Real-Time Live Broadcast News Subtitling System for Spanish

Alfonso Ortega, Jose Enrique Garcia, Antonio Miguel,
Eduardo Lleida

Aragon Institute for Engineering Research, University of Zaragoza, Spain
ortega,jegarlai,amiguel,lleida@unizar.es

Abstract

Subtitling of live broadcast news is a very important application to meet the needs of deaf and hard of hearing people. However, live subtitling is a high cost operation in terms of qualification human resources and thus, money if high precision is desired. Automatic Speech Recognition researchers can help to perform this task saving both time and money developing systems that delivers subtitles fully synchronized with speech without human assistance. In this paper we present a real-time system for automatic subtitling of live broadcast news in Spanish based on the News Redaction Computer texts and an Automatic Speech Recognition engine to provide precise temporal alignment of speech to text scripts with negligible latency. The presented system is working satisfactory on the Aragonese Public Television from June 2008 without human assistance.

Index Terms: Broadcast News, Subtitling, Speech Recognition

1. Introduction

Design for all principles and accessibility for deaf and hard of hearing people must be ensured by institutions being a very important field of research, where new technology has a huge potential to facilitate their lives. Efficient and low cost applications can be developed bringing down old barriers.

Broadcast TV Subtitling is a key point to allow free access to the Information Society for deaf and hard of hearing people. Beside, this feature can be also very useful not only for people with hearing impairments but for everyone in some scenarios such us railway stations, airports, restaurants or any other crowded place with high level of noise.

Closed-captions are available mainly for recorded programs, generally based on manual transcription operation. However, the difficulty for providing precise subtitles for live programs is much higher. This operations, usually implies the use of specialized stenographers (taking into account the shortage of skilled people in this area), the use of Automatic Speech Recognition Systems (ASR) used by shadow speakers [1] or some other sort of assisted systems [2]. The cost of these approaches can be excessive for small TV Companies and fully automatic tools can lead to a significant cost reduction. Some systems have been presented using real-time ASR without the need for shadow speakers [3] [4] but several problems that should be solved such as the need for low latency real-time recognition or the level of accuracy have been reported [5].

In order to obtain low cost fully-automatic subtitling, we propose a system based on speech-text alignment by using the news redaction computer systems (NRCS) texts and automatic speech recognition without decreasing the performance of the system compared with a human assisted system and negligible lantency.

This paper is organized as follows. A brief description of the system is presented in Section 2. In Section 3, the module responsible for the texts retrieval is described. The Speech-Text alignment system is described in Section 4, and a performance study is presented in Section 5. Finally, the conclusions are discussed in Section 6.

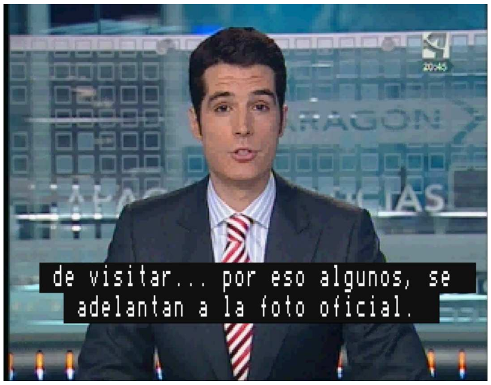
de visitar... por eso algunos, se
adelantan a la foto oficial.

Figure 1: *Output of the system with a two line subtitle.*

2. System Description

This section provides a general description of the presented system which is composed of several modules following a client-server architecture. A block diagram of the live broadcast news subtitling system can be found in Fig. 2.

The News Redaction Computer (NRC) contains the scripts of every piece of news that compose the news program. The Text Retrieval (TR) Module must locate them and send them out to the Speech-Text Alignment (STA) Module discarding all the information that is not useful for the caption creation. A piece detection system, responsible for the estimation of the current piece on air also receives the scripts of the pieces.

The STA Module, based on an Automatic Speech Recognition (ASR) engine, receives the texts that must be printed on the screen and performs the alignment with the audio signal coming from the production studio.

Finally, the subtitles are sent to the teletext server and then these captions are real-time displayed through the 888 page.

The whole system is assisted by the Continuity System which sends Start and Stop signals before and after the program to be subtitled begins and ends.

In order to control the height where subtitles will be shown, the system is also connected to the system responsible for the superimposition of text on the screen. This delivers a signal to the STA Module indicating that a text is now being shown and thus the STA module will raise the subtitles allowing the reading of the text being prompt.

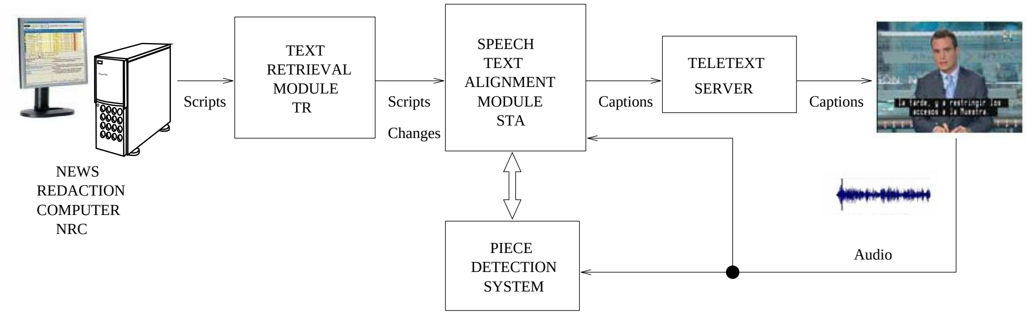

Figure 2: *Block Diagram of the Subtitling System.*

3. Text Retrieval Module

The Text Retrieval Module is responsible for obtaining all the information related to the current news program and sending it both, to the Speech-Text Alignment Module and to the piece detection system. It is also connected to the Continuity System receiving the Start and Stop commands from it.

All the information retrieved by the TR Module from the NRC is parsed discarding all the contents that are not useful for caption creation. The ramaining information is XML formatted and sent to the Speech-Text Alignment Module.

Once the news program has started, hundreds of modifications take place on the NRC. The content of the news program, the text of each piece or the order can vary during the program, just a few seconds before a specific piece starts. Hence, the TR module must be continuously monitoring the content of the NRC sending each modification in real-time to the corresponding modules.

4. Speech-Text Alignment

4.1. Caption Generation

The Speech-Text Alignment Module performs the submissions of the subtitles in a synchronized way with the audio signal to the teletext server. It receives the scripts from the TR Module and decodes the speech signal coming from the production studio by using an Automatic Speech Recognition (ASR) System.

4.1.1. Acoustic Models

The ASR engine uses continuous HMMs with context dependent acoustic units, where each unit is modelled with a 16 component Gaussian Mixture Model with diagonal covariance matrices. They are speaker-independent and gender independent.

The features vectors used are 12-order Mel Frequency Cepstral Coefficients and the normalised energy coefficient augmented by the corresponding delta and delta-delta coefficients.

Speech is captured and digitalized, if needed, in a 16 kHz, 16 bits per sample format.

4.1.2. Language Model

The language model is a finite state grammar, built using the text received from the NRC system. It also counts with a phoneme network linked to the script grammar every time a pause is inserted in the piece. The reason to use this phoneme network is because the pieces to be subtitled contain interview clips, 'pieces to camera', live reports,... for which the system doesn't have the corresponding text. Reporters, mark the place of insertion of those clips and the STA module stops the progression of the piece grammar. Hence, captions won't be sent to the teletext manager system during those periods. Thus, the phoneme network is used to model acoustically those parts of the pieces for which the system doesn't have the correct transcription.

4.2. Piece Detection

A very important problem that must be solved in this system is which piece of news is currently being uttered.

At the beginning of the program, the STA module receives the complete list of pieces that will be broadcast during the news program in the predefined order. However, the order of the pieces may vary during the program and there is no external signal that indicates the beginning or the end of every piece.

Thus, the STA module must ensure that a few seconds before a piece is going to start, everything is ready to perform its subtitling. Nevertheless, if the subtitling of a piece has not finished correctly, (the announcer has not uttered the script of the news correctly, a mistake in the order of the pieces, an error in the STA module, low audio quality, ...) the next piece will start and the STA module won't be ready to perform its subtitling.

To avoid this incorrect behaviour, a piece detector has been implemented. This module is responsible for retrieving the piece of news that is currently being uttered every time the STA module looses its synchronism.

This module is based on an ASR engine that estimates the piece currently on air by acoustic decoding of the audio stream coming from the broadcast news program. It decides that a piece of news is now being uttered if the acoustic decoding of one among the next few pieces is progressing correctly, that is, the ASR engine has one decoding path with more than a predefined number of words (around 10).

Anyway, there is another measure to avoid the system to remain clipped in a piece of news that uses the status information of a piece. If a piece of news with a video clip (not all the pieces behaves this way) is about to start its status field changes from CUED to PLAY. If this change is detected in a piece that is not the next one to be subtitled, the system automatically selects that piece as the next one to be subtitled.

5. Performance Evaluation

In order to evaluate the performance of the system two different studies have been carried out. On the one hand, several experiments has been made under controlled conditions in order to evaluate the ability of the system to overcome situations in which the transcriptions are highly imperfect. On the other hand, a statistical study has been carried out in order to evaluate the performance of the system working under real conditions. Over a one month period has been taken into account measuring the amount of pieces that are completely subtitled, partially subtitled or not subtitled at all.

5.1. Experimental Evaluation

During the utterance of a piece of news it is highly probable that the announcer does not follow exactly the text in the script.

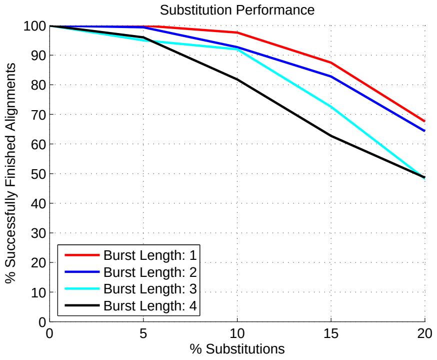

Figure 3: *Rate of successfully finished alignments when substitutions are present on the transcriptions.*

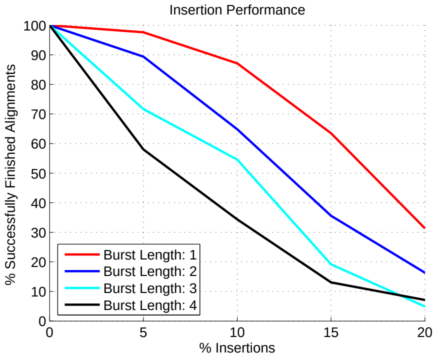

Figure 4: *Rate of successfully finished alignments when substitutions are present on the transcriptions.*

On the contrary, substitutions, insertions and deletions are very often found in almost every piece of news. In this context we consider insertion when a word is present in the text but not uttered by the speaker and a deletion when the speaker says a word not present in the script.

In order to evaluate the behavior of the system in those situations, several experiments have been made using part of the laboratoty corpus of the database AV@CAR [6]. Specifically, a group of phonetically-balanced sentences in Spanish uttered by 20 speakers. Substitutions, insertions and deletions were artificially added randomly to the transcriptions and the number of successfully finished sentences was evaluated. The length of the bursts of substitutions, deletions and insertions ranges from a single word to a group of four consecutive words.

Fig. 3 shows the ratio of successfully finished alignments when the rate of substituted words in the transcriptions varies. It can be seen how, the decrease in performance is not very fast and depends moderately on the length of the substitution burst.

The influence of insertions is much higher as can be seen in Fig.4. The ratio of sentences successfully aligned rapidly decreases from 100% for perfect transcriptions to 30% when transcriptions have 20% of words that have not been uttered by the speaker. The decrease is much faster when the length of the insertion burst is four, where only 50% of the alignments are correct with a 7% of inserted words.

Finally, deletions have a small influence on the performance of the system as shown in Fig. 5. Almost all the alignments finish successfully when deletions rate ranges from 0% to 20% regardless of the length of the deletion burst.

In summary, these experimental results shows that the proposed system is very robust to deletions. When the rate of substitutions is moderate it behaves satisfactory and it is more sensitive to insertions, specially when long burst of uttered words are not present in the text.

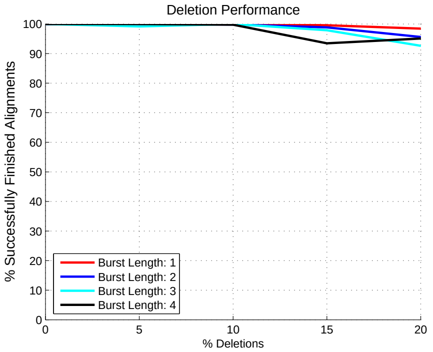

Figure 5: *Rate of successfully finished alignments when substitutions are present on the transcriptions.*

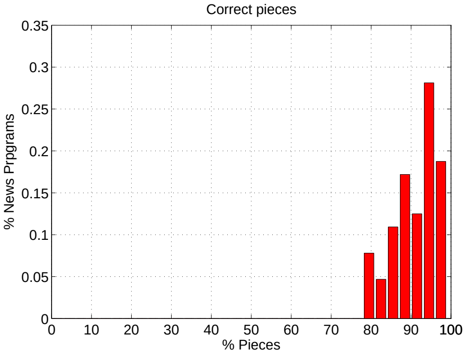

Figure 6: *Histogram of successfully completed pieces.*

5.2. Performance Evaluation Under Real Conditions

In Aragon Television, news program are scheduled in three editions. The main bulletin is broadcast everyday at 14:30 p.m. and is around 60 minutes long with more than 90 pieces to be subtitled. The main characteristic of this edition is that the number of changes in the order or content in the pieces are very high so the subtitling system must be adapting to these changes very often. The second edition is broadcast at 20:30 and is 45 minutes long. Changes occur not so often but there are still some of them during the program. The number of pieces to be subtitled is around 75, although Sundays edition is a reduced version lasting only 15 minutes and containing around 20 pieces to be subtitled. Finally the third edition is broadcast at around 0:00 and is around 30 minutes long, with around 50 pieces to be subtitled. The number of changes in order or content received for this edition is quite low.

To evaluate the system under real conditions, a one month period has been considered to obtain several performance measures. For that purpose, the number of pieces that were successfully subtitled were considered. Three different type of subtitled pieces were defined: Complete subtitled piece, when 100% of the captions were successfully sent to the teletext manager system, partially subtitled pieces, when some of the captions were not sent to the teletext manager system and missing pieces when none of the caption were sent.

In Fig.6 the histogram of correct subtitled pieces per news program is shown. It can be seen that none of the news program contain less than 75% of its pieces correctly subtitled. More than 55% of the news programs end with more than 90% of the pieces completed successfully.

Fig. 7 shows the rate of pieces that are partially subtitled,

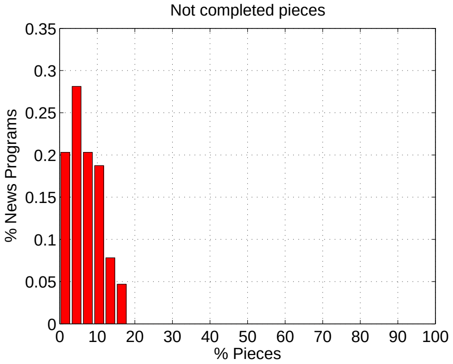

Figure 7: *Histogram of not completed pieces.*

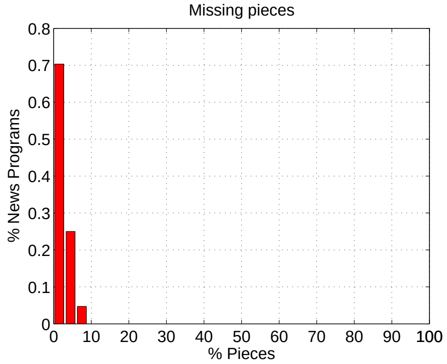

Figure 8: *Histogram of missed pieces.*

that is some of the captions are not sent to the teletext manager. None of the news program finishes with more than 20% of the pieces partially subtitled and near a 70% of the news programs end with less than a 10% of the pieces partially subtitled.

Finally, Fig. 8 shows the rate of pieces that are not subtitled at all. More than 90% of the news programs finishes with less than 5% of the pieces not subtitled at all.

In order to summarize the results obtained under real conditions, the degree of success for each news edition is presented in Table 1. As average, 91.52% of pieces in a news program are completed successfully, 6.7% are partially subtitled and 1.78% are not subtitled or missed. Since the fist edition is the most dynamic environment, the accuracy in the subtitling operation is lower than in the other editions, but still very high, obtaining an average of almost 90% of the pieces correctly subtitled. On the other hand, third edition is the most static environment and hence, the system achieves the higher performance with an average of almost 95% of the pieces successfully completed and less than 1% of missed pieces.

Regarding the latency of the system, due to the fact that the captions are preprocessed when the scripts are received and sent to the teletext system right after the first word in the caption is uttered, it can be concluded that the overall latency is negligible. Only a small delay is intentionally added at the beginning of each piece and right after the arrival of a period without transcription (video clip, interview, live report, ...). In those cases the sending of the caption is delayed around four words in order to avoid the appearance of false alarm submissions caused by the speech that has no corresponding text.

6. Conclusions

In this work, a fully automatic real-time subtitling system for broadcast news programs in Spanish has been presented. The system is based on an automatic speech recognition engine that receives the texts that should be sent to the teletext manager system and performs the temporal alignment of speech and text. To obtain the scripts that will be uttered by the announcer, a Text Retrieval module communicates with the News Redaction Computer System and sends them free of non-useful information to the Speech-Text Alignment Module. Then, the captions are sent to the teletext system and displayed in real-time through the 888 page. Since the STA module doesn't know exactly which piece will be uttered next, a piece detection system has been implemented. It is based on an ASR engine that estimates the piece currently on-air by acoustic decoding of the audio stream and it is assisted by lateral information coming from the NRC system. A statistical study performed over a one-month period shows that the accuracy of the system is high, reaching 91.5% of successfully subtitled pieces and less than 2% of missed pieces. In addition, the system is able to operate with almost zero latency since the captions are submitted to the teletext system right after the speech signal is acoustically decoded, a task that is able to run several times real-time. The subtitling system is working satisfactory on Aragon Television without human assistance from June 2008.

EDITION	COMPLETED	NOT COMPLETED	MISS
1st ED	88.64%	8.55%	2.81%
2nd ED	91.50%	6.90%	1.60%
3rd ED	94,42%	4.65%	0.93%
MEAN	91.52%	6.7%	1.78%

Table 1: Statistics of Completed, Not Completed and Missed Pieces

7. Acknowledgements

The authors would like to thank Corporacion Aragonesa de Radio y Television (CARTV) for its technical and financial support in this project.

This work has been partially funded by the national project TIN2008-06856-C05-04.

8. 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] J. Brousseau, J.F. Beaumont, G. Boulianne, P. Cardinal, C. Chapdelaine, M. Comeau, F. Osterrath and P. Ouellet, "Automated closed-captioning of live TV broadcast news in French", in Proceedings of Eurospeech 2003, Geneva, Switzerland, 2003.

[3] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Program", IEEE Transactions on Broadcasting, Vol. 46, No. 3, September 2000.

[4] 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.

[5] H. Meinedo, M. Viveiros, J. Paulo and S. Neto, "Evaluation of a Live Broadcast News Subtitling System for Portuguese", in Proc of Interspeech 2008, Brisbane, Australia, 2008.

[6] A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera, "AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition", in Proc. of 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.