# ASSISTED SUBTITLING: A NEW OPPORTUNITY FOR ACCESS SERVICES

C. Aliprandi[1], I. Gallucci[1], N. Piccinini[1], M. Raffaelli[1], A. del Pozo[2],
A. Álvarez[2], R. Cassaca[3], J. Neto[3], C. Mendes[3], M. Viveiros[3]

[1] Synthema, Italy; [2] Vicomtech-IK4, Spain; [3] VoiceInteraction, Portugal

## ABSTRACT

The demand for Access Services has quickly grown over the years, mainly due to National and International laws. This trend is expected to consolidate for subtitling in particular, as almost every broadcaster is nowadays working with digital content: large amounts of existing assets are going to be digitized in the near future. In terms of accessibility, digitalization is a very challenging task that can be turned into a profitable process if addressed with adequate technology.

In this paper we will focus on an emerging technique: Assisted Subtitling. Assisted Subtitling consists in the application of Automatic Speech Recognition (ASR) to generate transcripts of programs and to use the transcripts as the basis for subtitles. 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 and, evaluating its performances, we will present the promising results we have so far achieved.

## 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 Assisted Subtitling, another emerging trend which is raising a lot of expectations. Assisted Subtitling is the application of ASR to automatically generate transcripts of programs, to be used as the basis for subtitles. Despite the difficulties posed by the multitude of different voices and the variety of acoustic conditions, the accuracy achieved is 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 in the use of 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).

Figure 1: Respeaking of sport events

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

The main ASR engines are IBM ViaVoice (9), that nowadays has been discontinued from the market, Microsoft Windows Speech Recognition (10) and Nuance Dragon NaturallySpeaking (11). However, these ASR engines have some limitations. They are

Speaker Dependent, i.e. they have to be adapted to each user by dictation of training sentences. 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 they are not available for many languages, in particular for minor languages.

Less 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 (12) is one of such tools. Again in this context, the 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 (13) have revealed that such type of engine's high Word Error Rate (WER) make it unsuitable for fully automated subtitling. The adaptation of dictation engines to the domain has shown WER reduction and promising results, applicable to the automatic generation of draft transcriptions for post-editing (14).

Although the development of ASR technology has now moved towards transcription, there are still not many solutions available for subtitling in the market. The main reason is the amount of data required to train systems per domain and language. As a result, the commercially available transcription engines are widely scattered across languages and domains. Koemei (15), Vecsys (16) and Verbio (17) are companies offering transcription solutions for some languages and application scenarios. Synthema pioneered SpeechScribe (18), a subtitling solution for prerecorded content, and VoiceInteraction pioneered an online subtitling solution (19), 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 (20) and eCaption (21).

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 SUBTITLING ENGINES AND SYSTEMS

SAVAS is a novel Speaker Independent ASR technology specifically designed for Assisted Subtitling for 10 languages: English, Basque, Spanish, Portuguese, Italian, French, German, Swiss Italian, Swiss French and Swiss German.

In order to deliver quality subtitles, a number of challenging requirements has 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. Unlike most ASR technology available, we have specifically designed the SAVAS dictation and transcription engines for subtitling purposes. The SAVAS ASR engines have been trained for the news domain with large corpora of data: 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 a high quality ASR output, for each language. 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 satisfy challenging tasks like fast response, delay of less than 5 seconds and high accuracy. Also additional ASR features that may support subtitling have to be provided: for example, speaker identification may be useful to identify speaker changes. 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 provide useful operational capabilities required for online subtitling, such as:

- speech classification (speech, music, jingle detection)
- automatic capitalization and punctuation
- speaker change detection
- speaker identification
- subtitle formatting and normalization (splitting and timing)

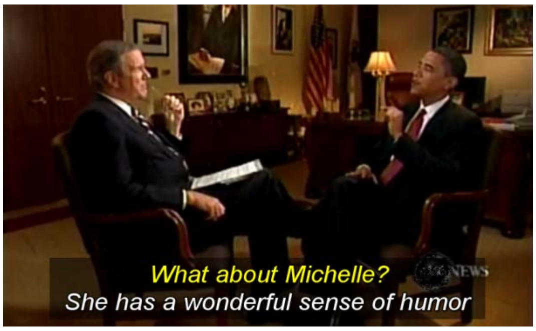


Figure 2: SAVAS speaker change detection

**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 gender and speaker identification. **S.Respeak!** is a 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.

**S.Scribe!**

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

The system receives as input an audio/video file, puts it in a processing list and notifies the user upon completion, so that the subtitle file can be downloaded. The most common and standard subtitling formats, like TTML or SRT, are supported.

S.Scribe! has 2 interfaces:

- HTML Interface: the system is available at a given web address. The user has to log in and can then submit an audio/video file 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 processing.

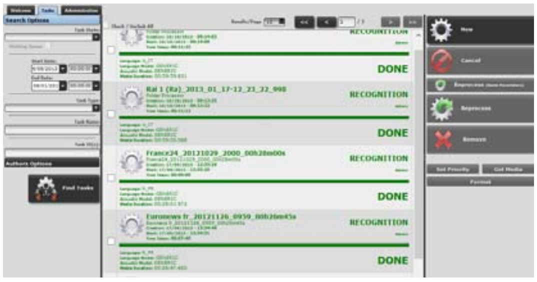

Search Options
Authors Options
Find Tasks
Rai 1 (Ra)_2013_01_17-12_23_22_098
France24_20121029_2000_00h28m00s
Euronews fr_20121126_0959_00h26m45s
RECOGNITION
DONE
New
Cancel
Reprocess
Remove
Format

Figure 3: Screenshot of S.Scribe!

### S.Live!

S.Live! is an online system, which receives an audio signal at the input audio board and produces subtitles at the output. Its output is directly connected to commercial software in charge of communicating with Teletext inserters. This system has a web interface for manual administration. The operation itself is automatic, with the system starting the subtitling process at the predefined time. The main operation is based on a scheduling process, where the programs to be automatically subtitled may be defined. This is done over the EPG of the corresponding channel.

The S.Live! TV Programs Scheduler allows the creation of a subtitling process in a specific time and with a specific duration. According to the defined schedule, the system will run automatically at the next occurrence. Figure 4 shows a screenshot of the S.Live! TV Programs Scheduler.

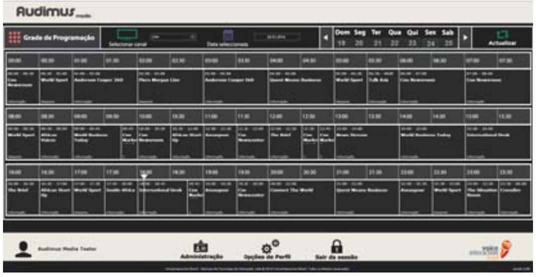
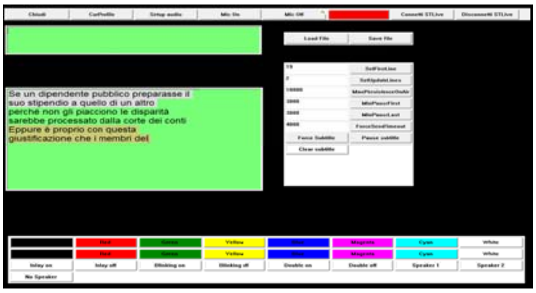


Figure 4: Screenshot of S.Live!

### 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.



Se un dipendente pubblico preparasse il
suo stipendio a quello di un altro
perché non gli piacciono le disparità
sarebbe processato dalla corte dei conti
Eppure è proprio con questa
giustificazione che i membri del

Figure 5: S.Respeak!

Besides traditional respeaking features, like correct on-screen subtitles persistence, management of subtitles formatting (colours, capitalization styles, ...), S.Respeak allows the use of domain-specific phrases and fast post-editing, leveraging on the output of the SAVAS engines. It also makes respeaking a more collaborative process, splitting the cognitive load among respeaker and editor, optionally coordinated by a supervisor.

### SAVAS ASR SUBTITLING

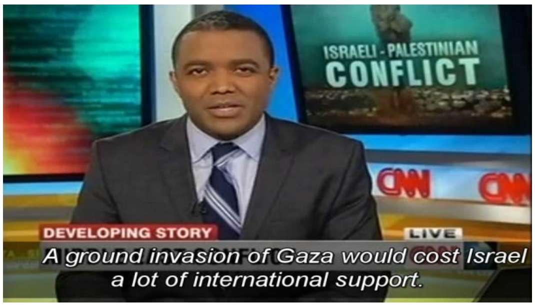
The output of the three SAVAS systems complies with the main subtitle layout, duration, punctuation and text editing constraints.

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

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 when names of entities (such as persons, locations or companies) are detected. Figure 6 shows a sample.



A ground invasion of Gaza would cost Israel a lot of international support.

Figure 6: online capitalization

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 is important to increase readability.

## EVALUATION METHODOLOGY AND RESULTS

A first evaluation of the systems has been carried out using the WER model, a traditional method for evaluating ASR accuracy (see Figure 7). WERs of the S.Live! system are shown in Figure 8, together with the WER evolution on training data for the SAVAS languages.

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

Figure 7: the WER model

WERs can be considered to be very good for Basque, Spanish, and Italian. The higher values for French and German can be attributed to a number of factors, one of which being the fact that the broadcast news programs in these languages have a higher amount of spontaneous speech than normal. The Swiss variations also present worse performances, mainly due to the higher amount of foreign speech and to the reduced amount of training data employed. Despite not all languages have the same level of WER, they all exhibit the same exponential decay behavior with the increase of training material.

| Language | WER | Language | WER |
|---|---|---|---|
| Basque | 15.79% | Spanish | 14.94% |
| Italian | 15.08% | Swiss Italian | 19.00% |
| French | 18.59% | Swiss French | 25.83% |
| German | 21.11% | Swiss German | 20.68% |
| Portuguese | 31.66% | | |

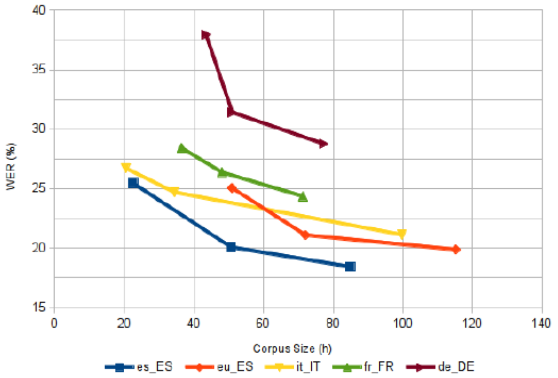


Figure 8: WERs and WERs vs. training data

Beyond ASR evaluation, we established a novel methodology for evaluating the quality and usefulness of the SAVAS systems for Assisted Subtitling. Leveraging on the NER model and the NERstar tool (22), we devised the **extended NER** (**eNER**) model. As a matter of fact, even if NER is a suitable model to evaluate subtitle quality, it focuses on respeaking and only considers recognition errors when applied to transcription. In order to evaluate the quality of the subtitles of the S.Scribe! and S.Live! systems, we have then extended NER to also consider other relevant types of features for Assisted Subtitling, namely: Splitting, Timing and Speaker Change Detection.

Figure 9 presents the eNER model: N is the total amount of subtitles; P is the number of features to be evaluated; R is the sum of the recognition errors, considering substitutions, deletions and insertions (no error [0], minor error [0.25], standard error [0.5], serious error [1]), S are the splitting errors (no error [0], error [1]), T are the timing errors (no error [0], error [1]) and SC are the Speaker Change Detection errors (no error [0], error [1]).

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

Figure 9: The eNER model

Figure 10 shows the overall quality of the S.Live! and S.Scribe! systems. As it can be appreciated, eNER values are around 75% on average for Basque, Spanish and Italian, without significant differences between the two systems. Although these values are far from the 98% NER values considered to correspond to top quality subtitles, eNER results are expected to reach relatively lower values because the extended formula considers a higher amount of quality features.

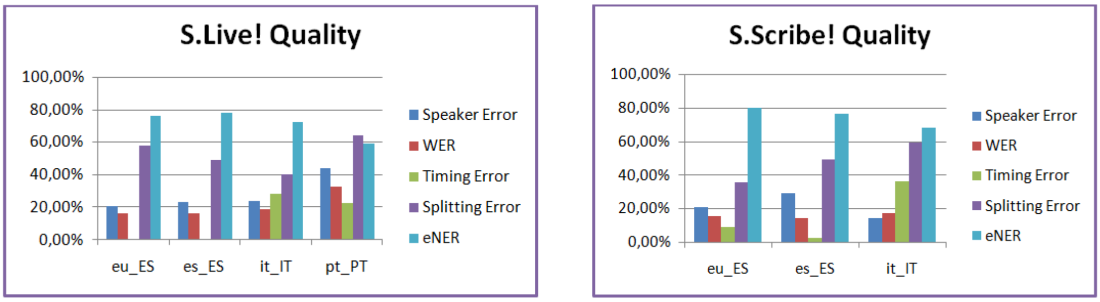


Figure 10: Final WER and WER vs. training data

If we look into the specific weight of each of the considered quality features on the overall eNER metric, we can see that splitting errors are the most frequent ones, followed by speaker change, WER and timing errors. This analysis gave us an useful feedback to improve the systems.

Finally, to assess the usefulness of the SAVAS systems for assisted subtitling, we evaluated the productivity gain. The aim was to test whether post-editing automatic subtitles generated by S.Scribe! is faster than manually creating them from scratch. Subtitling professionals were asked to post-edit automatic subtitles and to create them from scratch, using their usual quality standards. Figure 11 shows the productivity gains of the S.Scribe! system. All but one subtitler have managed to increase their productivity post-editing automatic pre-recorded subtitles when compared to creating them from scratch. Gains are highly subtitler dependent, ranging between 33% to 2% across post-editors. The S.Scribe! output has also been compared against the post-editing stenotype output. In this case, the post-editing stenotype output has achieved a higher productivity gain (22%), then the post-editing Scribe! output. This is not completely surprising, since stenotypists generate less text editing errors than state-of-the-art SAVAS technology, particularly in what capitalization and punctuation features are concerned. Consequently, the time devoted to correcting such kind of errors is reduced.

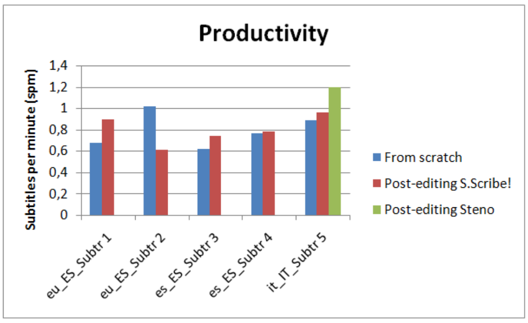


Figure 11: Productivity gain

Concluding, we consider that these results are good for Assisted Subtitling: the productivity gains achieved are very promising, suggesting that post-editing automatic subtitles is faster than creating them from scratch.

## CONCLUSIONS

This paper described recent advances in ASR, presenting emerging trends and new opportunities for Assisted 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.

We presented an overview of the tasks carried out to evaluate the performances of the SAVAS systems, and we introduced eNER, a novel method for evaluating their usefulness for Assisted Subtitling. eNER, unlikely other evaluation models, takes into consideration subtitling-specific features like Splitting, Timing and Speaker Change Detection.

The evaluation based on the WER model has shown very good results for Basque, Spanish and Italian. The evaluation based on the eNER model has shown good results for Assisted Subtitling: eNER values are around 75% on average for Basque, Spanish and Italian without significant differences between the S.Live! and the S.Scribe! systems. The productivity gains of the two systems, when compared to unassisted subtitling, are ranging between 33% to 2% across professional subtitlers, a very promising result suggesting that post-editing automatic subtitles is faster than creating them from scratch.

Concluding, the main advantage of Assisted Subtitling compared to traditional subtitling techniques is that it can actually produce similar results with less human effort, which helps reduce subtitling costs.

## 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-01.ibm.com/software/pervasive/viavoice.html

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

11.http://www.nuance.com/dragon/index.htm

12.http://www.screensystems.tv/products/wincaps-subtitling-software

13.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

14.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

15.https://www.koemei.com

16.http://www.vecsys-technologies.fr

17.http://www.verbio.com

18.http://www.synthema.it/index.php/en/Products/speechscribe/SpeechScribe.html

19.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

20.https://www.ubertitles.com

21.http://www.ecaption.eu

22.http://www.speedchill.com/nerstar/index.php/ nerstar-tool.html

## ACKNOWLEDGEMENTS

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