Project no. **296371**

SAVAS

'Sharing AudioVisual language resources for Automatic Subtitling'

Collaborative Project
Information and Communication Technologies

Annual Public Report 2012

Editor(s):	Arantza del Pozo
Contributor(s):	Miren Urteaga
Reviewer(s):	Consortium
Status-Version:	Final
Date:	15th November 2012

TABLE OF CONTENTS

1 Introduction

As a consequence of Article 7 of the Audiovisual Media Services Directive (AVMSD) approved by the European Parliament and Council in December 2007, member states have been taking the necessary measures to guarantee that the services of audiovisual providers under their jurisdiction are gradually accessible for persons with a visual or hearing disability.

Subtitling is one of the means to make audiovisual content accessible to the community of the deaf and hard of hearing or the ageing population. As a result of the new legal frameworks, the subtitling demand has grown fast in the past few years, to the extent that traditionally dubbing countries, such as Spain, France, Germany and Italy, as well as countries with a tradition in voice over, such as Poland and other Central and East European countries, are now also embracing subtitling.

Given the new situation, broadcasters and subtitling companies are seeking subtitling alternatives more productive than the traditional manual process to cope with the increasing subtitling demand in a cost-effective manner. Large Vocabulary Continuous Speech Recognition (LVCSR) is proving to be a useful technology for such a purpose. **Respeaking**[1] is consolidating as the main subtitling technique employed for live broadcast productions, quickly taking over traditional techniques, like stenotyping. It can also be employed to script pre-recorded programmes which can then be fed to assist subtitling applications. Another trend in use nowadays is the application of **speech recognition** to automatically generate a **transcript** of a programme's soundtrack without the need of a respeaker, and to use this as the basis of subtitles. The accuracy achieved by this technique can be good enough in bounded domains and systems of this kind are currently being employed by some broadcasters in the news domain.

Although **LVCSR** is the most powerful technology for automated subtitling, the **high cost of** its **development** has hindered its availability for many EU languages and application domains, limiting the coverage of the broadcasters' and subtitling companies' demand in the new audiovisual framework. This challenge will be addressed by the **SAVAS** project, which aims to **acquire**, **share** and **reuse audiovisual resources** of broadcasters and subtitling companies so that high-tech European ASR companies can use the shared data to **develop domain-specific LVCSRs** and/or **LVCSRs in new languages** to solve the automated subtitling needs of the audiovisual industry.

In practice, SAVAS will:

a) collect spoken and textual resources in six European languages (Basque, Spanish, Portuguese, Italian, French and German) from the broadcasters and subtitling companies acting as data providers within the consortium;

[1] a technique thanks to which a professional listens to the source audio and respeaks it, so that his/her vocal input is processed by a speech recognition engine which transcribes it, thus producing subtitles.

b) transcribe and annotate the collected corpora into a form suitable to train acoustic and language models of LVCSR systems using a combination of automatic and collaborative approaches;

c) build a local META-SHARE repository containing the collected and annotated SAVAS language resources to allow their reuse;

d) adapt and train dictation and transcription LVCSR systems with the SAVAS language resources;

e) integrate and evaluate the developed systems into several automated subtitling application scenarios in order to show the impact of audiovisual data sharing for automated subtitling.

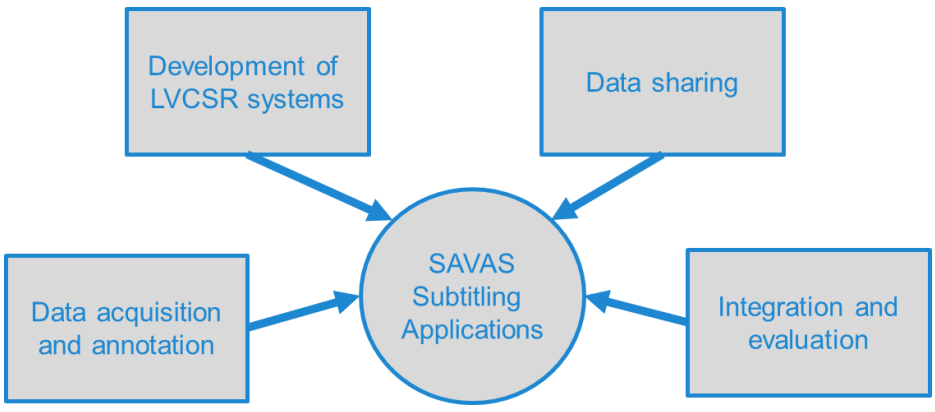
Table 1 summarizes the LVCSR systems and applications that will be developed for the different languages:

	SAVAS SUBTITLING APPLICATIONS	
	DICTATION	**TRANSCRIPTION**
BASQUE	sports	-
SPANISH	-	broadcast news
PORTUGUESE	-	interview/debate
ITALIAN	news	broadcast news
	sports	
FRENCH	news	broadcast news
GERMAN	news	broadcast news

Table 1. SAVAS Subtitling Applications

2 Summary of activities

The project is divided into the following four main activities and their supporting subtasks:



The SAVAS kick-off meeting took place in May at the premises of the European Commission in Luxembourg. For the first six months of the project, the work of the consortium has mainly been focused on data acquisition and annotation, together with the definition of the functional specifications of the LVCSR systems and the SAVAS applications to be developed.

Data acquisition and annotation. This activity is the one that will require most effort from all the partners within the project.

Data requirements and availability

The data requirements for each language depend on the type of LVCSR system to be built. However, the needed audio and text corpus for each language is considerably large. Table 2 summarizes the types of systems targeted per language and the amounts of audio and text ideally required to develop them:

	LVCSR SYSTEM TYPE		DATA REQUIREMENTS	
	Dictation	**Transcription**	**Audio**	**Text**
Basque	sports	-	200 + 20	1B+500k
Spanish	-	broadcast news	200	1B
Portuguese	-	interview / debate	20	500k
Italian	news	broadcast news	200	1B
	sports	-	-	500k
French	news	broadcast news	200	1B
German	news	broadcast news	200	1B

Table 2. Amounts of audio and text data required to develop the SAVAS LVCSR systems

The amount of data required does not only depend on the application scenario, but also on the development state of each system. While new LVCSR systems require 200 hours of audio for acoustic modelling and 1 billion words for language modelling, domain adaptation of an already existing system can be done with 20 hours and 500K words.

The type, amounts and quality of the speech and text resources made available by the consortium's broadcasters have been analysed and the ones that better suit the purposes of the project chosen. Additional sites for text crawling have also identified for each language and data collection and delivery plans defined.

The amounts and types of corpora already available to be collected for each language during the first phase of the project meet the targets, except for the amount of texts in Basque and the quality of the audio in Italian, French and German. Estimations have revealed that the quantity of Basque text material available within the consortium, plus that crawlable from the Internet, won't reach the ideally required 1B words for language modelling. In order to increase the

amount of Basque newswire text data, negotiations are ongoing to get an extra data provider involved. In addition, the audios archived by the consortium partners do not have the quality required to efficiently train the acoustic models of the Italian, French and German systems. In order to collect the necessary high quality audio material, an infrastructure to directly record newly produced broadcasts in those languages has been set up.

Annotation procedure

The availability of manual audio annotations is of primary importance for the development of the LVCSRs, since the resulting systems will be able to recognize the features contained in the annotations they have been trained with.

The annotations used in the project are composed of spoken utterance transcriptions combined with speaker turn and background noise segmentations. Utterance transcriptions are used for training robust acoustic models, as well as for building language models tailored to human daily conversations. Speaker turn segmentations provide training data for creating (gender-dependent) speaker clusters to be used in speaker diarization of audio recordings. Finally, the automatic annotation of audio contents is enriched by using background noise models derived from annotations at that level.

Because audio annotation is a very tedious and expensive task regarding time and economic resources, automatization approaches are planned to be implemented: the *Alpha* systems will be trained for each language with a first batch of manually annotated 50 hours. The output of these systems will be used as input for annotators, who will need to post-edit and correct potential errors. In the meantime, the subtitles and scripts available for the collected audios are being employed as support material for the annotation task.

Transcriber has been chosen as the annotation tool. It was developed for the creation and management of speech corpora following closely the Linguistic Data Consortium's (LDC) annotation recommendations defined for a type of data very similar to the one being compiled within SAVAS. Hence, the annotation method employed follows international conventions and is suitable for multilingual annotation.

In order to establish common annotation criteria, several annotation workshops have been carried out to teach the annotation methodology to the different annotators involved in the project.

Progress

The data collection and annotation tasks are ongoing. Table 3 shows the amounts compiled so far.

The audio acquisition targets have been reached for Portuguese and are almost complete for Basque and Spanish. Italian, French and German are a bit behind due to the extra recording efforts required.

Text collection is close to completion for Spanish and Italian. Negotiations to collect extra Basque texts are underway. The manual transcriptions of the acoustic data will be used for language modelling in Portuguese, since appropriate resources that contain the type of textual information required for the interview/debate domain have

not been identified. Crawling of newswire text for French is advancing and has just started for German.

Because the quality of the annotations will have a direct impact on the performance of the LVCSR systems, a revision procedure has been established to check the quality of the raw annotations produced by the different annotators. First, the consistency of the annotated tags is checked and the errors found are logged into a file and used to improve the labelling procedure and the training of the annotators. In addition, annotated .trs files are also corrected and quality checked by other experienced annotators.

As it can be seen in Table 3, audio annotation and correction is progressing fine. 50 hours have already been annotated and corrected for Basque, 20 for Italian, 14 for Spanish and 9 for Portuguese; besides, the annotation and correction process for French and German has just started. The annotation task for the Portuguese interview/debate content has turned out to be much more time consuming than expected. The high amount of disfluencies, repetitions, deletions, hesitations etc. characteristic of spontaneous speech present in the interview/debate type of content has raised the average time required for its annotation up to 50 hours/hour.

In addition, the first steps towards the automatization of Basque annotation are being taken: an *Alpha* system with the annotated 50 hours is being developed at the moment.

	DATA			
	COLLECTED		**ANNOTATED**	
			RAW	**QUALITY CHECKED**
	Audio	**Text**	**Audio**	**Audio**
Basque	200	150M	50	50
Spanish	188	800M	45	14
Portuguese	20	transcribing	13,5	9
Italian	98	700M	27	20
French	24	200M	2	ongoing
German	30	crawling	2	ongoing

Table 3. Data collected and annotated so far

Development of LVCSR systems. In this activity, we will develop the acoustic models, language models and lexicons of the LVCSR dictation and transcription systems in the selected languages and domains.

The development of the models for the LVCSR transcription systems will particularly be a challenge due to the number of new languages and the amount of resources involved, and because the real needs of the broadcast users must be fulfilled.

The already existing LVCSR broadcast news transcription system in Portuguese, which presently shows a good performance, will establish the base ground. The main objective is that the existing preliminary Spanish system and the new LVCSR systems to be developed for Basque, Italian, French and German will reach the same performance.

Data sharing. SAVAS aims to share the audiovisual language resources gathered within the project with other interested parties outside the consortium.

Instead of developing a proprietary platform from scratch, the consortium's approach will be to exploit an already existing data-sharing infrastructure: META-SHARE. A SAVAS META-SHARE repository will be built within the project. All the data collected will be available in such repository, with different licensing and fee arrangements according to the specificities – i.e. ownership and IPR, quantity, quality – of each particular dataset.

In general, the SAVAS consortium aims to compile the datasets, and at least, make them available only for research purposes, if IPR and exploitation plans allow it. In addition, the exploitation of the resources for commercial purposes using for-a-fee licensing schemes will also be analysed, discussed and considered in detail.

Within this activity we will deploy the infrastructure, metadata description and inventory, and legal and licensing arrangements required to share the acquired and annotated language resources through META-SHARE.

So far, a plan for managing the collection of legal and licensing issues for all the SAVAS language resources has been prepared, that details a timetable and actions for collecting, clearing and negotiating legal issues (like IPR and royalty) within the consortium.

Integration and evaluation. The consortium counts with existing commercial subtitling applications that will be adapted and extended to cover new languages and domains within the project.

SpeechTitle2.0 is a dictation solution which will be adapted to new domains for Italian, French and German. AUDIMUS.MEDIA and AUDIMUS.SERVER are automatic online and offline subtitling solutions which will also be extended to support new languages (Basque, Spanish, Italian, French and German) and domains (interview/debate for Portuguese).

The deployment plan will iterate in three-cycles fed by evaluation tests. The first set-up of the models and applications for each language will result in *Alpha* versions trained with the initial set of data available, that will be evaluated with standard metrics on initial test sets. *Beta* versions will then be trained with the full batch of data, evaluated on a new test set composed of a few hours of data as close as possible to the final testing conditions and integrated at the users' facilities. *Release* versions will include daily data adaptation and improvements derived from the end-users' feedback. A final evaluation cycle will then run comparisons of models and systems between languages in order to draw final conclusions.

3 Dissemination

A first version of the dissemination plan has been produced, in which the targeted audience and channels have been identified and various dissemination activities performed.

Broadcasters and the broadcasting industry, subtitling companies in the European market are the most relevant audience group for SAVAS. In addition, organisations and associations for disabled people will also be targeted, together with the R&D community and the general public.

SAVAS dissemination will be done through a number of different channels: website, META-SHARE, end-user panel, end-of-project showcase, dissemination material (press news, leaflets and poster) and participation to events and conferences. These are further detailed in the following subsections.

Website

The website has already been set up and is reachable via the following URL:

http://www.fp7-savas.eu/

It is intended to be a permanent information source about the project and its results and has been designed and implemented with a market-oriented look and feel, mainly devoted to dissemination from a commercial perspective.

META-SHARE

The SAVAS META-SHARE repository and the META-SHARE network (META-NET) itself will be used to create awareness of the project results, targeting a large international audience of language data and technology providers.

End-User Panel

The SAVAS End-User Panel accommodates advisors and experts from both the industrial and the scientific communities, bringing together broadcasters, subtitling companies, universities and organisations linked to the media sector, subtitling, accessibility and LVCSR technology.

Its aim is to give advice to the SAVAS development, supporting its promotion and dissemination and contributing to bridge the project results into the market. The role of the participants is to provide recommendations and feedback, to prioritise language/domain development, and to take part together with the consortium partners in the testing and evaluation of the SAVAS subtitling applications.

Members of the End-User Panel will participate in the four End-User Workshops that will be organised throughout the project lifetime. The first one is about to take place as a satellite event of the Languages & The Media conference, to be held on 21-23 November 2012 in Berlin.

End-of-Project Showcase

The purpose of the End-of-Project showcase is to give a good insight and overview of the project to the general public. It will include information about the project, presentations, videos, description of the main project outcomes and results and small demos, in the form of animations which will be made available through the project website.

Dissemination material

The SAVAS logo (see document cover and header) has been established to enable clear and easy recognition of the project. All the dissemination material and the technical material (deliverables, reports, etc.) will use the logo and will be produced according to the SAVAS branding guidelines.

A leaflet and a poster with details of the project have also been produced.

Events and conferences

SAVAS has attended the following events:

- **International Telecommunication Union (ITU) Workshop on Making Television Accessible, Tokyo** (May 2012)

  The aims of the project and the expectations of automatic subtitling were introduced to the ITU members.

- **IBC Exhibition, Amsterdam** (September 2012)

  SAVAS had a booth and gave a presentation.

- **Languages & The Media, Berlin** (November 2012) (forthcoming)

  SAVAS will be hosting its first End-User Panel Workshop as a satellite event of the conference on November 21st and will also be present at a stand.

4 Future work

The SAVAS future work will involve:

- finalizing the data acquisition and annotation task;
- developing LVCSR systems;
- creating the SAVAS META-SHARE repository;
- integrating and evaluating the SAVAS subtitling applications.

These tasks will follow the more specific time plan shown in the following diagram:

1.	Compilation and annotation of the audio and text data	May 2013
2.	Development of LVCSR systems	September 2013
3.	Creation of the SAVAS META-SHARE repository	March 2014

5 Further information

Please visit the SAVAS website at http://www.fp7-savas.eu/ for further information on the project and its progress.