Automatic Subtitling of the Basque Parliament Plenary Sessions Videos

Germán Bordel, Silvia Nieto, Mikel Penagarikano,
Luis Javier Rodríguez-Fuentes, Amparo Varona

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
german.bordel@ehu.es

Abstract

Subtitling of video contents offered in the web by Spanish administration agencies is required by law for allowing people with hearing impairments to follow them. The automatic bilingual video subtitling system described in this paper has been applied on the plenary sessions videos that the Basque Parliament posts in its web (http://www.parlamentovasco.euskolegebiltzarra.org/), and is running from September 2010. A specific characteristic of this system is the use of a simple phonetic decoder based on a joint selection of Basque and Spanish phone models, since it is not unusual for parliamentarians to make use of a mixing of the two languages. The system uses the manually transcribed Session Diaries (almost verbatim but containing some errors) as subtitles, synchronizing text and audio by means of an acoustic decoder, a multilingual orthographic-phonetic transcriber and a very-large-symbol-sequence aligner.[1]

Index Terms: Automatic Video Subtitling / Captioning, Text To Speech Alignment.

1. Introduction

An important area of activity for Speech Technology, and in particular for Automatic Speech Recognition (ASR) has to do with a social issue of great interest today: the accessibility for people with disabilities. A case in point is access to audiovisual materials for the deaf or hard of hearing. Probably, most efforts in this direction are focused on broadcasts, particularly news [1] [2], where many different approaches and tools are being explored to find an effective alternative to the hard and costly manual processing [3] [4] [5].

The speech recognition technology today is unable to solve the problem of automatic video subtitling with an acceptable quality, because these resources usually present difficult acoustics and speaker variability. Clearly, captioning might be an easier task if speech transcriptions were available, reducing the problem to synchronization, that is, to determine the proper text and audio time alignment. Nevertheless, this is not an easy task due to a couple of factors: first, the obvious way to apply ASR in this case is ”forced recognition”, but this is computationally unfeasible for signals longer than a few minutes; second, exact textual representations of audio contents are not commonly available because it would not be acceptable for captioning: speech is full of phonetic imperfections and different disfluences that captions should not display.

To cope with the first difficulty, a known approach consists in locating highly reliable intermediate segments and use them as anchors to divide the problem into smaller subproblems, relying on forced recognition when the length of the problem is short enough [6]. The second difficulty could be overcome by relaxing forced recognition to some extent [7].

Another approach consists in performing automatic recognition and then trying to find the best alignment to the text. In [8] ASR is performed at word level, and the resulting word sequence is directly aligned to the input text, so time stamping of words is straightforward. This approach requires quite high-quality ASR to have enough words to match in the alignment

In this paper, we deal with the automatic subtitling of the Basque Parliament plenary sessions videos and present a similar strategy to [8], but recognition is done at the phonetic level. This simplifies the decoding and the size of the managed set of units, which has a positive impact on the alignment procedure. Performing ASR at the phonetic level also allows to cope with the fact that many parliamentarians sometimes mix the Spanish and Basque languages.

The rest of the paper is organized as follows. Section 2 is devoted to explain the input data specificities. Section 3 explains the different aspects of the system relating to phonetics: the use of a bilingual set of phonemes, the automatic decoding and the text transcription. Section 4 focuses on the alignment of phoneme sequences. Section 5 describes the architecture of the system, where all the above mentioned components are combined. Finally, conclusions are presented in Section 6.

2. Data Issued by the Parliament

The website of the Basque Parliament is backed by a database and managed with a rigid procedure, so a premise to include subtitles in the videos was to use the materials as they are currently structured.

Due to limitations of video recording media, each video lasts a maximum of 3 hours, although the sessions are typically longer, and therefore they are recorded in two or three sections (exceptionally up to four). Recording is always done in high definition format using professional media but, for the web, videos are converted to RealMedia format (.rm), where the audio stream is downsampled to 22050 Hz, 16 bit/sample.

The Session Diary is available almost immediately as a single PDF document, because the speeches are transcribed on the fly by a team of manual transcribers. However, this document is made up of blocks associated with transcriber shifts (approximately 15 minutes per block) and there is some undefined overlap between them. These overlapped texts do not generally fully

[1] This work has been supported by the University of the Basque Country under grant GIU10/18, by the Government of the Basque Country under program SAIOTEK (project S-PE10UN87) and by the Spanish MICINN under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

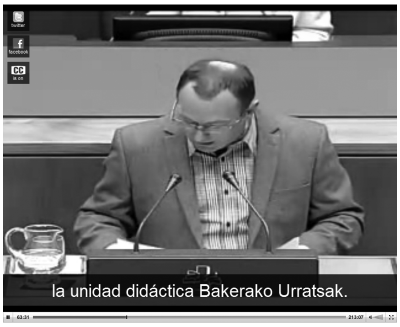

Figure 1: *A frame capture showing a mixed language close caption: "la unidad didáctica" -Spanish- (the didactic unit) "Bakerako Urratsak" -Basque- ("steps towards peace").*

match each other due to several reasons (the study of these overlaps is out of the scope of this paper, but gives some interesting hints about the difficulties of the transcription task.) Another important issue is that after each voting procedure, results are not transcribed verbatim as read by the president, but instead they are tabulated.

Starting from these two resources (video recordings and the session diary), the automatic subtitling system developed in this work generates the appropriate subtitle files (.smil and .rt). However, since including these files in the web application required some costly changes, the solution adopted by the Basque parliament was to overwrite the subtitles into the video frames.

3. Phonetic Processing

Once the peculiarities of the input data are treated to obtain an audio stream from the video and a text from the Session Diary (as tightly linked as possible to what is said in the audio), both resources must be transformed into phonetic sequences.

The application is intended to work with a language mix where speakers can switch between Spanish and Basque at any moment. Many speakers in the Basque Parliament use both languages, sometimes switching from one to the other at any point of the discourse, or even introducing isolated words from the language not used at that moment (see Figure 1). To cope with this issue, a set of phonetic units was selected to give a reasonable coverage of both languages. These units were modeled and jointly trained with the union of two databases, one for Spanish and one for Basque. As discussed below, these phonemes are used by an acoustic decoder to process the audio stream, and by a multilingual phonetic transcriber to obtain a phonetic transcription of the text .

3.1. Phonetic Inventory

Basque and Spanish share most of their phonetic units, but there are some differences. Looking for a reasonable set for both languages, and having in mind the planned use of it as a single set, we selected 26 units for Basque and 23 units for Spanish (see Table 1) which implies that just one foreign sound was added to Basque (θ in IPA coding) and four to Spanish (ʃ, ts, ts' and s' in IPA coding).

3.2. Acoustic Processing

The audio streams are converted to PCM and resampled at 16 KHz, 16 bit per sample. The resulting signals are processed to obtain a feature vector every 10 ms using a 25 ms Hamming window, first order preemphasis (0.97 coefficient), and a 26-channel Mel-scale filterbank, resulting 39-dimensional feature vectors, composed by 12-order Mel Frequency Cepstral Coefficients (MFCC) plus energy, and their delta and delta-delta coefficients.

Table 1: *Phoneme inventory for Spanish+Basque with examples. IPA coding is shown (Unicode), as well as our nearly readable 1-char coding (GTTS-ASCII). A numeric physiological codification (first column) is used by the phonetic transcriber as internal coding. The set is 27 units long (22 common units, 1 Spanish-only unit, and 4 Basque-only units).*

Physio CODE	Computational coding		Spanish		Euskera	
	IPA Unicode (HEX)	GTTS ASCII	Orthogr.	Example	Orthogr.	Example
111	i (0069)	i	i	pico	i	ipar
115	u (0075)	u	u	duro	u	umore
132	e (0065)	e	e	pero	e	hemen
135	o (006F)	o	o	toro	o	hori
173	a (0061)	a	a	valle	a	kale
21112	m (006D)	m	m	madre	m	ama
21142	n (006E)	n	n	nunca	n	neska
21172	ɲ (0272)	N	ñ	año	in	arraina
21211	p (0070)	p	p	padre	p	apeza
21212	b (0062)	b	b v	bolsa vino	b	begia
21241	t (0074)	t	t	tomo	t	etorri
21242	d (0064)	d	d	dónde	d	denda
21281	k (006B)	k	c qu k	casa queso kilo	k	ekarri
21282	g (0067)	g	g	gata	g	gaia
21321	f (0066)	f	f	fácil	f	afaria
21331	θ (03B8)	z	c z	cinco paz	--	--
21341	s (0073)	s	s	sala	s	hasi
21351	ʃ (0283)	x	--	--	x	xoxoa
21381	x (0078)	j	j	mujer	j	ijito
21624	r (0072)	R	r rr	rosa torre	rr	arrunta
21742	ɾ (027E)	r	r	puro	r	dirua
21942	l (006C)	l	l	lejos	l	lana
243	tʃ (02A7)	X	ch	mucho	tx	txikia
244	dʒ (02A4)	y	i y	hielo cónyuge	i dd	leoia onddo
24111	ts' (02A6 02BC)	C	--	--	tz	atzo
24122	ts (02A6)	S	--	--	ts	mahatsa
21342	s' (0073 02BC)	c	--	--	z	zoroa

The acoustic modeling is based on left-to-right non-contextual continuous Hidden Markov Models with three looped states and 64 Gaussian distributions per state. The system allows the use of either HTK [9] or Sautrela [10] as decoder. The material used to train the phonetic models was the union of the Albayzin [11] and Aditu [12] databases. Albayzin consists of 6800 read sentences in Spanish from 204 speakers, and Aditu is composed by 8298 sentences in Basque from 233 speakers.

Using these databases to train the HMM phonetic models, taking 4800/2000 and 5346/2952 sentences as train/test for Albayzin and Aditu respectively, and considering specific models for each language (23 for Spanish and 26 for Basque) we obtain the accuracy shown in the first column of Table 2. If we use the whole 27 units set, we obtain the accuracy shown in the second column. Obviously there is an accuracy loss due to the double effect of using a wider set of units and introducing out-of-language phonemes.

Table 2: *Loss in phonetic recognition rates due to the introduction of cross-language phonetic models is shown in the last column. Applied to Spanish (4 external phonemes) the recognition rate reduces by nearly 3.25 percentual points. And applied to Basque (1 external phoneme) the recognition rate reduces by nearly 1 percentual point.*

Test/Phone set	Specific	spa+eus (27 units)
Albyzin (spa)	80.69% (23 units)	77.43%
Aditu (eus)	83.94% (26 units)	82.88%

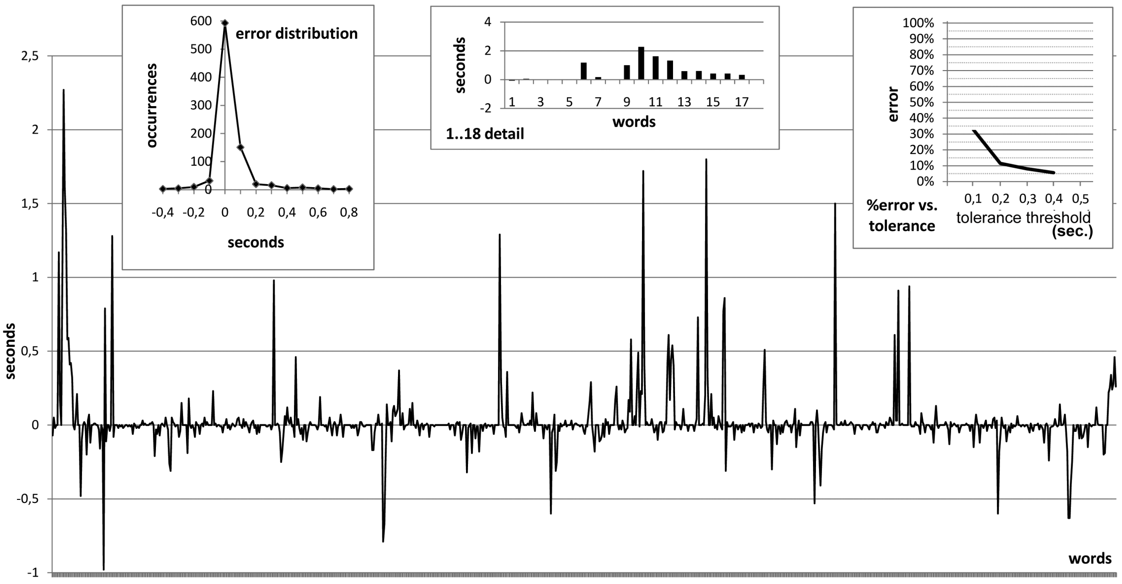

Figure 2: *Alignment performance on a speech segment including 876 words. The Y-axis shows the time deviation from a manual reference. The left-top figure shows the deviation distribution, which is intentionally skewed to positive values because it is better tolerated when watching captioned videos. The center-top figure shows the worst case, which occurs just at the beginning of the fragment and deviates up to 2 seconds. The right-top figure shows the error rate considering different thresholds for what could be regarded as a tolerable deviation.*

Table 3: *Phonetic recognition rates for a Plenary Session (5 hours long) using the original transcriptions and filtered transcriptions (discarding segments where the Session Diary is not correct).*

	Whole transcriptions	Correct transcriptions
Acc	61.27%	62.55%

Even though these accuracy figures (about 80%) correspond to an open test, when the mixed set of 27 units set is trained with all the materials in Albayzin plus Aditu, and applied to decode the audio of the plenary sessions, they yield recognition rates of about 60% due to task differences, background and speakers mismatch, etc. Table 3 shows recognition rates for one plenary session (lasting approximately 5 hours) taking into account the manual transcription as reference (61,27%) and discarding segments where the transcription does not exactly reflect the audio track (62,55%). We will see later that the subtitling task has proved to be not very sensitive to the recognition rate, so these figures are enough to obtain good alignment results and a more specific training is not necessary.

3.3. Phonetic Transcription

In parallel with the audio signal processing, the input text is translated to a phonetic representation using a multilingual phonetic transcriber. The output representation also allows recovering the original stream, which will be needed at the end of the process. This transcriber uses an internal coding for the phonemes that we call ”physio code” because it relates directly to the physiology of the production of each phoneme. Externally, this coding is mapped to IPA, SAMPA, etc. Since Spanish phonetics is very close to its orthography, we use a highly readable specific coding (GTTS).

The transcriber is able to work with any languages just by defining a specific module for that language. These modules are composed by tree elements: a dictionary of transcriptions, a rule-based transcriber and a number-to-text converter. For each word, the multilingual transcriber makes decisions about the corresponding language based on the response that each subsystem gives about having the word in its dictionary and, if necessary, considering the context. When a word is not in any dictionary, the most suitable language subsystem is called to provide a rule-based transcription in order to always produce an output, and a warning is issued to allow the continuous improvement of the dictionary and the rules. Usually rule-based transcriptions are correct, and each time a session is processed, the current transcriber generates a short number of warnings.

4. Phonetic Alignment

The alignment module receives two phonetic streams carrying information about words and times. The phonetic streams are isolated to be treated as symbol sequences by a kernel module implementing a variant of the Needleman-Wunsch algorithm [13], that finds the best global alignment between both sequences. The mentioned variation allows working with very large sequences, having a length of about 300.000 symbols in average. The consideration of ”best alignment” depends on the set of weights assigned to matches, substitutions, deletions and insertions. After some experimentation, the simple Levenshtein distance proved to give the best results. The algorithm can be configured to behave in different ways when there are more than one equivalent partial alignment. This was done so that the mismatched segments had a tendency to show the text before the audio because this is better tolerated by users when watching captioned videos.

Figure 2 shows, as a result from this alignment, a segment of 876 words where, for each word, the deviation from reference timestamps can be seen at the Y axis. Errors are distributed randomly along the time dimension and most of them are small enough to be negligible for the subtitling task. Table 4 shows the alignment error rate depending on a threshold value considered as a tolerable deviation. 95% of the words showed less than 0.5 seconds deviation, and nearly 70% less than 0.1 seconds.

5. System Architecture

Figure 3 shows the integration of the modules described above. There is a first block representing the specific treatment applied

Table 4: *Error rate considering different tolerance thresholds (in seconds) in the deviations from reference timestamps.*

Threshold	0.1	0.2	0.3	0.4	0.5
Error rate	32.31%	11.42%	7.99%	5.59%	4.57%

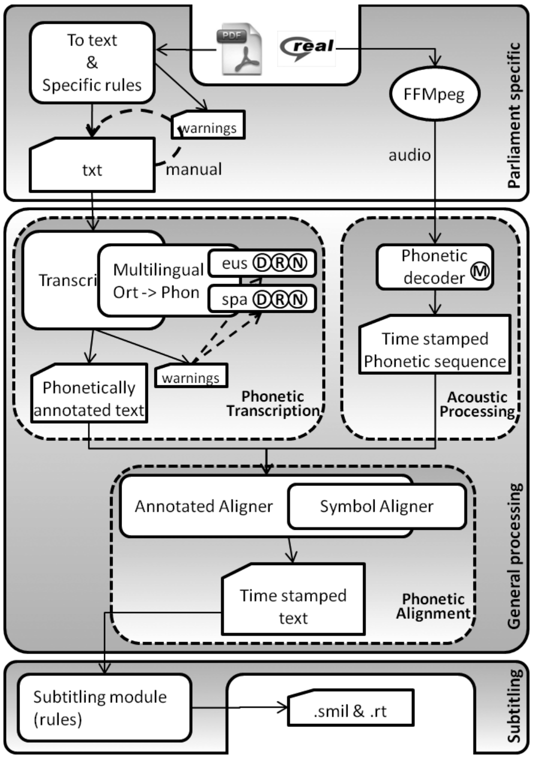

Figure 3: *System architecture.*

to the Basque Parliament materials. The audio stream is extracted from the video and converted using an off-the-shelf application (ffmpeg). The Session Diary is processed with a customized Java program that extracts and applies a set of rules to reconstruct overlaps, convert voting tables to text, etc., and warns when something goes wrong.

Once the audio and the text are ready, they are processed by the phonetic decoder described in section 3.2 and the phonetic transcriber described in section 3.3, respectively. The two phonetic streams include more information than the mere phoneme sequences, because once the aligner (described in section 4) relates both resources, the original text is reconstructed and words are marked with timestamps. This is the final result before a subtitling module exploits it to cut the whole text stream into suitable pieces for the video.

6. Conclusions

We have presented an automatic video subtitling system that is being applied since September 2010, taking the Session Diaries of the Basque Parliament as input to subtitle the video recordings of the Plenary Sessions. A single set of phonetic units is used for Spanish and Basque, to tackle the mixed use speakers make of both languages. It was the use of the aforementioned mixed language by the speakers what suggested that word-level ASR might not be suitable and just phonetic decoding should be used instead. So text an audio alignment is accomplished by means of searching the minimum edit distance after converting them to phonetic streams. Results show a random distribution of errors in the recognized phoneme sequences, which has a reduced impact on the aligner ability to match words. Subtitled videos can be seen following the links to the Plenary Sessions in: `http://www.parlamento.euskadi.net/cm_argic_dp/SDW?`

7. Acknowledgements

Authors want to acknowledge people from the Basque Parliament for their support and enthusiasm, in particular to Juan Luis Mázmela and Andoni Aia, who first started this collaboration, and also Juanjo Arruza. We also thanks the Bureau of the Basque Parliament, that approved and partially supported this initiative.

8. References

[1] A. Ortega, J. Garcia, A. Miguel, and E. Lleida, “Real-time live broadcast news subtitling system for Spanish,” in *Proceedings of the Interspeech*, Brighton, U.K., September 2009.

[2] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, “Broadcast news subtitling system in portuguese,” in *Proceedings of ICASSP*, Las Vegas, USA, 2008.

[3] G. Boulianne, F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet, and F. Osterrath, “Computer-assisted closed-captioning of live TV broadcasts in French,” in *Proceedings of the Interspeech*, Pittsburgh, USA, 2006.

[4] A. Alvarez, A. del Pozo, and A. Arruti, “Apyca: Towards the automatic subtitling of television content in Spanish,” in *IMCSIT*, Wisla, Poland, October, 18-20 2010, pp. 567–574.

[5] C. Cerisara, O. Mella, and D. Fohr, “JTrans, an open-source software for semi-automatic text-to-speech alignment,” in *Proceedings of Interspeech*, Brighton United Kingdom, 09 2009.

[6] P. Moreno, C. Joerg, J. Van Thong, and O. Glickman, “A recursive algorithm for the forced alignment of very long audio segments,” in *Proceedings of ICSLP*, 1998.

[7] P. Moreno and C. Alberti, “A factor automaton approach for the forced alignment of long speech recordings,” in *Proceedings of ICASSP*, I. C. Society, Ed., 2009, pp. 4869–4872.

[8] T. J. Hazen, “Automatic alignment and error correction of human generated transcripts for long speech recordings,” in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. O. andDave Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge, UK: CUED, Cambridge University Engineering Department, 2006.

[10] M. Penagarikano and G. Bordel, “Sautrela: A highly modular open source speech recognition framework,” in *Proceedings of the ASRU Workshop*, San Juan, Puerto Rico, December 2005, pp. 386–391.

[11] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Marino, and C. Nadeu, “Albayzin speech database: design of the phonetic corpus,” in *proceedings of Eurospeech*, Berlin, Germany, September 22-25 1993, pp. 175–178.

[12] Basque Government, “Aditu program,” 2005, (initiative to promote speech technologies in Basque). [Online]. Available: http://www.euskara.euskadi.net/r59-4572/es/contenidos/informacion/aurkezpena/es_8550/presentacion.html

[13] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443 – 453, 1970.