Guimarães - Portugal
acústica 2004
paper ID: 193 /p.1

An acoustic driven media watch system

I. Trancoso[a], J. Neto[a], H. Meinedo[a], R. Amaral[b], and D. Caseiro[a]

Spoken Language Systems Lab, INESC-ID Lisboa, R.Alves Redol, 9, 1000-029 Lisboa, Portugal
Isabel.Trancoso@l2f.inesc-id,pt
[a] Instituto Superior Técnico, Lisboa, Portugal
[b] Instituto Politécnico de Setúbal, Portugal

ABSTRACT: The goal of the media watch system described in this paper is to continuously monitor a TV channel, searching inside its news programs for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send indicating the occurrence of stories on that topic, enabling the user to follow the links to the corresponding video clips. This paper describes the three main blocks of the system: the Capture block, responsible for capturing the monitored programs, the Processing block, responsible for generating all the relevant markup information, and the Service block, responsible for the user and database management interface. We shall particularly emphasize the different stages of the acoustic driven Processing block (acoustic segmentation, speech recognition, story segmentation and indexing). The know-how gained in this project enabled us to deal with related topics, such as meeting and lecture transcription.

1. INTRODUTION

The development of the media watch system described in this paper was initiated in the scope of the ALERT European project [1] (2000-2002). Its goal is to continuously monitor a TV channel, searching inside the news programs for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program from the occurrence of the program jingles. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program stories. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories.
The use of advanced processing technologies for content-based indexing and management of multimedia information allows the user to select and receive only the information required, even when faced with an ever increasing range of heterogeneous sources, that may not be restricted to TV channels, as in the ALERT project. The potential of this type of media watch systems is thus enormous, and provides the motivation for continuing with this line of research. We are currently focusing on the issues that were considered more relevant for improving the performance, during the field trials conducted in the last stage of the project.

Section 2 briefly describes the 3 main blocks of our media watch system. The major part of this paper is devoted to the Processing block (Section 4). Before, however, we shall briefly describe the Broadcast News (BN) corpus collected during the initial stage of the ALERT project, which allowed the training and evaluation of the different modules of this block (Section 3). Besides evaluating each module separately, we also conducted some preliminary field trials to have an overall evaluation (Section 5). Section 6 presents some concluding remarks and our most recent research trends.

2. SYSTEM ARCHITECTURE

The system includes three main blocks: a Capture block, responsible for the capture of each of the programs defined to be monitored, a Processing block, responsible for the generation of all the relevant markup information for each program, and a Service block, responsible for the user and database management interface. A simple scheme of semaphores is used to control the overall process [2].
In the Capture block, we have access to a list of programs to monitor and their time schedule (expected starting and ending time). This time information is the input to a capture program that through a direct access to a TV cable network starts the recording of the specified program at the defined time. This is done using a normal TV capture board (Pinnacle PCTV Pro). This capture program generates an MPEG-1 video and audio encoding file, with the audio at 44.1KHz, 16 bits/sample and stereo. When the recording process is finished, an MPEG-1 file is generated, together with the corresponding semaphore signal that will initialize the next block.
In the Processing block, the audio stream extracted from the MPEG-1 file is processed through several stages that successively segment, transcribe and index it, compiling the resulting information into an XML file.
The Service block deals with the user interface, through a set of web pages, and database management of user profiles and programs. Each time a new program is processed, an XML file is generated and the database is updated. The matching between the new stories and the user profiles generates a list of alerts that are sent to the users through an e-mail service. The user can either sign up the service enabling him/her to receive alerts on future programs, or search on the current set of programs for a specific topic.
The system has been implemented on a network of 3 ordinary PCs running Windows 2000 (Capture and Service Blocks) and Linux (Processing Block). This process is running daily since May 2002 with success. The first implementation of the system was focused on demonstrating the usage and features of this system for the 8 o'clock evening news broadcasted by RTP (Radio Televisão Portuguesa), but it has already been expanded to cover other channels.
RTP, the Portuguese user partner in the ALERT project, is interested on indexing every story and not only the stories according to certain profiles. To accomplish this indexing task, we based our topic concept in a thematic thesaurus definition that was used at RTP in their manual daily indexing process. This thesaurus follows rules that are generally adopted within EBU (European Broadcast Union), and has a hierarchical structure with 22 thematic areas in

the first level. Although each thematic area is subdivided into (up to) 9 lower levels, we implemented only 3 in our system, due to the sparsity of topic-labeled training data in the lower levels. This structure, complemented with geographic (places) and onomastic (names of persons and organizations) descriptors, makes our topic definition. A user can also specify a free text string. The user profile definition results from a combination of "AND" or an "OR" boolean operators on these more or less specialized topics. The use of this hierarchically organized structure makes the Portuguese system significantly different from the one developed by the French [3] and German [4] partners of the project, and from the type of work involved in the TREC SDR Track.
Each alert email message (Fig. 1) includes the title of the news broadcast, the date, the news duration time and a percentage score indicating how well the story matched the chosen profile (e.g. **Telejornal + 2003-03-29 + 00:07:40 + 65%**)**.** It also includes the title of the story, a link to a URL where one could find the corresponding RealVideo stream, and the topic categories that were matched.

Figure 1. *Example of an alert e-mail.*

3. BROADCAST NEWS CORPUS

Our BN corpus has two main subsets: the speech recognition corpus (SR) and the topic detection corpus (TD). Prior to the collection of these two subsets, however, we collected a relative small pilot corpus of approximately 5h, including both audio and video, which was used to discuss and setup the collection process.
The SR corpus was collected from November 2000 to January 2001, including 122 programs of different types and schedules and amounting to 76h of audio data. The main goal of this corpus was the training of the acoustic models and the adaptation of the language models used in the large vocabulary speech recognition component of our system. The TD corpus was collected from March through October 2001, containing data related to 133 TV broadcast of the 8 o'clock evening news program. The purpose of this corpus was to have a broader coverage of topics and associated topic classification for training our topic indexation module.
All the audio data was first automatically transcribed. The orthographic transcriptions of the pilot corpus and the SR corpus were both manually corrected, using the Transcriber tool, and

following the LDC Hub4 (Broadcast Speech) transcription conventions. For the TD corpus, we only have the automatic transcriptions and the manual segmentation and indexation.

4. PROCESSING BLOCK

Figure 2 shows a functional diagram of the Processing block, based on successive processing stages that transform the MPEG-1 file generated by the Capture block into an XML file containing the orthographic transcription and associated.

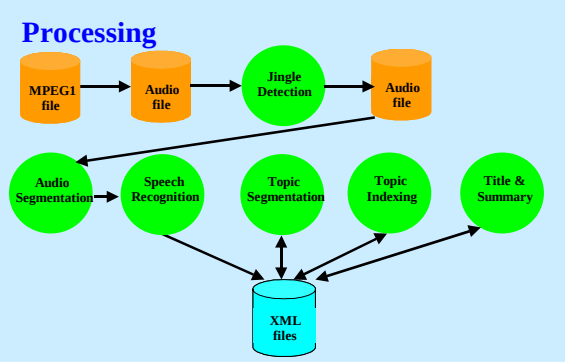

Figure 2 - *Functional diagram of the Processing block.*

The first stage extracts the audio file from the MPEG-1 stream downsampling it to 16kHz, disregarding, for the time being, any information that could be derived from the video stream. In the near future, we plan to integrate some image processing techniques in different processing stages, namely in terms of segmentation.
The resulting file is then processed by a ***Jingle Detector***, which tries to select the program's start and ending time, and cut the commercial breaks, based on the corresponding jingles. The output of this block is an MPEG-1 file, an audio file and the time slots corresponding to the relevant parts of the program.
The new audio file is then fed through an ***Audio Segmentation*** module [5]. The goal of this module is to select only transcribable segments and to generate a set of acoustic cues to speech recognition, story segmentation and topic indexing systems. The output is a set of segments that are homogeneous in terms of background conditions, speaker gender and special speaker id (anchors). Changes in these acoustic conditions are marked as segment boundaries. Each segment is passed through a classification stage in order to tag non-speech segments. All audio segments go through a second classification stage where they are classified according to background status. Segments that were marked as containing speech are also classified according to gender and are subdivided into "sentences" by an endpoint detector. All labeled speech segments are clustered by gender in order to produce homogeneous clusters according to speaker and background conditions. In the last stage, anchor detection is done, attempting to identify those speaker clusters that were produced by one of a set of pre-defined news anchors. This segmentation can provide useful information such as speaker turns, allowing for automatic indexing and retrieval of all occurrences of a

particular speaker. Grouping together all the segments produced by the same speaker, or by all speakers of the same gender, can also allow an automatic online adaptation of the speech recognition acoustic models to improve the overall system performance.
The audio segmentation module achieved a miss ratio of 14% in terms of segment boundary detection and an insertion ratio of 18%. The error rate for tagging speech segments as non-speech is 4.4%. The gender misclassification rate was 7.1%. Background classification turned out to be a hard task, namely because there are many segments in the training material with music plus noise. Our speaker clustering method achieves a cluster purity greater than 97%. The anchor identification module achieved a deletion rate of 9%, and an insertion rate of 2%.
Each transcribable segment is then processed by the ***Speech Recognition*** module. This module is based on AUDIMUS, a hybrid speech recognition system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs). The acoustic modeling of AUDIMUS combines phone probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. These probabilities are taken at the output of each MLP classifier and combined using an average in the log-probability domain. Currently our vocabulary is limited to around 60k words associated to a multi-pronunciation lexicon. The corresponding OOV rate is 1.4%. We use an interpolated 4-gram language model combining a model created from newspaper texts with a model created from BN transcriptions (45 hours). AUDIMUS presently uses a dynamic decoder that builds the search space as the composition of three Weighted Finite-State Transducers (WFSTs), the HMM/MLP topology transducer, the lexicon transducer and the language model transducer [6]. For the F0 focus condition (planned speech, no background noise, high bandwidth channel, native speech), the system achieved a word error rate (WER) of 14.8%. The average WER for all conditions was 26.5%. These results were obtained at 7.6 x Real Time (RT) on a Pentium III 1GHz computer running Linux with 1Gb RAM. The fact that our audio segmentation module frequently subdivides sentences has a negative impact in terms of language model errors near incorrect sentence boundaries. At the end of this module, an XML file is generated containing the audio segments, together with the corresponding markups, and the text transcript of each segment containing speech.
The ***Topic Segmentation*** module [7] is based on a very simple heuristic that assumes that all new stories start with a segment spoken by the speaker identified as anchor. Hence, it starts by defining potential story boundaries in every transition non-anchor transcript. Using the metric adopted in the 2001 Topic Detection and Tracking benchmark, and the same cost values of miss and false alarms [8], our algorithm achieved a normalized value for the segmentation cost of 0.84 for a priori target probability of 0.8. One of the reasons for boundary deletion is related to filler segments, introduced by the anchor and not distinguished from the following story also by the anchor. Another reason is the presence of multiple anchors, not yet supported by our simple heuristics. This new information of segmentation into stories is added to the XML file.
For each story the ***Topic Indexing*** module [7] generates a classification, according to the hierarchically organized thematic thesaurus. The indexation is performed in two steps. It starts by detecting the most probable topic, using the automatically transcribed text for each story. The decoder is based on a HMM methodology and the search for the best hypothesis is done

with the Viterbi algorithm. The n-gram statistical models for each of the 22 thematic domains were computed from automatically transcribed stories with manually placed boundaries. The corresponding text was post-processed in order to remove all function words and lemmatize the remaining ones. The second step finds for the detected domain all the relevant first and second level descriptors, by counting the normalized number of occurrences of the words corresponding to the tree leafs, and going up the tree accumulating in each node all the normalized occurrences from the nodes below. The decision of whether a node concept is relevant for the story is made by comparing with a pre-defined threshold. This processing is complemented with onomastic and geographic information extracted from the story. All this information is added to the XML file.

The performance of this module (ignoring filler segments) was 73.8% correctness for the first level of the thesaurus. For the second and third levels, the module achieved a precision of 76.4% and 61.8%, respectively, but the accuracy is rather low given the high insertion rate (order of 20%). After this classification stage, a post-processing segmentation step is performed, in order to merge all the adjacent segments classified with the same topics. The basic algorithm was improved in order to assign multiple topics per story, since most of the stories in the ALERT corpus have been classified into more than one topic. The detection is based on the log likelihood ratio between the topic likelihood and the non-topic likelihood and achieved a correctness of 94.1% for the higher-level topic, reducing the number of topic oscillations. The most recent modification is the inclusion of confidence measures that reflect the matching of the acoustic and language models for each decoded word. This brought a slight improvement in correctness (95.3%). The use of a named entity list in the onomastic and geographic classification also yielded a significant reduction of the false alarm rate.

Finally a module for generating a ***Title and Summary*** is applied to each story. The most frequent algorithms for summarization are based on sentence or word sequences extraction. However, having observed that the role of the newscaster is to give a brief introduction to the story, typically in the first sentence, we use this knowledge to generate a title based on the first sentence of the newscaster in each story. The final result was satisfactory in spite of being completely dependent on the story segmentation process.

At the end of this Processing block, the XML file generated contains all the relevant information that we were able to extract.

5. FIELD TRIALS

The evaluation of the different processing techniques was done individually during the development of the system and presented above in section 4. In order to evaluate the overall system we conducted a series of field trials, which can still be considered very preliminary since they involved only 12 in-house researchers. Each participant was asked to fill up a global evaluation form for each of the topics he/she selected, indicating hits or misses, the accuracy of the segmentation boundaries (on a 5-point scale where 0 represents the accurate timing), and false alarms. Altogether, the potential users filled short descriptions of 47 stories and reported only 1 false alarm. Table 1 shows our results.

Table 1 – *Global results.*

No. hits with:	**-2**	**-1**	**0**	**+1**	**+2**	**Total**
Start	2	5	19	4	1	31
End	2	4	18	5	2	31
No. subdivided stories						0

This evaluation allowed us to compare the performance of free string and thematic domain matching. As expected, free string matching is more prone to speech recognition errors, specially when involving only a single word. Likewise for the onomastic and geographic classification. Thematic matching is more robust, but the thesaurus classification using only the top levels is not self-evident for the untrained user. A frequent criticism concerned the merging of several stories into a single one. In fact, this occurred mainly with the topic *sports with ball* which proved to be very popular. Being a third-level topic, several news involving different teams were grouped in the same alert message. Since the amount of training material is significantly different from topic to topic, a different strategy, allowing lower level definition whenever a sufficient amount of training material is available, might be more useful. In fact, the large miss percentage is mostly due to the lack of training material for certain topics. Missing the anchor detection may have too much negative impact on the system, as this causes several stories to be grouped together. In fact, anchor detection has been fine tuned using a too small development corpus. We need a more elaborate segmentation strategy. The partial evaluation of each module was done using a test corpus that did not significantly differ in time from the training corpus. The global evaluation, however, revealed the inadequacy of the chosen lexicon and language models two years after.

6. CONCLUSIONS AND FUTURE WORK

The above evaluation results and subsequent discussion has hinted at a number of issues that must be dealt with in order to improve the performance of the system. We are currently working on extending lexical and language models, on using more robust jingle detection strategies, and on improving sentence boundary detection by using prosodic information. Recent attempts at building better topic models have not been successful due to the small amount of manually topic-annotated training data.

In spite of these problems, we would like to stress the fact that having a fully operational system is a must for being able to address user needs in the future in this type of service. Our small panel of potential users was unanimous in finding such type of system very interesting and useful, specially since they were often too busy to watch the full broadcast and with such a service they had the opportunity of watching only the most interesting parts.

The know-how gained in this project enabled us to deal with related topics in the area that is also known as computer enhanced human-to-human communication. Examples of applications in this area are the meeting browser and the lecture tracker. A major difference of these domains relative to broadcast news is the dominant presence of spontaneous speech. But many other issues are involved such as speaker diarization (marking the times corresponding

to speaker changes, and providing speaker identification information), inclusion of other modalities, multimodal structuring, summarization and information retrieval, etc. This motivates a worldwide effort to automatically produce what is denoted by “rich transcriptions”, including not only transcripts but also metadata. The accompanying metadata is useful in increasing the readability of the transcripts, not only in terms of marking speaker changes, but also in terms of disfluency recognition (marking verbal fillers such as filled pauses, discourse markers, and verbal edits) and semantic unit segmentation.

ACKNOWLEDGEMENT

This work was partially funded by FCT project POSI/PLP/47175/2002. Hugo Meinedo was also supported by a FCT scholarship SFR/BD/6125/ 2001. INESC ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

REFERENCES

[1] ALERT project web page http://alert.uni-duisburg.de/

[2] J. Neto, H. Meinedo, R. Amaral and I. Trancoso, *A system for Selective Dissemination of Multimedia Information Resulting from the ALERT project*, In Proc. ISCA ITRW on Multilingual Spoken Document Retrieval, Hong Kong, China, April 2003.

[3] Y. Lo and J. L. Gauvain, *The LIMSI Topic Tracking System for TDT 2002*, In Proc. DARPA Topic Detection and Tracking Workshop, Gaithersburg, USA, Nov. 2002.

[4] U. Iurgel, S. Werner, A. Kosmala, F. Wallhoff and G. Rigoll, *Audio-Visual Analysis of Multimedia Documents for Automatic Topic Identification*, In Proc. SPPRA 2002, Crete, Greece, June 2002.

[5] H. Meinedo and J. Neto, *Audio Segmentation, Classification and Clustering in a Broadcast News Task*, In Proc. ICASSP 2003, Hong-Kong, China, April 2003.

[6] D. Caseiro and I. Trancoso, *A Tail-Sharing WFST Composition for Large Vocabulary Speech Recognition*, In Proc. ICASSP 2003, Hong-Kong, China, April 2003.

[7] N. S. Group, *The 2001 topic detection and tracking (TDT2001) task definition and evaluation plan*, 2001.

[8] R. Amaral and I. Trancoso, *Indexing Broadcast News*, In Proc. NDDL'2003 - Third Int. Workshop on New Develop. in Digital Libraries, Angers, France, April 2003.