Automatic Live Subtitling: state of the art, expectations and current trends

Carlo Aliprandi,
Cristina Scudellari, Isabella Gallucci,
Nicola Piccinini, Matteo Raffaelli
Synthema
Pisa, Italy

Arantza del Pozo,
Aitor Álvarez, Haritz Arzelus
Vicomtech-IK4
San Sebastián, Spain

Renato Cassaca, Tiago Luis,
Joao Neto, Carlos Mendes,
Sérgio Paulo, Marcio Viveiros
VoiceInteraction
Lisbon, Portugal

Abstract - *The subtitling demand has grown quickly over the years. The path of manual subtitling is no longer feasible, due to increased costs and reduced production times. Assisted Subtitling is an emerging technique, consisting in the application of Automatic Speech Recognition (ASR) to automatically generate program transcripts.*

This paper will report on recent advances in ASR, presenting SAVAS, a novel Speaker Independent ASR technology specifically designed for Live Subtitling. We will describe the technology, presenting its features and detailing language and domain-specific tunings that we have carried out.

We will also introduce the S.Scribe!, S.Live! and S.Respeak! systems, which are based on SAVAS. S.Scribe! is a batch Speaker Independent Transcription system for subtitling. S.Live! is a first-of-a-kind Speaker Independent Transcription System, with real-time performances for online subtitling. S.Respeak! is a collaborative Respeaking System, for live and batch production of multilingual subtitles.

S.Respeak! has proven to be sufficiently robust for programs where the acoustic conditions are challenging and for spontaneous speech. Similar results are expected to be achieved also for S.Live! and S.Scribe!, which are currently being tested under real conditions at different broadcasters premises, to subtitle live programs, in both assisted and unassisted tasks. We will finally detail performances of the systems for 7 languages (English, Spanish, Italian, French, German, Portuguese and Basque).

INTRODUCTION

Subtitling is the process of producing transcriptions of audio, to be synchronously displayed with the video on a television, video screen or any other display device. If subtitles also include descriptive information of non-speech elements, like music or speaker names, they are usually referred to as captions. In this work we will refer to the general process of subtitling, as captions and subtitles are considered equivalent in many countries and cultures.

It is commonly agreed that subtitling was mainly conceived for television and for the benefit of deaf and hard of hearing people, hence the origin of the acronym SDH, Subtitles for the Deaf and Hard of hearing. Nevertheless subtitles are nowadays used in several new media and are spread for the benefit of all people.

Traditionally, the subtitling process is based on the manual production of time-aligned transcriptions of audiovisual content, a task which requires considerable effort. Manual production of high-quality subtitles has been reported to take between 8 to 10 times the length of the video material [1]. Although the use of dedicated subtitling software tools that facilitated the subtitling process among professionals, Automatic Speech Recognition (ASR) has only recently started to be adopted to increase its productivity.

Respeaking is a technique thanks to which a professional listens to the source audio and dictates it, so that his/her vocal input is processed by a speech recognition engine which transcribes it, thus producing subtitles. Respeaking has consolidated as the main subtitling technique employed for live broadcast productions, quickly taking over traditional techniques, like stenotyping. The reasons are two: on the one hand respeaking has a shorter learning and training process in comparison to stenotyping, i.e. two or three months vs. two or more years; on the other hand, the cost of a respeaker is lower than the cost of a stenotypist, i.e. one or two times less. In addition, the advancement of respeaking technology and respeaker expertise has so increased as to achieve results which are similar and even better than stenotyping and other reporting techniques, like typewriting and shorthand, as proven in the Intersteno championships [2].

Respeaking can also be employed to script pre-recorded programs, which can then be fed to assisted subtitling applications. These are tools which incorporate ASR technology capable of aligning the scripts to the spoken audio in order to automatically generate subtitle time-codes. Despite post-editing might still be required to adapt the transcriptions to the needs of the community of the deaf and hard of hearing, the use of respeaking for scripting and forced-alignment for automatic time-code assignment can still save a considerable amount of subtitle generation time.

In this paper we will focus on another emerging trend, which is raising a lot of expectations: the application of ASR to automatically generate transcripts of programs without using a respeaker, and to use the transcripts as the basis for subtitles. Despite the difficulties posed by the multitude of different voices and the variety of acoustic conditions, the accuracy achieved by this technique can be good enough in bounded domains. Systems of this kind are currently being employed by some broadcasters in the live news domain. The main advantage of this method compared to respeaking is that it can actually produce similar results without the need of a respeaker, which helps reduce subtitling costs.

ASR TECHNOLOGY AND ASSISTED SUBTITLING APPLICATIONS

The first experiments to use ASR for live subtitling were conducted when the technology was still in its preliminary stages. In [3] the use of speech input was proposed in conjunction to keyboard entry to control the formatting (like positioning, style or color) of live subtitles entered on a QWERTY keyboard, thus enabling the operator to focus maximum effort on text entry.

Once technology became available for Continuous Speech Recognition that let users dictate into applications, it was investigated as an application to deliver near real-time transcriptions for live subtitling. Production of acceptable subtitles became possible, with respeaking solutions like Synthema Voice Subtitle [4] and SysMedia SpeakTitle [5].

Today, respeaking tools are the most widely found Assisted Subtitling applications in the market. WINCAPS Q-Live [6], FAB Subtitler Live Edition [7], Miranda Softel Swift Create [8], Starfish Isis [9] and Ninsight Protitle [10] are examples of subtitling solutions which integrate commercial ASR engines specifically developed for dictation purposes.

FIGURE 1: RESPEAKING OF SPORT EVENTS

The main ASR engines are IBM ViaVoice [11], that nowadays has been discontinued from the market, Microsoft Windows Speech Recognition [12] and Nuance Dragon NaturallySpeaking [13]. However, these ASR engines have some limitations. They are Speaker Dependent, i.e. they have to be adapted to each user and require dictation of some training sentences to adapt recognition to the user speech. Since they have been designed for dictation applications, they sometimes do not perform well for spontaneous speech and in complex acoustic conditions. Finally, being developed to target languages for which training data is available (like English, Spanish, French, German, Italian) they are not available for many languages, in particular for minor languages.

Less subtitling solutions exist that allow respeaking of pre-recorded content and/or are capable of aligning (respoken) scripts to audio, for the automatic generation of subtitle time-codes. WINCAPS Qu4ntum [14] is one of such tools, which includes respeaking and automatic timing features. Again in this context, the underlying speech recognition technology is a dictation engine.

The lack of assisted subtitling tools allowing the automatic generation of subtitles from the audio, without the need of respeaking, has been limited by the unsuitability of the available dictation technology for audio transcription. Experiments directly applying dictation technology to transcribe audio [15] have revealed that such type of engine's high Word Error Rate make it unsuitable for fully automated subtitling. The mismatch of the acoustic conditions, vocabulary and linguistic domains between the data used to train the dictation engines and the tested audios are deemed responsible for the technology's underperformance. The adaptation of dictation engines to the acoustic condition, vocabulary and language domain of the audio to be transcribed has shown recognition accuracy improvements and more promising results, applicable to the automatic generation of draft transcriptions for post-editing [16].

Although ASR technology development has now moved towards transcription, there are still not many solutions of this kind available in the market. The main reason is the amount of audio and text data required to train high-quality transcription systems per domain and language. Several studies, like [17] and [18], state that at least 100 hours of annotated and transcribed audio are required for the adequate training of transcription engines and practical developments use as much data as possible. As a result, the commercially available transcription engines are widely scattered across languages and domains. Koemei [19], SailLabs [20], Vecsys [21] and Verbio [22] are companies offering transcription solutions for varying pools of languages and application scenarios such as lectures, open source intelligence or media. In the subtitling field, VoiceInteraction pioneered a transcription solution [23] capable of generating subtitles for Portuguese broadcast news, that was adopted and is currently in daily use by RTP – the public Portuguese broadcaster. More recently, internet services have arisen offering the generation of draft time-aligned subtitles for post-editing, from the alignment of original audio and scripts, like Ubertitles [24] and eCaption [25]. They offer services in different languages, based on proprietary transcription technology employed for the alignment task.

None of the transcription engines described above has yet been integrated in any of the main dedicated software tools employed by the subtitling industry, nor their performance and suitability for automatic subtitling has been formally assessed for the time being, especially for online processing of live programs.

SAVAS ASR SUBTITLING ENGINES AND SYSTEMS

In order to deliver quality live subtitles, a number of challenging requirements have to be satisfied, well beyond the performances that an ASR engine can provide. ASR technology has to be adequately improved to fit to more general quality and operational criteria.

There are three main tasks that have to be performed by a complete and efficient automatic subtitling system:

a) Raw subtitles production
b) Subtitles post-correction
c) Subtitles airing

The interactions between these tasks and the modalities of each of them may change depending on the type of subtitling required (e.g. Live or Semi-Live).

The most important task is the production of raw subtitles. Producing high quality raw subtitles means that less time (if any) has to be spent on correction and editing before airing. This naturally translates in reduced costs and better quality. Therefore, in contrast to most commercial speech recognition technology available, we have specifically designed the SAVAS dictation and transcription engines for subtitling purposes. They have been trained with data from the broadcast news, sports and interview/debate domains in 7 languages so far: English, Basque, Spanish, Portuguese, Italian, French, German, plus the Swiss variants of the latter three.

Table 1 shows the data that have been used to train the SAVAS engines for the different domains and languages.

Language	***Domain***	***Audio***	***Text***
Basque	Broadcast news	200h	350M
	Sports	20h	200k
Spanish	Broadcast news	200h	1B
Portuguese	Interview/debate	20h	200k
Italian	Broadcast news	165h	1B
	Sports	--	500k
Swiss Italian	Broadcast news	50h	100M
French	Broadcast news	150h	1B
Swiss French	Broadcast news	50h	100M
German	Broadcast news	150h	1B
Swiss German	Broadcast news	50h	100M

TABLE 1. TRAINING DATA PER LANGUAGE AND DOMAIN

As it can be seen, up to 200 hours of audio, coming from TV programs, and 1B words of text, mainly coming from scripts, subtitles and autocues, have been used to grant high quality ASR output, for most of the languages and domains.

Thanks to the large training data, it was possible to develop Speaker Independent engines: they can recognize different speakers without any training, and they work for different speaker accents, dialects and acoustic conditions.

For the production of Live subtitles, a Speaker Independent ASR engine is a requirement, but it may be not sufficient in several operational conditions. Live subtitling implies, besides real-time Speaker Independent ASR, an online operation, that may in particular support tasks b) and c) previously described. Transforming all the algorithms to online operation is not always a smooth and straight task. Putting a full subtitling system to work, requires a strong effort to develop the appropriate software and to be able to explore the specificity of the underlying hardware processing power, for which very good engineering skills are needed.

So we developed additional components for live subtitling, and we delivered three new subtitling systems based on the SAVAS engines: S.Scribe!, S.Live! and S.Respeak!.

All the three systems share the SAVAS engines and provide useful operational capabilities required for online subtitling, like:

- speech classification (speech, music, jingle detection)
- speaker diarization
- speaker change detection
- speaker identification (at least for anchor speakers)
- text normalization techniques (to process the ASR output result accordingly)
- subtitles formatting and editing
- subtitle generation component with TV broadcaster specifications

FIGURE 2: SAVAS ONLINE SPEAKER CHANGE DETECTION

S.Scribe! is a batch Speaker Independent Transcription and Subtitling system, capable of automatically transcribing audio and video files into time-aligned subtitles, detecting speech and non-speech audio, and giving information on speaker language and gender. **S.Live!** is a first-of-a-kind Online Subtitling system, capable of automatically transcribing speech into subtitles, detecting speech and non-speech, and giving information on speaker clustering, speaker gender and speaker identification. **S.Respeak!** is a remote Respeaking system for collaborative subtitling, with fast post-editing and automatic management of subtitle formatting, capable of producing subtitles with an acceptable delay and a correct on-screen persistence.

I. S.Scribe!

S.Scribe! is a client/server system, working offline: it can process a file of previously recorded audio/video and transcribes it, producing a subtitle file.

The system has an interface for administration and usage; it receives an audio/video file, puts it in a processing list and notifies the user upon completion, so that he/she can download the result. The most common and standard subtitling formats, like TTML or SRT, are supported.

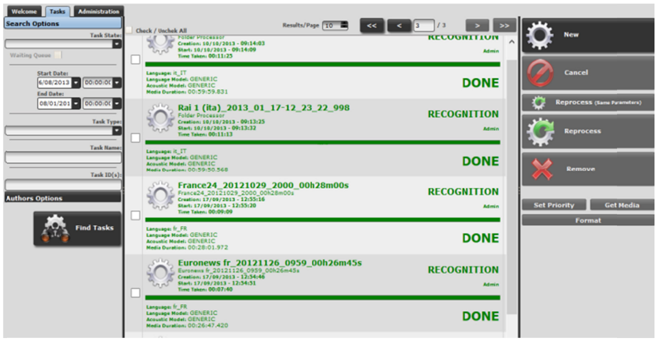

FIGURE 3: SCREENSHOT OF S.SCRIBE!

S.Scribe! has 2 operation modes:

- HTML Interface: the system is available at a given web address (URL). The user has to log in and then he/she can submit an audio/video files to be processed (see Figure 3);
- Webservice interface (SOAP/WSDL): the system is invoked through a webservice. The user specifies a URL where the audio/video file is expected to be available for downloading and processing.

II. S.Live!

Several modifications were introduced in the SAVAS engines for the S.Live system, in order to satisfy the operational conditions of online tasks.

It was necessary to implement a new efficient decoder based on Weighted Finite-State Transducers (WFST) and, in addition, to revise the Audio Pre-Processing module (see Figure 4), that allows for the production of meta-data, in order to adapt it to live performances and better characterize the audio content (e.g. speaker gender, background noise classification).

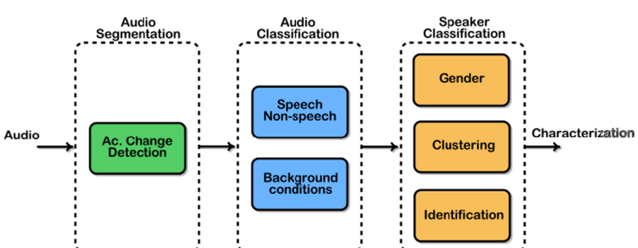

FIGURE 4: AUDIO PRE-PROCESSING BLOCK DIAGRAM

The modifications produced significant enhancements to the system, particularly to the Speaker Diarization component, that had positive consequences on the Speaker Change Detection (SD) component. SD is very important to properly identify different speakers in subtitles: a sample of SD is showed in Figure 2.

Language	*Old DER*	*New DER*
Italian	30.07 %	16.84 %
Portuguese	29.06 %	24.04 %
Spanish	41.30%	32.60 %

TABLE 2. DIARIZATION ERROR RATE IMPROVEMENT

Table 2 presents enhancements to Diarization Error Rates (DER) obtained for 3 languages. Figure 5 shows a screenshot of the S.Live! TV Programs Scheduler, used to schedule programs to be automatically live subtitled.

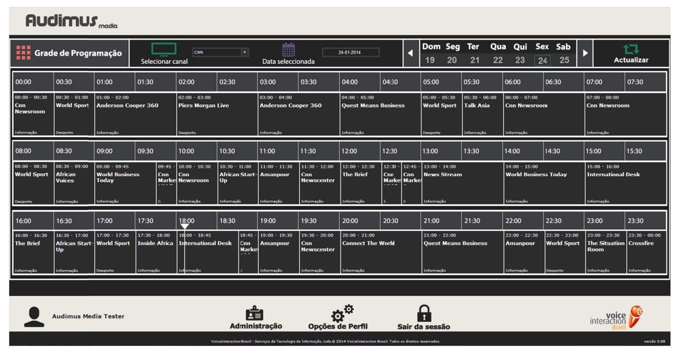

FIGURE 5: SCREENSHOT OF S.LIVE! TV PROGRAMS SCHEDULER

III. S.Respeak!

S.Respeak! is a client/server system and works both offline and online.

It is easily scalable from single user to complex collaborative workflows, even across the Internet. It has been designed to allow the best possible integration with the SAVAS ASR engines, so that either batch and online subtitling features could be integrated into a professional subtitling workflow.

S.Respeak! is available both as a standalone and a web client; the web client is presented in Figure 6.

FIGURE 6: S.RESPEAK! WEB CLIENT

Besides traditional respeaking features, like correct on-screen subtitles persistence, management of subtitles formatting (like colors and capitalization styles), S.Respeak allows the use of domain-specific phrases and fast post-speech editing, leveraging on the output of the SAVAS engines. It also introduces interesting new features, making respeaking a more collaborative process, thus splitting the cognitive load among several respeakers, correctors and subtitlers, optionally coordinated by a supervising operator.

SAVAS ASR SUBTITLING

The output of the three SAVAS systems complies with the main subtitle layout, duration, punctuation and text editing constraints.

Layout features such as the screen position of subtitles, the number of lines, text positioning, the number of characters per line, the font, background and speakers' colours and the transmission mode (block by block, line by line, word by word or scrolling) are configurable.

FIGURE 7: S.SCRIBE! ONLINE SUBTITLES

The persistence of subtitles on-screen can also be configured through features such as the average reading speed, the duration of short and single word subtitles, the average duration of one-line or two-line subtitles or the frame gap between subtitles.

In addition, the systems include statistical punctuation modules, trained on acoustic and linguistic features for each language, capable of inserting full stops and commas. The systems include also capitalization modules, that automatically capitalize words when necessary, like at the beginning of new subtitles or sentences, and when names of entities (such as persons, locations or companies) are detected. Figure 8 shows a sample.

FIGURE 8: S.SCRIBE! AUTOMATIC CAPITALIZATION AND PUNCTUATION

Finally, subtitle splitting rules based on punctuation, linguistic or geometrical features can also be applied, and abbreviations and numerals can be defined in order to reduce the amount of characters needed to represent them on the screen.

The optimal configuration of the duration and splitting features are important to increase the readability of the automatic subtitles.

EVALUATION METHODOLOGY

Once a language has reached the final training data target (see Table 1 for details), the final system for that language has been trained.

Preliminary evaluation of the systems has been carried out using the Word Error Rate (WER) model, a traditional metric for assessing ASR accuracy (see Figure 9).

$$\boxed{WER = \frac{S + D + I}{N}}$$

FIGURE 9: THE WER MODEL

S is the number of substitutions, D is the number of deletions, I corresponds to the number of insertions, and N indicates the number of words in the reference file.

Table 3 shows the WER achieved for the Italian, Basque, Portuguese and Spanish languages, for which the final S.Live! system is available.

Language	***WER***
Basque	15.79 %
Italian	15.80 %
Portuguese	17.72 %
Spanish	14.94 %
Swiss Italian	18.99 %

TABLE 3. S.LIVE! WER PER LANGUAGE

After this preliminary evaluation, a systematic evaluation phase is planned. The methodology to be followed in this systematic evaluation phase consists of two steps. The first one (*Test Cycle 1*) is an automatic evaluation, based on a corpus which has specifically been created for the evaluation, while in the second step (*Test Cycle 2*), human evaluators will manually assess the generated subtitles.

In Test Cycle 1, the S.Scribe! and S.Live! systems will be tested, automatically measuring the transcription quality in terms of Word Error Rate (WER) and Speaker Change Detection (SD). The type of content to be evaluated will be the same used to train the systems (news and debates). 5 hours per language will be evaluated (2.5 hours of pre-recorded content and 2.5 hours of live content). Currently, the reference files are being transcribed and annotated by professionals.

In Test Cycle 2, the S.Scribe, S.Live! and S.Respeak! systems will be evaluated, measuring not only the WER and SD, but also the quality of the automatically generated subtitles. The content to be evaluated in this phase (news and debates) adds up to a total of 5 hours per language (2 or 2.5 hours of pre-recorded content, 2 or 2.5 hours of live content and 1 or 0 hours of respoken content). Respoken content (1 hour) will be evaluated only for the languages for which a respeaking system was built (Basque and Italian). For these languages, 2 hours of pre-recorded and live content will be tested. The reference files are currently being annotated by professionals.

The quality of respoken subtitles will be evaluated using NERstar [26]. This system is based on the NER model, which calculates the accuracy in live subtitling through respeaking (see Figure 10).

$$Accuracy = \frac{N - E - R}{N} \; x \; 100$$

FIGURE 10: THE NER MODEL

N is the number of words in the respoken text, E is the edition errors caused by the strategies of the respeaker and R indicates the recognition errors.

During the manual evaluation, the evaluator will classify the errors as serious (1), standard (0.5) or minor (0.25). Using this model, quality subtitles are expected to reach 98% accuracy.

Even if this is a suitable model to evaluate subtitle quality, it only considers transcription errors. In order to evaluate the quality of the subtitles of the S.Scribe and S.Live! systems, we have extended the model (see Figure 11) to consider other types of features, like the accuracy of splitting, timing and Speaker Change Detection in the automatically created subtitles.

$$Quality = \frac{(N \times P) - \sum_{i=1}^{N}(R + S + T + SP)}{N} \; x \; 100$$

FIGURE 11: THE EXTENDED NER MODEL (ENER)

N is total amount of subtitles, P is the maximum punctuation per subtitle. R is sum of the recognition errors, considering substitutions, deletions and insertions per subtitle (*no error [0], minor error [0.25], standard error [0.5], serious error [1]*), S are splitting errors per subtitle (*no error [0], error [1]*), T are timing errors per subtitle (*no error [0], error [1]*) and SP are Speaker Change Detection errors (*no error [0], error [1]*).

The evaluation of subtitle quality will also be done by professionals during the manual correction. In addition to subtitle quality, the effort of correcting subtitles will also be measured, in order to compare it with the time needed to create subtitles manually from scratch.

The delay of the broadcasted subtitles is another important feature to be considered, since it directly affects the comprehensibility of the content. The time of the initial word of each broadcasted subtitle will be compared with the reference time codes, obtained by a forced alignment between the audios and their related transcriptions.

CONCLUSIONS

This paper described recent advances in Automatic Speech Recognition (ASR), presenting emerging trends and expectations for Live Automatic Subtitling. We focused our attention on SAVAS, a new Speaker Independent ASR technology, and on the three systems developed using this technology: *S.Scribe!*, a batch Speaker Independent Transcription system for pre-recorded subtitling, *S.Live!*, a first-of-a-kind Speaker Independent Transcription system, with real-time performances for online subtitling, and *S.Respeak!*, a collaborative Respeaking System for live and batch production of multilingual subtitles.

An evaluation of the SAVAS technology, based on the Word Error Rate (WER) model, has shown very promising results for Italian, Basque, Portuguese and Spanish. We expect to achieve similar results for other languages, such as English, French and German (plus the Swiss variants of the latter two), which are currently under final training. In particular, S.Respeak! has proven to be sufficiently robust for programs where the acoustic conditions are challenging and for spontaneous speech. Same results are expected to be achieved also for S.Live! and S.Scribe!, which are currently being tested at different broadcasters premises, to subtitle live programs in both assisted and unassisted tasks.

ACKNOWLEDGEMENT

This work is supported by EC within FP7 under Grant 296371 with the acronym SAVAS.

REFERENCES

[1] M. Flanagan, "Human Evaluation of Example-Based MT of Subtitles for DVD", Dublin City University, 2009

[2] C. Aliprandi and F. Verruso, "Human Language Technologies and Real Time Subtitling: State of the Art in Italy and Experience at the last Intersteno Speech Recognition Championships", *Proceedings of the First International Seminar on Real-time Intralingual Subtitling*, InTRAlinea 8, Special Issue on Respeaking, 2006

[3] R. Damper, A. Lambourne and D. Guy, "Speech input as an adjunct to keyboard entry in television subtitling", In Shackel, B., Eds. *Proceedings Human-Computer Interaction---INTERACT'84*, 1985, pp. 203-208

[4] C. Aliprandi et al., "RAI Voice Subtitle: how the lexical approach can improve quality in Speech Recognition systems", *eAccessibility by Voice:*

VOICE Recognition supporting people with disabilities, available at http://www.voiceproject.eu, 2003

[5] A. Lambourne et al., "Speech-Based Real-Time Subtitling Services", *International Journal of Speech Technology,* Vol.7 No.4, 2004, pp. 269-280

[6] http://www.screensystems.tv/products/wincaps-q-live

[7] http://www.fab-online.com/eng/subtitling /production/subtlive.htm

[8] http://www.miranda.com/pdf/datasheets/Swift CreateLive.datasheet.en.pdf

[9] http://www.starfish.tv/captioning-and-subtitling/index.php

[10] http://www.ninsight.fr/w2/index.php?Itemid=72&id=54&lang=en&option=com_content&view=article

[11] http://www-01.ibm.com/software/pervasive/viavoice.html

[12] http://windows.microsoft.com/en-us/windows7/dictate-text-using-speech-recognition

[13] http://www.nuance.com/dragon/index.htm

[14] http://www.screensystems.tv/products/wincaps-subtitling-software

[15] M. Obach, M. Lehr and A. Arruti, "Automatic Speech Recognition for Live TV Subtitling for Hearing Impaired People", *Challenges for Assistive Technology*, AAATE 07, G. Eizmendi et al. (Eds), IOS Press, 2007, pp. 286-291

[16] A. Alvarez and A. del Pozo, "APyCA: Towards the Automatic Subtitling of Television Content in Spanish", *Proceedings of the International Multiconference on Computer Science and Information Technology*, 2010, pp. 567-574

[17] S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News", *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006

[18] J. Loof, Ch. Gollan, S. Hahn, G. Heigold, B. Hoiffmeister, Ch. Plahl, D. Rybach, R. Schluter and H. Ney, "The RWTH 2007 TC-STAR Evaluation System for European English and Spanish", *Interspeech 2007*

[19] https://www.koemei.com

[20] http://www.sail-labs.com

[21] http://www.vecsys-technologies.fr

[22] http://www.verbio.com

[23] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a Broadcast News Speech Recognition System for the European Portuguese Language", *Proceedings of PROPOR'2003*, Faro, Portugal, 2003

[24] https://www.ubertitles.com

[25] http://www.ecaption.eu

[26] http://www.speedchill.com/nerstar/index.php/nerstar-tool.html