D3.3 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 :	M24
Actual date of delivery	December 24th 2015
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	M24
Actual date of delivery	December 24th 2015
Deliverable name	D3.3. Progress report on rich audio transcription
Type	R – Report
Status	Submitted
Version number	1
Number of pages	20
WP / Task responsible	WP3/LIUM
Author(s)	Yannick Estève
Other contributors	
EC Project Officer	Mrs. Alina Lupu Alina.LUPU@ec.europa.eu
Abstract	Progress report on rich audio transcription: speech recognition in English and German. Architecture, training, data, performances. Error detetection
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

This deliverable describes the last evolutions of the automatic speech recognition (ASR) systems developed by the LIUM under the framework of the EUMSSI project. At the beginning of the project, the LIUM planned 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. During the first year, competitive systems were produced for French and English languages (cf. deliverable D3.1.). After the first review of the project on January 2015, it has been decided to follow the reviewer's remarks and to focus on only two languages, considered as the most relevant ones to the project. English and German language were retained.

For this second year, strong effort was produced in order to get the fastest possible ASR system, without reducing accuracy, for both English and German languages. One can notice that it was a real challenge for the LIUM partner to develop a so competitive ASR system in German language, because this language was never processed by this partner before, and because linguistic resources for German language necessary to develop a such system are very rare with a reasonable cost.

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 two international evaluation campaigns on speech recognition:

- the ASR task of the MGB challenge for English language;
- the ASR task of the IWSLT 2015 campaign for German language.

These participations were successful: LIUM reached the second position at the ASR track of the MGB campaign (in collaboration with the CRIM laboratory from Montreal, Quebec, Canada) and LIUM won the ASR task of IWSLT 2015 for German. Moreover, the 2015 ASR system is more than ten times faster than the 2014 ASR one.

In addition to these works, LIUM has started a study on ASR error detection. This task can be very useful in the framework of the EUMSSI project for several reasons:

- to filter misrecognized words to reduce false alarms when looking for automatic transcriptions containing some requested words;
- to help natural language processing applied on automatic transcriptions (like name entity recognition);
- to improve the ASR performances by injecting confident automatic transcriptions into the training corpus of acoustic model: larger amount of training data improves the quality of acoustic models.

This preliminary study has outperformed state-of-the-art approaches.

2 ARCHITECTURE OF THE 2015 LIUM ASR SYSTEM

2.1 Evolutions of the LIUM ASR system

The LIUM ASR system built for the EUMSSI project has evolved during the second year of the project. The engine core, based on the Kaldi Speech Recognition Toolkit [15], is the same, but the multi-step architecture has changed in order to reduce the computation time. In the 2014 architecture, two successive acoustic decoding processes were needed before rescoring word-graphs. The first one, using GMM/HMM[1], was used in order to exploit its outputs to compute a fMLLR matrix transformation. This fMLLR matrix was applied to the acoustic features in order to make the second acoustic decoding process, based on DNN/HMM[2], more adapted to the speaker and to the acoustic conditions.

Among the different steps of the entire recognition process, acoustic decoding processes are largely the most time and computation power consuming ones.

We have decided to keep only one acoustic decoding process in the new architecture of the ASR system instead of two previously. This means that no fMLLR matrix is built, no fMLLR adaptation is applied, and so no loud speaker adaptation is realized: only a cepstral mean normalization is applied on a speech segments labeled to the same speaker. In addition, acoustic features are now different: TRAP features replaces PLP/LDA features. Figure 2.1 illustrates these changes.

2.2 The 2015 ASR system: main language-independent features

This section presents the common features of the 2015 ASR systems for English and German languages.

2.2.1 Speaker segmentation

The speaker diarization system used within the ASR process is the same as the one integrated in the 2014 LIUM ASR system: 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 Speech recognition

The LIUM ASR system can be still considered as a multi-pass system, even if only one acoustic decoding process is now done. It is based on the Kaldi system for acoustic decoding and on

[1] GMM: Gaussian Model Mixture. HMM: Hidden Markov Model.
[2] DNN: Deep Neural Networks.

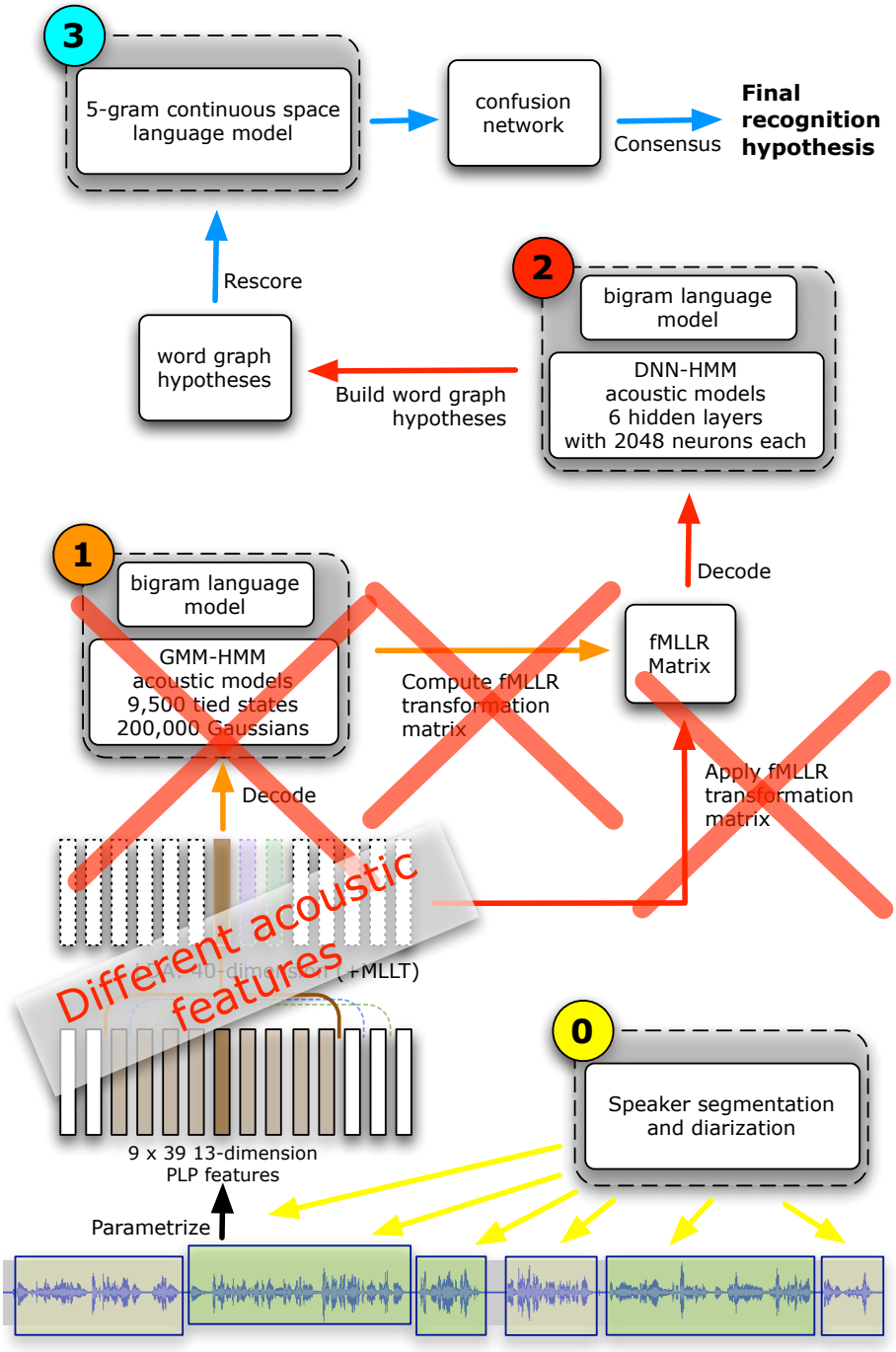

Figure 1: Main changes from 2014 LIUM ASR system to 2015 LIUM ASR system developed in the framework of the EUMSSI project

LIUM tools built from the CMU Sphinx project for linguistic rescoring [3]. Some parts of the source codes were modified in order to accelerate the decoding processing, for instance by improving the multi-threading management in order to better exploit the computation power available on a machine.

The first pass produces word-graphs by using the DNN acoustic models combined with a 2-gram language models. Acoustic models are based on DNN. For each frame, DNN inputs are composed of 368 TRAP coefficients computed on a sliding window of 31 frames. To compute TRAP features, the 23-dimensional filterbank features are normalized to zero mean per audio file. 31 frames of these 23-dimensional filterbank features (15 frames on each side of the current frame) are spliced together to form a 713-dimensional feature vector. This 713-dimensional feature vector is transformed using a hamming window (to emphasize the center), passed through a discrete cosine transform and the dimensionality reduced to 23x16 or 368-dimensional feature vector per frame. As written above, speaker adaptation is trivial: it only consists on mean subtraction applied on the filterbank features of all the frames associated to a speaker. This choice was retained because internal experiments showed that the use of TRAP features in combination with DNN provides similar results, in terms of accuracy, to our former architecture using MLP/LDA features and fMMLR adaption. In the same time, this new solution divides by more than two the computation time needed for the speech recognition process. The DNN was built following the approach described in [19] and it was composed of six hidden layers with 2048 units, while the output softmax layer had several thousands outputs depending on the language (4627 for English).

Next passes consists on expanding and rescoring the word-graphs by using 3-gram, then 4-gram back-off LMs, then a 5-gram neural network model (including the 5-gram back-off LM) [17].

At the end, an accelerated version of the consensus approach [9], which takes into account temporal information to speed up the processing, is applied on the confusion networks built from the 5-gram rescored word-graphs.

Figure 2 presents the general architecture of the 2015 LIUM ASR system in the framework of the EUMSSI project.

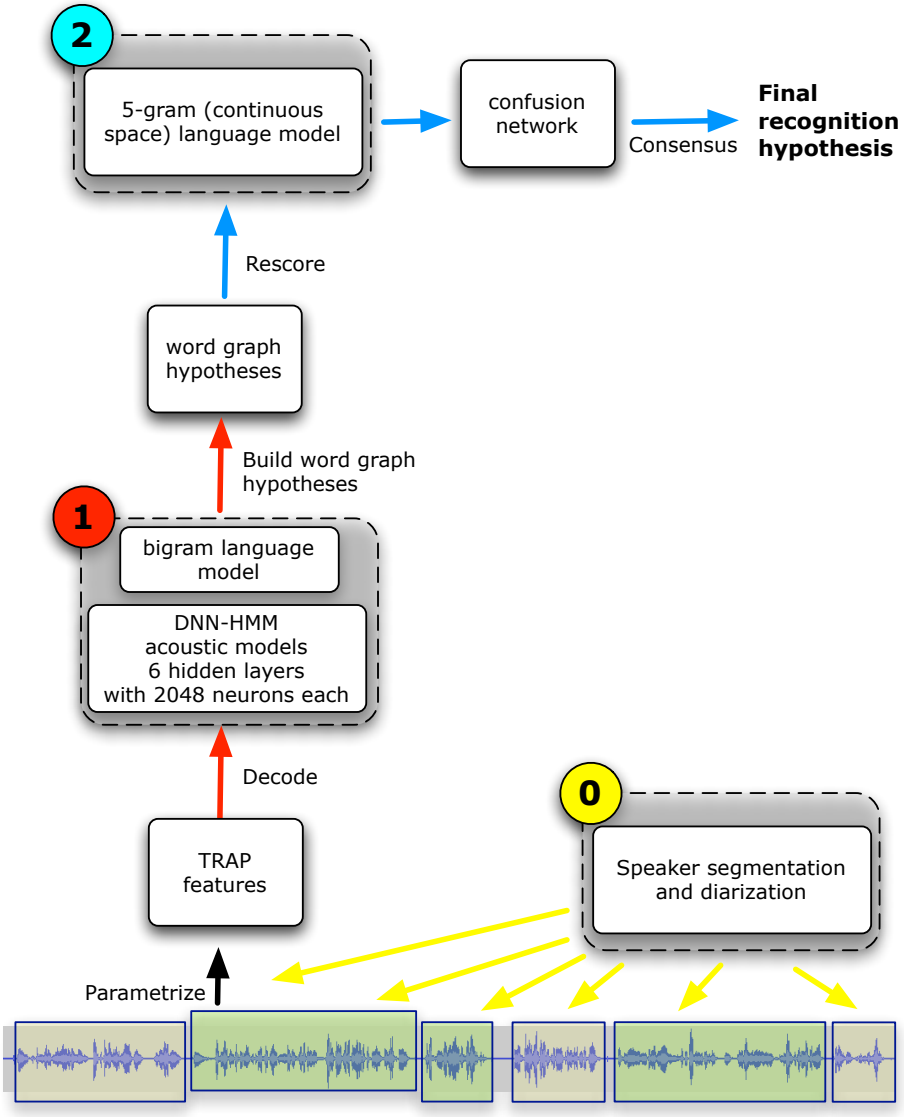

Figure 2: General architecture of the 2015 LIUM ASR system developed in the framework of the EUMSSI project

3 EVALUATION OF THE ASR SYSTEMS

3.1 English language: participation to the ASR task of the MGB 2015 Challenge

The Multi-Genre Broadcast Challenge at ASRU 2015 is a controlled evaluation of speech recognition, speaker diarization, and lightly supervised alignment using BBC TV recordings. This challenge was an official event of the IEEE workshop on Automatic Speech Recognition and Understanding.

LIUM participated to two submissions: one is the fast system developed in the framework of the EUMSSI project, tuned to be as fast as possible with a high accuracy, and the another one, made in collaboration with the Center of Research on Informatics of Montreal (CRIM), does not take into account the computation time, and aims to get a the lowest possible word error rate.

3.1.1 Language models

The LIUM ASR system involved in the MGB challenge uses 2-gram, 3-gram, 4-gram back-off LMs, and a 5-gram back-off used in combination with a 5-gram feed forward neural network model: it is the one described in section 2.2.2. Back-off LMs were estimated through the SRILM toolkit, while the neural network language model (NNLM) was estimated by using the CSLM toolkit, developed at LIUM and distributed under LGPL license [17]. All these language models created by the LIUM were estimated on the entire normalized data provided by the organizers. No LM adaptation was applied.

For this campaign, LIUM's vocabulary contains 152K words, the most frequent ones in the normalized training data. Classical back-off n-gram models were trained by using the modified Kneser-Ney smoothing, without cutoff nor pruning. The 5-gram LM is composed of 152K 1-grams, 25M 2-grams, 125M 3-grams, 254M 4-grams, and 330M 5-grams. The 5-gram NNLM is composed of a projection layer of 640 units, corresponding to 160-dimensional word embeddings, two hidden layers of 1024 units each, and an output layer providing probabilities for a short-list composed of the 16384 most frequent words.

The impact of the use of each LMs presented in section 3.1.2.

3.1.2 Acoustic models: using imperfect transcripts to build a training corpus for acoustic models

To train the acoustic models, participants to the MGB project could only used the imperfect transcripts of about 1600h of TV shows. Imperfect transcripts were both subtitles made manually with very rough timecodes and automatic transcripts provided by a baseline system owned by the organizers.

LIUM investigated its own approach to extract relevant audio/text alignments to train acoustic models. First, ASR outputs and pronunciation dictionary provided by the organizers were used to train DNN acoustic models. Then, all the audio files provided by the organizers as part of the training corpus were processed by using the LIUM internal tool for speaker diarization [10]. Each produced speech segment was transcribed by using the first DNN acoustic models, combined with a 2-gram language model presented in section 3.1.1. This

processing generated a word-graph for each speech segment. Each word-graph was aligned with subtitles made by human annotators and provided with the audio files. Word-graph alignment consists of searching a path within the word-graph that matches with the subtitles, accepting that rough timecode values from subtitles and precise timecode values within the word-graph could be delayed by 20 seconds max. Only long speech segments with no more than one word mismatch (insertion, substitution, or deletion) between subtitles and the closest path in the word-graph were selected. The text associated with a selected speech segment is the one coming from the closest path in the word-graph in regards with the subtitles. The training alignments generated by LIUM result in 700 hours of training audio.

3.1.3 Word error rate and computation time

Table 1 presents the official results of the MGB campaign. They will be published during the IEEE ASRU workshop on December 2015.

System	Global WER
Sys1	23.7%
CRIM/LIUM	26.6%
Sys 3	27.5%
Sys 4	27.8%
Sys 5	28.8%
Fast LIUM	30.4%
Sys 7	30.9%
Sys 8	31.2%
Sys 9	35.0%
Sys 10	38.0%
Sys 11	38.7%
Sys 12	40.8%

Table 1: Official results of the MGB Challenge.

The fast LIUM ASR system reaches an interesting rank: while this system is designed to be as fast as possible, it reaches the 6th rank on 12 participants. This system was also integrated into the ASR system combination built with the CRIM, which reaches the second rank of the challenge. The five first systems are built on a such architecture based on ASR system combination: this implies a computation time highly more important than the one needed for the fast LIUM system.

This computation time of the fast LIUM system was analyzed in details on the development corpus. Each step has been studied, and table 2 presents these results. Speed is computed in terms of real time. **The entire decoding process needs 0.17 times real time, which means that 17 minutes are necessary to process 100 minutes (1h40) of speech**. This can be reduced to 0.07 times real time if the CSLM rescoring is not applied. This would imply an increase of the word error rate (1 point). Other internal experiments, not described here, has shown that similar word error rates were reached by the 2014 system and the fast 2015 one. The most interesting difference comes from the computation time, since the 2015 ASR system is more than ten time faster than the 2014 ASR one.

Step	Comment	WER	Comput. time
1	DNN + 2-gram	-	0.065 x RT
2	3-gram rescoring	31.4%	0.0015 x RT
3	4-gram rescoring	30.4%	0.002 x RT
4	CSLM 5-gram rescoring	29.6%	0.1 x RT
5	consensus	29.4%	0.001 x RT
Total	Full process (official submission)	29.4%	0.17 x RT

Table 2: WER and computation time (in Real Time) on the Dev set (dev.full+dev.longitudinal) of the Fast LIUM system which has participated to the MGB Challenge.

The MGB dataset is composed of very heterogeneous data. This implies a high variability of acoustic conditions and spoken languages (fluent, disfluent, spontaneous, prepared, familiar, unfamiliar, ...). Table 3 presents detailed results of the LIUM ASR system for each kind of show in the MGB test evaluation dataset. One can notice that the ASR obtains good results on documentaries or political news, which are data close to the ones provided by the Deutsche Welle partner. Performances are degraded when sketch comedies or series are processed. A similar behaviour is observed with the CRIM/LIUM ASR system and with all the other ASR systems participating to the MGB campaign.

3.2 German language: participation to the ASR task of the IWSLT 2015 evaluation campaign

Last year, LIUM has participated to the ASR task of the IWSLT 2014 evaluation campaign for the English language, in the framework of the EUMSSI project, and also to the ASR task for the Italian language in the framework of a partnership with an industrial partner. This year, we aimed to participate to the ASR task of the IWSLT 2015 campaign for the German language, to evaluate our ASR system dedicated to German language and developed for the EUMSSI project.

A crucial issue in developing such a system is the access to available training corpus for acoustic models in the focused language. These training data are ideally audio recordings with manual transcriptions. For the German language, such data are very rare at a reasonable cost. Hopefully, the industrial partner which co-participated with LIUM in the IWSLT 2014 ASR task for Italian language has nicely accepted to provide a little more than one hundred of hours of audio recordings in German with their manual transcriptions. This agreement between LIUM and its industrial partner, which is not a EUMSSI partner, has been signed for research purpose only, limited to the framework of the EUMSSI project.

The fast LIUM ASR system described above was adapted to German language: a vocabulary of 300K words (twice more than the English one to deal with German compound words) has been built, with language models trained on the data described in Table 4. A data selection was made by using an internal tool [16], XenC (distributed under open source license), based on cross-entropy [12]. The data selection permitted us to focus on the IWSLT 2015 topics (TED conference talks). Acoustic models were trained on the 125 hours provided by the LIUM's industrial partner.

Show	Global WER	Comments
Dragons' Den	14.1%	Reality television featuring entrepreneurs pitching their business ideas
Daily Politics	14.5%	Current affairs and politics, interviews with leading politicians and political commentators
Magnetic North	14.9%	Documentary
Athletics London	19.4%	
Eggheads	19.8%	Quiz show
Point of View	23.4%	
Syd Barrett	25.7%	
Top Gear	29.3%	Motoring Entertainment
Blue Peter	30.4%	
Legend of the Dragon	31.7%	
The North West 200	34.4%	
Holby City	41.7%	
The Wall	43.8%	
One Life Special Mum	43.8%	
Goodness Gracious Me	45.1%	Sketch comedy
Oliver Twist	52.2%	Miniseries: British television adaptation of Charles Dickens' novel Oliver Twist

Table 3: Detailed official results of the fast LIUM ASR system on the test set of the MGB Challenge.

Following the same strategy as our joint participation to the MGB Challenge on English with the CRIM, we have also developed an second ASR system, based on the previous LIUM ASR system, developed in 2014, and described in the report D3.1, in order to obtain better performances in terms of accuracy. Table 5 presents the results obtained by the two single ASR systems and their combination on the IWSLT 2015 development corpus. The gap between the results of the 2014 and 2015 systems must not be interpreted as an improvement of the accuracy between the 2014 and 2015 systems: actually, we introduced also some changes (LM weights, DNN training, heuristics, ...) in the 2014-based system which could degrade its results: our goal was to produce two sufficiently different systems in order to get some complementarities useful to the system combination. Combining the 2015 fast ASR system and the 2014-based one allowed us to improve the performances of the 2015 ASR system.

Official performances of the ASR system were evaluated by the organizers of the IWSLT 2015 campaign. Official results are presented in Table 6.

This shows that a WER of 17.8% was reached by the combined LIUM ASR system on the evaluation corpus. With this performance, **LIUM won this competition**. The second position was reached by the German Karlsruhe Institute of Technology, which won the competition in 2014.

As a success indicator of the EUMSSI document of work, **the WP3 had to produce**

Corpus	Original # of words	Selected # of words	% of Orig.
manual transcriptions of speech	2.85M	2.85M	100.00
Common Crawl	48.04M	4.24M	8.82
Europarl	47.40M	3.20M	6.74
News Crawl	1.4G	130.60M	9.26
News-Comm.	5.06M	0.62M	12.25
Total (w/o IWSLT14)	1.5G	138.66M	9.18

Table 4: Characteristics of the text data used to train the language models for the German ASR systems.

System	WER
2015 Fast ASR system	15.8%
2014-based ASR system	16.8%
ASR combination	15.1%

Table 5: Word error rate of the LIUM ASR systems on the IWSLT 2015 development corpus on German language.

an ASR system able to reach the Word Error Rate obtained by the best ASR system during IWSLT 2013 in German language. **In 2013, the best ASR system reached 25.2% of WER on the tst2013 data**. **Our 2014-based ASR system reaches 23.7% of WER on the tst2013 data** which is now a part of the IWSLT 2015 development corpus. In addition, it is very probable that we can reduce more this WER by using our 2015 system, and more again by combining both the 2014 and 2015 systems: we can estimate that our goal in terms of accuracy in German has been more than reached.

System	WER
LIUM	17.8%
KIT (Karlsruhe Institute of Technology)	20.3%
MLLP	43.3%

Table 6: Official results of the ASR German track of the IWSLT 2015 test corpus on German language.

4 ASR ERROR DETECTION

In the framework of the EUMSSI project, automatic transcripts are exploited to retrieve specific video documents, and also to get a fast access to information supported by speech. With the state-of-the-art technology, errors are unavoidable, like this can be observed in the results presented in section 3. In WP3, we started one year ago working on ASR error detection, in the framework of the Mrs Sahar Ghannay's Ph. D., partially funded (50%) by the EUMSSI project. Very good preliminary results have been already obtained on French language, outperforming recent state-of-the-art approaches based on the use of Conditional Random Fields (CRF).

In this work, we have investigated the use of word embeddings as input features of a neural network-based error detection system. We experimented the use of three different types of word embeddings and propose to combine them with an auto-encoder in order to take advantage of their complementary.

4.1 Related work

For two decades, many studies have focused on the ASR error detection task. Recently, the best proposed approaches were based on the use of CRF. In [13], authors have focused on detecting error regions generated by Out Of Vocabulary (OOV) words. They proposed an approach based on Conditional CRF tagger, which takes into account contextual information from neighboring regions instead of considering only the local region of OOV words. A similar approach for other ASR errors was presented in [1], which proposes an error detection system based on CRF tagger using various ASR, lexical and syntactic features.

In the work made in the framework of the EUMSSI project, we compare the performances of the state-of-the-art CRF-based ASR error detection system proposed in [1] to our proposition based on a neural network architectures and the use of an effective combination of word embeddings built on a huge text corpus.

4.2 Set of features

An error detection system has to attribute the labels correct (c) or error (e) to each word. This attribution is made by analyzing each recognized word within its context. A set of relevant features must be selected to capture the good information to get a precise classification.

4.2.1 ASR, lexical and syntactic features

In this work, we nearly use the same features as the one presented in [1], which are detailed as follows:

- ASR features: posterior probabilities generated from the ASR system.
- Lexical features: length of the current word and three binary features indicating if the three 3-grams containing the current word have been seen in the training corpus of the ASR language model.
- Syntactic features: POS tag, dependency labels and word governors, which are extracted from the transcriptions by using the MACAON NLP Tool chain [3].

[3] http://macaon.lif.univ-mrs.fr

- Word: orthographic representation in CRF approaches, as used in [1]. With our neural approach, we will use word-vectors, which permit us to take advantage of some generalizations extracted during the construction of these word embeddings.

4.2.2 Word embeddings

Word embeddings are vector representations of words that have been successfully used in several natural language processing tasks [2]. This representation is a vector space can be trained through different methods and is computed from a textual corpus.

In our study, we have tested different kinds of word embeddings coming from different available implementations: Collobert and Wetson word embeddings revisited by Turian in [18], continuous bag of words (CBOW) and skip-ngrams proposed by Mikolov in [11], global vectors (GloVe) introduced in [14], and word embedding extracted from a neural network language model similar to the one used in our ASR system [17]. Our goal was to build complementary word embeddings for the ASR task detection. For this task, we need to capture syntactic information in order to use them to analyze sequences of recognized words, but we also need to capture semantic information to measure the relevance of co-occurrences of word in the same ASR hypothesis.

100-dimensional word embeddings were computed from a large textual corpus, composed of about 2 billions of words. This corpus was built from articles of the French newspaper "Le Monde", from the French Gigaword corpus, from articles provided by Google News, and from manual transcriptions of about 400 hours of French broadcast news.

In order to take advantage of their complementary, we propose to combine the word embeddings by investigating the use of an auto-encoder [20] or the use of a classical Principal Component Analysis (PCA). The auto-encoder is composed of one hidden layer with 100 or 200 hidden units. It takes as input a concatenation of the different embedding vectors and outputs a vector with the same size as the input vector. The auto-encoder is trained in order to get in output the same vectors as the ones presented as inputs. For each word, the vector of numerical values produced by the hidden layer will be used as the combined word embedding.

4.3 Neural network architecture

Neural networks accepting only digital data vectors, features must be represented as numerical values. We identify some non-numeric features (POS tags, dependency labels and word governors), we need to convert them to a digital representation. We propose to use a one-hot representation to replace the POS tags and the dependency labels. For instance, as we use 25 POS tags, we represent the i^{th} POS tag by a 25-dimensional vector, with all its elements equal to 0, except for the i^{th} one, which is equal to 1.

The word governors and the current words are represented by their word embeddings. Figure 3 presents an example of a 252-dimensional feature vector for one word. An input is the concatenation of 5 word feature vectors.

We propose to extend classical multilayer perceptron classifier (MLP) by using the multi stream strategy for training the network. An MLP multi stream (MLP-MS) architecture is used in order to better integrate the contextual information from neighboring words. This

current word Embed vec 100 dim	word length	PAP	3 3-grams features	Pos tag vec 25 dim	dependency labels vec 22 dim	word governor Embed vec 100 dim

Figure 3: Neural network input features vector format.

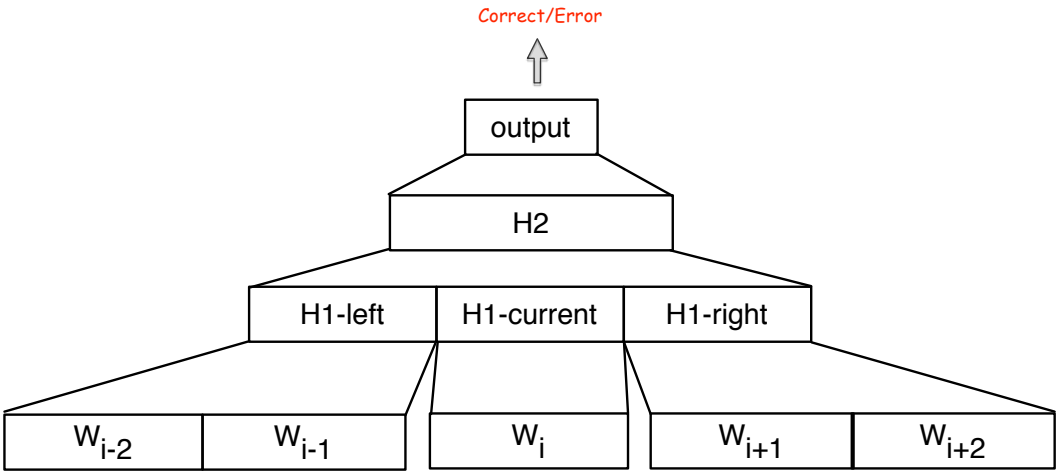

Figure 4: MLP-MS architecture for ASR error detection task.

architecture is inspired by [4] where they integrate word and semantic features for theme identification in telephone conversations. The training of the MLP-MS is based on pre-training the hidden layers separately and then fine tuning the whole network. The proposed architecture, depicted in Figure 4, is detailed as follows: three feature vectors are used as input to the network. These vectors are respectively the feature vector representing the two left words (L), a feature vector representing the current word (W) and a feature vector for the two right words (R). Each feature vector is used separately in order to train a multilayer perceptron (MLP) with a single hidden layer. The output layer has two nodes corresponding to the labels **c** and **e**. Details are described in [5, 6].

4.4 Experiments

The performance of the CRF and neural network approaches is evaluated and compared by using f-measure (based on recall and precision measures) for the erroneous word prediction, and by using global Classification Error Rate (CER) defined as the ratio of the number of misclassifications over the number of recognized words.

4.4.1 Experimental data

Experimental data are based on the entire official ETAPE corpus [8], composed by audio recordings of French Broadcast News shows with manual transcriptions.

Name	#words	WER	err Ferti.
Train	316K	25.9	3.29
Dev	50K	25.2	3.65
Test	53K	22.5	3.58

Table 7: Description of the experimental corpus.

Corpus	Approach	f-measure	CER
Dev	CRF / state-of-the-art baseline	61.08	10.38
	MLP best single word embedding	59.47	10.06
	MLP word embedding combination	63.38	9.79
	MLP word embedding combination + prosody	66.23	9.52
Test	CRF / state-of-the-art baseline	60.53	8.56
	MLP word embedding combination	63.23	8.07
	MLP word embedding combination + prosody	65.55	7.96

Table 8: Error detection results on ASR transcriptions.

This corpus was enriched by automatic transcriptions generated by an ASR system, which is the multi-pass LIUM ASR system existing before EUMSSI. This system is based on the CMU Sphinx decoder, using GMM/HMM acoustic models. This ASR system won the ETAPE evaluation campaign in 2012. A detailed description is presented in [3].

The automatic transcriptions have been aligned with reference transcriptions using the sclite [4] tool. From this alignment, each word in the corpora has been labeled as correct or incorrect (error). Size, WER and ASR error fertility of the corpora are described in Table 7. The fertility of an ASR error specifies the number of contiguous errors, including insertions and substitutions, observed in the automatic transcriptions for a misrecognized word in the reference transcriptions.

4.5 Results

Our experiments showed that using auto-encoder to combine different word embeddings provides some significant improvements in terms of ASR error detection. Details are provided in [5, 6].

Table 8 presents final results in terms of CER and f-measure of CRF and our neural network approach.

These experimental results shows that our neural network approach, using a combination of word embeddings (the best single ones: skip-ngram, GloVe, and CBOW), outperforms the state-of-the-art CRF approach for ASR error detection.

In [7], we have also proposed to add some prosodic features in addition to the ones used in the input vector illustrated in Figure 3. This results in a slight reduction of the classification error rate.

Finally, on the test corpus, our approach based on neural networks and word embeddings combination reduces the CER of 7% in comparison to the previous state-of-the-art

[4] http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

CRF approach presented in [1] , which has been computed as being a statically significant reduction.

5 CONCLUSION AND PERSPECTIVES

This year, effort was produced in order to get the fastest possible ASR systems, with good accuracy, for both English and German languages. We have accelerated a lot our 2015 EUMSSI ASR system in comparison to the 2014 EUMSSI ASR system: we divided by ten the computation time needed to process speech. This improvement is more important if we compare the 2015 system to the 2013 LIUM ASR system, built before the beginning of the EUMSSI project: in two years, we divided the computation time by 25.

In the same time, we have improved the accuracy of our ASR system: our very good results in both international evaluation campaigns for English and German languages illustrate this.

Moreover, we are exploring new neural approaches in order to detect ASR errors. This task can be very useful in the framework of the EUMSSI project for several reasons: (i) to filter misrecognized words to reduce false alarms when looking for automatic transcriptions containing some requested words, (ii) to help natural language processing applied on automatic transcriptions (like name entity recognition), and (iii) to improve the ASR performances by injecting confident automatic transcriptions into the training corpus of acoustic model: larger amount of training data improves the quality of acoustic models. Preliminary results outperforms the state-of-the-art based on CRF approaches. Next year, we will continue this study and we will integrate this information in the data provided in the EUMSSI demonstrators. Last, we have now to integrate the technology we developed during the two first years into the EUMSSI workflow, and to process all the videos provided by our partners, and especially Deutsche Welle.

References

[1] F. Béchet and B. Favre. ASR error segment localisation for spoken recovery strategy. In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference, 2013.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res., 12:2493--2537, 2011.

[3] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? In Interspeech, Brighton, Royaume-Uni, 2009.

[4] Y. Estève, M. Bouallegue, C. Lailler, M. Morchid, R. Dufour, G. Linarès, D. Matrouf, and R. D. Mori. Integration of word and semantic features for theme identification in telephone conversations. In 6th International Workshop on Spoken Dialog Systems (IWSDS 2015), 2015.

[5] S. Ghannay, Y. Esteve, and N. Camelin. Word embeddings combination and neural networks for robustness in asr error detection. In European Signal Processing Conference (EUSIPCO 2015), Nice, France, volume 31, 2015.

[6] S. Ghannay, Y. Estève, N. Camelin, C. Dutrey, F. Santiago, and M. Adda-Decker. Combining continuous word representation and prosodic features for asr error prediction. In Statistical Language and Speech Processing, pages 84--95. Springer, 2015.

[7] S. Ghannay, Y. Estève, N. Camelin, C. Dutrey, F. Santiago, and M. Adda-Decker. Combining Continuous Word Representation and Prosodic Features for ASR Error Prediction. In 3rd International Conference on Statistical Language and Speech processing (SLSP 2015), 2015.

[8] 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.

[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] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Proceedings of Workshop at ICLR, 2013.

[12] R. C. Moore and W. Lewis. Intelligent selection of language model training data. In Proceedings of the ACL Conference Short Papers, pages 220--224, Juillet 2010.

[13] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek. Contextual information improves oov detection in speech. in North American chapter of thes Association for Computational Linguistics (NAACL), 2010.

[14] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), volume 12, 2014.

[15] 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.

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

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

[18] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semisupervised learning. In ACL, pages 384--394, 2010.

[19] K. Veselý, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In Interspeech 2013, Lyon, France, 2013.

[20] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, 2008.