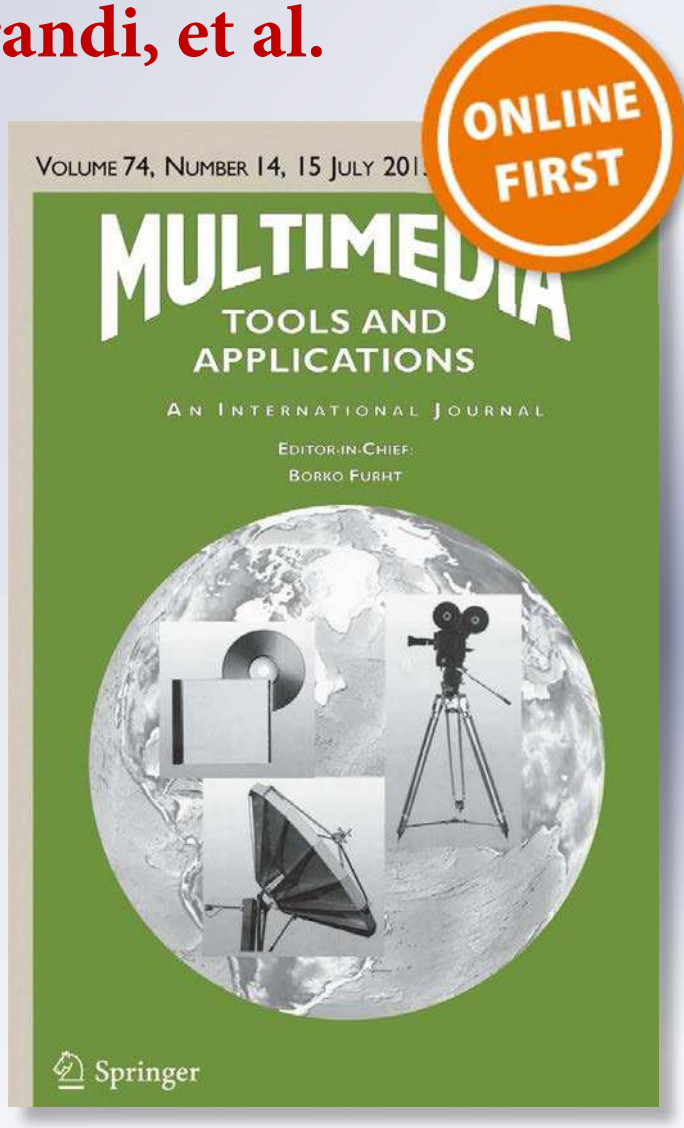
Automating live and batch subtitling of multimedia contents for several European languages

Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, Tiago Luís, Sérgio Paulo, Nicola Piccinini, Haritz Arzelus, João Neto, Carlo Aliprandi, et al.

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-015-2794-z



Springer

Multimed Tools Appl
DOI 10.1007/s11042-015-2794-z

Automating live and batch subtitling of multimedia contents for several European languages

Aitor Álvarez[1] · Carlos Mendes[2] · Matteo Raffaelli[3] · Tiago Luís[2] · Sérgio Paulo[2] · Nicola Piccinini[3] · Haritz Arzelus[1] · João Neto[2] · Carlo Aliprandi[3] · Arantza del Pozo[1]

Received: 22 December 2014 / Revised: 22 June 2015 / Accepted: 29 June 2015

Abstract The subtitling demand of multimedia content has grown quickly over the last years, especially after the adoption of the new European audiovisual legislation, which forces to make multimedia content accessible to all. As a result, TV channels have been moved to produce subtitles for a high percentage of their broadcast content. Consequently, the market has been seeking subtitling alternatives more productive than the traditional manual process. The large effort dedicated by the research community to the development of Large Vocabulary Continuous Speech Recognition (LVCSR) over the last decade has resulted in significant improvements on multimedia transcription, becoming the most powerful technology for automatic intralingual subtitling. This article contains a detailed description of the live and batch automatic subtitling applications developed by the SAVAS consortium for several European languages based on proprietary LVCSR technology specifically tailored to the subtitling needs, together with results of their quality evaluation.

Keywords Multimedia communication · Multimedia systems · Automatic speech recognition · Automatic subtitling · Subtitling quality · Access services

✉ Aitor Álvarez
aalvarez@vicomtech.org

[1] Department of Human Speech and Language Technologies, Vicomtech-IK4 Foundation, San Sebastian-Donostia, Spain

[2] VoiceInteraction-Speech Processing Technologies, SA, Lisbon, Portugal

[3] Synthema-Language and Semantic Technologies, Pisa, Italy

Published online: 11 July 2015

1 Introduction

The subtitling demand of multimedia content has grown quickly over the last years, especially after the adoption of the new European audiovisual legislation (Article 7 of the Audiovisual Media Services Directive). This law regulates the right of persons with disabilities and elderly people to participate and be integrated in the social and cultural life of the Community, through accessible multimedia services including aspects such as sign-language, subtitling, audiodescription and easily understandable menu navigation.

As a result of this new legal framework, public and private TV channels have been moved to produce subtitles for a high percentage of their broadcast content. However, the subtitling process is traditionally 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 multimedia material [14]. Hence, broadcasters and subtitling companies are looking for solutions that can help them cope with the increasing subtitling volumes and demand.

The large effort in research and development of Large Vocabulary Continuous Speech Recognition (LVCSR) over the last decade has resulted in significant improvements on multimedia data transcription, retrieval and indexation [16, 23, 43], making it the most powerful technology available to increase productivity in several automated intralingual subtitling tasks. In the last few years, respeaking - a technique in which a professional listens to the source audio and dictates it so that his/her voice input can be processed by a speech recognition engine which transcribes it, thus producing subtitles - has consolidated as the most widely adopted live subtitling technique. Another trend in use today is the application of LVCSR to automatically generate transcripts from the source audio as the basis for subtitles. The main advantage of this method compared to respeaking is that it can actually produce similar results in bounded domains without the need of a respeaker, which reduces costs.

In order to comply with the new legal requirements, broadcasters started focusing their increased subtitling effort on quantity, considering quality a secondary issue. However, an increasing demand to improve the quality of automatic subtitles has arised recently. The quality of subtitles involves several features linked to subtitle layout, duration and text editing. Layout parameters include: the position of subtitles on screen; the number of lines and the amount of characters contained in each line; the typeface, distribution and alignment of the text; the front and background colors; speaker colors; and transmission modes, i.e. blocks or scrolling/word-by-word. Duration features involve delay in live subtitling and the persistence of subtitles on screen. Finally, text editing parameters are related to capitalization and punctuation issues, segmentation or the use of acronyms, apostrophes and numerals.

This article contains a detailed description of the automatic live and batch subtitling applications developed by the SAVAS consortium[1] for several European languages based on proprietary LVCSR technology tailored to the specific needs of the subtitling industry, together with results of their quality evaluation. Applications have been developed for Portuguese, Spanish, Basque, Italian, French, German and the Swiss variants of the latter three, and trained and tested on several domains such as broadcast news, sports, interviews and debates. Their performance has been evaluated against a variety of metrics, including

[1] http://fp7-savas.eu

both standard LVCSR and subtitle quality metrics. In Section 2, an overview of the state-of-the-art of LVCSR and the existing assisted subtitling applications is presented. Section 3 details the SAVAS technology and the developed live and batch subtitling applications. The methodology followed to evaluate them is then described in Section 4. Finally, Section 5 presents the evaluation results and the main conclusions are summarized in Section 6.

2 Related work in automatic speech recognition and assisted subtitling applications

LVCSR technology is employed to transcribe speech into text for further linguistic processing. Despite progress in the last decade, LVCSR is still highly task- and domain-dependent due to its statistical nature. In terms of accuracy, LVCSR system performance varies with the task [3]: clean read speech transcription achieves better performance than TV and radio news broadcasts, telephone conversations, lectures or plenary sessions of the European Parliament. Although comparable performance has been achieved for several languages, English is still the most developed one today.

LVCSR technology has been exploited commercially, mainly for dictation and command-based interaction applications in specific domains, like the HealthCare or the Parliamentary domain [20, 30, 32]. The main LVCSR engines of this type available are IBM ViaVoice [20], now discontinued from the market, Microsoft Windows Speech Recognition [30] and Nuance Dragon Naturally Speaking [32]. Many subtitling tools currently employed by the industry (e.g. WINCAPS Q-Live [35], WINCAPS Qu4ntum [36], FAB Subtitler Live Edition [12], Grass Valley captioning and subtitling solution [19], Starfish Isis [37]) support the Nuance Dragon Naturally Speaking dictation engine for respeaking purposes. However, there are no tools on the market that allow generating automatic intralingual subtitles from the source audio without respeaking.

This has been limited by the unsuitability of the available dictation engines for audio transcription [33] and by the absence of more sophisticated LVCSR technology for transcription of multimedia contents. Dictation engines have several limitations. First, they are speaker dependent; that is, they have to be adapted to each user. Second, they do not perform well on multimedia material containing complex acoustic conditions (e.g. background music or noise) or spontaneous speech, because they have been designed for dictation purposes. Finally, they have only been developed for languages with a high market potential (e.g. English and Spanish) and are not available for many other languages. On the other hand, training high-quality LVCSR transcription engines requires huge amounts of audio and text per language and specific domain. Several studies (see [15] and [24]) state that at least 100 hours of annotated and transcribed audio are necessary to adequately train the acoustic models of such kind of LVCSR engines. Regarding language modelling, [29] have estimated that ideally one billion words of texts are required.

Recently, few internet services offering the automatic generation of time-aligned subtitles from a source audio and its transcript have arisen. Ubertitles [39], eCaption [11] or SyncWords [38] offer such kind of service in different languages, based on proprietary audio and text alignment technology. However, they are not capable of producing subtitles automatically from the audio source and carry out the transcription step manually.

On the other hand, Koemei [21], SailLabs [34], Vecsys [40] or Verbio [41] are companies that commercialise automated transcription solutions for varying pools of languages and application scenarios such as lectures, open source intelligence or media, but do not produce subtitles. Since recently, Google [17] supports the automatic generation of time-aligned

draft transcriptions and subtitles from the videos uploaded to Youtube - serving multiple languages and allowing their automatic translation through Google Translate [18]. Nevertheless, for the moment Youtube transcriptions do not include punctuation and capitalization nor follow the standard professional subtitling practices (see Section 4).

In the more specific subtitling field, VoiceInteraction pioneered a transcription solution [27] capable of generating subtitles for Portuguese broadcast news, which was adopted and is currently in daily use by the public Portuguese broadcaster RTP. The SAVAS automatic live and batch subtitling applications described in the next section have followed this approach, being specifically designed for subtitling purposes and taking into account the most relevant features of quality subtitles.

3 SAVAS technology, development and applications

In this article, full systems for automatic subtitling and transcription of audiovisual contents are presented. The systems were trained over Broadcast News, sports and interview/debate domains, which contain a great variety of speakers (journalists and citizens), topics (economy, politics, sports), acoustic conditions (studio and outside news), and types of speech (planned and spontaneous). These aspects thus turned this type of content into an optimal resource to train the more robust and flexible LVCSR systems as possible.

The systems were developed for several European languages, including Portuguese, Spanish, Basque, Italian, French, German and the Swiss variants of the latter three. All the systems were initially trained over Broadcast News contents. In the case of Portuguese, this system was then extended and adapted to a more complex domain, i.e. interview and debate domain, containing repetitions, hesitations, disfluences, unfinished sentences, overlapping speech etc. All of these issues make the recognition process more difficult. Moreover, the Basque and Italian dictation systems were also adapted to the Sports domain for Respeaking purposes.

In addition, three type of applications were built over the systems described above for several subtitling and transcription purposes. The first application is a batch Speaker Independent Transcription and Subtitling application (S.Scribe!), capable of automatically transcribe pre-recorded audio and video files into time-aligned enriched subtitles. The second application corresponds to an Online Subtitling System (S.Live!) and it is able to automatically transcribing live audio into configurable and well-formatted subtitles. The final application involves a Respeaking engine (S.Respeak!) for dictation, which can be easily integrated into any commercial subtitling solution and capable of producing subtitles with an acceptable delay. Table 1 summarizes the LVCSR based SAVAS systems and applications that we developed per language and domain.

As it can be seen in Table 1, the three applications were developed for all the languages involved. Unlike the rest of the languages, Portuguese S.Scribe! and S.Live! applications were developed for both Broadcast News and Interview/debate domains. It implied an adaptation process of the base Portuguese models trained on Broadcast News contents to a more specific domain, in which more complicated issues related to spontaneous speech had to be considered.

Regarding S.Respeak! dictation system, all the languages were covered by engines trained on the news domain. In addition, these engines were also adapted to the Sports domain in the case of Basque and Italian languages. The S.Respeak! application is composed by an engine to produce transcriptions for dictation and respeaking purposes. It is not a complete solution for respeaking, since it has to be integrated with a subtitling software to

Table 1 LVCSR based SAVAS systems and applications per language

Language	S.Scribe!	S.Live!	S.Respeak!
Portuguese	Broadcast News	Broadcast News	News
	Interview/debate	Interview/debate	–
Spanish	Broadcast News	Broadcast News	News
Basque	Broadcast News	Broadcast News	Sports and News
Italian	Broadcast News	Broadcast News	Sports and News
Swiss Italian	Broadcast News	Broadcast News	News
French	Broadcast News	Broadcast News	News
Swiss French	Broadcast News	Broadcast News	News
German	Broadcast News	Broadcast News	News
Swiss German	Broadcast News	Broadcast News	News

create subtitles. Synthema's Voice Subtitle [4] and SysMedia's SpeakTitle [22] are examples of these subtitling solutions.

In the next subsections, the data resources we compiled to train the systems, the development of the main technological components of the systems, in addition to the type of applications, are described in more detail.

3.1 Compiled data resources

The development of robust LVCSR systems for automatic transcription and subtitling in the audiovisual domain requires considerably large audio and text corpora for the acoustic and language modeling. Based on previous experience [26, 29], we estimated that the development of good performance transcription systems would ideally require at least 200 hours of audio and 1000 million words of text. The same data could also be exploited to develop dictation systems. Besides, the adaptation of an already existing transcription or dictation system to a new domain was estimated to be achievable with 20 hours of audio with at least 500k words.

With regard to audio data for acoustic modeling, in this work the most of the data were gathered from programs produced by broadcasters. The TV programs were then manually annotated using the Transcriber tool,[2] following internal transcription conventions. Since manually annotating 200 hours of audio data is a highly costly task, we followed an incremental automation approach. The first 50 hours per language were annotated manually from scratch, followed by several stages of manual annotation with draft automatic transcriptions. At each stage, new acoustic models were trained to decrease the amount of errors produced by automatic recognition and thus to speed up the manual transcription process.

On the other hand, the text sources for language modeling and vocabulary creation were a mix of autocue scripts and subtitles provided by the broadcasters and subtitling companies, plus newswire and sports text crawled from the Internet. In addition, the transcriptions of the collected audio content were also used as text data. Table 2 shows the final amounts of audio and text corpora collected for each language.

As it can be seen from the Table 2, most of the targeted amounts were almost reached, except for Basque text corpora in the broadcast news and Portuguese in the interview/debate

[2] http://trans.sourceforge.net/en/presentation.php

Table 2 Collected audio and text corpora per language and domain

Language	Variant	Domain	Audio	Text
Portuguese	European	Broadcast News	113H	1012M
		Interview/debate (adaptation)	20H	200K
Spanish	European	Broadcast News	200H	1009M
Basque	Standard Basque	Broadcast News	200H	329M
		Sports (adaptation)	20H	500K
Italian	Italian	Broadcast News	162H	950M
		Sports (adaptation)	–	500K
	Swiss Italian	Broadcast News	50H	100M
French	European	Broadcast News	150H	932M
	Swiss French	Broadcast News	50H	100M
German	European	Broadcast News	151H	808M
	Swiss German	Broadcast News	51H	100M

domain for adaptation purposes. As a minority language, the availability of Basque text corpora in the news domain was limited. With Portuguese, the difficulty was to find text resources containing the type of spoken information common in the interview/debate domain.

Although an exact 1000 million word text corpus was not achieved for all languages, this cannot be considered as critical, since the purpose of having such large text corpora was to use pruning techniques in the final language models in order to reduce noise and texts particularly out of domain. With regard to Italian, French, German and their Swiss variants, the originally targeted amounts were distributed according to previous experience on dialect adaption [26].

All the audio and text data collected from broadcasters and subtitling companies were shared through the META-SHARE[3] repository, once the corresponding permissions were granted by to the contents owners. In addition to the raw audio and text, three transcribed audio test sets were also shared per language. These test sets will allow other LVCSR technology developers to compare the performance of their systems with that of the SAVAS engines. The commercial license established for the sharable resources is the META-SHARE Commercial-NoReDistribution-For-a-Fee (C-NoReD-FF) license. On the other hand, the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license was established for research purposes. All the data were shared through the SAVAS META-SHARE specific repository, which has become one of the biggest available audio and data sources exploitable for LVCSR development.

Further information about the collected data resources can be found in the work presented in [9].

3.2 Development of the systems

Automatic Subtitling and Transcription of audiovisual contents is a highly complex task that requires several modules of functionalities to provide useful operational capabilities. Although the LVCSR is the most important component, there are other technologies

[3] http://www.meta-net.eu/meta-share

involved in this process. The SAVAS systems can thus be represented as a pipeline of processing blocks, which represent the different components, as shown in Fig. 1.

In global terms, the Audio Pre-Processing block receives the program audio, discriminates between Speech and Non-speech and sends the audio to the Large Vocabulary Continuous Speech Recognition in case of speech. Additionally, it gives information on speaker clustering, speaker gender and speaker identification in case of relevant speakers. The Large Vocabulary Continuous Speech Recognition block transcribes the audio input stream according to a vocabulary and a language model. This component is the most important and critical one since its performance will be reflected directly in the final result. The Output Normalization block converts sequences of words representing digits, connected digits, and numerals into numbers. It also capitalizes the names and introduces the punctuation marks. Finally, the Subtitling Generation block creates the subtitles according to each broadcaster subtitling rules and specifications.

The overall system works in a pipeline and asynchronous operation mode, where each block is responsible for fulfilling its own task and send the results to the next block. In the following subsections, we give a more detailed description of the system components.

3.2.1 Audio pre-processing (APP)

The full operation of the APP block is intended to provide a complete description of the input audio, including speech/non-speech segmentation (SNS), gender classification (male/female), background classification (clean, noise, music) and speaker diarization, which performs speaker clustering and speaker identification (in case of relevant speakers as pivots).

The technology integrated for acoustic change detection and classification, background conditions classification and gender classification is described in the work presented in [28]. However, new algorithms were developed in this work for a more efficient Speaker Clustering and Speaker Identification, described in the following two subsections.

Speaker clustering Our previous work [45] in speaker clustering based on the Bayesian Information Criteria (BIC) obtained low performance mainly caused by the audio segmentation component (SNS), which sometimes produced small Speech segments. The new

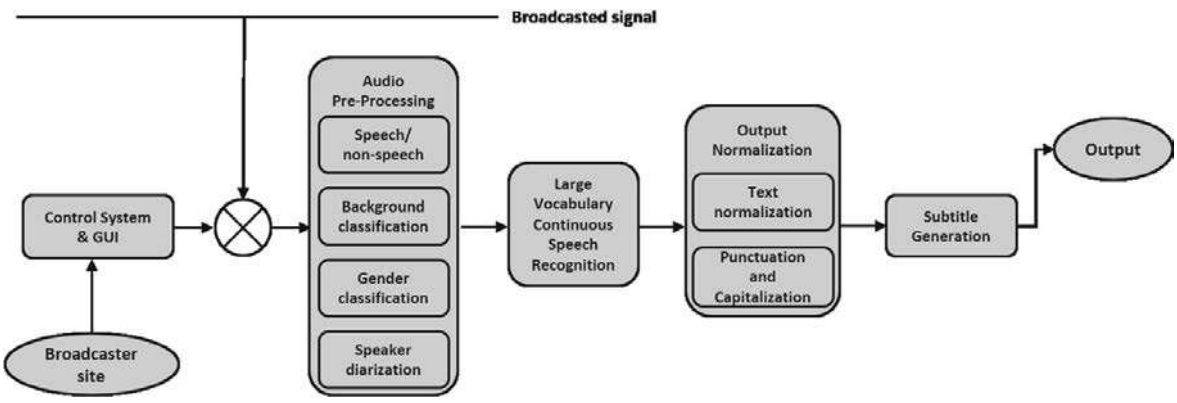


Fig. 1 Pipeline of the SAVAS subtitling systems

Table 3 Improvement on DER with the new algorithm

Language	Previous Algorithm	Improved Algorithm
Portuguese	29.06 %	24.04 %
Basque	41.30 %	32.60 %
Italian	30.07 %	16.84 %

algorithm adds a BIC based speaker turn detection before the BIC clustering to overcome this problem.

The algorithm starts by detecting speaker turns using BIC, where change points are detected through generalized likelihood ratio (GLR), using Gaussians with full covariance matrices. SNS segments are also modeled with Full Gaussian and compared with the current speaker Full Gaussian.

$$BIC_{i,j} = \frac{n_i + n_j}{2} log\,|\Sigma| - \frac{n_i}{2} log\,|\Sigma_i| - \frac{n_j}{2} log\,\left|\Sigma_j\right| - \lambda P \tag{1}$$

Equation (1) gives the BIC score of the similarity of two segments/clusters, where $|\Sigma_i|$ and $\left|\Sigma_j\right|$ are the determinants of the Gaussian associated to segment/cluster i and j respectively, $|\Sigma|$ is the determinant of the Gaussian associated to segment/cluster i plus j and P is a penalty factor. If the BIC score is lower than 0, then the two segments/clusters are merged together as one, otherwise a new speaker is detected and new statistical information is gathered for the new speaker Full Gaussian.

The hierarchical clustering algorithm is an adaptation of the algorithm described in [25]. In our hierarchical clustering algorithm, the current speaker cluster, provided by turn detection and modeled with Full Gaussian, is compared with the clusters obtained so far. This comparison differs from the implementation in [25], where all clusters are compared. This difference allows on-line processing of the clusters.

Table 3 presents results obtained for 3 languages, where significant reduction in Diarization Error Rate (DER) can be observed comparing the old BIC algorithm with the new one. DER metric defines the ratio of incorrectly detected speaker time to total speaker time, as it is described in [13].

Speaker identification Total Variability has emerged as one of the most powerful approaches to the problem of speaker verification. This technique jointly models speaker and channel variabilities as a single low rank space. Our Speaker Identification component uses the low-dimensionality total variability factors, known as identity-vectors (i-vectors), produced by the Total Variability technique to model known speaker identities. The i-vectors are extracted, as depicted in Fig. 2, using an Universal Background Model (UBM) and

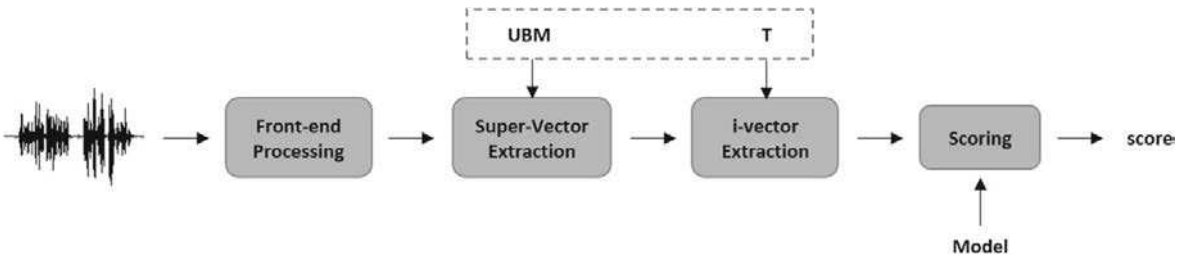


Fig. 2 Total Variability i-vector extraction

the Total Variability matrix (T). This technology is based on previous work conducted on Language Identification [1].

The Speaker Identification component works after the Speaker Clustering, once all the speakers have been grouped into clusters. Since this component works on-line, every time an unseen speaker starts talking, the component is not able to know the speaker identity immediately. To overcome this problem, we produce a first estimate for its identity (if it is a known speaker) after 10 seconds of speech and a final identity estimation after 30 seconds. Since the zero- and first-order sufficient statistics (and the respective i-vectors) from Total Variability are associated with the cluster, the speaker information is immediately available whenever a cluster with a known identity appears.

For this work, we trained Total Variability models with the same acoustic features and parameters used in [1], but modeling speakers instead of languages. Regarding the scoring, we also use Linear Logistic Regression (LLR), but only using the information from the i-vectors.

3.2.2 Large vocabulary continuous speech recognition (LVCSR)

The LVCSR engine named Audimus [31] is based on a hybrid speech recognition structure combining the temporal modeling capabilities of Hidden Markov Models (HMMs), with the pattern discriminative classification capabilities of Multilayer Perceptrons (MLPs). The processing stages are represented in Fig. 3.

The system uses and combines phone probabilities generated by several MLPs trained on distinct feature sets, resulting from different feature extraction processes in order to better model the acoustic diversity. This is relevant in the recognition of TV programs and multimedia contents, with a high diversity of speakers and environments. These probabilities are taken at the output of each MLP classifier and combined using an appropriate algorithm.

The decoder is based on the Weighted Finite-State Transducer (WFST) approach [7]. In Audimus systems, the search space is a large WFST, which results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model. 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

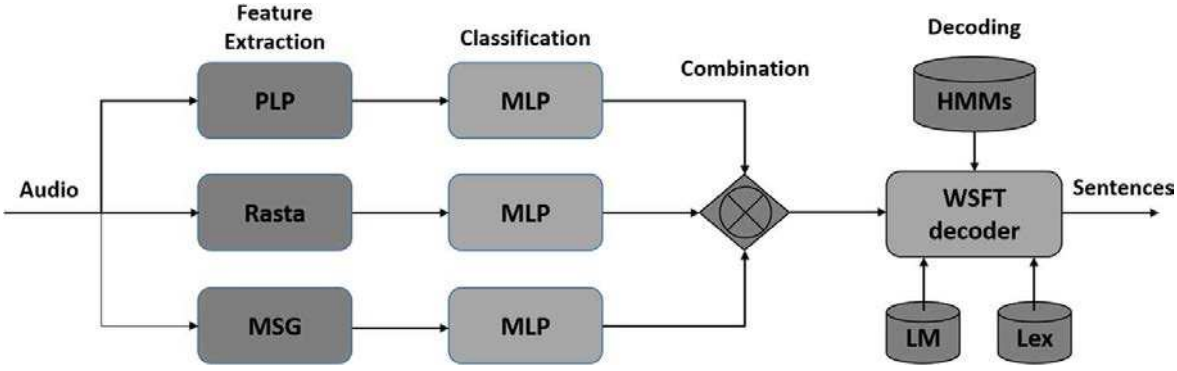


Fig. 3 Audimus processing pipeline

entropy classifier, whose output represents the probability of each word being correct [7]. Confidence measures for the recognized text are necessary to filter the output text in the subtitling composition stage.

Acoustic modelling The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches. Different feature extraction and classification branches effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, commonly present in multimedia content. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the third uses 28 MSG (Modulation SpectroGram) coefficients for each audio frame. Each MLP classifier incorporates local acoustic temporal context through an input window of 13 frames (the MSG branch uses 15 frames) and two fully connected non-linear hidden layers. The number of units on each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characteristic phonetic set of each language.

The acoustic model training was carried out with the same techniques used in [26]. For the first stages of the annotation process, monophone based systems were trained for each language. The acoustic models were trained in a series of stages to aid the process of manual annotation. Figure 4 presents the Word Error Rate (WER) against the amount of annotated material for the initial stages of the monophone based systems. As it can be observed in Fig. 4, not all languages have the same level of WER, but they all exhibit the same exponential decay behavior with the amount of training material.

Once a language reached all the training material, diphone systems were built using alignments and labels generated by the monophone systems. Since there is much less data for the Swiss variations, the Italian, French, German and their Swiss variations were trained with the combined data of both counter parts to enrich the models. Specifically, Italian was trained in conjunction with Swiss Italian, and the same for the other pair of languages.

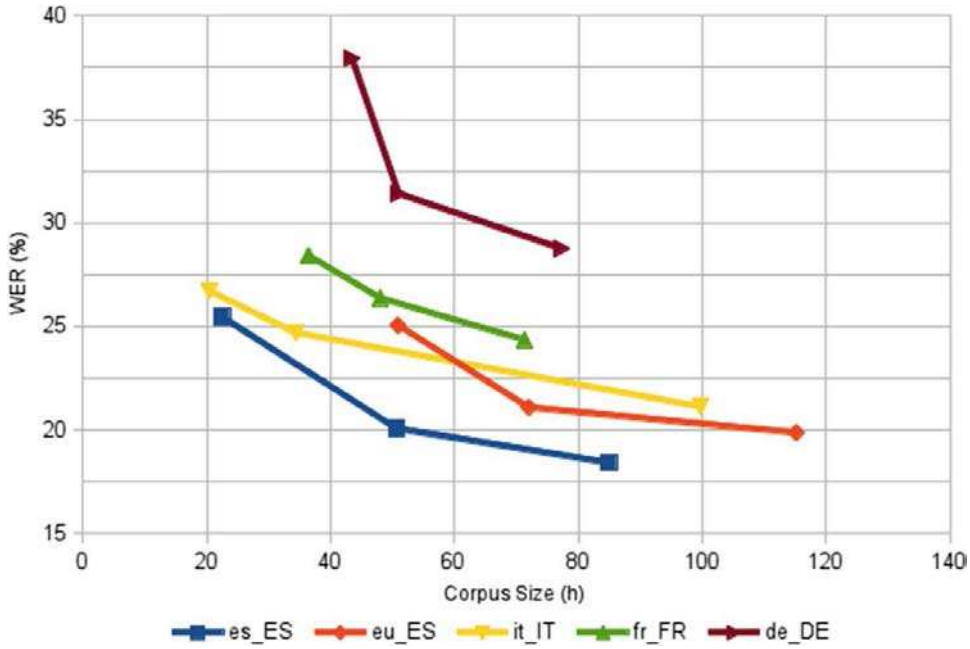


Fig. 4 WER vs annotation material

Language modelling The techniques used in the construction of the language models are the same as in [26]. The language models were created using statistical backed-off N-gram models. N-grams are probabilistic models which exploit the ordering of words predicting the next word from the previous N-1 words. In a bit of terminological ambiguity, the term N-gram is usually used to refer to either the word sequence or the predictive model.

In SAVAS systems, N-gram models resulted from the interpolation of several specific language models. The number of these specific language models varies across languages and depends on the availability of text data from different sources. For instance, in the case of the Spanish system, the first specific language model (LM) was a 4-gram LM trained on data from several online ES newspapers texts ranging from 2000 to 2014 totalizing 900M words. The second one was a 3-gram LM estimated on the BN training transcriptions which has 1.2M words. The third model was a 4-gram LM estimated on autocue scripts toting up 200M words. These three language models were then linearly interpolated with optimization of the weights on the automatic transcription texts. The final interpolated LM for Spanish was a 4-gram LM, with Kneser-Ney modified smoothing, with 100k words (1-gram), 8.0M 2-gram, 14.9M 3-gram, 9.8M 4-gram, with a perplexity value of 88.

In Table 4, the obtained number of N-grams (N-gram counts) and perplexity (PPL) values are presented for each model and language. Perplexity is the most common evaluation metric for N-gram language models and it is a function of the probability that the language model assigns to a data set. The smaller perplexity is, the better is the model.

Pronunciation lexicons In TV and multimedia contents large variety of topics are discussed over time. Additionally, in order to guarantee the performance of the LVCSR systems, the vocabularies have to be limited to a fixed amount of words, typically to the most frequent ones. Both constrains imply that Out-Of-Vocabulary (OOV) words cannot be avoided. The regular approach is to use a vocabulary containing at least 60K words for English and even more for other inflectional languages.

With the aim of having a reasonable coverage of the languages, while maintaining performance, we used the 100K most common words for the vocabularies. Even if agglutinative languages, like Basque and German, may require larger vocabularies, increasing the language coverage beyond the 100K word vocabulary can dramatically degrade performance, mainly in on-line systems.

For the pronunciation lexicons we used lexica developed for each language. For unavailable words we used grapheme-to-phoneme systems to generate the corresponding

Table 4 Language models n-gram counts and perplexities (PPL)

Language	2-gram	3-gram	4-gram	PPL
Portuguese	8.1 M	11.2 M	7.3 M	144
Spanish	8.0 M	14.9 M	9.8 M	88
Basque	11.5 M	13.3 M	5.3 M	261
Italian	10.4 M	14.1 M	9.2 M	151
French	9.2 M	12.1 M	6.8 M	88
German	10.2 M	12.9 M	7.5 M	174
Swiss Italian	10.4 M	14.1 M	9.2 M	198
Swiss French	9.2 M	12.1 M	6.8 M	130
Swiss German	10.2 M	12.9 M	7.5 M	214

pronunciations. Table 5 presents the number of pronunciations obtained for each language with the 100k vocabularies.

3.2.3 Output normalization

This component involves a set of actions to convert the raw output of the LVCSR into normalized text suitable for subtitles. These actions aim to reduce the dimensionality of the text and to improve the readability of subtitles.

The normalization operation is two-fold. First, numbers, numerals, dates and amounts (e.g. money and percentage) are converted to their digit representation through rule-based functions developed for each language. Besides, this block also punctuates the text and capitalizes the acronyms and proper names. The most common acronyms were included in the vocabularies, and techniques based on maximum entropy models are used for the automatic Punctuation and Capitalization of named entities [6]. These techniques are based on the information provided by the preceding APP and LVCSR components, such as pauses, speaker changes, Part-Of-Speech (POS) information of the present, previous and following words, in addition to the confidence measure associated to each word.

3.2.4 Subtitle generation

After the normalization operation, words are organized with the aim of composing the final subtitles, according to a series of options configurable by the user.

The SAVAS systems allow the configuration of the most common layout features related to subtitling, such as the position of subtitles on screen, the number of lines per subtitle, the amount of characters per line, the typeface, the distribution and alignment of the text, the transmission modes (i.e. blocks or scrolling), and the colors linked to different speakers. For this last feature, the information given by the APP component about the speaker gender is used to change the colors of subtitles.

Regarding features related to the duration of subtitles on screen, automatic pre-recorded subtitles created by the S.Scribe! application can be configured either to be synchronized to the audio or to follow minimum and maximum duration and speed rules to improve readability.

3.3 SAVAS applications

As mentioned before, three type of applications were developed for different subtitling purposes: batch transcription and subtitling system (S.Scribe!), online subtitling system

Table 5 Amount of pronunciations in the lexicons

Language	Pronunciations
Portuguese	120 K
Spanish	104 K
Basque	155 K
Italian	138 K
French	159 K
German	169 K

(S.Live!) and a respeaking and dictation engine (S.Respeak!). In the following subsections, these applications are described in more detail.

3.3.1 S.Scribe!

S.Scribe! is a client/server system, working offline: it is capable of processing a file of previously recorded audio or video and transcribe it, producing a subtitle file.

The application has an interface for administration and usage; it receives an audio/video file, adds it to 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. The results can also be downloaded in text (TXT) and meta-data (XML) formats. S.Scribe! includes 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 audio/video files to be processed.
- 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.

3.3.2 S.Live!

S.Live! is a speech transcription system which operates in both online and real time modes. It receives input streaming (audio or video in digital or analogue format) from the broadcaster or a multimedia content from the web, producing live captioning and broadcasting live subtitles in both analogue or digital formats.

The system can be used for captioning television programmes as well as available multimedia content on the web. It offers a Language Model Adaptation Service, which allows daily adaptation to new and different topics. It works in conjunction with an Acoustic Segmentation Module for separating the relevant acoustic areas for captioning. The S.Live! application was prepared to be integrated with the most commercial subtitling softwares, such as Screen or FAB subtitling systems, thorugh an IP based protocol.

Nowadays, the S.Live! system is used in several television channels in a number of countries like Portugal and Brazil.

3.3.3 S.Respeak!

The main engine of the S.Respeak! application is VOXControl, a software for dictation and re-speaking. Its use is foreseen in audiovisual contents with a high degree of background noise, music and spontaneous speech, where S.Live! or S.Scribe! applications underperform and a human operator is required to re-speak the relevant information. With this operation, the difficulties associated to the automatic transcription of those kinds of programs are overcome. This application allows to adapt the acoustic model to a specific user, in order to increase performance.

VoxControl can be integrated with re-speaking applications in two different modes:

- The first solution is to put the cursor on top of a window where the output text is needed. This is the normal operation of a dictation system. However, this implies that the user is not allowed to move the cursor and that only one user can operate the application. For re-speaking, this option could be limiting.

– Due to the limitations posed by the first solution, a second mode was implemented in which the application runs in the background. In this mode, the software communicates with the commercial subtitling software through the same mechanism and protocols of the S.Live! application.

4 Evaluation methodology

The evaluation methodology had the objective of measuring the performance of the developed SAVAS applications regarding the subtitling quality features accepted by the industry. With this aim, a varied set of metrics were employed, some of which were designed specifically for this work.

A number of guidelines and good practice codes for subtitling have been published over the last years. Among well-known ones are: Ofcoms Guidance on Standards for Subtitling;[4] BBCs Online Subtitling Editorial Guidelines;[5] ESISTs Guidelines for Production and Layout of TV Subtitles;[6] the Spanish UNE 153010 norm [2] on subtitling for the deaf and hard of hearing and a reference textbook on generally accepted subtitling practice published in 2007 by Jorge Diaz-Cintas et al. [10]. Standard guidelines cover the various aspects of subtitle quality, such as subtitle layout, duration and text editing, which are shared among subtitling companies and broadcasters.

Concerning layout features, the most widely accepted subtitling practice uses two centered lines at the bottom of the screen, of 37 characters each one, with white font over a black background and in block mode, which is much easier to read than scrolling. In order to highlight different speakers, colours such as yellow, cyan or green are usually employed.

With regard to duration, recommendations span features related to the delay or the persistence of subtitles on screen. While high latencies have a negative impact on the perceived quality of subtitles, short persistence on screen has shown to decrease their readability. Standard guidelines recommend a maximum delay of 3 seconds and a maximum speed of 160-180 words per minute. Although these recommendations are generally followed for pre-recorded programs, the difficulties posed by live subtitling currently result in median latencies of around 6 seconds, with spikes of up to 24 seconds.

Finally, the standard subtitling practice related to text editing employs mixed letter case as in printed material, splits subtitled text at the highest possible syntactic nodes and makes use of the most common and recognizable acronyms, apostrophes and numerals to save character space.

With the aim of measuring all these parameters, we employed and defined several metrics, which are described in more detail in the following subsections.

[4] http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/standards_for_subtitling/subtitling_1.asp.html

[5] http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf

[6] http://www.translationjournal.net/journal/04stndrd.htm

4.1 Metrics

4.1.1 Word error rate (WER)

WER is a common metric used to measure the performance of speech recognition systems. It is computed by comparing reference annotations against automatic transcriptions, which in our case correspond to automatic subtitles. WER is calculated through the following formula:

$$WER = \frac{Substitutions + Deletions + Insertions}{Words\ in\ the\ reference} \times 100 \tag{2}$$

Substitutions refer to words which are replaced. *Deletions* are related to words which are missed out and *Insertions* are words incorrectly added by the recognizer. *Words in the reference* is the number of total words in the reference annotation.

4.1.2 Speaker change detection (SD)

The SD metric was computed thorough the F1-measure metric, which combines the harmonic mean of Precision and Recall metrics as follows:

$$F1 - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision refers to the fraction of retrieved instances that are correct, while the Recall metric describes the fraction of correct instances that are retrieved.

4.1.3 Capitalization and punctuation

The performance of the capitalization and punctuation features was measured using the Precision, Recall and F1-measure metrics.

4.1.4 Timing

As mentioned before, two duration features are linked to subtitle timing: delay and persistence.

Delay The delay of automatic live subtitles is composed by (1) the latency of the LVCSR technology, (2) the time needed to compose each subtitle and (3) the time length to insert it into the audiovisual signal to be transmitted. This feature was measured by the broadcasters, once the SAVAS technology for automatic subtitling was integrated in their premises. The delay was computed comparing the time of the broadcasted subtitles against word-level time-codes synchronized to the audio.

The technology for Basque and Spanish was evaluated at Euskal Telebista (ETB, Basque Television), the Basque Countrys public broadcast service, while the Portuguese Radio and Television (RTP) integrated the technology for Portuguese. The Italian Public Service Broadcaster (RAI, Radio Televisione Italiana) tested Italian and SWISS Teletext (SWISS TXT) was in charge of evaluating the rest of the languages. Each of the broadcasters used a different insertion software

Persistence Subtitle persistence on screen was measured by subtracting the time-in from the time-out of each subtitle and averaging across entries. Because the number of words or characters per subtitle also has an impact in their readability, metrics such as words per minute (wpm) or characters per second (cps) were employed.

4.1.5 Splitting

It is difficult to measure subtitle splitting objectively, since there is generally more than one way of segmenting correctly a particular subtitle. Thus, the approach adopted to measure splitting quality involves asking subtitling experts to manually rank the quality of both inter- and cross-subtitle splitting.

4.1.6 Overall quality

The NER model[7] has been used since some years to measure live respoken subtitle errors. The model uses the following formula to determine the quality of live respoken subtitles:

$$NERvalue = \frac{N - E - R}{N} \times 100 \tag{4}$$

where N is the number of words in the respoken text, E corresponds to the edition errors caused by the respeaker's strategies, and R is the errors committed by the recognizer. Computing the formula, a NER value of 100 indicates that the content was subtitled entirely correctly. Good quality live subtitles are expected to go beyond 98 % accuracy according to this.

Since the NER model was devised for quality assessment of respoken subtitles, it considers only recognition and respeakers edition errors. However, speaker colour, timing and splitting information is also relevant in automatic subtitling and helps establishing the quality of subtitles. In this work, the NER model was extended to also consider errors related to those features.

The extended eNER formula is as follows:

$$eNERvalue = \frac{(N \times P) - \sum_{i=1}^{N}(R + SD + T + S)}{(N \times P)} \times 100 \tag{5}$$

where N is the number of test subtitles, P is the number of parameters to be evaluated, R corresponds to the recognition errors, SD represents the speaker change errors, T is the timing persistence errors scoring 0 (no error) or 1 (error) values, and S represents the splitting errors, scoring 0 (no error), 0.5 (inter- or cross- subtitle error) or 1 (inter- and cross-subtitle errors). All the parameters has a maximum value of 1.

The recognition (R) and speaker change (SD) errors were calculated using the WER and F1-measure metrics respectively. The timing (T) and splitting (S) parameters were evaluated manually by subtitling experts.

The evaluation of the S.Respeak! applications for Basque and Italian was carried out using the NER model through the NERstar tool.

[7] http://www.speedchill.com/nerstar/

4.1.7 Productivity gain

The aim of the productivity gain evaluation was to test whether post-editing automatic subtitles is faster than creating them manually from scratch. Subtitling professionals in all languages were asked to post-edit automatic subtitles and to create them from scratch, using their usual subtitle editing software and quality standards. The productivity gain was measured using the Subtitles per minute (spm) metric.

5 Evaluation and results

5.1 Test Set description

The Test Set for the evaluation of the systems was composed of a total amount of 30 hours of news, interview/debate and sports TV programs broadcasted in 2014. This material was annotated to compare it against the outputs generated by each application. Table 6 details the amount of data collected for each application type per language.

For Basque and Spanish, 5 hours were compiled per language. In Basque, 2 hours were used to test the S.Live! subtitling application and other 2 hours to evaluate the S.Scribe! application in the news domain. Another hour was compiled to evaluate the S.Respeak! application in the sports domain. In Spanish, the test set was divided in two parts and employed to evaluate S.Live! and S.Scribe! applications. The data for Basque and Spanish were gathered from news and sport programs broadcasted by ETB.

For Portuguese, 2 hours were compiled from a debate program to test the Live subtitling application in the interview/debate domain. Overall, 6 hours were compiled for the Italian Test Set, including 4 hours for Italian and 2 hours for Swiss Italian. The Italian data were gathered from news programs broadcasted by RAI.

Regarding French, German and their Swiss variations, they follow the same distribution as Italian. The French contents were gathered from news programs broadcasted by Euronews, France24 and the Swiss French television. The German contents were collected from news programs broadcasted by DasErste, ZDF and the Swiss German television.

5.2 Word error rate

As shown in Fig. 5, WERs follow similar distribution for Basque, Spanish, Italian, French and German, achieving performances around 15 %. The Swiss variants of French and German languages goes up to 20 % and the more challenging Portuguese case reaches 30 %. There is no significant difference between live and pre-recorded mode and performance variations are most probably due to the use of different testing contents.

Table 6 Test Set data amounts per language

Application	PT	ES	EU	IT	FR	DE	IT_CH	FR_CH	DE_CH
S.Live!	2 H	2.5 H	2 H	1.5 H	2 H	2 H	1 H	1 H	1 H
S.Scribe!	-	2.5 H	2 H	1.5 H	2 H	2 H	1 H	1 H	1 H
S.Respeak!	-	-	1 H	1 H	-	-	-	-	-

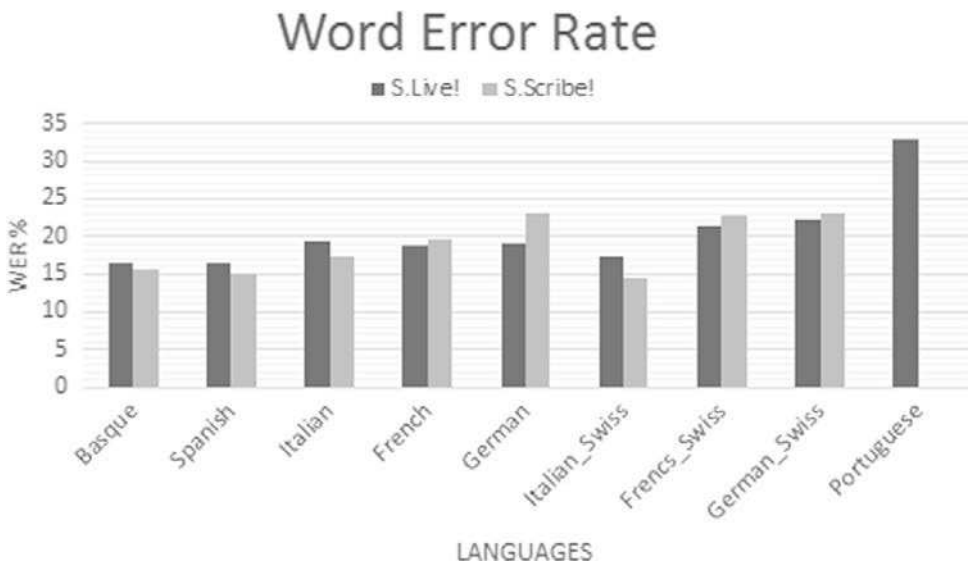


Fig. 5 WERs per application and language

5.3 Speaker change detection

Speaker Change Detection (SD) performance is shown in Fig. 6 per language and application type. The results are around 80 % with slightly worse performance going down to 60 % for Swiss French due to the acoustic conditions of these language on the test set. In the case of Portuguese, the particularity of the test material, which was composed by debate programs including many overlapped turns, degrades the accuracy for this language.

5.4 Capitalization and punctuation

Regarding Capitalization and Punctuation, Figs. 7 and 8 show that average F1-measures are around 85 % and 50 % respectively. Even if the results for Capitalization are promising, the accuracy obtained on automatic Punctuation reflects the difficulty posed by these type of contents, containing topic and speaker-dependent emotional pronunciations and intonations, in which acoustic pauses usually do not correspond to the real ends of phrases and sentences.

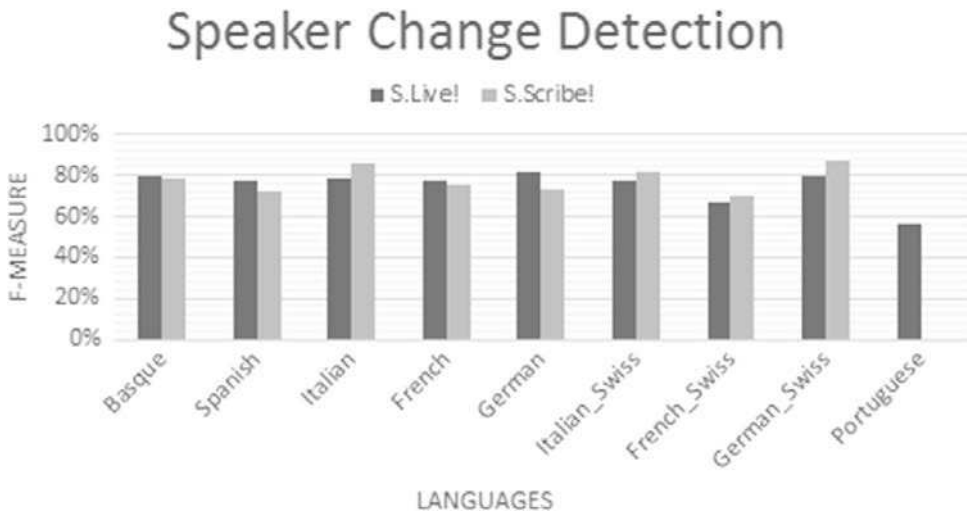


Fig. 6 Speaker Change Detection accuracies per application and language

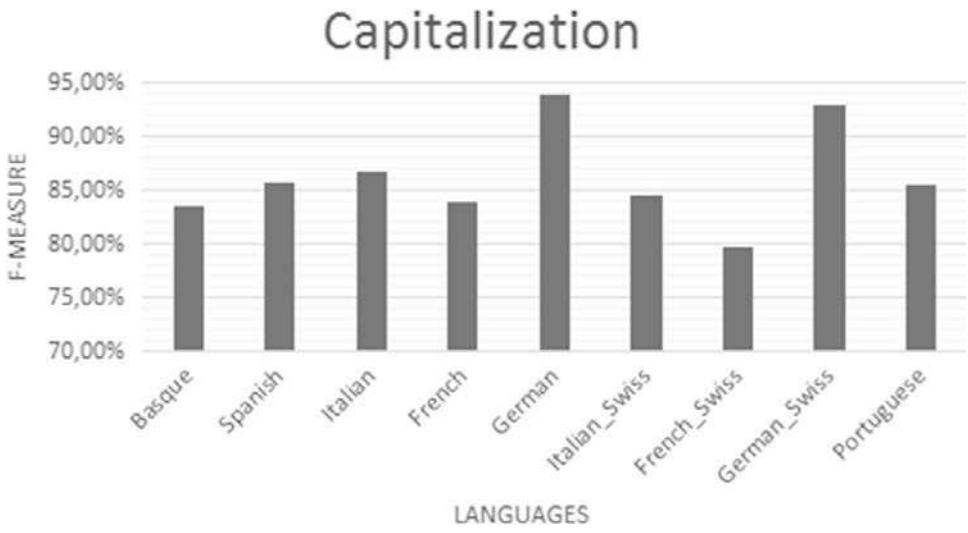


Fig. 7 Capitalization accuracies per language

5.5 Timing

5.5.1 Delay

The delay of the S.Live! applications is presented in Fig. 9. The results suggest that there could be a clear dependence between the delay and the software employed by each of the broadcasters to compose and insert the automatic subtitles into the broacasted signal. Each of the broadcasters used a different subtitling software: ETB employed WinCAPS for Basque and Spanish, RTP and SWISS TXT integrated FAB for Portuguese and Swiss variants respectively, and RAI used Speech Title for Italian. Nevertheless, if we leave the Italian outlier aside, the average delay results in 7 seconds which can be considered state-of-the-art performance of live respoken subtitling.

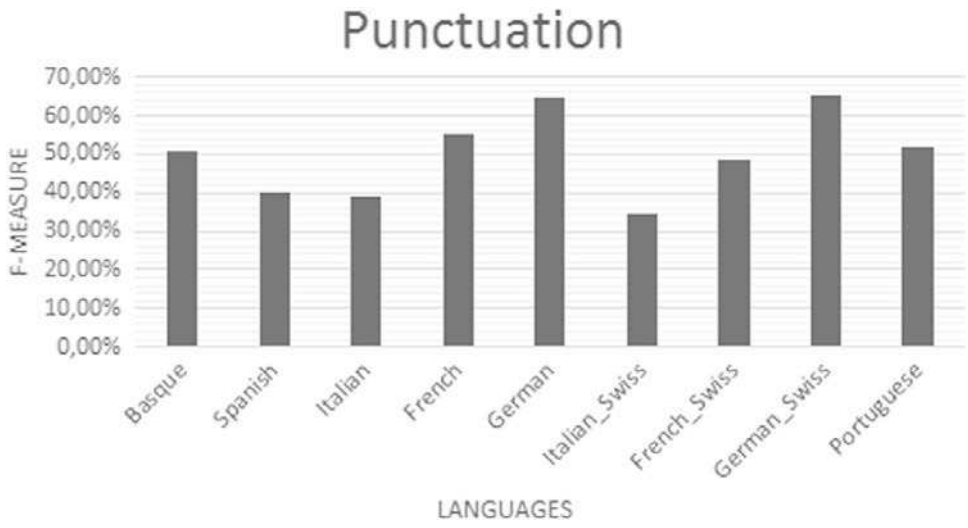


Fig. 8 Punctuation accuracies per language

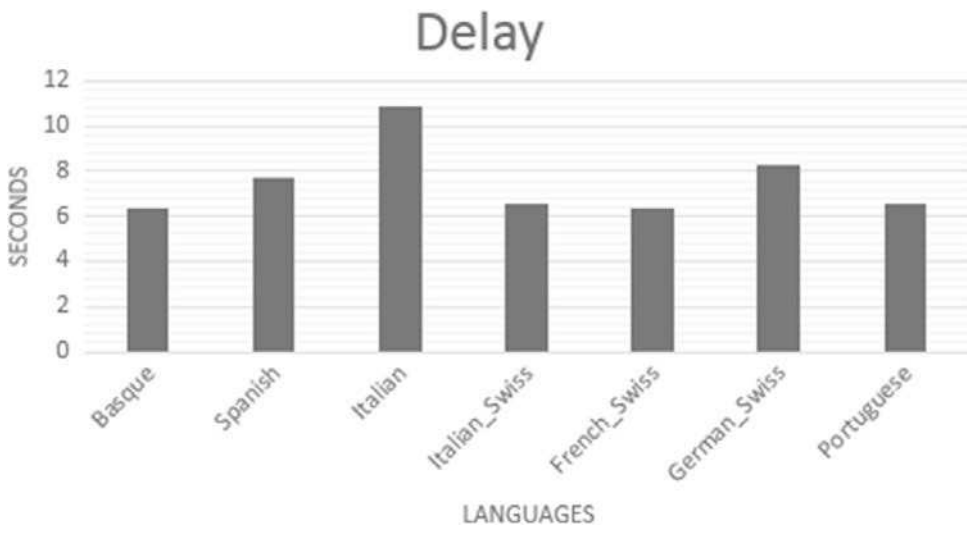


Fig. 9 Delay in seconds of the S.Live! applications

5.5.2 Persistence

Regarding persistence, Fig. 10 shows the average characters per second (CPS) computed per language in live and pre-recorded mode. Overall, average automatic subtitle persistence is below the maximum thresholds of 17 and 19 accepted by the subtitling community for pre-recorded and live programs respectively.

Results suggest persistence to be language specific, with Basque and Spanish automatic subtitles having the slowest speed, followed by Portuguese, Italian and the Swiss variants of Italian, French and German. Further inspection of the content also showed positive correlations between the amount of spontaneous sections in the audiovisual material, in which speech rate is usually faster, and their higher CPS values. The Swiss news programs in particular have higher proportions of interviews and spontaneous interactions than the rest. The reason why the live persistence of automatic Basque and Spanish subtitles is lower than their pre-recorded counterparts is that the WinCAPS insertion software employed at ETB had been configured to force subtitle speed output to 10 cps. Similarly, the persistence of the Swiss languages was also configured in the insertion software during live operation while

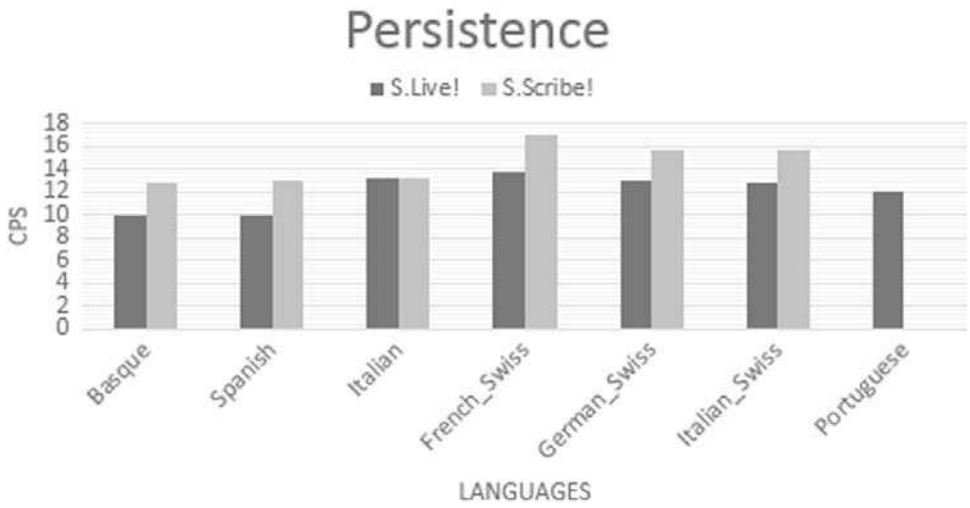


Fig. 10 The persistence in CPS per language and applications

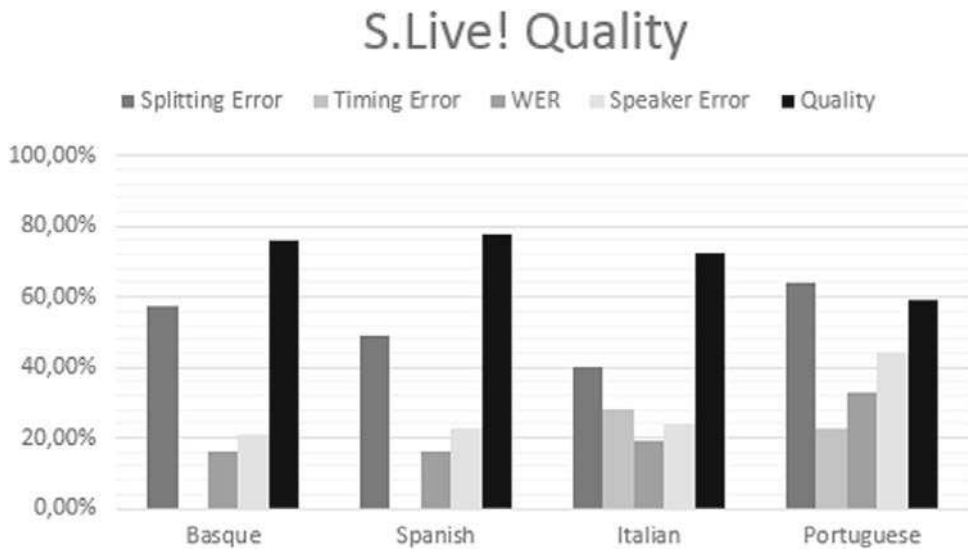


Fig. 11 S.Live! overall quality results per language

pre-recorded timing was simply set to be synchronized with the audio. These results show that configuring automatic subtitles just to be synchronized to the audio reduces delay but worsens persistence and, as a consequence, readability.

5.6 Overall quality

5.6.1 eNER of S.Live! and S.Scribe!

Figures 11 and 12 show the overall quality of the S.Live! and S.Scribe! applications. It was computed for Basque, Spanish, Italian and Portuguese languages. As it can be appreciated, eNER values are around 75 % on average for the broadcast news domain in Basque, Spanish and Italian without significant performance differences across operation modes and 60 % for the interview/debate domain in Portuguese. Although these values are far from the 98 % NER values considered to correspond to good quality subtitles, eNER results are expected

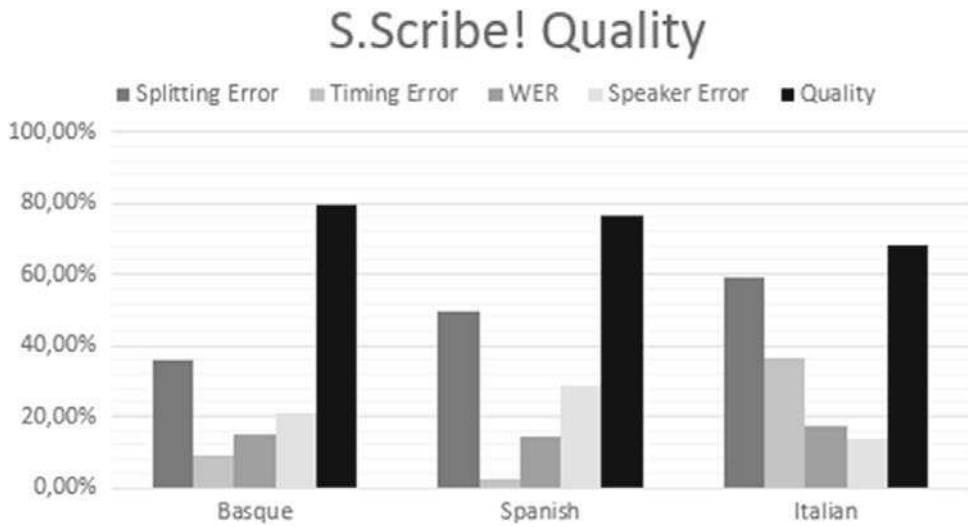


Fig. 12 S.Scribe! overall quality results per language

to reach relatively lower values because the extended formula considers a higher amount of quality features.

If we look into the specific weight of each one of the considered quality features on the overall eNER metric, we can see that splitting errors are the most frequent ones for all the languages.

Generally speaking, only 10 percent of the examined subtitles were free of splitting errors, about half of them contained at least one of the two possible splitting errors, a third both of them. And as far as the error-free subtitles were concerned, one needed in most cases to link them with other subtitles in order to guarantee readability. This can be considered as reasonable, since technology for automatic splitting was not trained and developed within this work. The splitting of subtitles was done just counting up the characters and controlling not to exceed the maximum length of each subtitle line.

Manual evaluation of timing errors resulted zero for S.Live! in Basque and Spanish because the WinCAPS insertion subtitle software was configured to force subtitle speed output to 10 cps during broadcasting.

In Italian, timing errors outweight those related to speaker change and WER for both applications. Even if the average persistence achieved for Italian was in the 13 cps range, a further study of the results demonstrated a high fluctuation between low and high cps values. Regarding timing errors for S.Live! in Italian, the high delay presented above was the main reason for these discrete results.

5.6.2 NER of S.Respeak!

The S.Respeak! applications were evaluated for Basque and Italian in the sports domain. Due to the effort required by the only respeaker available, the respeaking task for Basque was divided into two parts of 20 and 30 minutes. The NER values achieved for each part were 86.55 % and 85.05 % respectively. Most of the errors were due to minor edition mistakes related to the speakers' strategies, while recognition errors were mostly caused by substitutions.

In Italian, a 84,64 % NER value was obtained. In this case, the great majority of errors was due to minor recognition mistakes mainly classified as deletions. Edition errors committed by the respeaker were smaller and minor.

We believe that the presented results could be improved by more proficient respeakers for both languages.

5.7 Productivity gain

Finally, Figure 13 summarizes the productivity gains achieved in the post-editing task. 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. We believe post-editing training and practice should help increase them.

For Italian, post-editing S.Scribe! output has also been compared against post-editing stenotype output. As it can be appreciated in the Fig. 13, the latter has achieved higher productivity gains. Stenotypists probably generate less text editing errors than state-of-the-art LVCSR technology, particularly in what capitalization and punctuation features are concerned and, thus, the time devoted to correcting such kind of errors is reduced. However, stenotyping requires personnel resources that automatic transcription does not, which is an important factor to be considered.

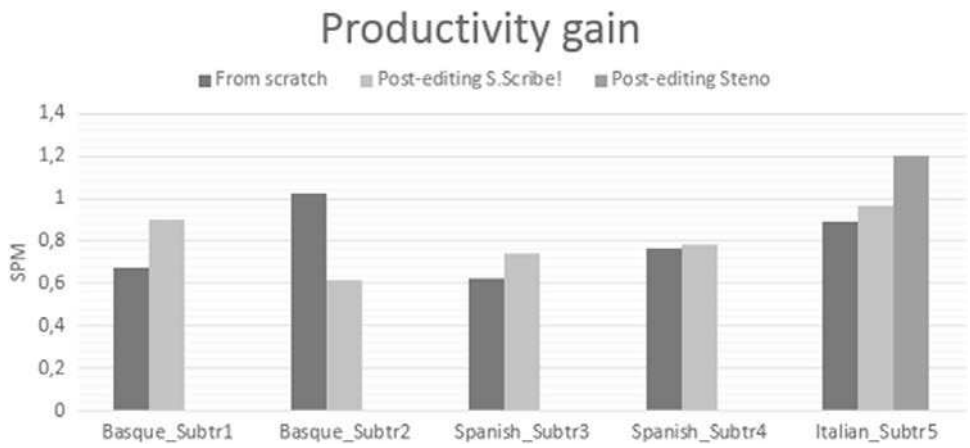


Fig. 13 Productivity gain results

5.8 Comparative evaluation

As we detailed in Section 2, Google is currently the only company which provides publicly the automatic generation of time-coded transcriptions and subtitles. This is a service available to the videos uploaded by the users to Youtube platform and it works in batch mode. However, the automatically generated transcriptions and subtitles from Youtube do not include punctuation and capitalization at the moment. Moreover, subtitles do not fulfill the standard subtitling practices defined by professionals. Thus, SAVAS systems could be compared to Google technology only at word error rate level.

For this comparative evaluation, we employed contents from 4 languages of the SAVAS project, including Spanish, Italian, French and German. We uploaded contents to Youtube and obtained the related transcriptions. These transcriptions were then compared to the results obtained with the SAVAS S.Scribe! application for the same audiovisual contents. The comparison was done using the WER metric explained in Section 4. Table 7 shows the results obtained for each language using Youtube and SAVAS S.Scribe! application.

As it can be seen in Table 7, S.Scribe! application outperforms the results obtained by Google technology through Youtube platform for all the languages in the test set. However, it could be expectable since the domain of these test contents was the same employed to train SAVAS systems; that is, broadcast news domain. Google offers technology which contains models trained on general domain data. The most remarkable differences are appreciated in Italian and French languages with improvements of around 10 and 12 points respectively.

Table 7 WER metrics for Youtube and S.Scribe! application

Language	Duration	Youtube	S.Scribe!
Spanish	2 H	24.91 %	16.50 %
Italian	2.5 H	27.80 %	17.81 %
French	2 H	37.61 %	25.81 %
German	2 H	31.10 %	26.92 %

6 Conclusions

In this article, new LVCSR based subtitling systems for both batch and live multimedia contents have been presented for several European languages. As it was described in Section 2, SAVAS systems can be considered pioneers offering a solution for full transcriptions and subtitles generation for different types of applications and languages. A survey of the literature in the field actually provides no references on such full automatic subtitling systems. Besides supporting different European languages, SAVAS systems include LVCSR engines trained over a huge amount of annotated data, technology for automatic punctuation and capitalization, speaker clustering and identification, in addition to modules for text normalization and subtitles generation following configurable standard subtitling rules. The systems were developed considering the needs of the subtitling companies and market, including the integration of the systems with the main subtitling softwares. Furthermore, SAVAS systems include three types of applications per language; S.Scribe!, a batch Speaker Independent Transcription system for offline subtitling; S.Live!, a Speaker Independent Transcription System, with real-time performances for live subtitling; and S.Respeak!, a dictation engine for live and batch production of subtitles.

The SAVAS systems were further evaluated using several metrics related to LVCSR technology and subtitling quality specific features. Although the developed automatic subtitling applications do not perform as well as professional subtitlers, they have achieved favorable results. The WER both in live and pre-recorded mode can be considered promising since it performed much better than other reference systems like Google for this specific domain. The delay of S.Live! subtitles is perfectly consistent with the recommendations. The speaker change detection technology works well even if it can be refined to work better with spontaneous speech. Automatic punctuation is error-prone considerably often due to the difficulty posed by the TV contents, and the splitting algorithm shall look into making use of syntactic information to achieve better results. Finally, productivity gain experiments suggest that post-editing automatic subtitles is faster than creating them from scratch.

The future work will be focused on the improvement of the technologies involved in the SAVAS systems. Regarding LVCSR technology, recent studies have demonstrated that Deep Neural Networks (DNNs) models have driven significant improvements on a variety of speech recognition benchmarks and data sets [42, 44]. With regard to SAVAS systems, further work will include research and development of a hybrid DNN-HMM recognition system for a more efficient offline and specially online subtitling. With the aim of improving the automatic punctuation module, new features will be included for classification, including prosodic and speaker related information. Furthermore, recent advances in Natural Language Processing with Neural Networks [8] has shown promising results that could be useful in sentence boundaries detection and punctuation marks prediction. The automatic splitting of subtitles should be also improved considering syntactic information to create linguistically coherent line-breaks, which is the preferred and most adopted solution in the community. To this end, the literature offers solutions based on the use of machine learning algorithms [5]. Finally, future work will also include the expansion of the SAVAS systems to more European and Asian languages.

Acknowledgments This work was funded by the FP7-ICT-2011-SME-DCL project 296371 - SAVAS (Sharing Audiovisual contents for Automatic Subtitling). http://www.fp7-savas.eu

References

1. Abad A (2007) The L2F language recognition system for NIST LRE 2011. In: The 2011 NIST language recognition evaluation (LRE11) workshop
2. AENOR (2003) Spanish Technical Standards. Standard UNE 153010:2003: Subtitled Through Teletext. http://www.aenor.es
3. Ajot J, Fiscus J (2009) The rich transcription 2009 speech-to-text (STT) and speaker attributed STT results. Tech. rep., NIST - National Institute of Standards and Technology, Rich Transcription Evaluation Workshop, Melbourne, Florida
4. Aliprandi C, et al. (2003) RAI voice subtitle: how the lexical approach can improve quality in Speech Recognition Systems. https://www.voiceproject.eu/
5. Álvarez A, Arzelus H, Etchegoyhen T (2014) Towards customized automatic segmentation of subtitles. In: Advances in speech and language technologies for Iberian languages. Springer, pp 229–238
6. Batista F, Caseiro D, Mamede N, Trancoso I (2008) Recovering capitalization and punctuation marks for automatic speech recognition: case study for Portuguese broadcast news. Speech Comm 50(10):847–862
7. Caseiro D, Trancoso I (2006) A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Trans Audio Speech Lang Process 14(4):1281–1291
8. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537
9. Del Pozo A, Aliprandi C, Álvarez A, Mendes C, Neto J, Paulo S, Piccinini N, Raffaelli M (2014) SAVAS: collecting, annotating and sharing audiovisual language resources for automatic subtitling. In: LREC 2014. Proceedings of the 9th international conference on language resources and evaluation
10. Díaz-Cintas J, Orero P, Remael A (2007) Media for all: subtitling for the deaf, audio description, and sign language, vol 30. Rodopi
11. eCaption: http://www.ecaption.eu/
12. FAB - Teletext & Subtitling Systems: FAB Subtitler Live Edition. http://www.fab-online.com/eng/subtitling/production/subtlive.htm
13. Fiscus J, Garofolo J, Ajot J, Michet M (2006) Rt-06s speaker diarization results and speech activity detection results. In: NIST 2006 spring rich transcrition evaluation workshop, Washington DC
14. Flanagan M (2009) Recycling texts: human evaluation of example-based machine translation subtitles for DVD. Ph.D. thesis, School of applied language and intercultural studies. Dublin City University, Dublin
15. Galliano S, Geoffrois E, Gravier G, Bonastre JF, Mostefa D, Choukri K (2006) Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In: Proceedings of LREC, vol 6, pp 315–320
16. Gauvain JL, Lamel L, Adda G (2001) Audio partitioning and transcription for broadcast data indexation. Multimedia Tools Appl 14(2):187–200
17. Google: Automatic captions in youtube. https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html (2009)
18. Google: Translate youtube captions. https://www.mattcutts.com/blog/youtube-subtitle-captions/ (2009)
19. Grass Valeey: Subtitle and Caption Creation. http://www.grassvalley.com/products/subcat-subtitle_and_caption_creation
20. IBM: Viavoice. http://www-01.ibm.com/software/pervasive/viavoice.html
21. Koemei: https://www.koemei.com/
22. Lambourne A, Hewitt J, Lyon C, Warren S (2004) Speech-based real-time subtitling services. Int J Speech Technol 7(4):269–279
23. Lan ZZ, Bao L, Yu SI, Liu W, Hauptmann AG (2013) Multimedia classification and event detection using double fusion. Multimedia Tools Appl 1–15
24. Lööf J, Gollan C, Hahn S, Heigold G, Hoffmeister B, Plahl C, Rybach D, Schlüter R, Ney H (2007) The RWTH 2007 TC-STAR evaluation system for european English and Spanish. In: INTERSPEECH, pp 2145–2148
25. Meignier S, Merlin T (2010) LIUM SpkDiarization: an open source toolkit for diarization. In: CMU SPUD workshop, vol 2010, Dallas
26. Meinedo H, Abad A, Pellegrini T, Trancoso I, Neto J (2010) The L2F broadcast news speech recognition system. Proc Fala 93–96
27. Meinedo H, Caseiro D, Neto J, Trancoso I (2003) Audimus.media: a broadcast news speech recognition system for the european portuguese language. In: Computational Processing of the Portuguese Language. Springer, pp 9–17

28. Meinedo H, Neto JP (2005) A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models. In: INTERSPEECH. Citeseer, pp 237–240
29. Meinedo H, Viveiros M, Neto JP (2008) Evaluation of a live broadcast news subtitling system for portuguese. In: INTERSPEECH, pp 508–511
30. Microsoft: windows speech recognition. http://www.windows.microsoft.com/en-us/windows7/dictate-text-using-speech-recognition
31. Neto J, Meinedo H, Viveiros M, Cassaca R, Martins C, Caseiro D (2008) Broadcast news subtitling system in portuguese. In: IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE, pp 1561–1564
32. Nuance: Dragon Naturally Speaking. http://www.nuance.com/index.htm
33. Obach M, Lehr M, Arruti A (2007) Automatic speech recognition for live TV subtitling for hearing-impaired people. Challenges for Assistive Technology: AAATE 07 20:286
34. Sail Labs: http://www.sail-labs.com/
35. Screen Systems: WinCAPS Q-live for live and news subtitling and captioning. http://www.screensystems.tv/products/wincaps-q-live/
36. Screen Systems: WINCAPS QU4NTUM subtitling software. http://www.screensystems.tv/products/wincaps-subtitling-software/
37. Starfish Technologies: Subtitling and closed captioning systems. http://www.starfish.tv/captioning-and-subtitling/
38. SyncWords: https://www.syncwords.com/
39. Ubertitles: http://www.ubertitles.com/
40. Vecsys: http://www.vecsys-technologies.fr/en/
41. Verbio: https://www.verbio.com/
42. Vu NT, Imseng D, Povey D, Motlicek P, Schultz T, Bourlard H (2014) Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 7639–7643
43. Woodland PC (2002) The development of the HTK broadcast news transcription system: an overview. Speech Comm 37(1):47–67
44. Zhang X, Trmal J, Povey D, Khudanpur S (2014) Improving deep neural network acoustic models using generalized maxout networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 215–219
45. Zibert J, Mihelic F, Martens JP, Meinedo H, Neto J, Docio L, García-Mateo C, David P, Zdansky J, Pleva M et al (2005) The COST278 broadcast news segmentation and speaker clustering evaluation: overview, methodology, systems, results. In: 6th Annual conference of the international speech communication association (Interspeech 2005); 9th European conference on speech communication and technology (Eurospeech), vol 2005. International Speech Communication Association (ISCA), pp 629–632

Aitor Álvarez He works as staff Researcher of the Human Speech and Language Technologies group at VICOMTECH. He studied Computer Science at the University of the Basque Country (2005). He carried out his final year project at the Department of Architecture and Computer Technology of the same university, where he continued working as a scholar. He is currently a PhD student, has completed the Diploma of Advanced Studies and is working on his thesis on advanced audio and speech processing techniques

for the media. In July 2007 he joined VICOMTECH, where he has been working as project manager and researcher involved in speech related R&D projects. He has been involved in several European projects and has published several papers in relevant international and national conferences.

Carlos Mendes He works as chief Researcher of the Research and Development department at VOICEINTERACTION and was previously a Researcher at INESC-ID Lisbon. He received his MSc in Electrical and Computer Engineering by the Institutio Superior Técnico (IST), Lisbon, in 2008. As a Researcher at the INESC-ID Lisbon, he participated in European projects, such as EU FP6 E-Circus; national projects, such as TECNOVOZ; and was also actively involved in the development of the new implementation of the DIXI Speech Synthesis System. As a researcher at VOICEINTERACTION, he participated in two EU FP7 projects: SAVAS and CAPER; and has conducted research activities in Speech Synthesis and Automatic Speech Recognition with emphasis on Language and Acoustic Modeling. Currently, he is a PhD student from the same University and is working on his thesis on Methods of Acoustic Modeling for Automatic Speech Recognition Systems.

Matteo Raffaelli He gained his PhD in Philosophy from the University of Bamberg (Germany). R&D Project Manager at Synthema, he has extensive experience in research projects. His main research interests are Artificial Intelligence, Multimodal Analytics, Natural Language Processing (both Text and Speech Mining) and Semantic Technologies. Several scientific publications in these areas.

Tiago Luís He is a researcher at VOICEINTERACTION and was previously a researcher at INESC-ID Lisbon. He received a MSc degree in Information Systems and Computer Engineering in 2008 from Instituto Superior Técnico (IST), Lisbon. As a researcher at INESC-ID, he participated in European projects such as PT-STAR (Speech Translation Advanced Research to and from Portuguese). As a researcher at VOICEINTERACTION, he participated in two FP7 projects: SAVAS and CAPER. His activities have been in the areas of Speech Recognition, Speaker Clustering and Natural Language Processing.

Sérgio Paulo In 2008 he co-founded VoiceInteraction and holds a position in the board of the company. He was previously at INESC-ID Lisbon as Senior Researcher. He graduated in Electrotechnical and Computers Engineering by the Instituto Superior Técnico (IST), Lisbon, in 2001, and received his PhD degree in Electrotechnical and Computer Engineering from the same University in 2009 with a thesis on Speech Synthesis. As a Researcher at the INESC-ID Lisbon, he participated in European projects, such as EU FP6 E-Circus; national projects, such as TECNOVOZ, in which he was involved in the new implementation of the DIXI speech synthesis system, and was also in the ECESS (European Center of Excellence in Speech Synthesis) Actions. More recently, he has been actively involved in two EU FP7 projects: SAVAS and CAPER. He has conducted research activities in speech synthesis, speech processing at large, and also in methods for automating the creation of speech corpora.

Nicola Piccinini Researcher. Degree in Physics in 2008 at the University of Pisa. Joined Synthema as a software programmer in 2000; since 2008 he is working as a researcher in the R&D Department. His research interest are in NLP tools and resources development, particularly for Automatic Speech Recognition. He is a Product Manager for the ASR Business Unit in Synthema, having been involved into several commercial projects.

Haritz Arzelus He is a researcher in the Human Speech and Language Technologies group at Vicomtech-IK4. He studied Computer Engineering at the Faculty of Informatics of the University of the Basque Country in San Sebastian (2009), where he was a fellow as part of the team of laboratory technicians during his last year. Since October 2009, he has been working in Vicomtech-IK4 as a researcher in speech and language processing technologies on local, national and European projects. He has published several papers in relevant international and national conferences.

João Neto He is Assistant Professor at Instituto Superior Tecnico (IST) and CEO of VOICEINTERACTION. He graduated in Electrotechnical and Computers Engineering by the Instituto Superior Técnico (IST), Lisbon, in 1987, and received his MSc and PhD degrees in Electrical and Computer Engineering, from the same University in 1991 and 1998, respectively. His PhD thesis was on the topic of speaker-adaptation for continuous speech recognition systems. In 1991 he started as lecturer and since 1998, he has held a position of Assistant Professor at IST. He participated in several European projects as PYGMALION, WERNICKE, SPRACH, ALERT, VIDIVIDEO, IDASH, LIREC and CAPER as well as several national projects. His activities have been focused in the areas of neural networks and speech recognition, with a recent emphasis on subtitling and dialogue systems. He has also written a significant number of scientific papers. He is a member of IEEE and ISCA. In 2008 he cofounded VOICEINTERACTION where currently holds the CEO position.

Carlo Aliprandi He received a degree in Computer Science in 1992 at the University of Milan. In 1997 he joined SYNTHEMA, where he is currently International R&D Manager, mainly promoting the participation of SYNTHEMA on European and international collaborative projects. He also leads the SSR team, whose activities have pioneered a number of techniques to accelerate human tasks such as speech recognition for subtitling and reporting, text entry, and fast typing. Since 2003 he teaches Assistive Technology at the University of Trieste. He is author of scientific papers, published in international journals and conference proceedings.

Arantza del Pozo Head of the HSLT Group at VICOMTECH. She graduated from the Electronics and Telecommunications Engineering Department at the University of Deusto, Bilbao (2003). After graduation, she attended Cambridge University to complete an M.Phil. in Computer, Speech, Text and Internet Technologies (2004) and then continued on to complete a PhD (2008) under the supervision of Prof. Steve Young. Since May 2008, she has been working as a principal researcher at VICOMTECH, where she leads the HSLT team and manages speech and language related R&D projects.