Project no. **296371**

SAVAS

'Sharing AudioVisual language resources for Automatic Subtitling'

Collaborative Project
Information and Communication Technologies

Deliverables: D4.1, D4.2, D4.3 and D4.4

Acoustic models, Language models, Lexicons and Evaluation report

Due date of deliverable: 31/01/2014
Actual submission date: 31/01/2014

Start date of project: 01/05/2012 Duration: 24 Months

Contact person responsible for this deliverable: João P. Neto
Organisation name responsible for this deliverable: VOICEINTERACTION

Revision [1.0]

Project co-funded by European Commission within the Seventh Framework Programme		
Dissemination level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	X

DOCUMENT INFO

Authors

Author	Company	E-mail
João P. Neto, Sergio Paulo, Carlos Mendes	VOICEINTERACTION	Joao.Neto, sergio.paulo, carlos.mendes@voiceinteraction.pt
Arantza del Pozo	VICOMTECH	adelpozo@vicomtech.org

Document Control

Document version #	Date	Change
V0.1	28/01/2014	First draft
V0.2	30/01/2014	Revised contributions
V0.3	31/01/2014	Final conclusions

Document Data

Point of Contact	
	Name: João P. Neto Partner: VOICEINTERACTION Address: Rua Alves Redol, 9 – 1000-029 Lisboa Phone: +351 213100315 E-mail: Joao.Neto@voiceinteraction.pt

TABLE OF CONTENTS

1 Executive Summary

This deliverable is related to Work Package 4 and includes a link to the repository of the acoustic models, language models and pronunciation lexica developed for the targeted languages in the project.

In addition, it also includes a report with the results of their objective evaluation against Test Set I, compiled from a subset of the transcribed and annotated speech corpora for each language.

2 Acoustic models, language models and pronunciation lexica

Audimus is the large vocabulary speech recognition engine used in SAVAS. Like common speech recognition engines, its operation builds on a series models:

- acoustic models, used for producing posterior probabilities for all phonetic segments given each input acoustic frame
- language models, to estimate the probabilities that words occur in given contexts
- pronunciation lexica, that are responsible for mapping chains of phonetic segments into words

2.1 Acoustic models

The Audimus acoustic models are hybrid and combine the temporal modeling capabilities of Hidden Markov Models (HMMS) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs).

Their acoustic features combine posterior phone probabilities produced by the following three distinct phonetic classifiers:

1) 26 PLP (Perceptual Linear Prediction) coefficients
2) 26 Log-RASTA (log-RelAtiveSpecTrAl) coefficients
3) 28 MSG (Modulation Spectrogram) coefficients

Each MLP classifier incorporates local acoustic temporal context via an input window of 13 frames.

Acoustic models have been trained on the speech corpora described in the updated version of deliverable D2.3. Model training has been incremental. As annotated data was made available by the partners, new acoustic models have been trained leading to improvements in the recognition accuracy. Initial models were monophone-based. However, as soon as sufficient annotated data was produced for a particular language, diphone-based acoustic models have been trained, which has resulted in improved recognition accuracies.

At the current stage, diphone-based acoustic models are available for Basque, Spanish, Portuguese and Italian. The diphone versions of the French and German acoustic models are still under development, thus, only monophone models are

available for use in these languages. They can be downloaded from the following URL locations:

	URL
Basque	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/eu.ES.daily.news.bn.generic.16k.diphones.am-1.3.0-0.0295e.zip
Spanish	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/es.ES.daily.news.bn.generic.16k.diphones.am-1.1.0-4.93040.zip
Portuguese	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/pt.PT.daily.news.bn.generic.16k.diphones.am-1.1.0-2.56685.zip
Italian	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/it.IT.daily.news.bn.generic.16k.diphones.am-1.1.0-0.82f47.zip
French	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/fr.FR.daily.news.bn.generic.16k.monophones.am-1.1.0-4.a88d8.zip
German	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/de.DE.daily.news.bn.generic.16k.monophones.am-1.1.0-3.19c81.zip

Table 1. Acoustic model locations

2.2 Language models and pronunciation lexica

Each Audimus language model (LM) is a statistical 4-gram model that results from the interpolation of the following three specific LMs:

1) Backoff 4-gram LM estimated on the SAVAS text corpora, described in the updated version of D2.3;
2) 3-gram LM estimated on the manual transcriptions of the SAVAS speech corpora, also described in the updated version of D2.3;
3) 3-gram LM estimated on adaptation data composed of articles crawled from the online editions of reference newspapers in the last 7 days;

The following table presents a detailed description of the composition of the language models for each of the SAVAS languages.

Language	2-gram	3-gram	4-gram
Basque	11.5M	13.3M	5.3M
Spanish	8.0M	14.9M	9.8M
Portuguese	8.1M	11.2M	7.3M
Italian	10.4M	14.1M	9.2M

French	9.2M	12.1M	6.8M
German	10.2M	12.9M	7.5M

All technological partners provided reference pronunciation lexica for their respective languages. Such lexica were then used to automatically derive language-specific grapheme-to-phone conversion rules to generate the pronunciations of the out-of-lexicon words of the recognition vocabularies. Vocabulary size is of 100k words.

In order to optimise the recognition task, the language models were combined with the respective pronunciation lexica and are available for download at the following locations:

	URL
Basque	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/eu.ES.daily.news.bn.generic.daily.diphones.lm.LoG-1.0.0-2.0b2a9.zip
Spanish	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/es.ES.daily.news.bn.generic.daily.diphones.lm.LoG-1.0.0-25.4f937.zip
Portuguese	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/pt.PT.daily.news.bn.generic.daily.diphones.lm.LoG-1.0.0-189.e233f.zip
Italian	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/it.IT.daily.news.bn.generic.daily.diphones.lm.LoG-1.0.0-1.8ec89.zip
French	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/fr.FR.daily.news.journalistic.monophones.lm.LoG-1.0.0-6.933da.zip
German	http://services.voiceinteraction.pt/Downloads/Projects/FP7/SAVAS/repository/de.DE.daily.news.journalistic.monophones.lm.LoG-1.0.0-3.ca6b6.zip

Table 2. Language model and lexica locations

3 Evaluation report

The SAVAS speech corpus has been divided in three for each language: 80% for training, 10% for development and 10% for evaluation. The latter is referred as Test Set I, since two additional evaluation sets are being compiled to evaluate the SAVAS applications after Beta and Release integration at the broadcasters' facilities.

The performance of the developed acoustic models, language models and pronunciation lexica has been evaluated calculating perplexities (PPL) and word error rates (WER) against Test Set I for each language. The following table shows the results achieved by the recognition engines for the multiple languages.

Language	Monophones (WER)	Diphones (WER)	Perplexity
Basque	17.99%	15.79%	261
Spanish	15.79%	14.94%	88
Portuguese	21.57%	17.72%	114
Italian	21.12%	17.21%	200
French	24.35%	-	124
German	28.76%	-	242

Table 3. Evaluation results

As it can be seen from the table, not all languages have the same level of WER. The higher WER in German is justified by its agglutinative nature and by its high phonetic variability. Interestingly enough, the low phonetic variability of Basque has contributed greatly to a lower WER.

Nevertheless, performance has proven to improve across languages with the amount of training material and the exploitation of diphone acoustic models. Thus, word error rates of the final French and German systems under development are expected to decay further.

4 Conclusions

This deliverable has presented the acoustic models, language models and pronunciation lexica developed within the project so far, and the performance of the corresponding engines against a test set derived from the compiled SAVAS corpora.

System development is not complete yet. Due to delays in data annotation delivery, final acoustic models for French and German are still under training. In addition, experiments linked to optimal model combination for the Swiss versions of Italian, French and German are yet undergoing.

Nevertheless, results achieved so far present good expectations. It is our belief that the final systems will be able to deliver quality automated subtitling across languages overall.