BROADCAST NEWS SUBTITLING SYSTEM IN PORTUGUESE

J. Neto[1,2], *H. Meinedo*[1], *M. Viveiros*[1], *R. Cassaca*[1], *C. Martins*[1], *D. Caseiro*[1,2]

(1) L2F – Spoken Language Systems Lab / INESC-ID
(2) Instituto Superior Técnico / Technical University of Lisbon
Joao.Neto@inesc-id.pt
http://www.l2f.inesc-id.pt

ABSTRACT

The subtitling of broadcast news programs are starting to become a very interesting application due to the technological advances in Automatic Speech Recognition and associated technologies. However, to build this kind of systems, several advances are necessary both in terms of the technological components and on main blocks integration. In this paper, we are presenting the overall architecture of a subtitling system running daily at RTP (the Portuguese public broadcast company). The goal is to integrate our components in a system for the subtitling of RTP programs. The global system includes the subtitling of recorded and direct programs.

Index Terms— Speech processing, speech recognition, multimedia systems.

1. INTRODUCTION

The subtitling of broadcast news (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.

The TV broadcasters have been supplied close-captioning to recorded programs based on manual transcription operation. The live programs are the most difficult to achieve, since it implies specialized stenography or the use of real-time Automatic Speech Recognition (ASR) systems. Current systems are based on shadow speakers operation [1] using user adapted acoustic models and thematic language models.

Over the last decade, the speech research community spent a large effort in the research and development of broadcast news (BN) systems [2], initially for English, and after for a variety of other languages. The development of a system for a new language is a challenging task due to the need of new acoustic training data, vocabulary definition, lexicon generation and language model estimation. Despite the good results very few systems were used for subtitling. The subtitling operation implies, besides real-time, an online operation. Transforming all the algorithms to online operation is not always a smooth and straight task. Putting a full subtitling system to work, needs a lot effort to develop the appropriate software and to be able to explore the specificity processing power of the machines, in which are needed very good engineering skills.

Fig. 1 – Output of current system with two subtitling lines on top of the screen.

In our lab, we have been working for the development of a BN system for the European Portuguese language. We used our knowledge to develop an hybrid ASR named AUDIMUS.MEDIA [3]. Simultaneously, an APP block was developed, incorporating several components [4]. The system presented here was the result from a close cooperation with RTP (Rádio Televisão de Portugal), the Portuguese public broadcast company. The goal was to integrate our components in a system for the subtitling of RTP programs. The global system includes the subtitling of recorded and direct programs.

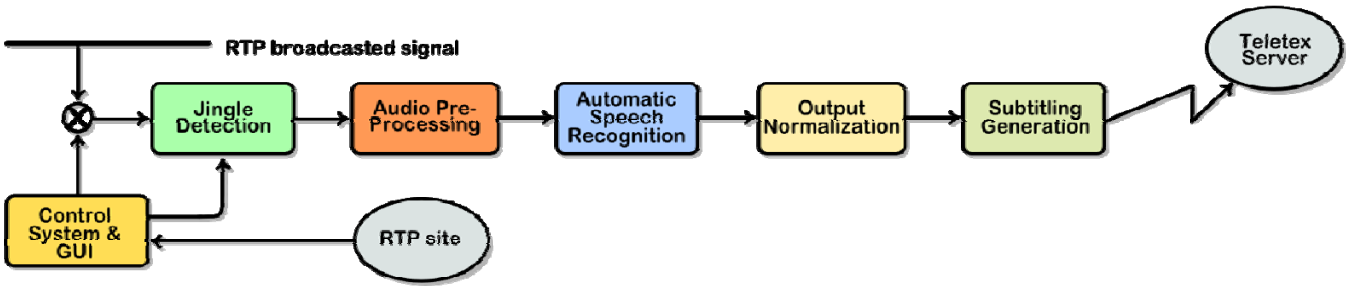

Fig. 2 – Block diagram of the subtitling system

2. GLOBAL SYSTEM DESCRIPTION

The overall system could be represented as a pipeline of processing blocks, as shown in Fig. 2.

The **Control System & GUI** block receives information from the administrator about the program to subtitle, specifically name and periodicity. This block checks from the RTP internal site the precise schedule of the program and at the specified time, it starts the operation.

At system operation, the **Jingle Detection** block searches for the beginning jingle of the program. After the jingle's end, it feeds the audio to the Audio Pre-Processing block. The Jingle Detection block also filters the commercials and the end jingle.

The **Audio Pre-Processing** block receives the net program, without jingles and commercials. This block discriminates between Speech and Non-speech, only sending the audio to the Automatic Speech Recognition in case of speech. Additionally, it gives information on speaker clustering, speaker gender and speaker ID (in case of relevant speakers). Doing a pre-processing task, the performance of this block is fundamental due to the filter role and the generation of a complete set of information that will be used by later blocks on their normal operation.

The **Automatic Speech Recognition** block transcribes the audio input stream according to a vocabulary and a language model. This block operation is critical due to the specific task of transcribing the speech whose performance will be reflected in the final service.

The **Output Normalization** block converts sequences of words representing digits, connected digits, and numerals in numbers. It also capitalizes the names and tries to introduce the punctuation, a difficult task on spontaneous speech.

The last block, **Subtitling Generation,** creates, from the output of the previous block, the subtitles according to the definitions of a standard subtitling and the teletext restrictions, and sends it to the Teletext Server to be broadcasted.

The overall system works in a pipeline and asynchronous operation mode, where each block is responsible for fulfilling is own task and propagate the results to the next block. Next, we shall give a brief description of main blocks.

3. JINGLE DETECTION (JD)

The goal of this block is to identify, in the audio stream, specific acoustic patterns. These patterns are known as "jingles" and are used in Broadcast News shows for drawing the listener's attention to important events like the start and the end of the show. The news program that we are processing has basically three different types of jingles: jingles to mark the beginning and the end of the news program, jingles that mark a commercial break during the program, and jingles for filler/headlines sections. These filler jingles appear when the news anchor is giving emphasis to some news story that will be developed later, or when he is summarizing the news stories that were covered during the program. In either situation, these filler sequences do not convey relevant information. Fig. 3 represents a possible time sequence for the news show we are processing.

Fig. 3 – A news show time sequence

The block diagram of the Jingle Detection is represented in Fig. 4 and includes 5 main components: feature extraction, pattern classification, median filter to smooth the classifier output, threshold operation and a finite state machine to represent the events transitions of a particular news show (Fig. 5).

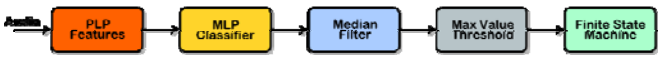

Fig. 4 – Jingle Detector block diagram

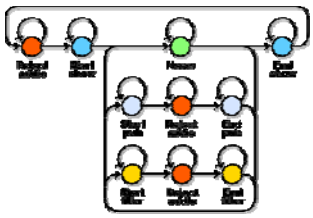

Fig. 5 – Jingle Detector block diagram

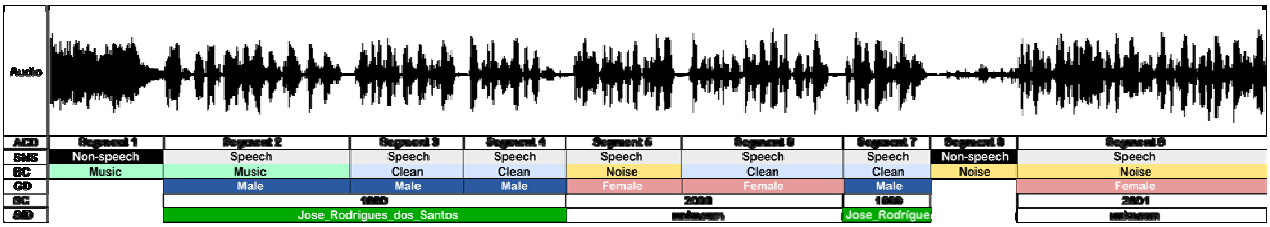

Fig. 6 – News show jingle time sequence

4. AUDIO PRE-PROCESSING (APP)

The operation of the APP block is two-fold: to filter the non-speech parts and to give additional information to the following blocks: gender classification, background classification, speaker clustering and speaker identification (in case of relevant speakers as pivots).

In Fig. 6 we illustrate the APP operation starting by audio segmentation, in terms of acoustic change detection, classification in speech/non-speech, classification of background conditions (clean, noise, music), gender classification (male/female), speaker clustering and finally speaker identification. The full operation of this block is complex but intended to provide a complete description of the input audio.

In Fig. 7 we group together the different classifiers in three main operations: the audio segmentation, the audio classification and speaker classification. The different classifiers share the same architecture based on Multilayer Perceptrons (MLP)[4] with special emphasis to work online.

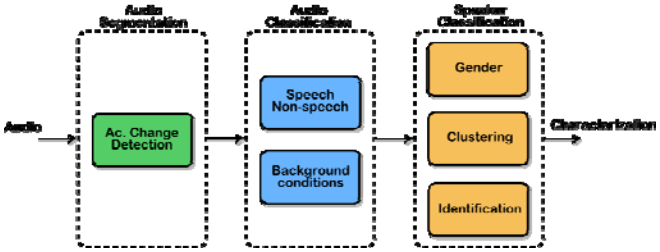

Fig. 7 – Diagram block of APP

The information provided by APP block characterizing the audio signal is relevant for the blocks that follow. As an example, 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.

5. AUTOMATIC SPEECH RECOGNITION (ASR)

This block receives an audio input stream that was previously filtered by JD and APP blocks. The processing in the previous blocks was made to facilitate the ASR block operation since it receives a "clean" audio with some additional categorization information. This block has to generate the most correct transcript at the output working in a real time and online mode.

This ASR block is based on the AUDIMUS.MEDIA [3] system. It is based on a hybrid speech recognition structure combining the temporal modeling capabilities of Hidden Markov models (HMM), with the pattern discriminative classification capabilities of MLPs. The processing stages are represented in Figure 8.

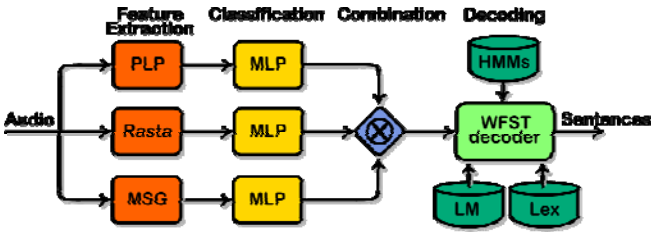

Fig. 8 – AUDIMUS.MEDIA block diagram.

The acoustic modeling of AUDIMUS.MEDIA combines phone probabilities generated by 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 BN, 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 [5]. 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, to filter out the output text in the subtitling composition stage.

ASR system vocabulary design principles intend to achieve a high and specific coverage of the domain. In a BN 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 [6] and a re-estimation of the language model [7].

6. OUTPUT NORMALIZATION AND SUBTITLING GENERATION

On these final blocks there are a set of actions transforming the output text of the ASR block in a sequence of subtitles. These actions comprise a normalization stage, in order to reduce the dimensionality of the text to present and to increase the readability of the subtitling. Other stage is to capitalize the names and acronyms and attempting to organize the ASR text output in a set of sentences. After these two stages, the sequence of words is organized in order to compose the subtitles, according to a series of options. Since our APP gives information about the speaker gender, we use that information to change the color of the subtitling.

The normalization operation converts numbers, numerals, dates and amounts (mainly money and percentage) in their digit representation.

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 [8].

Both these steps improve the readability of the subtitles. A lot of effort was put in order to understand the best way to present the information. Presently, we are using 2 lines on the top of the screen with some timing control in order to give to the users enough time to read the information and keep the normal flow of information in real time without any delays, only a small latency.

8. CONCLUSIONS

This system is the result of several years of research and development in the BN area for the Portuguese language. There are very few examples of BN subtitling systems and less working online and in real time. All these developments provide us with a unique platform where we are using results from research to improve the products of companies and using those products into improving the people's quality of life.

Also, this product could be used on other applications as a contribute to the generation of semantic concepts associated with multimedia documents and on transcriptions of lectures for e-learning applications [9].

ACKNOWLEDGMENTS

This paper presents an extensive work only possible with the collaboration of several people and institutions. First we would like to thank RTP and their collaborators, namely João Sequeira, Teotónio Pereira and their Departments at RTP. This work was partially funded by PRIME National Project TECNOVOZ number 03/165, European program project VidiVideo FP6/IST/045547 and FCT project POSC/PLP/58697/2004.

REFERENCES

[1] G. Boulianne, F.-F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet, F. Osterrath, "Computer-assisted closed-captioning of live TV broadcasts in French", in Proceedings Interspeech 2006, Pittsburgh, USA, 2006.

[2] P. Woodland, "The development of the HTK Broadcast News transcription system: an overview", Speech Communication, vol. 37, pp. 47-67, 2002.

[3] H. Meinedo, D. Caseiro, J. Neto and I. Trancoso, "AUDIMUS.MEDIA a Broadcast News speech recognition system for the European Portuguese language", in Proceedings PROPOR'03, Faro, Portugal, 2003.

[4] H. Meinedo, J. Neto, "A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models", in Proceedings of Interspeech 2005, Lisboa, Portugal, 2005.

[5] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition", IEEE Transactions on Audio, Speech and Language Processing, vol. 14, nº4, pp. 1281-1291, July 2005.

[6] C. Martins, A. Teixeira, J. Neto, "Vocabulary selection for a Broadcast News Transcription System using a Morpho-Syntatic Approach", in Proceedings Interspeech 2007, Antwerp, Belgium, 2007.

[7] C. Martins, A. Teixeira, and J. Neto, "Language Models in Automatic Speech Recognition", in Magazine of DET-UA, Aveiro, vol. 4, nº 4, 2005.

[8] F. Baptista, D. Caseiro, N. Mamede, I. Trancoso, "Recovering Punctuation Marks for Automatic Speech Recognition", in Proceedings Interspeech 2007, Antwerp, Belgium, 2007.

[9] I. Trancoso, R. Nunes, L. Neves, C. Viana, H. Moniz, D. Caseiro, A. Mata, "Recognition of Classroom Lectures in European Portuguese", In Proc. INTERSPEECH 2006, Pittsburgh, USA, September 2006.