baseline, not trained on broadcast data	triphones	31780	79,7%
2. baseline with broad-cast language model	triphones	31780	72,3%
3. acoustic models trained on broadcast data	monophones	1722	54,3%
4. acoustic models opti-mized on broadcast data	triphones	96417	22,8%

Table 1: Word error rates for the German broadcast speech recognition system in different configurations

From this table, it can be seen that a system trained on spontaneous speech (system #1, which produces around 20% WER if tested with spontaneous speech in speaker-independent mode) is not at all very suitable for broadcast speech recognition. Furthermore, the influence of a language model specialized on broadcast news seems to have only a limited impact on the error rate (#2). Training acoustic models on real broadcast data results in a huge improvement (#3) and can be further perfected if context-dependent models are trained (#4). This system has a quite respectable word error rate of 22% and should be one of the most advanced German broadcast speech recognition systems.

The labels resulting from the speech recognition process are used both directly for topic detection and evaluation, as well as in information retrieval and data preparation in order to send it to the user. For instance, knowing the exact cut boundaries for a specific video shot makes it easier to send the user a video clip containing exactly the required scene(s), without including additional superfluous and potentially disturbing video material. Combining the output of the audio and video tracks can also help in avoiding errors: if the boundary located in the audio channel between a speech segment and a music segment is misplaced, it can be revised if a cut is detected in the video signal between an announcement and a following report, which starts with introductory music. Conversely, an astute viewer may notice that the displayed images do not necessarily align temporally with the speech. This is because the display is programmed in advance, and it is up to the presenter to keep within the designated timing. The audio partitioning will reflect the true story span.

The partitioning result (based on the audio and video segmentation) and the enhanced transcription of the audio data serve as input to the topic detection modules. These modules analyze the enhanced transcription (word lattice with time markers for words, confidence scores), and produce a list of topics that can be assigned to a segment representing a specific audio or video sequence. In this way, a link is established between the topic list and the raw audio or video data which has been previously subdivided into meaningful segments, using the video and audio processing techniques described above.

Statistical methods for topic detection appear to be extremely promising when applied to automatically generated transcriptions, as they are more robust with respect to the characteristics of spoken language and to transcription errors than approaches requiring a full linguistic analysis. Standard text-based topic detection techniques often make use of punctuation markers and syntactic boundaries and therefore cannot be directly applied to automatically generated transcriptions in which these markers are often not present. Statistical methods are also less dependent upon the a priori definition of topic-specific keywords. This can be explained by the fact that there are many different ways to talk about a given topic, and generally speaking there are no keywords that absolutely have to occur for a topic to be present. The terms describing the topic itself do not necessarily have to occur in the text. Therefore, it is necessary to exploit the probability distribution of certain words in the segment for each topic of interest. A straightforward approach is the use of mixture models for topic detection as outlined in [2], similar to the approach followed in classical HMM-based speech recognition. However, there are substantial differences with classical speech recognition algorithms, ranging from the presentation of the input to the statistical models to architectural issues and details of the estimation procedure.

As already mentioned above, we investigate topic detection from word lattices, where confidence scores are associated with each word. Given that the word error rate in the word lattice will be significantly lower than that of the best word hypothesis output by the recognizer, this should improve topic detection accuracy. Probabilistic topic spotting methods that can exploit the information in confidence-scored word lattices will need to be developed. These will be achieved by extending the algorithms available for topic spotting on the basis of best hypothesis transcriptions. It has been recently shown in [2] that similar precision can be obtained by using neural network models for topic detection as are obtained using Gaussian mixtures. By varying the architecture of the neural network, a robust approach capable of adjusting the number of parameters to the available training data can be obtained.

Two basic functionalities of the demonstrator are supported: on-line processing and batch processing of multimedia data. For on-line processing, new audio and video data will be analyzed, producing audio and video-based segmentations and transcriptions. These will be input to the topic detection module, which will identify any story segments corresponding to a pre-specified topic list provided by the client. When a topic is detected, the user (in this case the customer service representative at the media monitoring company) will be informed via an alert message. At the user organizations in the ALERT project, humans currently transcribe breaking news, highlighting important words, which are then scanned to detect related stories. A brief keyword list is generated for on-topic segments, and the clients are contacted either by email or telephone. The goal of the ALERT system is to reduce the manual workload by automatically producing alerts.

For the purpose of batch processing, information is retrieved from a multimedia database. The customer will initiate a search for a specific topic or set of topics (typically via a user-interface). This information is conveyed to the search engine, and segments with matching topic indexes and content class labels are retrieved, along with the original audio or video data. It should be noted that most distributors of selective information do not maintain long term archives of the audio and video material due to the enormous storage requirements, as they monitor hundreds of radio and TV channels. Thus, batch mode processing will only be able to retrieve multimedia data for the last few weeks or months. Typically, older transcriptions are archived, so text-based retrieval can extend further into the past.

5. CONCLUSION

We presented a survey of the ALERT system, which is developed using the most advanced techniques in speech recognition and video processing in order to become one of the most powerful multilingual systems for broadcast speech recognition, monitoring of broadcast news services and retrieval of multimedia information resources.

6. ACKNOWLEDGEMENT

The author would like to thank all his colleagues from Germany, France and Portugal in the ALERT consortium, and especially his collaborators at the Department of Computer Science at Gerhard-Mercator-University Duisburg, **A. Kosmala, U. Iurgel, S. Werner, and D. Willett**, for their numerous contributions to the ALERT project during the last 2 years.

7. REFERENCES

[1] J. Rottland, Ch. Neukirchen, D. Willett, G. Rigoll. Speaker Adaptation Using Regularization and Network Adaptation for Hybrid MMI-NN/HMM Speech Recognition, Eurospeech'99, Budapest, Hungary, Sep. 1999.

[2] Ch. Neukirchen, D. Willett, G. Rigoll. Experiments in Topic Indexing of Broadcast News using Neural Networks, ICASSP, Phoenix, 1999.

[3] D. Willett, A. Worm, Ch. Neukirchen, G. Rigoll. Confidence Measures for HMM-based Speech Recognition, ICSLP, Sydney, 1998

[4] S. Eickeler, S. Müller. Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models, ICASSP, Phoenix, 1999.

[5] F. Wallhoff, D. Willett, G. Rigoll. Frame Discriminative and Confidence-Driven Adaptation for LVCSR. In IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, June 2000.

[6] D. Willett, Ch. Neukirchen, G. Rigoll. Ducoder - the Duisburg University LVCSR Stackdecoder. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, June 2000.

[7] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker. Transcribing Broadcast News Shows. ICASSP-97, Munich.

[8] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker. Transcription of Broadcast News. Eurospeech'97, Rhodes.

[9] G. Adda, M. Adda-Decker, J.L. Gauvain, L. Lamel. Text Normalization and Speech Recognition in French. Eurospeech'97, Rhodes.

[10] J.L. Gauvain, L. Lamel, G. Adda. Partitioning and Transcription of Broadcast News Data. Proc ICSLP'98, Sydney, 1998.

[11] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino. The LIMSI 1998 Hub-4E Transcription System. DARPA Broadcast News Workshop, Morgan Kaufmann Publishers, Hernon, VA, 1999.

[12] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino. Recent Advances in Transcribing Television and Radio Broadcasts. Eurospeech'99.

[13] U. Iurgel, R. Meermaier, S. Eickeler, G. Rigoll. New Approaches to Audio-Visual Segmentation of TV News for Automatic Topic Retrieval. Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, USA, May 2001.