D3.1 PROGRESS REPORT ON RICH AUDIO TRANSCRIPTION

Grant Agreement nr	611057
Project acronym	EUMSSI
Start date of project (dur.)	December 1st 2013 (36 months)
Document due Date :	M12
Actual date of delivery	December 8th 2014
Leader	LIUM
Reply to	yannick.esteve@univ-lemans.fr
Document status	Submitted

Project co-funded by ICT-7th Framework Programme from the European Commission

Project ref. no.		611057
Project acronym		EUMSSI
Project full title		Event Understanding through Multimodal Social Stream Interpretation
Document name		EUMSSI_D3.1 Progress report on rich audio transcription_20141209
Security (distribution level)		PU - Public
Contractual date of delivery		M12
Actual date of delivery		December 8th 2014
Deliverable name		D3.1. Progress report on rich audio transcription
Type		R – Report
Status		Submitted
Version number		1
Number of pages		16
WP /Task responsible		WP5/LIUM
Author(s)		Yannick Estève, Anthony Rousseau
Other contributors		
EC Project Officer		Mrs. Aleksandra WESOLOWSKA Aleksandra.WESOLOWSKA@ec.europa.eu
Abstract		Progress report on rich audio transcription: speech recognition in English and French. Architecture, training, data, performances. Transcriptions of video files.
Keywords		Speech recognition on video document.
Circulated to partners		Yes
Peer review completed		Yes
Peer-reviewed by		IDIAP
Coordinator approval		Yes

Table of Contents

1 INTRODUCTION

A general goal of the EUMSSI project is the integration of information from different social and media modalities and the investigation of their usefulness in different applications. As such, information extracted from audio files or videos may contribute to the overall data flow, providing journalists complementary visual meta-data that might not be evident to obtain from text alone, or that could be used to further filter documents retrieved from more semantic information and re-rank them to promote sufficient visual content diversity in the recommendation.

Speech processing of the EUMSSI project has to be the most accurate possible. Because of the large number of audio and audiovisual documents to process, speech processing has to be the fastest possible. These two goals are the main objectives addressed by the scientific propositions that will be developed in the EUMSSI project. To make speech processing outputs exploitable in a such project, speed and accuracy are two major issues.

This deliverable describes the automatic speech recognition (ASR) systems developed by the LIUM under the framework of the EUMSSI project. In this project, the LIUM has to develop competitive ASR systems in four European languages: English, French, German, and Spanish. System performances are related to automatic transcription accuracy conjointly to computation time.

For this first year, strong effort was produced in order to get the best ASR system possible for both English and French languages. Since the same software will be shared by the ASR systems developed for the four different languages, this effort will be capitalized in the development of ASR systems for other languages.

In order to compare our ASR system with other state-of-the-art ASR systems, and also to get an independent evaluation of our ASR systems, we have decided to participate to the ASR task of the international evaluation campaign of the International Workshop Spoken Language Translation (IWSLT) 2014. The system developed to participate to this campaign was the one used to automatically transcribe thousands of video files provided by Deutsche Welle, in the framework of the EUMSSI project. More details about these files are given in section 4.

2 ARCHITECTURE OF THE 2014 LIUM ASR SYSTEM AND ITS APPLICATION TO ENGLISH LANGUAGE

This section presents the architecture of the 2014 LIUM ASR system. It has been developed for two languages: English and French. This section focuses on the English version while the specificities of the French one are presented in section 3.1. Differences between these systems come from the acoustic and linguistic models.

First, training of acoustic and linguistic models for English is presented, then the general architecture of the ASR system. Last, experimental results are presented, in order to measure the performances of this system. Next sections present how we selected the training data for acoustic and language models, in order to get a competitive ASR system.

An ASR system is based on the search of the word sequence $\hat{W}$ such as:

$$\hat{W} = argmax_W P(W)P(X|W)$$

where:

- W is a word sequence;
- X is a sequence of acoustic observations;
- $P(W)$ the probability the word sequence W occurs in the targeted language;
- $P(X|W)$ the probability to observe the sequence of acoustic observations X when the word sequence W is uttered.

To make this computation, an ASR system needs two major kinds of models: acoustic models, which provide $P(X|W)$, and language models, which provide $P(W)$.

In order to make acoustic models more robust to variability, for instance to adapt them to speakers, two passes of acoustic decoding are usually needed in an ASR system: the first one is processed by using generic acoustic models, and the second pass analyses the automatic transcriptions provided by the first pass to adapt the acoustic models and make a new decoding process with better acoustic models. This is summarized in Figure 1 (step 1 in orange, and step 2 in red).

To build acoustic models, training data is necessary. For very competitive models, a large amount of training data is needed, and this data must be as much as possible close to the applicative data (for instance, acoustic models trained by using data containing TV shows will be better to transcribe TV shows than telephone conversations): data collection and selection to train acoustic models are very crucial.

For language models, the same principles exist: we need a large amount of data, if possible as close as possible to the applicative data.

2.1 DATA SELECTION

Performance of statistical Natural Language Processing (NLP) systems like ASR systems can often be enhanced using various methods, which can occur before, during or after the actual system processing. Among these, one of the most efficient pre-processing method is data selection, i.e. the fact to determine which data will be injected into the system we are going to build. As we aimed to get the best result possible in the framework of the IWSLT14 campaign, data selection processing was done in order to construct large training corpora containing data as close as possible to the targeted application. This target focused on TED (Technical, Entertainment, Design: www.ted.com) conferences in which native and non-native speakers occur. Since the major part of the available textual training data are related to news (newspaper, web archives of press agencies, ...), we assume that the resulting language models will be relevant to the video files to process in the EUMSSI project. For the acoustic part of the ASR system, we aimed to build robust models by injecting broadcast news data.

2.1.1 DATA SELECTION FOR ACOUSTIC MODELS TRAINING

For our acoustic modeling we used as a primary source the TED-LIUM corpus release 2 [14], removing from it all talks recorded after December 31st, 2010. This removal is due to some

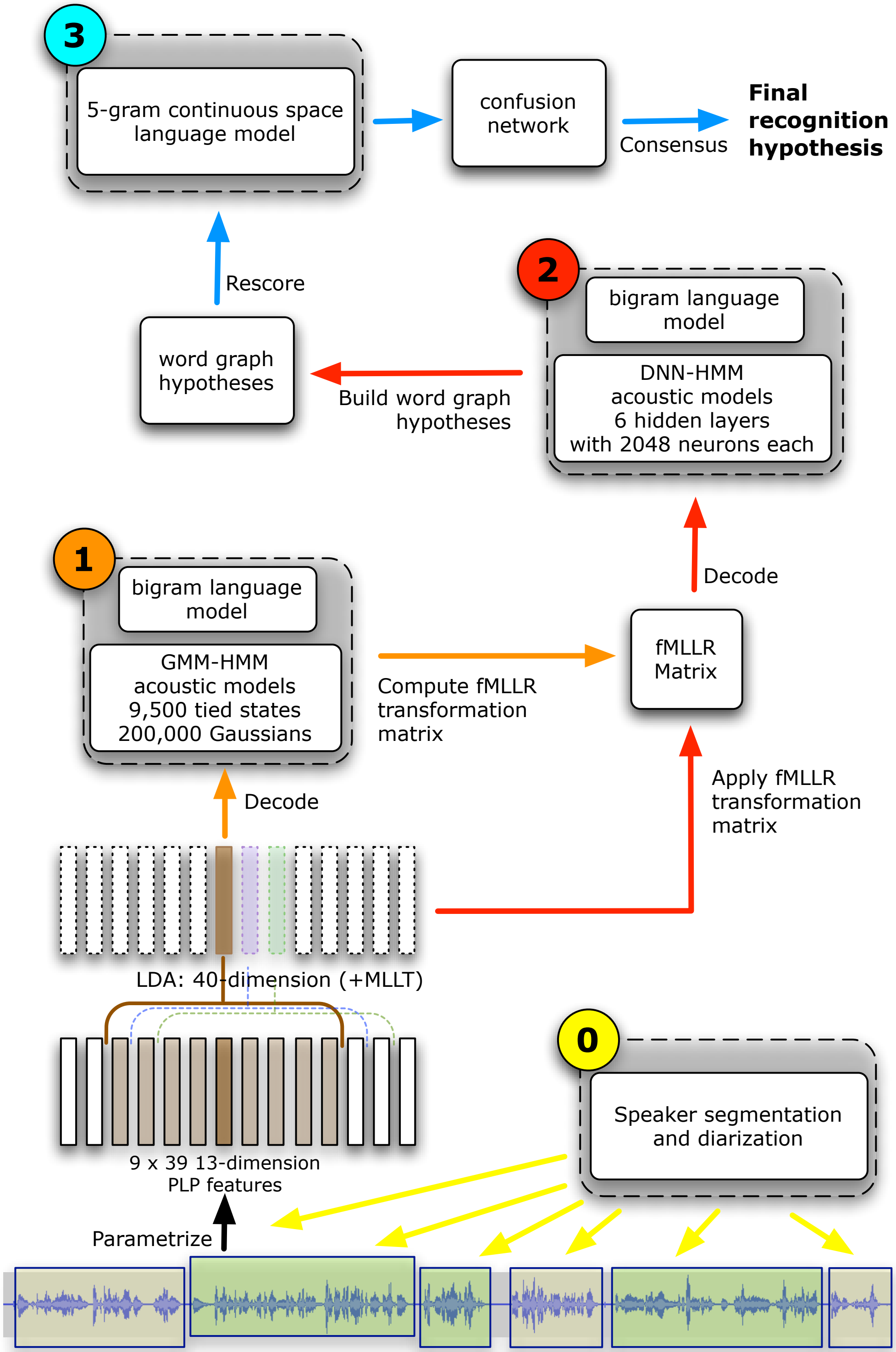
3
5-gram continuous space language model
confusion network
Consensus
Final recognition hypothesis
Rescore
2
bigram language model
DNN-HMM acoustic models 6 hidden layers with 2048 neurons each
word graph hypotheses
Build word graph hypotheses
Decode
1
bigram language model
GMM-HMM acoustic models 9,500 tied states 200,000 Gaussians
Compute fMLLR transformation matrix
fMLLR Matrix
Apply fMLLR transformation matrix
Decode
LDA: 40-dimension (+MLLT)
0
Speaker segmentation and diarization
9 x 39 13-dimension PLP features
Parametrize

Figure 1: Simplified architecture of the Automatic Speech Recognition system developed by the LIUM in the framework of the EUMSSI project

limitations from the IWSLT14 campaign. In order to strengthen this base, as seen above, we first added data from the Euronews corpora [8] distributed by the IWSLT campaign organizers and from the 1997 English Broadcast News Speech (HUB4) [3]. Then, from the MediaEval 2014 evaluation campaign Search and Hyperlinking Task data transcripts (BBC recordings from 2008 which were decoded by the LIUM) [2], we applied a threshold on our confidence measures to select the best possible segments for our system within a limit of 50 hours of speech. The table 1 summarizes the characteristics of the data included in our ASR system acoustic models.

Corpus	Duration	Segments	Words
TED-LIUM	130.1h	61 796	1 447 022
Euronews	68.2h	33 686	817 649
1997 HUB4	75.0h	20 652	852 517
MedialEval 14	50.0h	46 713	368 118
Total	323.3h	162 847	3 485 306

Table 1: **Characteristics of the acoustic data used in the LIUM ASR system acoustic models for English language.**

2.1.2 DATA SELECTION FOR LANGUAGE MODELS

Since language models training data was constrained for the ASR task of the IWSLT14, we applied our data selection tool XenC [12] on each allowed corpus at our disposal: basically all of publicly available WMT14 data (WMT14 refers the Workshop on Machine Translation, which organized a evaluation campaign for machine translation), a provided TED Talks closed-captions corpus and the LDC Gigaword. Based on cross-entropy difference from a corpus considered as in-domain and out-of-domain data, our tool allows to perform relevant data selection by scoring out-of-domain sentences regarding their closeness to the in-domain data. The table 2 summarizes the characteristics of the monolingual text data used to estimate our system language models.

2.2 DETAILED ARCHITECTURE OF THE LIUM ASR SYSTEM

The LIUM ASR system built for the EUMSSI project is based on the Kaldi Speech Recognition Toolkit, which uses Finite State Transducers (FSTs) for decoding (the general approach is described in [11]). A first step is performed with the Kaldi decoder by using a bigram language model and classical Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) models to compute a feature space Maximum Likelihood Linear Regression (fMLLR) matrix transformation. A second step is performed by using the same language model and deep neural network acoustic models. This pass generates word-lattices: an in-house tool, derived from a rescoring tool from the CMU Sphinx project, is used to rescore word-lattices with a 5-gram continuous space language model [15] before applying a variant of the consensus approach described in [9]. The entire architecture is illustrated in Figure 1.

Corpus	Original # of words	Selected # of words	% of Orig.
IWSLT14	0.1M	0.1M	100.00
Common Crawl	195.4M	13.6M	6.98
Europarl v7	56.4M	1.8M	3.22
Gigaword LDC	4 985.3M	168.2M	3.37
10^9 FR-EN	649.4M	11.9M	1.83
News Crawl	1 503.1M	44.8M	2.98
News-Comm. v7	4.7M	0.7M	14.04
UN 200x	360.1M	1.8M	0.50
Yandex 1M	24.1M	4.6M	19.01
Total (w/o IWSLT14)	7 778.5M	247.4M	3.18

Table 2: **Characteristics of the text data used in the LIUM ASR system language models for English language.**

In this section we will first present the training data used to estimate the LIUM models, then describe how the system was built using this toolkit.

2.2.1 SPEAKER SEGMENTATION

To segment the audio recordings and to cluster speech segments by speaker, we used the LIUM_SpkDiarization speaker diarization toolkit [10]. This speaker diarization system is composed of an acoustic Bayesian Information Criterion (BIC)-based segmentation followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. A Viterbi decoding re-segments the signal using GMMs with 8 diagonal components learned by EM-ML, for each cluster. Segmentation, clustering and decoding are performed with 12 MFCC+E, computed with a 10ms frame rate. Gender and bandwidth are detected before transcribing the signal.

More details about the speaker segmentation are given in the deliverable D3.2.

2.2.2 ACOUSTIC MODELING

The GMM-HMM models are trained on 13-dimensions Perceptual Linear Predictive (PLP) features with first and second derivatives by frame. By concatenating the four previous frames and the four next frames, this corresponds to $39*9 = 351$ features projected to 40 dimensions with Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). Speaker Adaptive Training (SAT) is performed using fMLLR transforms. Using these features, the models are trained on the full 323.3 hours set, with 9 500 tied triphone states and 200 000 gaussians.

On top of these models, we train a Deep Neural Network (DNN) based on the same fMLLR transforms as the GMM-HMM models and on state-level Minimum Bayes risk (sMBR) as discriminative criterion. Again we use the full 323.3 hours set as the training material.

The resulting network is composed of 7 layers for a total of 36.8 millions parameters and each of the 6 hidden layers has 2 048 neurons. The output dimension is 7 296 units and the input dimension is 440, which corresponds to an 11 frames window with 40 LDA parameters each. Weights for the network are initialized using 6 Restricted Boltzmann Machines (RBMs) stacked as a deep belief network (DBN). The first RBM (Gaussian-Bernoulli) is trained with a learning rate of 0.01 and the 5 following RBMs (Bernoulli-Bernoulli) are trained with a rate of 0.4. The learning rate for the DNN training is 0.00001. The segments and frames are processed randomly during the network training with stochastic gradient descent (SGD) in order to minimize cross-entropy between the training data and network output. When these training steps are done, the last step of training is processed, by applying the minimum Bayes risk criterion, as indicated above. To speed up the learning process, we use a general-purpose graphics processing unit (GPGPU) and the CUDA toolkit for computations.

2.2.3 LANGUAGE MODELING

For language modeling, we rely on the SRILM language modeling toolkit [16] as well as the Continuous Space Language Model toolkit. The vocabulary used in the LIUM ASR system is composed of 165 371 entries. The bigram language model (2G LM) used during the Kaldi decoding part is trained on the data presented in section 2.1.2.

With the SRILM toolkit, one 2G LM is estimated for each corpus source, with no cut-offs and the modified Kneser-Ney discounting method. These 2G LM are then linearly interpolated to produce the final 2G LM which will be used in the final system, using the IWSLT 2011 development and test corpora. To rescore the word-lattices produced by Kaldi, a trigram and a quadrigram language models (3G and 4G LM) are estimated with the same toolkit, again by training one LM by corpus source and then linearly interpolating them. A 5G continuous-space language model (CSLM) is also estimated for the final lattice rescoring, with no cut-offs and the same discounting method as for the bigram language model. Last, a consensus approach is applied after generating a confusion network from the rescored final word-lattice. The table 3 details the interpolation coefficients for the 2G, 3G and 4G language models as well as the final perplexity for each final model.

2.3 PERFORMANCES

Performances of the ASR system were evaluated by the organizers of the IWSLT14 campaign. Official results are presented in Table 4: WER means "word error rate" and evaluates the transcription quality, while NCE means "Normalized Cross Entropy" and evaluates the quality of the confidence measure values provided by the ASR system. We can notice that there is a large variability of performances depending on the speaker. The evaluation data are TED conference talks, and there is only one speaker by file. This high variability can be explained by the heterogenous origins of the English native and non-native speakers, in addition to the heterogeneousness of the topics addressed in each talk.

Depending of the talk, the word error rate has values between 4.2% of WER and 32.7%, for **a global word error rate of 14.1%**. Considering that **the computation time needed by the ASR system was $2 \times RT$**, these results can be considered as very competitive. RT

Corpus	Coefficients		
	2G	3G	4G
IWSLT14	0.36353	0.23963	0.19110
Common Crawl	0.14404	0.26584	0.34979
Europarl v7	0.00272	0.00244	0.00277
Gigaword LDC	0.30076	0.27450	0.24411
10^9 FR-EN	0.02709	0.02882	0.02701
News Crawl	0.13535	0.14751	0.14241
News-Comm. v7	0.00173	0.00254	0.00220
UN 200x	0.00300	0.00411	0.00391
Yandex 1M	0.02179	0.03461	0.03670
Perplexity	209.31	134.38	107.72

Table 3: **Interpolation coefficients and perplexities for the bigram, trigram and quadrigram language models used in the LIUM ASR system for English language.**

is for "real time", which means that 2 hours of computation time are needed to process 1 hour of speech.

In the official IWSLT 2014 campaign, we have combined the outputs provided by the system described above with a variant of this system, in order to reduce the word error rate. This combined approach needs more computational time, so it is not relevant to process huge volume of data, as targeted in EUMSSI. As official results with this combined approach, we reached 13.0% of word error rate on the official test set. At this time, we have no information about the results reached by the other participants, but we expect to be among the competitive ones.

3 NON-ENGLISH ASR SYSTEMS

3.1 FRENCH LANGUAGE

At the beginning of the EUMSSI project, the LIUM ASR system was mainly built around the open source CMU Sphinx project and was evolving to the architecture described in figure 1, based on the open source Kaldi toolkit. This evolution was accelerated thanks to the EUMSSI project and our ASR system for French language was the first of our systems to be operational in this new architecture.

Thanks to this evolution, the LIUM reaches an word error rate of 16.0% on the test data of the ASR task of the REPERE evaluation campaign. This word error rate dropped to 13.5% by combining the LIUM outputs with the outputs of a Canadian partner of the LIUM, the Centre de Recherche en Informatique de Montréal (CRIM). This LIUM ASR system for French language and its combination with the CRIM ASR system are described with details in [13]. These results are not the official results obtained by the LIUM on January 2014: at this date, the ASR system developed in the framework of the EUMSSI project was not achieved.

In this section, the REPERE context is presented, with also the data used to train the

audio file	# words	WER (%)	NCE
ted.tst2014.talkid1443	3083	8.7	0.310
ted.tst2014.talkid1477	836	13.0	0.079
ted.tst2014.talkid1650	1955	10.3	0.217
ted.tst2014.talkid1733	878	5.6	0.327
ted.tst2014.talkid1736	783	27.7	-0.152
ted.tst2014.talkid1741	2240	8.9	0.179
ted.tst2014.talkid1755	755	14.6	0.124
ted.tst2014.talkid1781	686	26.4	-0.023
ted.tst2014.talkid1829	1440	11.6	0.113
ted.tst2014.talkid1835	1270	6.0	0.208
ted.tst2014.talkid1852	2003	16.4	0.157
ted.tst2014.talkid1854	1352	32.1	-0.130
ted.tst2014.talkid1858	1271	3.9	0.307
ted.tst2014.talkid1864	1311	14.7	0.038
ted.tst2014.talkid1898	1934	17.8	0.129
GLOBAL	**21797**	**13.4**	**0.156**

Table 4: **Official results of the LIUM ASR system for English language on the test set data of the ASR task within the IWSLT14 campaign**.

acoustic and linguistic models, and the final results of the LIUM ASR system for French language on the official REPERE test set. Since the LIUM ASR system for French language shares the same architecture as the LIUM ASR system for English language, please refer to section 2.2 for details about it.

3.1.1 THE REPERE CHALLENGE

REPERE is an evaluation project in the field of people recognition in television documents [4], funded by the DGA (French defense procurement agency) and ending in 2014. Several evaluation tasks were organized, including an evaluation of automatic speech recognition systems on French TV shows.

3.1.2 TRAINING DATA

The training set used to build the LIUM system for French language consists of 145,781 speech segments from several sources: the radiophonic broadcast ESTER [5] and ESTER2 [6] corpora, which accounts for about 100 hours of speech each; the TV broadcast ETAPE corpus [7], accounting for about 30 hours of speech; the TV broadcast REPERE train corpus, accounting for about 35 hours of speech and other LIUM radio and TV broadcast data for about 300 hours of speech, which have been segmented using the speaker diarization system described in section 2.2.1. The training dictionary has 107,603 phonetized entries. Table 5 summarizes the characteristics of each dataset.

Sources	Speech	Segments
ESTER	100h	12,902
ESTER2	100h	15,162
ETAPE	30h	8,378
REPERE	35h	10,269
LIUM v8	300h	99,070
Total	565h	145,781

Table 5: **Characteristics of the training data for acoustic modeling for French language.**

For language modeling, the training data is composed of the manual transcriptions from the training corpus used to estimate the acoustic models, of articles extracted from of TV websites, of articles extracted from Google News, of the French Gigaword corpus, of articles from newspaper 'Le Monde'. All of these data were collected before January 2013.

Table 6 presents of the number of words in each corpus in the training corpus used to estimate language models.

Sources	Number of words
Manual transcriptions from the training corpora used to train the acoustic models	8M
Articles from TV websites (≤2012)	5M
Google News (≤2012)	204M
French Gigaword (≤2012)	1015M
Newspapers (≤2012)	366M
Subtitles of TV Newspaper (≤2012)	11M
Total	1609M

Table 6: **Characteristics of the training data for language models for French language.**

3.1.3 PERFORMANCES

A first evaluation on the REPERE ASR task was organized in 2013, in which the LIUM ASR system ranked first, on similar but different test data. This system appears in Table 7 under the name old 2013 LIUM system: it can be used to measure improvements achieved since last year.

The old 2013 LIUM ASR system was based on the CMU Sphinx toolkit, with some improvements, for instance the use of hybrid MLP/HMM acoustic models. A variant of this system is described in [1].

The main difference between the new ASR system developed by LIUM and the old one comes from the use of DNN acoustic models and the use of the finite state machine paradigm.

These functionalities are both offered by the Kaldi toolkit. Notice that the linguistic rescoring tool is the same one in both LIUM ASR systems. With the same language models and the same training data for the estimation of acoustic models, the word error rate (WER) of the LIUM ASR system is reduced of 2.6 points (14%) to 16.0%.

ASR system	WER
2013 LIUM	18.6%
2014 LIUM	16.0%

Table 7: **Word error rates on REPERE test corpus (TV shows) for French language.**

3.2 GERMAN AND SPANISH LANGUAGES

During the first year of the EUMSSI project, we have developed a new ASR system with a new architecture. We have focused on English and French languages, and the software developed for these systems will be used for German and Spanish languages in the next months.

Even if we don't have developed ASR systems for German and Spanish languages, we have started collecting data to train acoustic and language models. Such collect is easy for Spanish, but German data for acoustic models are very hardly accessible. Through our participation to the IWSLT campaign, we were able to access data distributed by the organizers [8], including close to 100 hours of data in German language from the Euronews website. These data will be used as a bootstrap to train our acoustic models for German language.

4 AUTOMATIC TRANSCRIPTION OF EUMSSI VIDEO FILES

Currently, more than 3,000 hours of videos in English language have been transcribed using the LIUM ASR system in the framework of the EUMSSI project. These files come from various sources:

- Deutsche Welle news videos from their website
- YouTube videos from:
  - Deutsche Welle channel
  - The Guardian channel
  - various other sources

4.1 STATISTICS

In this section we report some statistics on the data which has been already transcribed for the EUMSSI project.

4.1.1 DEUTSCHE WELLE NEWS VIDEOS

The DW transcriptions have been uploaded to the EUMSSI platform. The content from the DW website account for a total of 2 142 hours of speech. Table 8 summarizes the statistics about these transcriptions.

Characteristic	Statistic
Number of videos	18 039
Total duration (hours)	2 142
- Male	1 237
- Female	905
Mean duration	7m 07s
Number of segments	989 148
Number of words	21 175 027

Table 8: **Deutsche Welle news videos and transcriptions statistics.**

4.1.2 YOUTUBE VIDEOS

The table 9 summarizes the statistics about the transcriptions of the YouTube content for each main source. These transcriptions will be uploaded to the EUMSSI platform in a very near future. They are automatic transcriptions of 1,282 hours of speech.

Characteristic	Deutsche Welle	The Guardian	Other	Total
Number of videos	1 725	871	5 004	7 600
Total duration (hours)	180	60	1 042	1 282
- Male	115	37	746	898
- Female	65	23	296	384
Mean duration	6m 16s	4m 05s	12m 29s	10m 06s
Number of segments	82 999	27 126	478 876	589 001
Number of words	1 738 476	480 376	8 608 491	10 827 343

Table 9: **YouTube sourced videos and transcriptions statistics.**

4.2 IMPORT FORMAT

This section present the format of data provided by the ASR module to the Solr EUMSSI platform.

4.2.1 ORIGINAL TRANSCRIPT FORMAT

The transcriptions produced by our system are in the NIST CTM format (see `http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm#ctm_fmt_name_0`). The CTM file format is a concatenation of time mark records for each recognized word in a waveform. The records are separated with a newline. Each word token has a waveform id, channel identifier [1 | 2] or [A | B], start time, duration, confidence measure and word text. Each record follows this BNF format:

<F> <C> <ST> <DUR> word <CNF>

Where:

- <F> is the waveform name.
- <C> is the waveform channel. Either [1 | A] or [2 | B].
- <ST> is the start time (in seconds, ss.SS format) of the word, measured from the start time of the file.
- <DUR> is the duration (in seconds, ss.SS format) of the word.
- <CNF> is the confidence score, expressed with a value between 0.00 and 1.00.

4.2.2 EUMSSI PLATFORM IMPORT FORMAT

Per requested by the Solr EUMSSI platform requirements, the analysis result must be in the JSON format. This result must be sent with a POST request, using the parameters specified in the API. Here is a minimal example of a CTM file converted to the JSON format:

```
{
    "id": "482645",
    "httpHigh": "dwtv_video/flv/age/age20120306_gesamt_sd_avc.mp4",
    "date": "2014-11-10 11:19:18.926774",
    "version": "LIUM_EN_ASR-201409",
    "content": [
        {
            "type": "word",
            "start": "13.55",
            "end": "14.11",
            "item": "hello",
            "conf": "1.00"
        },
        {
            "type": "word",
            "start": "14.12",
            "end": "14.45",
            "item": "and",
            "conf": "1.00"
        },
        {
            "type": "word",
            "start": "14.46",
            "end": "14.94",
            "item": "welcome",
            "conf": "1.00"
        }
    ]
}
```

Where:

- "id" is the identifier of the content, here the DW video ID.
- "httpHigh" is the content source, here the DW video download link.
- "date" is the JSON file generation date.
- "version" is the LIUM ASR system version.
- "content" is the list of records from the CTM file.

For each "content" item, an entry is described as follows:

- "type" is the type of the entry, "word" or "filler".
- "start" indicates the starting point of the word from the beginning of the transcript.
- "end" indicates the ending point of the word, also from the beginning of the transcript.
- "item" indicates the text content, whether the actual word or the filler.
- "conf" is the confidence measure.

The content is then encapsulated in a HTTP POST request to the EUMSSI server via the documented API.

5 CONCLUSION AND PERSPECTIVES

The EUMSSI project allows the LIUM partner to develop a more competitive ASR system: the LIUM ASR system reached 16.0% of word error rate by transcribing French TV shows during the REPERE evaluation campaign, while the previous (2013) version of the LIUM ASR system reached 18.0% of word error on the same data. The LIUM ASR system for English data is also competitive: it reached 13.4% of word error rate on the last evaluation campaign for ASR in English language, IWSLT14.

This LIUM ASR system for English language was used to process a very large amount of video files provided by the Deutsche Welle partner, but also to process a very large amount of video files coming from YouTube and proposed by the L3S partner.

For the next months, the LIUM will develop ASR systems for Spanish and German languages using the same software architecture. Then, new solutions will be proposed in order to significantly speed up these ASR system, without increase the word error rate.

References

[1] F. Bougares, P. Deléglise, Y. Esteve, and M. Rouvier. LIUM ASR system for Etape French evaluation campaign: experiments on system combination using open-source recognizers. In Text, Speech, and Dialogue, pages 319--326, 2013.

[2] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. Jones. The search and hyperlinking task at MediaEval 2014. In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, october 2014.

[3] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett. 1997 English broadcast news speech (HUB4) LDC98S71. Web Download, 1998.

[4] O. Galibert and J. Kahn. The first official REPERE evaluation. In First Workshop on Speech, Language and Audio in Multimedia (SLAM), pages 43--48, Marseille, France, 2013.

[5] S. Galliano, E. Geoffrois, G. Gravier, J. f. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the Ester evaluation campaign for the rich transcription of French broadcast news. In 5th international Conference on Language Resources and Evaluation (LREC), pages 315--320, 2006.

[6] S. Galliano, G. Gravier, and L. Chaubard. The Ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In Interspeech, 2009.

[7] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In Eighth International Conference on Language Resources and Evaluation (LREC), pages 114--118, Istanbul, Turkey, 2012.

[8] R. Gretter. Euronews: a multilingual speech corpus for ASR. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may 2014.

[9] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Computer Speech & Language, 14(4):373--400, 2000.

[10] S. Meignier and T. Merlin. LIUM SpkDiarization: an open source toolkit for diarization. In CMU SPUD Workshop, Dallas, Texas, USA, 2010.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burge, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, december 2011.

[12] A. Rousseau. XenC: An open-source tool for data selection in natural language processing. The Prague Bulletin of Mathematical Linguistics, 100:73--82, 2013.

[13] A. Rousseau, G. Boulianne, P. Deléglise, Y. Estève, V. Gupta, and S. Meignier. LIUM and CRIM ASR System Combination for the REPERE Evaluation Campaign. In Text, Speech and Dialogue, pages 441--448. Springer, 2014.

[14] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may 2014.

[15] H. Schwenk. CSLM - a modular open-source continuous space language modeling toolkit. In Interspeech, pages 1198--1202, august 2013.

[16] A. Stolcke. SRILM - an extensible language modeling toolkit. In Proceedings of Interspeech, pages 901--904, September 2002.