APyCA: Towards the Automatic Subtitling of Television Content in Spanish

Aitor Álvarez, Arantza del Pozo
Vicomtech Research Centre
Mikeletegi pasealekua, 57
Miramon Teknologia Parkea
20009 Donostia-San Sebastian, Spain
Email: {aalvarez, adelpozo}@vicomtech.org

Andoni Arruti
The University of the Basque Country
Dept. of Computer Architecture and Technology
Manuel de Lardizabal Pasealekua 1
20018 Donostia-San Sebastian, Spain
Email: andoni.arruti@ehu.es

***Abstract*—Automatic subtitling of television content has become an approachable challenge due to the advancement of the technology involved. In addition, it has also become a priority need for many Spanish TV broadcasters, who will have to broadcast up to 90% of subtitled content by 2013 to comply with recently approved national audiovisual policies. APyCA, the prototype system described in this paper, has been developed in an attempt to automate the process of subtitling television content in Spanish through the application of state-of-the-art speech and language technologies. Voice activity detection, automatic speech recognition and alignment, discourse segment detection and speaker diarization have proved to be useful to generate time-coded colour-assigned draft transcriptions for post-editing. The productive benefit of the followed approach heavily depends on the performance of the speech recognition module, which achieves reasonable results on clean read speech but degrades as this becomes more noisy and/or spontaneous.**

I. Introduction

SUBTITLING plays an important role in the increasingly multimedia and globalised world we live in. Its usefulness extends from the enrichment of TV content – in order to make it more accessible for people with hearing difficulties or to facilitate audiovisual information retrieval – to its application in noisy environments such as airports and transit stations, where it is not possible to hear TV broadcasts. In addition, subtitling has also become a priority need for many Spanish TV broadcasters, who will have to broadcast up to 90% of subtitled content by 2013 to comply with recently approved national audiovisual policies[1].

However, subtitling is a labor-intensive and economically costly process. As a general rule, manual production of high-quality subtitles can be assumed to take between 8 and 10 times the length of the video material [1]. Nevertheless, mainly due to the higher demands, the time allotted to production of the subtitled material has decreased in recent years [1], [2].

Experienced professionals currently employ dedicated subtitling software tools to help them generate subtitles faster. However, these tools simply display the subtitles on the computer screen as they will appear on the television or movie screen and facilitate purely mechanical functions, such as cueing the subtitles, spell-checking and other basic text processing functions [3]. Only recently speaker-dependent automatic speech recognition has become popular for live subtitling through re-speaking, a technique in which a professional subtitler is trained to dictate live subtitles as the programme happens. Products such as Protitle Live® (NINSIGHT)[2] and WinCAPS® (Sysmedia)[3] allow trained speakers to dictate live subtitles into trained ASR engines. Nevertheless, there is still no ASR-based system in use for fully automated subtitling.

The application of the following state-of-the-art technologies can also contribute to making the subtitling process more automatic and productive:

A. Voice Activity Detection (VAD)

TV content presents a wide range of acoustic conditions: e.g. music, clean speech, outdoor speech, speech with background music, sound effects, noise, etc. However, only those segments that contain speech are to be subtitled. In addition, the different acoustic conditions might require different kinds of processing.

VAD technology can be used to automatically detect the audio segments containing speech. VAD segmentations can also be used to automatically classify and group audio segments with similar acoustic characteristics for further processing.

B. Automatic Speech Recognition (ASR) and alignment

ASR can be employed to obtain automatic transcriptions of the spoken information. Even though ASR can potentially save a lot of time, it is a difficult task mainly due to the high variability of the spoken environments, speakers and speech types present in TV content. Spoken environments vary from clean (studio recordings) to noisy (outdoor recordings, speech mixed with background music or sound effects). The type of speech may differ from dictation (newsreader) to spontaneous (debate or interview). The combination of these

[1] S. Government, "Spanish Audiovisual Law on Subtitles. http://www.cesya.es/es/normativa/legislacion/Financiacion_Radio_TV," 2008.

[2] http://www.ninsight.fr/FR/

[3] http://www.sysmedia.com/

possibilities seriously challenges ASR technology, which also needs to deal with speaker independence and the uncontrolled vocabulary of TV programs.

The time-stamps output by the ASR system can also be employed to align the recognised transcripts to the audio signals. In cases where the transcripts already exist, forced alignment can be used instead of recognition to obtain more accurate synchronizations between audio and text.

C. Discourse segment detection (DSD)

The detection of entities, relationships or individual events of speech and its segmentation into sentences and phrases is a crucial step for the transition from speech recognition to its full understanding. Unless explicitly dictated, speech recognisers output strings of words without a right segmentation of the output into discursive segments. As a result, ASR transcriptions consist of raw text that is quite difficult to understand for the reader.

DSD techniques can be used to automatically segment ASR transcriptions into segments which contain whole meaning, in order to make them more readable.

D. Speaker diarization (SD)

SD is the task of segmenting a multi-speaker audio signal into homogeneous parts and clustering them into different groups, each containing the voice of a single speaker.

In the context of subtitling, SD can be employed to automatically assign a specific color to the subtitles spoken by each speaker.

APyCA, the prototype system described in this paper, integrates the four technologies described above in a unique application, whose aim is to facilitate the manual production of subtitles by experienced professionals, reducing as a result the high cost of subtitle production.

The paper is structured as follows. Section 2 describes the state-of-the-art of the technologies involved and Section 3 presents the resources and tools developed and integrated within the project. Section 4 then describes the implemented prototype. Evaluation of the different modules is presented in Section 5 and finally, Section 6 discusses the main conclusions and further work.

II. State of the art

Much work has been made on the four main technologies involved in APyCA: Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Discourse Segments Detection (DSD) and Speaker Diarization (SD).

A. Voice Activity Detection (VAD)

With increasing demand for voice interfaces, the ability to distinguish human speech from other sounds is becoming crucial. Many works have attempted to discover characteristic features of human voices that are present only in speech. Since such characteristic features have not yet been discovered, short-time energy, zero crossing rate (ZCR), low-variance spectrum (LVS), spectral entropy (SE), periodicity, and so on have been used instead [4], [5]. While it is true that speech has such characteristics, the problem is that they can also be present in some non-speech sounds. This leads to a high false acceptance rate for specific kinds of noise. For example, loud white noise can also have high energy and ZCR.

For these reasons, statistical pattern classification approaches such as Gaussian Mixture Models (GMMs) have gained wider acceptance [6], [7]. In statistical VAD methods, both speech and noise models are trained via corresponding training data. Then, log likelihood ratio tests are applied to input data for speech and noise discrimination. These VAD methods have been shown to exhibit superior performance than the previous approach.

B. Automatic Speech Recognition (ASR)

There have been several projects focused on the development of ASR technology for the automatic transcription of Broadcast News (BN). However, most of them were developed for languages other than Spanish, such as English [8], French [9], Portuguese [10] or German [11]. As a result, there is not much data available in Spanish to train a robust speech recogniser for the automatic transcription of broadcast content. Several studies [9], [12] state that at least 100 hours of annotated and transcribed data is required for the adequate training of BN ASR engines and practical development works tend to use as much data as possible. For example, [13] uses up to 1000 hours of training speech data for Persian while [14] employs 81 hours for English, 52 for Portuguese and, in comparison, only 15 for Spanish. This lack of data is the main reason why we decided to use a commercial ASR engine within the APyCA prototype, and to explore adaptation of its default models to improve performance.

Despite the improvement of automatic speech recognisers, developing a system for the automatic transcription of content broadcasted in radio or television is still a challenge for many research groups. A system aimed at the automatic transcription of Portuguese BN, working in a real application scenario currently is [10]. It is based on a hybrid acoustic modelling approach that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multilayer Perceptrons (MLPs). Such acoustic modelling combines phoneme probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. The feature extraction methods are PLP, Log-RASTA and MSG. The training of the language model is done using both, Portuguese newspaper texts combined with the transcriptions used for acoustic model training.

With regard to the performance of the ASR systems developed for the different languages on the broadcast domain, the resulting error rates reflect in general the varying level of the acoustic and linguistic complexity of the recordings [11]. WERs range from 16.1% to 64.5%. For Spanish, [14]

achieved a mean WER of 18.9%, while [12] managed to decrease it up to 10% by restricting the recognition domain.

The vast majority of the previous studies consider classifying, labeling and structuring the acoustic signal into homogeneous segments essential to optimise the training of the acoustic models of the recogniser, for its subsequent proper operation [15].

C. Discourse Segments Detection (DSD)

The most widely investigated two sources of information to resolve the problem of detecting discursive segment boundaries are word transcriptions (what the speakers say) and prosody (how they say it). It is common to use two statistical models: language and prosodic models. In general, the language model gives the probability of a segmental boundary occurring in a context, while the prosodic model expresses the relationship between prosodic features and segmental boundaries.

Most previous works on discourse segment detection, e.g. [16], are based on the combination of these two information sources. [17] presents a system for punctuation generation which combines both prosodic and linguistic information, in addition to acoustic models. [18] and [19] use a general HMM framework that allows the combination of lexical and prosodic information to recover punctuation marks. A similar approach was used to detect sentence boundaries in [20] and [21].

D. Speaker Diarization (SD)

The varied and wide applicability of speaker diarization technology has led different research groups to develop several systems. The SD process often consists of three main phases: front-end acoustic processing, initial segmentation and final speaker clustering and refinement. The pre-processing step has two main goals. The first one is to normalise the signal in order to remove corrupting noise. In [22], for example, Wiener filtering is applied on each audio channel with that purpose. The second one is to parameterise the signal. Mel Frequency Cepstum Coefficients (MFCC) and Linear Frequency Cepstrum Coefficients (LFCC) are commonly used parameter features, in vectors of several dimensions which often include deltas and/or deltas-deltas.

The initial segmentation phase aims to provide an approximate speaker turn labeling to initialise and speed-up the subsequent segmentation and clustering stages. Several distance criterions can be used in this step. While [23] applies a classical GLR speaker turn detection criteria, [24] uses a segmentation similar to the KL2 metric, measuring the maxima of a local Gaussian divergence between two adjacent sliding windows of five seconds.

The most common clustering method employed in the speaker clustering and refinement phase is the Bayesian Information Criteria (BIC) or a variation called ΔBIC [22, 24]. Initial clusters are generally modelled by single Gaussians with full covariance matrices estimated on the acoustic frames of each segment output by the initial segmentation step. The BIC or ΔBIC metrics are commonly used both, to measure inter-cluster distances and as stop criterions.

III. Resources, tools and applications developed

The main tools integrated in APyCA are: (1) a VAD module; (2) a large vocabulary continuous speech recognition module for recognition and alignment and modules for (3) the detection of discursive segment boundaries and (4) speaker diarization.

A. Voice Activity Detection (VAD)

In order to feed the speech recogniser with audio segments containing speech, a previous segmentation and classification of the audio signal is required. This classification should be as comprehensive as possible, to ensure that no misclassified speech segments are lost.

APyCA segments and classifies the input audio into four different acoustic types: speech, speech plus noise, noise and silence, based on the speech detection functionality of the open source LIUM_SpkDiarization tool [25]. Such segmentation is obtained through Viterbi decoding of one-state Hidden Markov Models (HMMs) trained for the different acoustic conditions on the ESTER broadcast news corpus.

B. Automatic Speech Recognition (ASR) and alignment

APyCA employs the Windows Speech Recogniser (WSR) 8.0 as its ASR engine, integrated through the SAPI 5.3 functionality on the .NET Framework 3.5 environment.

In order to improve its performance, default models have been adapted with acoustically similar (i.e. clean speech vs. noisy speech) and/or TV genre-specific data by feeding the system with the corresponding audio recordings and text transcripts.

As well as for generating textual transcriptions of the spoken information, the ASR module is also used to obtain word-level time-stamps to align the audio and the text. In those cases where transcriptions already exist and the recognition step is not required, audio and text synchronization is computed by an alignment module developed using the HTK Toolkit [26]. The alignment module is a monophone recogniser trained on the Albayzin corpus [27], which extracts 39-dimensional feature vectors containing MFCC, delta and delta-delta coefficients on 25ms windows every 10ms and uses the Spanish version of SAMPA as its phoneme set, plus silence and short pause models. Each monophone (except from the short pause model) consists of non-emitting start and end states plus three emitting states, connected left-to-right with no skips and modelled by a single Gaussian. Viterbi is used for decoding.

C. Discourse Segment Detection (DSD)

Any subtitling platform integrating a speech recognition engine requires the development of algorithms for the automatic segmentation of the recognised output into dircursive segments.

APyCA has four different ways to automatically predict discourse segment boundaries: two of them are related to the acoustic and prosodic processing of the speech signal, another one is based on the linguistic analysis of the transcribed text and the last one combines the previous three approaches. The different techniques employed are presented in more detail in the following sections.

1) DSD based on Acoustic Information

Acoustic pauses are detected by analysing word start and end time-stamps produced by the recogniser or the alignment module during the recognition and alignment processes respectively. Whatever the difference, any non-coincidence in time between the end of a word and the start of the next has been taken as a potential acoustic pause.

It is important to emphasise at this point that even if acoustic pauses do not always correspond to discursive breaks, their relationship is evident in many cases.

2) DSD based on Prosodic Information

Acoustic pauses are not always grammatically correct as they may coincide with breathings, stops or speech difluencies that are not always related to true discursive boundaries. Discourse segment detection based on prosodic information can help resolve this problem.

The implemented algorithm detects discursive segment boundaries based on CART classifiers trained with the Waikato Environment for Knowledge Analysis (WEKA) tool [28]. Three different classes have been used: "silence", "question" and "nothing" - corresponding to the cases where a word is followed by silence, question mark or nothing, respectively. Each class is trained on prosodic features extracted for each word of the Multext Prosody corpus [29] using the Purdue Prosodic Feature Extraction Tool (PPFE) [30]. 232 prosodic features are extracted around each word. These features are mainly related to:

•***Duration:*** the duration and normalised duration of each word and word boundary are extracted. In addition, the duration and normalised duration of the last vowel and rhyme before a word boundary are also measured.

•***Pitch:*** several different types of F0 features are computed, based on the stylized pitch contour.

◦ *Range features:* these include the minimum, maximum, mean, and last F0 values of each word and reflect its pitch range.

◦ *Movement features:* measure the movement of the F0 contour within the voiced regions of the words preceding and following a boundary. The minimum, maximum, mean, first and last stylized F0 values of each word are computed and compared to those of the following word, using log differences and ratios.

◦ *Slope features:* the last slope value of a word preceding a boundary and the first slope value of a word following a boundary are also calculated.

•***Energy:*** similar to the F0 features, a variety of energy related range features, movement features, and slope features are computed, using various normalization methods.

Each word in the training corpus was manually labeled to belong to one of the three classes defined above.

3) DSD based on Linguistic Information

The linguistic algorithm has the same purpose as the prosodic and acoustic algorithms, i.e. estimating discourse segmentations of the transcribed text. The philosophy used to develop this module has been based on two types of heuristics: grammatical and structural.

On the one hand, a probabilistic part-of-speech (PoS) tagger based on Hidden Markov Models (HMM) has been developed in order to grammatically categorise each word. It has been trained on a proprietary lexical database that includes thousands of grammatical categories of words. In addition, heuristic rules have been developed to detect combinations of grammatical categories, before or after which it is more likely to have segment discourse boundaries.

On the other hand, the most frequent and meaningful structural elements present in the Multext Prosody corpus before or after which it is highly likely to have a discourse segment boundary have been identified. Based on this information, heuristic rules to detect discursive boundaries have been designed.

The latter approach has been found to be more robust than the former one in the automatic subtitling scenario, since recognition errors can lead to grammatical miscategorisations which weaken the designed grammatical heuristic rules.

4) DSD based on Combined Information

The global APyCA DSD system is modular in nature. This means that the input text can be independently segmented into discourse segments using any of the modules designed: acoustic, prosodic and/or linguistic.

A combined model has also been developed, which takes the predictions and confidence measures provided by the three modules described above and gives a final result based on their weighted combination. In general, if two modules detect a pause in a word boundary with enough confidence, we take it as a real pause. This approach exploits the complementarity of the three very different sources of information used for the detection of discourse segments.

D. Speaker Diarization (SD)

This module aims at segmenting the acoustic signal according to the speaker identities, so that each speaker can be assigned a different subtitle colour.

APyCA uses the LIUM_SpkDiarization open source tool [25] to solve the task of speaker diarization. Signals are parameterised using 13 Mel Frequency Cepstral Coefficients (MFCC) including coefficient C0 as energy, computed with the Sphinx 4 tools [31]. 20 ms windows are employed with an overlap of 10 ms. Cepstral Mean Normalisation (CMN) is

not applied, due to its tendency to increase the error rate of the diarization task.

The diarization process consists of three main phases. Instantatenous signal change points corresponding to segment boundaries are detected first, using distance-based segmentation metrics which combine the Generalised Likelihood Ratio (GLR) and Bayesian Information Criterion (BIC). GLR is computed using full covariance Gaussians estimated on sliding windows of five seconds and followed by a second ΔBIC pass, which also uses full covariance Gaussians, to fuse consecutive segments of the same speaker.

Then, nonadjacent segments of the same nature and speaker are brought together in clusters using a hierarchical agglomerative clustering algorithm with a stopping criterion based on the ΔBIC metric.

Finally, Viterbi decoding is performed to generate improved segmentations. Each cluster is modeled by a one-state HMM, represented by a GMM with 8 components and a diagonal covariance matrix learned by Expectation-Maximization Maximum-Likelihood (EM-ML) over the segments of the cluster. The log-penalty between two HMMs is fixed experimentally.

IV. Integration of Components into a Demo Application. Description of the Prototype

APyCA is a prototype oriented towards the automatic transcription of TV content in Spanish which integrates the technologies described in the previous sections of the paper and aims to serve the professional subtitlers as a tool to facilitate the creation and editing of subtitles. The following sections describe the main features and architecture of the developed demo application.

A. Features

Its input is TV content in the form of video or audio. It supports many different formats, including the main standards used by television producers, such as *mpeg2, h.264, aac or wav.*

Its output is a well-formed STL (binary) or SRT subtitle file, which respects the maximum number of characters allowed per line and includes colours to differentiate speakers. Time-spotting is based on the estimated word time-stamps and discourse segment boundaries. If needed, these subtitle files can be easily edited further using commercial software for subtitle generation, e.g. WinCAPS, FAB Teletext and Subtitling, Subtitle Workshop, etc.

The technologies involved have been grouped into three automatic functionalities: transcription, time-spotting and

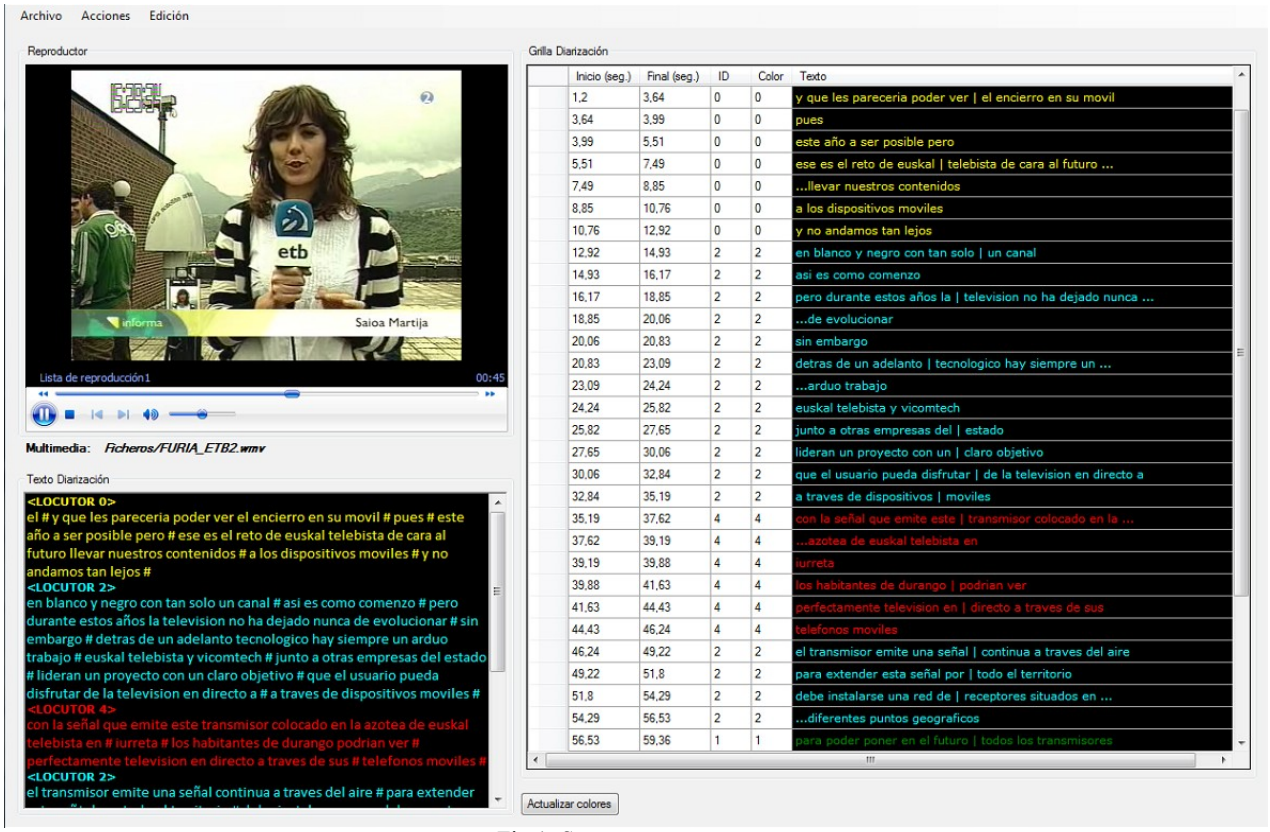

Fig 1: System screen capture

speaker diarization – which can be applied and edited independently through dedicated graphical user interfaces.

•The **Automatic Transcription** screen shows the raw text returned by the ASR engine for those segments that contain speech. It also allows playback and editing of the transcriptions, so that subtitlers can manually correct the errors of the recogniser. The fewer the transcription mistakes, the better the time-spotting will be.

•The **Automatic Time-Spotting** functionality chunks the transcribed text into discoursive segments and aligns them with the audio. The start times, end times and text of each subtitle can be edited manually.

•The **Speaker Diarization** screen automatically assigns different colours to the subtitles spoken by different speakers, also allowing their manual edition.

These three functionalities can be combined to suit the needs of the professional subtitlers. It is possible, for example, to skip the automatic transcription step and upload already transcribed audiovisual content. Or to generate the subtitle files without speaker diarization information.

The prototype has been developed entirely using the Microsoft .NET platform, the C# programming language and several Perl scripts for text processing.

A screen capture of the system is shown in Fig. 1.

B. *Architecture*

Fig. 2 illustrates how the different modules interact within the system and with the user.

The system supports the input of TV content in video or audio formats, as well as with or without its corresponding textual transcription. The FFmpeg [32] tool is used to extract the audio from the video in different formats and configurations. If the transcription does not exist, the audio will be recognised. If the transcription exists, forced alignment will be applied instead to obtain word time-stamps. In any case, time-stamps are required for the discourse segment detection and speaker diarization modules. Output subtitle files can be generated after recognition/alignment, after discourse segment detection or after speaker diarization.

The modular architecture of the system will allow simple integration of additional modules providing new functionalities in the future.

V. Evaluation

A. *Data*

The main modules of the APyCA prototype have been evaluated individually on two different genres of TV content: weather forecasts and political interviews.

Weather forecasts do not present difficulties related to the the spoken environment, the type of speech used, the number of speakers involved or the quantity and quality of their interventions. In fact, they contain just one anchor presenter following a previously written and rehearsed script in a noise-free recording studio. However, the employed vocabulary is very specific of the meteorological domain. Political interviews involve many different types of spoken environments (studio, parliament, street), types of speakers and speech (presenters following pre-prepared scripts and/or spontaneous interviewees) and a very domain specific vocabulary, with many mentions to names of politicians.

Ten programs of each type were recorded and used for training (8) and testing (2) the different modules of the system. Their reference transcripts were obtained by manually correcting the transcriptions output by the WSR 8.0 with default models.

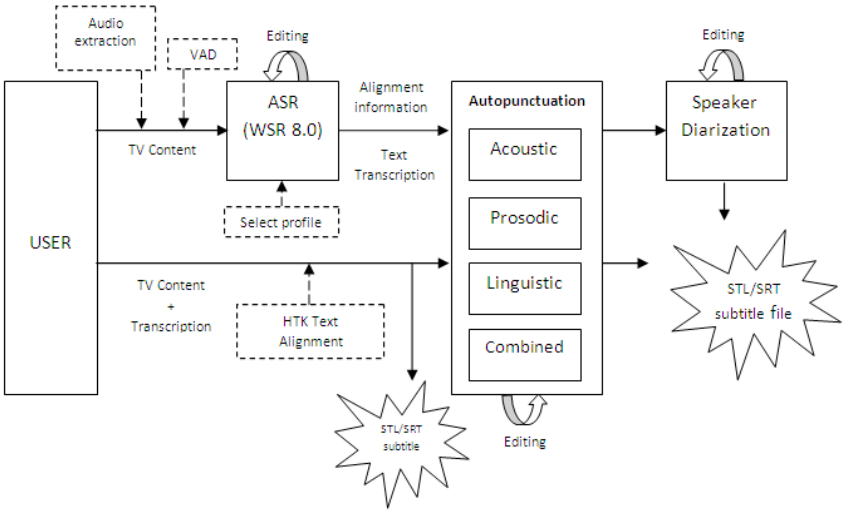

Fig 2: System architecture

B. Automatic Speech Recognition (ASR)

The performance of the WSR 8.0 recogniser was tested in three different conditions: (i) using the default recogniser models, (ii) using clean and noisy speech profiles adapted to the clean and noisy acoustic conditions found in each corpus, and (iii) using TV genre-specific profiles trained for the weather forecast and political interview domains.

Results for *the weather forecast* corpus are shown in Table 1. The average percentage of words correctly recognised overall is especially promising. The column labeled *Baseline* shows the performance of the default profile of the commercial WSR 8.0 engine. The column labeled *TV-genre profile* corresponds to the recognition rate achieved using a profile trained with all the training content of the weather forecast corpus. The *Acoustic profiles* column shows the results obtained after applying the recognition profiles trained with clean and noisy speech. Contrary to expectations, acoustic profiling does not achieve the best results probably due to the loss of context caused by the more detailed audio segmentation involved.

TABLE I.
AVERAGE RECOGNITION RATE IN THE WEATHER FORECAST CORPUS

Baseline	***TV-genre profile***	***Acoustic profiles***
81.3 %	96.65 %	92.34 %

Results for the *political interview* corpus are shown in Table 2. Less satisfactory recognition rates were obtained with this corpus overall, due to the inherent difficulty of the content type. It is remarkable that the application of acoustic profiling did not improve the results obtained by the default recognition profiles. On the other hand, TV-genre profiling only improves baseline results slightly.

TABLE II.
AVERAGE RECOGNITION RATE IN THE POLITICAL INTERVIEW CORPUS

Baseline	***TV-genre profile***	***Acoustic profiles***
79.54 %	79.80 %	78.60 %

C. Discourse segment detection (DSD)

Results concerning the evaluation of the different DSD modules are shown in Tables III, IV and V.

Each module was evaluated individually, against manually labeled reference test files. Acoustic labels take into account acoustic silences and short pauses. Prosodic labels are based on the intonation of the related sound files. Linguistic labels consider syntactic information of the associated text.

According to the followed evaluation methodology, "*Matching breaks*" refers to the percentage of breaks that match the reference file, while "*Unassigned breaks*" relates to the percentage of breaks present in the reference labels which have not been assigned by the different modules. The percentage of extra breaks assigned by the DSD modules that do not appear in the reference files is counted as "*Extra breaks*".

TABLE III.
RESULTS OF THE ACOUSTIC MODULE

Matching breaks	***Unassigned breaks***	***Extra breaks***
92.67 %	7.33 %	16.60 %

TABLE IV.
RESULTS OF THE PROSODIC MODULE

Matching breaks	***Unassigned breaks***	***Extra breaks***
64.49 %	35.51 %	63.64 %

TABLE V.
RESULTS OF THE LINGUISTIC MODULE

Matching breaks	***Unassigned breaks***	***Extra breaks***
51.92 %	48.08%	1.44 %

Results show that in 92.67% of the cases, acoustic segmental boundaries were assigned correctly, 7.33% of the acoustic pauses were not detected and 16.80% were wrongly assigned, particularly those matching breathing stops and speech disfluencies. Spontaneous speech was the main enemy of the prosodic module, mainly trained under a database of read speech. Nevertheless, it achieved a non negligible 64.49% accuracy rate. As for the linguistic module, its performance was penalised by the recognition errors which affect the designed heuristic rules. Overall, the acoustic module has proved to be the most efficient to detect discourse segment boundaries, due to its high speed and hit rate.

D. Speaker Diarization (SD)

The speaker diarization module achieved very good performance. Even in the rich acoustic environment of the political interview corpus, results achieved 87% success rate. Errors were mainly due to background acoustic changes, which caused the same speaker to be classified as two in some cases where the background acoustic environment was different, since the BIC criterion employed in APyCA for speaker diarization was actually designed to classify those segments as different.

VI. CONCLUSIONS AND FURTHER WORK

Voice activity detection, automatic speech recognition and alignment, discourse segments detection and speaker diarization technologies have been developed, customized and integrated in a prototype to support the subtitle generation process of Spanish TV content.

Objective evalutation of the different modules has shown that the proposed approach is feasible and applicable to generate automatically time-coded and colour-assigned draft transcriptions for post-editing. The commercial WSR 8.0 engine has shown adequate performance for the task. Adaptation of the default profiles to each TV-genre has shown to improve recognition accuracy. However, transcription performance degrades overall as the input speech becomes more noisy and/or spontaneous. Acoustic discourse segment detection has been found to be very efficient in terms of high speed and hit rate for time-spotting. The LIUM_SpkDiariza-

tion tool has also shown good results in the colour assignment task.

However, there is still quite a lot of room for improvement. Techniques to enhance ASR accuracy in noisy and/or spontaneous environments could be integrated. The prosodic DSD module could be trained on spontaneous and/or emotional speech corpora to better match intonation patterns of certain TV content. Finally, feature normalization techniques could be added to the speaker diarization module to obtain one-to-one relationships between clusters and speakers.
Although some positive informal usability tests have been done with professional subtitlers, a more comprehensive assesment should be carried out in order to verify the feasibility and quantify the time and money savings which could be provided by a software tool similar to the developed prototype.

Acknowledgment

This work has been partially funded by the Basque Government. The authors would like to thank Mixer, Irusoin and ETB for providing the corpora and giving usability feedback.

References

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

[2] M. Carroll, “Subtitling: Changing standards for new media? LISA Newsletter Global Insider, XIII, 3.5. 2004. http://www.lisa.org/globalizationinsider/2004/09/subtitling_chan.htm,”

[3] L. Bowker, Computer-aided Translation Technology: A Practical Introduction, Ottawa: University of Ottawa Press, 2002.

[4] J.L. Shen, J.W. Hung, and L.S. Lee, “Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Environments”, Proc. Int. Conf. Spoken Language Process., paper 0232, 1998.

[5] I.D. Lee, H.P. Stern, S.A. Mahmoud, “A Voice Activity Detection Algorithm for Communication Systems with Dynamically Varying Background Acoustic Noises,” Proc. Veh. Technol. Conf., 1998.

[6] J. Sohn, N.S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection”, IEEE Signal Process. Lett., vol. 6, no. 1, pp. 1-3, 1999.

[7] A. Davis, S. Nordholm, R. Togneri, “Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold” , IEEE Trans. on Signal Proc., vol 14, no 2, pp. 412-424, 2006.

[8] J.S. Garofolo, J.G. Fiscus, W.M. Fisher, “Design and preparation of the 1996 hub-4 broadcast news benchmark test corpora,” in Proceedings of the DARPA Speech Recognition Workshop., pp. 15–21, 1997.

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

[10] H. Meinedo, D. Caseiro, J. Neto, I. Trancoso. “AUDIMUS.MEDIA: a broadcast news speech recognition system for the European Portuguese language”. In Proceedings of PROPOR 2003, Portugal, 2003.

[11] D. Baum, B. Samlowski, T. Winkler, R. Bardeli, Schneider: “DiSCo - a speaker and speech recognition evaluation corpus for challenging problems in the broadcast domain”. Proceedings of the GSCL Symposium 'Sprachtechnologie und eHumanities' 2009.

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

[13] C. Gollan, H. Ney, “Towards automatic learning in LVCSR: Rapid development of a Persian broadcast transcription system,” Interspeech' 08.

[14] F. Batista, I. Trancoso, N. J. Mamede. "Comparing Automatic Rich Transcription for Portuguese, Spanish and English Broadcast News". In Automatic Speech Recognition and Understanding Workshop, 2009.

[15] J.-L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. de Kercadio, “The Limsi SDR systemfor TREC-9,” in Proc. 9th Text Retrieval Conference, TREC-9, pp. 335–341, Gaithersburg, Md, USA, 2000.

[16] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, B. Peskin, and M. Harper. “The ICSI-SRI-UW Metadata Extraction System”. ICSLP 2004, International Conf. on Spoken Language Processing, Korea. 2004.

[17] J.H. Yim. “Named Entity Recognition from Speech and Its Use in the Generation of Enhanced Speech Recognition Output”. Darwin College, University of Cambridge and Cambridge University Engineering Department. 2001.

[18] H. Christensen, Y. Gotoh, and S. Renals, “Punctuation annotation using statistical prosody models,” in Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 35–40, 2001.

[19] J. Kim, P. C. Woodland, “The use of prosody in a combined system for punctuation generation and speech recognition,” Proc. Eurospeech' 01.

[20] Y. Gotoh and S. Renals, “Sentence boundary detection in broadcast speech transcripts,” in Proc. of the ISCA Workshop: ASR-2000.

[21] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “Prosody based automatic segmentation of speech into sentences and topics,” Speech Communications, vol. 32, no. 1-2, pp. 127–154, 2000.

[22] T. L. Nwe, H. Sun, H. Li, S. Rahardja, "Speaker Diarization in Meeting Audio", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, April 19-24, 2009.

[23] J. Huang, E. Marcheret, K. Visewswariah, G. Potamianos, “The IBM RT07 Evaluation Systems for Speaker Diarization on Lecture Meetings”, in Multimodal Technologies for Perception of Humans, Springer, 2008.

[24] C.Wooters, M. Huijbregts. “The ICSI RT07s Speaker Diarization System”. In Rich Transcription 2007 Meeting Recognition Workshop.

[25] S. Meignier, T. Merlin. “LIUM_SpkDiarization: An Open Source Toolkit For Diarization”. CMU Sphinx Workshop 2010, Dallas, 2010.

[26] Hidden Markov Model Toolkit (HTK) 3.2, Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/, 2002.

[27] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio: “Development of Spanish Corpora for Speech Research (Albayzin)”. Workshop on International Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods, Italy, 199.1

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. “The WEKA Data Mining Software: An Update”; SIGKDD Explorations, Volume 11, Issue 1. 2009.

[29] E. Campione, (Ed.) Multext-Prosody. A multilingual prosodic database. CD-ROM Distributed by ELRA/ELDA. 1999.

[30] Z. Huang, L. Chen, M. Harper. “Purdue Prosodic Feature Extraction Toolkit on Praat”. Spoken Language Processing Lab, Purdue University. 2006.

[31] Sphinx-4. “A speech recognizer written entirely in the Java programming language”. http://cmusphinx.sourceforge.net/sphinx4/

[32] FFmpeg. “A complete, cross-platform solution to record, convert and stream audio and video”. http://www.ffmpeg.org/