SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling

Arantza del Pozo[1], Carlo Aliprandi[2], Aitor Álvarez[1], Carlos Mendes[3], Joao P. Neto[3], Sérgio Paulo[3], Nicola Piccinini[2], Matteo Raffaelli[2]

[1]Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
[2]Synthema, Pisa, Italy
[3]VoiceInteraction, Lisbon, Portugal

[1]{adelpozo, aalvarez}@vicomtech.org, [2]{carlo.aliprandi, nicola.piccinini, matteo.raffaelli}@synthema.it,[3]{carlos.mendes, joao.neto, sergio.paulo}@voiceinteraction.pt

Abstract

This paper describes the data collection, annotation and sharing activities carried out within the FP7 EU-funded SAVAS project. The project aims to collect, share and reuse audiovisual language resources from broadcasters and subtitling companies to develop large vocabulary continuous speech recognisers in specific domains and new languages, with the purpose of solving the automated subtitling needs of the media industry.

Keywords: audio and text corpora, speech recognition, automatic subtitling

1. Introduction

Due to recently approved European and National directives and laws, the subtitling demand has grown fast in the past few years throughout Europe. As a result, broadcasters and subtitling companies are seeking for subtitling alternatives more productive than the traditional manual process.

Large Vocabulary Continuous Speech Recognition (LVCSR) is proving to be a useful technology for such a purpose. Respeaking is consolidating as the main subtitling technique employed for live and pre-recorded broadcast productions. Another trend in use today is the application of speech recognition to automatically generate a transcript of a programme's soundtrack without the need of a respeaker [1], and to use this as the basis of subtitles in order to increase subtitling efficiency and reduce costs. Unfortunately, the high expenses associated to the collection and annotation of the audio and text corpora required to train each LVCSR system for respeaking or automatic transcription has hindered the development of new languages and application domains.

Within the SAVAS project[1], data is being collected and annotated and LVCSR technology for automated subtitling is being developed for the languages and domains shown in Table 1.

The consortium is formed by a mix of broadcasters, subtitling companies and technology developers with expertise in the targeted languages. Given the participation of a Swiss partner, Swiss Italian, Swiss French and Swiss German data is also being collected and annotated and the development of Swiss variants of the Italian, French and German systems is being explored.

In addition, the consortium is striving to share the collected language resources for which intellectual property rights can be cleared, without compromising the business plan of the main results of the project. In this paper, we describe the data collection, annotation and sharing activities carried out within the project.

System type	Domain	Language(s)
Transcription	Broadcast news	Basque, Spanish, Italian, French and German
	Interview/debate	Portuguese
Dictation	Broadcast news	Italian, French, German
	Sports domain	Basque, Italian

Table 1. SAVAS systems, domains and languages

2. Data collection

The development of robust LVCSR systems for automatic subtitling, capable of producing transcriptions with word error rates (WER) below 20%, requires considerably large audio and text corpora for acoustic (AM) and language modeling (LM).

2.1 Data targets

Table 2 summarizes the targeted amounts of audio and text data aimed to be collected within the project.

Based in previous experience [2,3], we estimated that the development of transcription systems for automatic subtitling in new languages within the broadcast news domain would ideally require at least 200 hours of audio and one billion words of text. The same data could also be exploited to develop dictation systems in the same domain for the considered languages. On the other hand, the adaptation of an already existing transcription system for automatic subtitling to a new domain was estimated to be achievable with 20 hours of audio and 500k words. Finally, 20 hours and 500k words of audio and text were deemed enough to adapt existing transcription systems for automatic subtitling to a different dictation domain.

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

System type	Data	
	Audio for AM	**Text for LM**
Transcription/Dictation in the broadcast news domain	200 hours	1B words
Transcription in the interview/debate domain	20 hours	500k words
Dictation in the sports domain	[20 hours]	500k words

Table 2. Targeted audio and text data

For existing dictation systems with robust enough acoustic models, the text alone was estimated sufficient for adaptation.

2.2 Collected data

Most of the required audio data has been gathered from programs produced by the broadcasters in the consortium. The text sources are a mix of autocue scripts and subtitles provided by the broadcasters and subtitling companies in the consortium, plus newswire and sports text crawled from the Internet. In addition, the transcriptions of the collected audio content have also been used as text data for language modeling. Table 3 shows the final amounts of audio and text corpora collected for each language and domain.

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

Table 3. Collected audio and text corpora per language and domain

As it can be seen from the table above, most of the targeted amounts have been reached, except for Basque and Portuguese text corpora in the broadcast news, sports and interview/debate domains, respectively.

As a minority language, the availability of Basque text corpora in the news and sports domains is limited.

With Portuguese, the difficulty has been to find text resources containing the type of spoken information common in the interview/debate domain: repetitions, hesitations, disfluencies, unfinished sentences, etc. Thus, the text corpus from the conversational domain has been compiled based on the transcriptions of the corresponding 20 audio hours.

For Italian, French, German and their Swiss variants, the originally targeted amounts have been distributed according to previous experience on dialect adaptation [3].

3. Data annotation

The annotations used in the SAVAS project are composed of spoken utterance transcriptions combined with speaker turn and background noise segmentations.

3.1 Tools & methodology

Transcriber 1.5.1 [4] has been chosen as annotation tool, since it has been developed for the creation and management of speech corpora closely following the Linguistic Data Consortium's [2] annotation conventions and recommendations which SAVAS follows.

The methodology employed has aimed at making the annotation process as productive as possible, following an incremental automation approach. The first 50 hours per language have been annotated manually from scratch, with the support of autocue scripts as a basis for transcription when available. Such manual annotations have then been used to develop a set of automation tools described in Section 0, for the automatic generation of transcriptions and annotations that can be imported into Transcriber. From then on, annotators only needed to correct the errors produced by the automatic tools instead of transcribing and annotating audio content from scratch.

3.2 Quality assurance

The consistency and accuracy of the annotations has been ensured through personalized training and a centralized review methodology. An annotation core team has been established per language, responsible for training the rest of annotators and reviewing the consistency and quality of their annotations.

Training courses have been organised for annotators to learn the SAVAS annotation guidelines and carry out their first annotation tasks in a supervised manner. Each annotator's initial set of annotations were then thoroughly reviewed by the core annotation team in each language. This intensive review-and-feedback process was repeated with each annotator until the core team considered that the quality of their annotations was good and consistent. After that, core teams kept reviewing all annotations and reporting on repeated mistakes annotators may have produced.

3.3 Automation tools

3.3.1 For transcriptions

The first batch of manually annotated 50 hours was used, together with the text material available at the time, to develop full LVCSR systems for each language. Annotators started employing the output of these systems as draft timed-transcriptions to be post-edited with Transcriber. As more annotated audio and text became available, updated versions of the LVCSR systems were

[2] http://www.ldc.upenn.edu/

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	AVG	STD
P1	45	50	60	50	70	30	60	30	50	28	60	48.45	14.08
P2	32	40	50	35	65	20	45	24	40	21	48	38.18	13.72
P3	28	35	40	30	55	18	35	18	28	18	32	30.63	11.02
P4	26	30	40	25	43	16	26	18	18	16	22	25.45	9.18

Table 4. Annotators' effort across phases: P1=first annotations from scratch; P2=use of autocue scripts; P3=experienced; P4=use of automation tools

trained with more data. This kept improving transcription accuracy.

Automatic punctuation and capitalization modules were also iteratively trained with the data available in each cycle, which further improved the quality of the automatic transcriptions for post-edition.

3.3.2 For annotations

Audio segmentation and speaker diarization modules were implemented to allow the creation of draft background noise and speaker turn labels.

In addition, already annotated named entity labels were exploited to allow their automatic tagging in the remaining data.

The development of these modules was also carried out in an iterative manner, based on the annotated data available per language and cycle.

3.3.3 Achieved productivity gain

Table 4 shows the effort, measured in terms of reported average hours needed to annotate one hour of content, required by a control group of 11 annotators, A1-A11, across several languages and annotation phases.

Annotation effort has resulted to be heavily linked to each particular annotator. As shown in the table, the quickest annotator, A10, is almost three times faster than the slowest one, A5.

The numbers in the table also show that all annotators managed to reduce the required effort with automation and experience. On average, productivity increased 21% from P1 to P2, which suggests that the use of autocue scripts as draft transcriptions markedly speed up the annotation process. The experience gained between P2 and P3 further improves productivity in 16% on average, so we can conclude that the more content an annotator annotates, the more productive he/she will become. Finally, the use of automatic tools for transcription and annotation also increased productivity from P3 to P4 in 11% on average. The overall productivity gain achieved on average from P1 to P4 was thus a considerable 48%.

4. Data sharing

Most of the resources present in the existing data infrastructures and repositories such as ELDA[3], LDC[2] or META-SHARE[4] are textual (i.e. written corpuses, lexical resources and treebanks). In comparison, oral resources such as those required to build LVCSR systems are less common [5], since their collection is in general more costly.

In order to leverage the audio and text data compilation work carried out within the project, the consortium has made an effort to clear intellectual property rights (IPR) and maximize the amount of resources to be shared with the rest of the community without compromising its business plan. It is worth to note that the call[5] under which the SAVAS project is funded seeks among other things for the financed consortia, in particular SMEs, to commercially exploit project results.

A SAVAS META-SHARE repository that hosts the project data has been developed and can be accessed through http://metashare.synthema.it:8000. The legal status of the shared audiovisual resources has been cleared and their licensing foundations have been established.

Table 5 summarizes the type, amounts and license schemes of the data shared for each language. As it can be seen, all audio and text data collected from the consortium broadcasters and subtitling companies is shared. In addition to raw audio and text, two transcribed audio test sets are also shared per language. These test sets will allow other LVCSR technology developers compare the performance of their systems with that of the SAVAS engines, which we plan to publish in other relevant conferences.

Because the consortium SMEs are looking into the market exploitation of automated subtitling and transcription applications trained on the compiled annotations, these will not be made available in the repository.

In those cases in which data sources external to the consortium such as French and German audio or Internet text crawls have had to be exploited, data sharing permission has been requested to the respective owners. Unfortunately, this process has resulted highly time-consuming and little productive. More than 20 broadcasters and newspapers have been contacted following an approach based in [6]. Among those, only the main Basque newspaper, Berria[6], has agreed to clear its copyright for sharing purposes. Our experience has shown that most data owners fear undue competition and brand damaging derived from misuse of their data and, in general, rather not risk. Although data sharing negotiations with some IPR owners are still pending, we do not expect big changes from the figures reported in Table 5 by the end of the project.

The commercial license established for the SAVAS sharable resources is the META-SHARE C_NoReD_FF[7]

[3] http://www.elda.org/

[4] http://www.meta-share.eu/

[5] FP7-ICT-2011-SME-DCL

[6] http://www.berria.info/

[7] Commercial_NoReDistribution_For-a-Fee

Basque	Language Resource		Amount	Commercial license	Research license
	Audio	Broadcast news	200h	C-NoReD-FF	CC BY NC SA
		Sports	20h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	143M	C-NoReD-FF	CC BY NC SA
		Crawled news	176M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	5h+5h	C-NoReD-FF	CC BY NC SA
Spanish	Audio	Broadcast news	200 h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	178M	C-NoReD-FF	CC BY NC SA
		Crawled news	17M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	5h+5h	C-NoReD-FF	CC BY NC SA
Portuguese	Audio	Interview/debate	20h	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Italian	Audio	Broadcast news	150h	C-NoReD-FF	C-NoReD-FF
	Text	Crawled news	9M	C-NoReD-FF	C-NoReD-FF
	Transcribed audio	Test sets I and II	4h+4h	C-NoReD-FF	C-NoReD-FF
Swiss Italian	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	6M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Swiss French	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	31M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Swiss German	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	32M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA

Table 5. SAVAS META-SHARE repository data

license[8]. On the other hand, the Creative Commons CC BY NC SA license has been established mainly for research purposes. In a nutshell, language resources with the C-NoReD-FF licensing schema cannot be redistributed, require the payment of a fee and allow derivatives which can be used for commercial purposes. Resources with CC BY NC SA schemes cannot be redistributed either, require attribution, are free and allow derivatives which can only be used for non-commercial purposes and need to be shared under the same terms.

5. Conclusions and future work

This paper has described the data collection, annotation and sharing activities of the SAVAS project. A considerable amount of audio and text data has been collected for each of the targeted languages. In addition, the followed annotation methodology has managed to ensure high quality annotations and improve the productivity of the task. Finally, the consortium has worked to share the greatest number of compiled resources with the rest of the community. The compiled SAVAS META-SHARE repository can be currently considered one of the biggest available multilingual audio and text data sources exploitable for LVCSR development.

The final types, amounts and license schemes of the shared resources do not match the total collected for two main reasons. On the one hand, data can be considered a commercial asset by owners looking into its exploitation. On the other hand, data owners fear competition and damage from its misuse. We believe considerable work remains to be done to inculcate the data sharing culture among data owners. Future data collection, annotation and sharing approaches of this kind aiming to achieve higher impact in data compilation for open technology research and development should devote special efforts to negotiate and clear copyright issues with the corresponding data owners.

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

6. Acknowledgements

The authors wish to thank all the data owners who have agreed to share the resources described in this article. SAVAS is funded through the EU FP7 SME-DCL Programme (2012-2014), under grant agreement 296371.

7. References

[1] C. Aliprandi, C. Scudellari, I. Gallucci, N. Piccinini, M. Raffaelli, A. del Pozo, A. Álvarez, H. Arzelus, R. Cassaca, T. Luis, J. Neto, C. Mendes, S. Paulo and M. Viveiros, "Automatic Live Subtitling: state of the art, expectations and current trends", in Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies, Las Vegas, April 2014

[2] H. Meinedo, M. Viveiros and J. Neto, "Evaluation of a Live Broadcast News Subtitling System for Portuguese", in Proceedings of Interspeech, Brisbane, Australia, 2008.

[3] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso and J. Neto, "The L2F Broadcast News Speech Recognition System", in Proceedings of Fala, Vigo, Spain, 2010

[4] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production", Speech Communication special issue on Speech Annotation and Corpus Tools, vol. 33, no. 1-2, January 2000

[5] C. Parra, N. Bel, V. Quochi, "Survey and assessment of methods for the automatic construction of LRs", deliverable D6.1a, FlaReNEt, September 2009

[6] O. De Clercq and M. Montero Perez, "Data Collection and IPR in Multilingual Parallel Corpora. Dutch Parallel Corpus", in Proceedings of LREC, Malta, 2010