A MEDIA MONITORING SOLUTION

J. Neto
Spoken Language Systems Lab
INESC-ID
Instituto Superior Técnico
Technical University of Lisbon
Joao.Neto@inesc-id.pt

H. Meinedo
Spoken Language Systems Lab
INESC-ID
Hugo.Meinedo@inesc-id.pt

M. Viveiros
Speech Processing Technologies, SA
Marcio.Viveiros@voiceinteraction.pt

http://l2f.inesc-id.pt

ABSTRACT
There are large amounts of information as video and audio not searchable. In a time where Business Intelligence is fundamental for all areas doing this kind of analysis only on text sources is a limiting factor. The use of large vocabulary speech recognition systems with increasing performance is giving rise to different applications. Despite the diversity, these applications share the extensive use of the contents of the transcription. In this paper we describe the results of a development project between a startup company and a research lab to build a full automatic system for monitoring TV and Radio channels. This system is composed by three main blocks: a recording block (records the selected channels in high and web streaming quality and broadcast to the next block), a processing block (generate metadata information) and a storage and accessing block (make available the metadata and videos). There is an optional block, the delivering, that could be customized according to the use and client needs. The processing block receives the video from the recording block time filtered by a scheduling interface. For that video the goal is to generate a full annotation with metadata describing the content and semantic information. This metadata make possible a filtering process for selective dissemination of information.

Keywords
Multimedia systems, audio processing, speech recognition, semantic contents.

1. INTRODUCTION
There are large amounts of daily produced video and audio information, from TV and Radio channels, that is not searchable and indexed. Also the availability of a video on the web is problematic because there is no way to access it through content. A search engine is blind to this kind of data. Certain efforts have been made to manually tag these videos, in an autonomous process or through entertainment games. However these paths conduct only to partial annotations. Video providers start looking to large vocabulary continuous speech recognition (LVCSR) systems as a way to automatically index that information.

Transcribing a video could be a very difficult task depending of acoustic conditions, speaker variability and information contents. In case of specific needs to access broadcast news (BN) information a large effort in the last decade was made by the speech community [1], mainly for English, but spreading for other languages, with some particular focus on Arabic and Mandarin, for evident reasons. Large campaigns for TREC\(^1\), first based on transcription contents and later based on video annotation techniques, address this problem.

We have been working on this area for a long time with special focus on the development of speech recognition techniques for online subtitling applications [2]. The subtitling of BN programs are starting to become a very interesting application due to the technological advances in Automatic Speech Recognition (ASR) and associated technologies as Audio Pre-Processing (APP). Also, there is a generic request from society and governments that are pushing the TV broadcasters to increase the amount and diversity of TV programs subtitled. In the front line, there are the people with special needs, mainly the hearing handicapped and elderly people, which are requesting full subtitling coverage of TV programs. The broadcast media plays an important role on the lives of these people by providing access to news, information and entertainment. Also, there are some situations as noisy places, airports, restaurants ... where this feature is very useful and requested. Additionally to these direct situations other applications could take advantage from subtitling as content search, selective dissemination of information and machine translation, among others.

In terms of selective dissemination of information several companies are providing systems [3] using different technological solutions. However very few have the capacity to use LVCSR working on a online mode with capacity to a direct transcription, preferring language independent phone recognition and searching.

The system presented here was the result from a close cooperation between a research laboratory with a SME company responsible for the development and solution implementation. Being a full product working daily in several clients for different languages gives precious feedback about systems errors and users needs.

In Section 2 we present the global system description and in section 3 the contribution of the different processing blocks to the metadata generation. We finish with some conclusions.

\(^1\) www.trecvid.com
2. GLOBAL SYSTEM DESCRIPTION

The overall system could be represented as a pipeline of different blocks accessible through different interfaces, as shown in Fig. 1.

There are three main blocks: a recording block (records the selected channels and broadcast to the next block), a processing block (generate metadata information) and a storage and accessing block (make available the metadata and videos). There is an optional block, the delivering, that could be customized according to the use and client needs. There are two main interfaces: the scheduling (define the programs to be annotated) and the editing (make possible to correct the metadata automatically generated).

2.1 Recording block

This block needs high performance hardware both in terms of audio/video signal acquisition and storage. Worldwide there are several companies with different solutions in terms of video capturing, transcoding options, video formats, recording software and interface software for accessing videos. There are solutions that propose a server with several channels (several simultaneous cards) or single channel solutions. The acquisition cards could be tuner based (single or dual tuner) to select the TV channel or just transcoding cards (pre-selected channel at the input).

In our architecture we use a server with 5 dual tuner cards giving a total of 8 channels (leaving one card as a redundancy system). Also we have solutions for 1, 2, or 4 channels in a single server. Mixed solutions are possible depending on the number of desired channels. We developed a SDK for accessing the dual tuner cards that allows the recording to disk of native format (normally high quality MPEG2 configurable from 6 to 2 Mbps), transcoding options to one of WMV, FLV or H.264 formats followed by recording to local disk and simultaneously streaming through a IP port. The recordings on disk are made of 1 hour blocks in both formats (high quality and web streaming). The recordings on local disk are of short duration (normally 3 days). After each 1 hour block be completed is transferred to a long term storage system. The server on this system is connected to the external network making available the videos through a webservice.

2.2 Processing block

This block receives input from the scheduling interface to process relevant programs chosen by system administrator. At the defined time the system connects to the IP port where the recording block broadcast that channel and starts processing the program. It is on this block that metadata is generated using different processing technologies including automatic speech recognition. This block will be explored in detail in section 3.

2.3 Storage and accessing block

After processing each program an XML file including all the generated metadata is produced and imported to a database. This database will accumulate the metadata from all the programs during a certain amount of time (client setup). This database is accessible through a webservice. From the user point of view he could import a WSDL description on his own application accessing the database information and videos.

2.4 Delivering block

This block could be tailored to the client needs and business models. In a base version includes a web searching interface (basic and advanced searches on database information) and delivery by web or email. Several plugins are available as SMS, MMS and RSS feeds.

2.5 Interfaces

There is an Administrator interface for programs scheduling that filters the access to the processing block. The metadata generated in the processing block makes sense for certain kind of programs as news, interviews, documentaries, ... This interface uses web information and/or EPG (Electronic Programming Guide) to create the channel scheduling.

The editing interface allows an human assistant to visualize the metadata generated and correct the segmentation, transcription and indexing information automatically generated.
3. METADATA GENERATION

The processing block receives the video from the recording block time filtered by the scheduling interface. For that video the goal is to generate a full annotation with metadata describing the content and semantic information. The audio stream is extracted from the video and fed into the system.

We are following three different parallel threads on this process: semantic description of audio contents [4], through blocks for Audio Pre-Processing, Speaker Clustering and Identification, Automatic Speech Recognition and Segmentation and Indexation; audio acoustic events, through the description of specific events [5]; and video information, as banner texts, shots segmentation and shots classification [6]. They are synchronized at the end in an XML file that collects all the information from the different threads. In this paper we will describe only the first of these threads, the semantic description of audio contents.

3.1 Audio Pre-Processing

An hierarchical data annotation (as shown in Fig. 2) is generated. The operation of this block is to filter the non-speech parts and to generate homogeneous acoustic segments. It starts by audio segmentation, in terms of acoustic change detection, classification in speech/non-speech and classification of background conditions (clean, noise, music). The non-speech segments are filtered out and only the speech segments will be forward to the next block. Optionally we use a block for Jingle Detection in order to help to define a model to the news identifying the start of the program, the commercial breaks, and the end. The different classifiers share the same architecture based on Multilayer Perceptrons (MLP) [7] with special emphasis to work online.

3.2 Speaker Clustering and Indexation

This block starts by classifying each speech segment in terms of gender (male/female), using the same structure as previous classifiers. After that classification the segments are grouped together in speaker clusters.

For some specific speakers there is an additional processing of identification. The news anchors identification could give access to personalized acoustic models in the ASR component. They introduce the news and provide a synthetic summary for the story. Normally, this is done in studio conditions (clean background) and with the anchor reading the news. This means that a very large portion of the news show is spoken by very few (recurrent) speakers. We use the collected information to create speaker models for the main station speakers, improving speaker clustering diarization performance and collecting information to speaker adapted acoustic models.

3.3 Automatic Speech Recognition

This block receives an audio input stream that was previously filtered by the Audio Pre-Processing block. The processing in the previous blocks was made to facilitate the ASR block operation since it receives only speech segments with some additional categorization information. This block has to generate the most correct transcript at the output working in a real time with emphasis on online mode.

This ASR block is based on the AUDIMUS.MEDIA [2] system. The processing stages are represented in Figure 3.

The acoustic modeling combines phone probabilities from several MLPs trained on distinct feature sets, resulting from different feature extraction processes in order to better model the acoustic diversity. This is more relevant in the recognition of broadcast news, where in each program there are a diversity of speakers and environments. These probabilities are taken at the output of each MLP classifier and combined using an appropriate algorithm.

Our decoder is based on the Weighted Finite-State Transducer (WFST) approach, where the all search space is a large WFST [8]. In our case, the search space results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one. Our decoder composes and optimizes the various components of the system in runtime.

Resulting from the decoder, a series of values describing the complete recognition process are generated. These values will act as features for a maximum entropy classifier that will output a value of a confidence measure. This value is fundamental to define the confidence on the recognized text, which could be helpful in the searching and filtering news processing.

ASR system vocabulary design principles intend to achieve a high and specific coverage of the domain. In a broadcast news task a large variety of topics are discussed over time. Additionally, a highly flexional language as the Portuguese needs larger vocabularies to achieve a similar cover of the domain. Both
constrains will imply that out-of-vocabulary (OOV) words cannot be avoided. The regular approach is to use a static large vocabulary, around 60K typically for English and larger for other inflectional languages. In our case, we are using a baseline vocabulary of 100K words combined with a daily modification of the vocabulary and a re-estimation of the language model [9], performing better than a static large vocabulary (200K).

The recognized text suffer a transformation in order to capitalize the names and acronyms and attempting to organize the ASR text output in a set of sentences. To both capitalizations, a problem of Named Entity Retrieval, and punctuation we are using a technique based on maximum entropy models. This technique is based on information from the APP and ASR modules as pauses, speaker change, previous, present and next words, as the grammatical class of each word and the confidence measure associated to each word [10].

3.4 Segmentation and Indexation

A final block that transforms a continuous stream of sentences, that outputs from the ASR, in a set of homogeneous segments according to their contents. This news segmentation process use all the metadata generated to create a news model. From this metadata we use the information about the newscaster, the pauses, timing information, selection of relevant words and max entropy algorithms to decide a change of news.

The indexation process it is not so relevant for the searching but more in terms of news presentation. In this process for each news segment we classify in one or several of the topics. The topics could be a broad class as Economic, Politics, Sports, .... or something more specific. This broad classification it is very interesting making possible a news presentation very simple and intuitive.

4. Evaluation

The goal of a Media Monitoring System is to make available for search and retrieval all the information delivered by TV and Radio channels. Since the full transcription is available for searching all the contents from individual word expressions, to more complex queries are possible. An evaluation of this system was made at three different levels. In terms of monitoring information volume, in terms of individual news segments and in terms of word error rate (WER).

There are two kind of TV channels: generic channels and news channels. A generic channel produce daily around 8 hours of news programs and a news channel around 16 hours daily. We are working with different channels and several languages (English, Spanish, European Portuguese and Brazilian Portuguese).

On one hour news block in mean we have 30 news segments. In the generic channels our performance in mean are 98% dropping to 78% in a news channel. This is due to several very short news where the change of contents it is difficult to define.

Our system work online, meaning just one recognition passage, with different WER% for the different languages: 20.4% for English, 21.2% for Spanish, 18.4% for European Portuguese and 20.8% for Brazilian Portuguese. These systems were evaluated in two hours test sets (mean).

CONCLUSIONS

This system results from the research and development work of a research lab combined with the application development of a SME. The use of speech recognition systems will introduce new opportunities to make available more information giving rise to new areas as Business Intelligence. The combination of a complete system being able to monitoring several channels in different languages with advance techniques for speech transcription, segmentation and indexation is presented on this paper. Also the combination of the best techniques in terms of video recording and transcoding gives to this system a high value.

5. ACKNOWLEDGMENTS

This work was partially funded by projects QREN SI IDT 2525, SI IDT 5108, FCT PTDC/PLP/72404/2006. Support by FCT (INESC-ID multiannual funding) through the PIDDAC Program.

6. REFERENCES