Evaluation of a Live Broadcast News Subtitling System for Portuguese

Hugo Meinedo, Márcio Viveiros, João Neto

L^2F - Spoken Language Systems Lab, INESC-ID
Instituto Superior Técnico
{hugo.meinedo,marcio.viveiros,joao.neto}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

Broadcast news play an important role in our lives providing access to news, information and entertainment. The existence of subtitles is an important medium for inclusion of people with special needs and also an advantage on noisy and populated environments. In this work we will describe and evaluate a system for subtitling live broadcast news for RTP (Rádio Televisão de Portugal) the Portuguese public broadcast company. Developing a fully automatic subtitling system is a huge breakthrough which results from the convergence of different research models and software developments to create a working system. Our online system has 12% word error rate for the displayed subtitles working under real time with an average latency of just 6.5 seconds.

Index Terms: Speech processing, Speech Recognition, Speaker Diarization, Online processing, Subtitling

1. Introduction

As a generic request from society TV broadcast companies display an increasing interest on the subtitling of TV programs. In the front line are people with special needs mainly the hearing handicap and elderly people which are requesting full subtitling coverage of TV programs. Broadcast media plays an important role in the lives of these people by providing access to news, information and entertainment. Also there are some situations such as noisy places, airports, shopping malls and restaurants where this feature would be very useful and demanded by users. Additionally other applications can take advantage from subtitling like content search, selective dissemination of information and machine translation among others. TV broadcasters have been supplying close-captioning to recorded programs based on manual transcription operation. Live programs are the most difficult to subtitle and since their weight on the emission is increasing TV broadcasters are paying more attention to this issue. Currently subtitling live programs requires specialized stenography or the use of real-time Automatic Speech Recognition (ASR) systems. These 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]. Despite these good results, these developments did not have a strong impact on subtitling systems. The subtitling operation implies not only real time but also an online operation. Transforming all the features of the algorithms to an online operation is not always a smooth and straight task. Also a fully working subtitling system takes a lot of effort to program and tune appropriate software in order to be able to explore the specificity processing power of the computers.

In our laboratory we have been working on the development of a BN system for the European Portuguese language. 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. We are using our knowledge to develop a BN speech recognition engine named AUDIMUS.MEDIA [3] which has an hybrid MLP/HMM acoustic model and uses an efficient decoder approach based on Weighted Finite-State Transducer (WFST) models [4]. Simultaneously an BN Audio Pre-Processing (APP) module has been developed to characterize and enrich the audio stream with metadata [5]. Also developed were an automatic capitalization and punctuation module [6] and an ASR output normalization module to further enrich the stream and to improve human readability.

The system presented here results from a close cooperation between our laboratory and RTP (Rádio Televisão de Portugal) the Portuguese public broadcast company. The main goal is to integrate our components in a system for subtitling RTP's programs. The global system includes subtitling of live and prerecorded programs. In addition to the fully automatic system there is the possibility of using re-speaking to further enhance the overall performance. This system has been running for several months at RTP and went public on March 7, 2008, RTP's 51st anniversary.

This paper focus manly on the details of the work done to develop a fully automatic subtitling system for the 1 o'clock and 8 o'clock pm (prime time) news shows. Section 2 gives an overall description of the subtitling system followed by a detailed description and evaluation of the main components, Audio Pre-Processing (section 3) and Automatic Speech Recognition (section 4). The transcription output normalization and subtitling generation is described in section 5. In section 6 the evaluation of the overall system is presented and finally in section 7 some conclusions are drawn.

2. Subtitling System Description

The overall system is represented as a pipeline of processing blocks, shown in Figure 1.

The Control System & GUI block receives specific information from the administrator about the program to subtitle, such as name and periodicity. This block checks in RTP web site the schedule of the news show and at the specified time it starts the audio streaming to the subtitling system. The Jingle Detection block searches for the beginning jingle of the program. At the end of this jingle the audio if fed to the Audio Pre-Processing block. The Jingle Detection block also filters out the commercial breaks and streams the audio signal to the

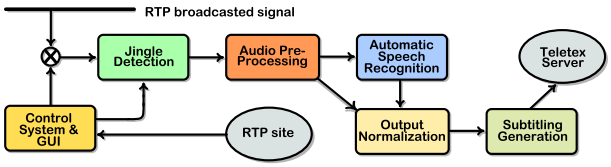

Figure 1: *Block diagram of the subtitling system.*

next block until the occurence of the news show end jingle. The APP block receives the net audio without jingles and commercials. Next the Audio Pre-Processing block segments the audio into smaller homogeneous chunks and discriminates between speech and non-speech sending only the audio to the ASR in case of speech. Additionally it outputs information regarding speaker gender, speaker clustering and true speaker identification (in case of relevant speakers). The Automatic Speech Recognition block transcribes the audio input stream according to specific vocabulary and language model provided on a daily basis. The Output Normalization block converts sequences of words representing digits, connected digits, numerals and percentages in numbers. It also capitalizes the names and introduces punctuation marks (commas and end of sentences). The last block, the Subtitling Generation creates from the output of the previous blocks the actual subtitles according to the definitions of a standard subtitle and teletext restrictions. It then sends the subtitle to the Teletext Server to be broadcasted. The overall system works in a pipelined asynchronous operation mode where each block is responsible for fulfilling its own task and for propagating the results to the next block.

2.1. Jingle and Filler Detection

This block [7] identifies in the audio stream specific acoustic patterns ("jingles") which are used in BN shows for drawing the listener's attention to important events like the start and end of the show. The 1 o'clock and 8 o'clock pm news shows that we are subtitling have basically four different jingles for marking the begin and end of the news show, for marking commercial breaks and for marking filler/headlines sections.

The block diagram of the Jingle Detection includes 5 main components [5, 8]. First component extracts 26 PLP features followed by an MLP pattern classifier trained to estimate the probability of the input frame being a certain jingle. Afterwards the outputs from this classifier are smoothed by a median filter with a small window ($t_{median} = 0.5$ seconds) and thresholded. To represent the transition events of a particular news show an appropriate Finite State Model (FSM) was developed. The last component uses the smoothed and thresholded frame probabilities and the FSM diagram for deciding whether to reject or stream the input audio signal into the Audio Pre-Processing block. In terms of performance the Jingle detection block uses a strategy where it operated with negligible "zero latency" (equal to $t_{median}/2$) when streaming the audio into the APP block and operates with $t_{median}/2 + t_{Min}$ (being $t_{Min} = 0.8$ seconds the analysis window) latency when rejecting the input audio. When a new jingle occurs signaling the start of a commercial break or the end of the news show the Jingle Detection has to reject the input audio. In this situation due to the "zero latency" mode there will be a small portion of audio with duration t_{Min} belonging to that jingle that will be incorrectly passed to the APP block. This is not relevant because it will be easily filtered out by the Speech/Non-speech component of the APP block. This "zero latency" mode was an important breakthough because it permits the Jingle Detection block to operate without delay when it is most relevant (during the useful portions of the news show). In terms of speed, the Jingle Detection block operates under 0.01 xRT. In terms of detection performance our subtitling system was evaluated using a test set named RTP07 composed by six 1 hour long news shows recorded during 2007 (ranging from May until October). This test set has 12 start/end news show jingles, 6 commercial breaks (12 start/end commercial break jingles) and 17 filler sections. Our Jingle Detection block identified correctly all these events except one of the filler sections. These results ilustrate the excelent performance of this block.

3. Audio Pre-Processing

The operation of the APP block is two-fold: filter out the non-speech parts and give additional information to the ASR and subsequent blocks such as gender classification, background conditions classification, speaker clustering and speaker identification for relevant speakers (news anchors).

The APP module [5, 8] includes six separate components: one for audio segmentation (Acoustic Change Detection), four components for classification (Speech/Non-speech, Background, Gender and Speaker Identification) and one for Speaker Clustering. These components are mostly model-based, making extensive use of feed-forward fully connected MLPs trained with the back-propagation algorithm. These components share a similar architecture that first extracts 26 PLP features and does a per frame classification using an MLP model with two hidden layers.

Despite the Acoustic Change Detection and Speech/Non-speech (SNS) blocks being conceptually different they were implemented simultaneously in the SNS component considering that a speaker turn is preceeded by a small non-speech segment. The idea was to take advantage of the probabilistic output of this Speech/Non-speech component to determine the start of speech and non-speech segments. To acomplish this the SNS MLP classifier output is smoothed using a median filter with a small window (t_{median}). This smoothed signal is thresholded and analyzed using a time window t_{Min} by a Finite State Machine (FSM). This FSM uses 4 possible states (Probable Non-speech, Non-speech, Probable Speech and Speech). If the input audio signal has a probability of speech above $Threshold_{Hi}$ the Finite State Machine is put into "Probable Speech" state. If after t_{Min} interval the average speech probability is above an average given confidence value the FSM goes to "Speech" state. Otherwise it goes to "Non-speech" state. The FSM generates segment boundaries for non-speech segments larger than t_{median} (related with the resolution of the median window). Additionally non-speech segments larger than t_{Min} are discarded. The t_{Min} value is an open parameter of the system and was optimized in the development test set so as to maximize the non-speech detected. This ACD/SNS component uses $t_{Min} = 0.85$ seconds window for decisions and a median window of $t_{median} = 0.25$ seconds. Additionaly there is a segment extend time parameter for preventing the start/end of a speech segment to close to the actual start/end of the speech which tipically induces recognition errors. This value was set to $t_{extend} = 0.2$ seconds. Boundaries placed inside short non-speech segments (between 0.25 and 0.85 seconds) are placed exactly in the middle of the non-speech segments to prevent possible speech recognition mistakes. Given this, the maximum delay is $t_{median}/2 + t_{Min} + t_{extend} = 1.175$ seconds.

When a speaker change is detected the first $t_{sum} = 300$ frames (equivalent to 3 seconds) of that segment are used to calculate gender (male or female) , background conditions (clean, noise or music) and speaker identification (anchors) classifications. Each classifier computes the decision with the highest average probability over all the t_{sum} frames. To compensate for the ACD/SNS component segment extend time the first t_{extend} frames of the new speech segment are not used for classification since they probably will contain non-speech. Given this the delay for the classification components is 3.2 seconds.

Finally, the Speaker Clustering (SC) component which uses an online leader-follower strategy tries to group all segments uttered by the same speaker. The first t_{sum} frames (at most) of a new segment are compared with all the same gender clusters found so far. SC depends on GD and SID decisions. Two SC components are used in parallel (one for each gender). Final decision of a cluster tag is only taken after GD, SID, SC male and SC female components have finished computing. Creating and updating clusters is also done afterwards. This is necessary in order to prevent stalls in the pipeline and consequently keep the latency as small as possible. A new speech segment is merged with the cluster with the lowest distance provided if it falls bellow a predefined threshold. This SC component uses 12th order PLP plus energy coefficients as features. The distance measure used for merging clusters is a modified version of the Bayesian Information Criteria (BIC) [8].

	CER / DER		
t_{sum} delay	2 sec	3 sec	4 sec
Speech/Non-Speech	4.7	4.7	4.7
Gender	2.9	2.7	2.4
Speaker Clustering (anchors)	6.2	4.8	4.1
Speaker Clustering (all)	29.2	26.3	26.0

Table 1: *Audio Pre-Processing evaluation results.*

In terms of performance the whole APP block operates in 0.014 xRT which is excelent. Table 1 summarizes the evaluation results conducted in the RTP07 test set for Speech/Non-speech, Gender Detection and Speaker Clustering for 3 different t_{sum} delay values. The results are represented by frame Classification Error Rate (CER) for SNS and GD blocks and by Diarization Error Rate (DER) for SC. Comparable results to state of the art algorithms were obtained by SNS and GD components. Speaker clustering performance for news anchors exhibits excelent results due to the SID models. For the totality of speakers the results are acceptable although still not perfect. In part these DER results can be accounted to the long duration of the RTP07 news shows which have an average of 64 (!) different speakers per news show, 23 of them being reporters. There are 3 different news anchors. Additionally these 1 o'clock and 8 o'clock pm news shows have a large percentage of speech with background noise. This reflects the current European Portuguese news shows style where field reports and street interviews are favored. This represents challenging conditions for both the APP and ASR blocks.

4. Automatic Speech Recognition

This block receives an audio input stream previously filtered by JD and APP blocks. The processing done in previous blocks facilitates ASR operation since it receives speech segments with some additional categorization information. This block has to output the most correct transcript in a real time and online mode.

Our BN speech recognition engine named AUDIMUS.MEDIA [3, 7] has an hybrid MLP/HMM acoustic model combining posterior phone probabilities generated by several MLP's trained on distinct input features. Different feature extraction and classification streams effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, present in BN data. The 1st stream extracts 26 PLP features, the second 26 Log-RASTA features and the 3rd uses 28 Modulation Spectrogram (MSG) coefficients for each audio frame. Each MLP classifier incorporates local acoustic context via an input window of 13 frames (MSG uses 15 frames). The resulting network has two fully connected non-linear hidden layers with 2,000 units each and 39 softmax output units (corresponding to 38 Portuguese phones plus silence).

Initially this acoustic model was trained with BN data collected from RTP and manually annotated in a total of 46 hours. Currently automatically collected and transcribed data is being reused to perform unsupervised training. Recognized words that have a confidence measure above 91.5% are chosen for new training data. This is an iterative and never ending process while we get better performance with more data. Table 2 summarizes the acoustic model unsupervised training improvements. Currently we are using 378 hours of training data, 332 of which were automatically annotated using word confidence measures. This represented a significant 8.5% relative improvement in Word Error Rate (WER).

Training data	Hours			% WER	
	Raw	Usefl	Sum	F0	All
training set	—	46 h	46 h	11.3	23.5
" + 2 months 05	60 h	33 h	79 h	11.0	22.7
" + 6 months 06	166 h	100 h	179 h	10.8	22.1
" + 1 year 06-07	405 h	199 h	378 h	10.5	21.5

Table 2: *Acoustic model unsupervised training.*

AUDIMUS.MEDIA decoder is based on the Weighted Finite-State Transducer (WFST) approach, where the search space is a large WFST that results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one [4]. This decoder uses a specialized WFST composition algorithm of the lexicon and language model components in a single step. Furthermore it supports lazy implementations where only the fragment of the search space required in runtime is computed. Besides the recognized words the decoder outputs a series of values describing the recognition process. In order to generate a word confidence measure these features are combined through a maximum entropy classifier whose output represents the probability of each word being correct [4]. Confidence measures for the recognized text are fundamental not only to select new acoustic training data but also to filter the output text in the subtitling composition stage.

Our ASR system uses a 100k word vocabulary adapted in a daily base to reflect the new words that appear in web newspaper texts [9]. This daily modification of the vocabulary implies a re-estimation of the language model and retraining of the word confidence measures classifier. In order to validate the new vocabulary and language model generated, a benchmark test was created, running after the daily adaptation process. This validation data is then used to retrain the confidence measure classifier. Figure 2 presents ROC curves for the original and adapted confidence measures. The most relevant aspect is the linearization of the confidence threshold after the adaptation.

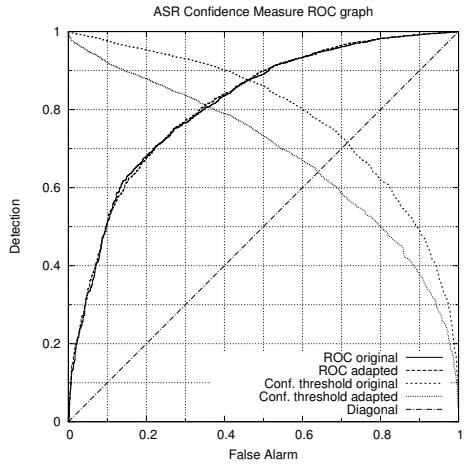

Figure 2: *Confidence measure classifier adaptation.*

5. Normalization and Subtitling generation

The final two processing blocks transform the output information from the ASR block into subtitles. This comprises the normalization stage which converts numbers, numerals, dates and amounts (mainly money and percentages) in their digit representation. The other block capitalizes the names and acronyms and recovers the punctuation on the recognized text with commas and end of sentence marks. This technique is based on information from the APP and ASR modules such as pauses, speaker changes, previous present and next words, the grammatical class of each word and the confidence measure associated to each word [6]. Both blocks improve the readability of the subtitles. A considerable effort was put into understanding the best way to present the information. Presently subtitles are shown using 2 lines on the top of the screen with speaker gender information being used to change the color of the subtitled text.

Threshold	100 %		82 %	
Speakers	% Time	% WER	% Discard	% WER
Anchors	24.5	9.4	5.9	8.1
Reporters	44.1	14.1	20.3	11.1
Other	31.4	37.4	61.4	22.1
Total	100.0	20.3	29.7	12.0

Table 3: *Subtitling results in RTP07 test set.*

6. Global Evaluation

Table 3 represents the evaluation results in the RTP07 test set. The first two columns of results represent the % of time for each speaker category and the WER when all recognized text is displayed (Threshold at 100%). Overall WER is slightly above 20% which is very good considering the dificult conditions present in the news shows. The average word confidence measures in a sentence are used to filter out sentences that are poorly recognized. The last two columns of Table 3 represent the results in terms of % of discarded data from each category and the WER in the remaining data when the Threshold is set to 82%. We can see that anchors and reporters are much less affected than other speakers. WER for the filtered subtitles is 12% which is surprisingly good considering that 70% of the total recognized words are being displayed. In terms of processing time the ASR block works under 0.82 xRT. Latency was another important aspect for which an exhaustive evaluation was made. JD + APP + ASR blocks have an average of 3.5 seconds delay. Then the use of 2 lines for subtitling introduced an additional latency since it is necessary to fill the subtitle before displaying it. The complete delay for all blocks in the subtitling system is in average 6.5 seconds.

7. Conclusions

This subtitling 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 even less working online and in real time. All these developments provided a unique research and development platform that we will continue to explore in the future reducing the gap between the BN systems for Portuguese and for other languages. The evaluation results presented on this paper are at similar level with state-of-the-art systems for other more developed languages, and resulting from a system with several and innovative differences. We will continue in the near future to explore these differences, mainly introducing new speaker adaptation techniques, improving language modeling and reducing vocabulary size to better accommodate the word types generated each day, in order to further improve the system performance and reduce the latency time.

8. Acknowledgements

This paper represents an extensive work only possible with the cooperation of several persons and institutions. We would like to thank RTP and their collaborators, specially João Sequeira and Teotónio Pereira. This work was partially funded by PRIME National Project TECNOVOZ number 03/165 and European program project VidiVideo FP6/IST/045547 and FCT National project PTDC/PLP/72404/2006.

9. References

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

[2] M. Gales, D. Kim, P. Woodland, H. Chan, D. Mrva, R. Sinha, and S. Tranter, "Progress in the cu-htk broadcast news transcription system," *IEEE Transactions on Audio, Speech and Lang. Proc.*, 2007.

[3] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus.media: a broadcast news speech recognition system for the european portuguese language," in *Proc. PROPOR '2003*, Faro, Portugal, 2003.

[4] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[5] H. Meinedo and J. Neto, "A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models," in *Proc. Interspeech '2005*, Portugal, 2005.

[6] F. Baptista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering punctuation marks for automatic speech recognition," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[7] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. ICASSP 2008*, Las Vegas, USA, 2008.

[8] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, IST, Lisbon, Portugal, 2008.

[9] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *Proc. ASRU 2007*, Kyoto, Japan, 2007.