FBK @ IWSLT 2014 - ASR track

B. BabaAli[1], R. Serizel[2], S. Jalalvand[2], D. Falavigna[2], R. Gretter[2] and D. Giuliani[2]

[1]College of Science, University of Tehran, Tehran, Iran
[2]HLT research unit, Fondazione Bruno Kessler (FBK), Trento, Italy
babaali@ut.ac.ir (giuliani, falavi, gretter)@fbk.eu

Abstract

This paper reports on the participation of FBK in the IWSLT 2014 evaluation campaign for Automatic Speech Recognition (ASR), which focused on the transcription of TED talks. The outputs of primary and contrastive systems were submitted for three languages, namely English, German and Italian.

Most effort went into the development of the English transcription system. The primary system is based on the ROVER combination of the output of 5 transcription subsystems which are all based on the Deep Neural Network - Hidden Markov Model (DNN-HMM) hybrid. Before combination, word lattices generated by each sub-system are rescored using an efficient interpolation of 4-gram and Recurrent Neural Network (RNN) language models. The primary system achieves a Word Error Rate (WER) of 14.7% and 11.4% on the 2013 and 2014 official IWSLT English test sets, respectively. The subspace Gaussian mixture model (SGMM) system developed for German achieves 39.5% WER on the 2014 IWSLT German test sets. For Italian, the primary transcription system was based on hidden Markov models and achieves 23.8% WER on the 2014 IWSLT Italian test set.

1. Introduction

This paper describes the English, German, Italian FBK large vocabulary continuous speech recognition systems developed for the IWSLT 2014 evaluation campaign (*http://workshop2014.iwslt.org*). As the IWSLT 2013 evaluation campaign [1], the ASR track of the IWSLT 2014 evaluation campaign focused on the transcription of TED talks (*http://www.ted.com*). The main challenges for automatic transcriptions of TED talks include: variability in acoustic conditions, large variability of topics (hence a large, unconstrained vocabulary), presence of non-native speakers and a rather informal speaking style.

Most effort went into the development of the English transcription system. The primary system for English is based on the ROVER combination [2] of the output of 5 transcription sub-systems. Most of the progress demonstrated for English, w.r.t. the FBK participation into the IWSLT 2013 campaign [3], is due to the switching from the Hidden Markov Model - Gaussian Mixture Model (HMM-GMM) approach to DNN-HMM hybrid systems, the use of an improved n-gram language model, and an N-best list rescoring strategy based on an interpolation of n-gram and RNN Language Models (LMs). In addition, we took advantage by using the Kaldi open source toolkit for system development [4].

In this paper, more details are reported for the experiments conducted for English than for German and Italian.

The rest of this paper is organized as follow. Section 2 describes the speaker diarization module, while Section 3 describes the ASR systems developed for English and Section 4 describes the ASR systems developed for German and Italian. Section 5 presents the automatic transcription results achieved on the TED talk data for all languages. Finally, some conclusions are reported in Section 6.

2. Speaker diarization

The input audio signal is first processed by a speaker diarization module which performs: start-end point detection, speech segment classification and segment clustering based on Bayesian information criterion [5]. At the end of this process, each audio file has assigned a set of temporal segments, each having associated a label that indicates the cluster to which it belongs (e.g. female_1, male_1, etc). This processing is common to all transcription systems presented in this paper and was not changed since the IWSLT 2013 evaluation [3].

3. English Transcription System

3.1. Acoustic data selection

Acoustic Model (AM) training was performed using in-domain data. To this end, TED talk videos released before the cut-off date, 31 December 2010, were downloaded with the corresponding subtitles which are not a verbatim transcription of the speech. Subtitles are, in fact, content-only transcriptions in which anything irrelevant to the content is ignored, including most non-verbal sounds, false starts, repetitions, incomplete or revised sentences and superfluous speech by the speaker. A simple but robust automatic procedure was implemented to select only audio data with an accurate transcription. The approach adopted is that of se-

This work was done while Bagher BabaAli was at FBK as a Visiting Researcher.

lecting only those portions in which the human transcription and an automatic transcription agree [6]. For details on the speech data selection procedure adopted the reader can refer to [3].

The collected data consisted in 820 TED talks, for a total duration of ∼216 hours, with ∼166 hours of actual speech. The speech data selection procedure resulted in ∼144 hours of transcribed speech effectively used for AM training. This year, for acoustic model training we used only this in-domain data, while the previous year, in-domain data was augmented with HUB4 training data [3].

3.2. LM training

Text data used for training the LMs are those released for the IWSLT2013-SLT Evaluation Campaign. Before training, texts were cleaned, normalized (punctuation was removed, numbers and dates were expanded) and double lines were removed. Training documents come from the following three sources:

- **giga5** GIGAWORD 5-th edition. Contains documents stemming from seven distinct international sources of English newswire. It is released from the Linguistic Data Consortium (see *http://www.ldc.upenn.edu*). In total it contains about 4G words.

- **wmt13** Formed by documents in WMT12 news crawl, news commentary v7 and Europarl v7 (see IWSLT2013 official web site for some more details about these corpora). In total it contains about 1G words.

- **ted13** An in-domain set of texts extracted from TED talks transcriptions. It contains about 2.7M words.

Three 4-gram LMs, namely **giga5**, **wmt13** and **ted13** were independently trained on the three sources using the modified shift-beta smoothing method as supplied by the IRSTLM toolkit [7]. Then, two additional "mixture" LMs were trained using the "mix" adaptation method implemented in the IRSTLM toolkit [7]. The **two-mix** LM is built mixing the smoothed (with the modified shift-beta approach) n-grams of both **wmt13** and **ted13** collections, the **all-mix** LM is obtained mixing the smoothed (with improved Kneser-Ney method [8]) n-grams of all of the three collections aforementioned: **giga5**, **wmt13** and **ted13**. We point out that in the case of **all-mix** LM training no pruning of singleton 4-grams was applied.

A further 4-gram LM, namely **sel172M**, was trained on 172M words automatically selected from **giga5** collection in order to match the in-domain set of documents **ted13**. Also in this case the IRSTLM toolkit was employed, together with the modified shift beta method for smoothing probabilities of the n-grams not seen in the training set. The method used for automatically selecting documents from the **giga5** collection is based on "term frequency inverse documents frequency" (TFIDF) coefficients and uses the **ted13** collection as seed corpus. Details can be found in [9].

Finally, two different RNN LMs (namely **RNNLM1** and **RNNLM2**) were trained, using the toolkit described in [10], on the **ted13** collection and on a text corpus including both the **ted13** collection and a subset of documents (containing around 10M words) automatically extracted from the **giga5** collection, respectively. Hence, the **RNNLM2** LM was trained over around 12.7 M words, mapping the singletons into the "<unk>" symbol. The RNNLM1 LM has 450 hidden neurons in its hidden layer and the RNNLM2 LM has 500 hidden neurons.

Note that, the **wmt13** LM is the LM used by ASR systems developed for the IWSLT 2013 evaluation, while the **two-mix** LM is used by all ASR systems developed for the IWSLT 2014 evaluation. Perplexity (PP) and out-of-vocabulary (OOV) rates measured on the reference transcriptions of the IWSLT English 2010 development data set (containing 44505 words) are reported in Table 1. We can see that the **two-mix** LM exhibits a significant lower perplexity that the **wmt13** LM. Column "Interp." in the table reports PP and OOV rate obtained by linearly interpolating the **all-mix**, **ted13**, **sel172M**, **RNNLM1** and **RNNLM2** LMs: interpolation weights are estimated in order to minimize the overall perplexity on the transcriptions of the English 2010 development set. Interpolation of these LMs is applied at recognition time for N-best list rescoring, as it will be detailed in Section 3.4.2.

LM	giga5	wmt13	ted13	two-mix	Interp.
PP	495	461	223	378	289
%OOV	0.4	1.7	7.5	1.6	0.3

Table 1: Perplexities and % OOV rates measured with several LMs on transcriptions of IWSLT English 2010 development data set.

3.3. Lexicon

Word pronunciations in the English lexicon are based on a set of 45 phones. They were generated by merging different source lexica for American English (LIMSI '93, CMU dictionary, Pronlex). In addition, phonetic transcriptions for a number of missing words were generated by using the phonetic transcription module of the Festival speech synthesis system. The lexicon did not change with respect to the previous year.

3.4. ASR system development using Kaldi

In the open source software Kaldi [4], there are two separate setups for neural network training implementation, namely Dan's and Karel's setups or recipes [11, 12]. In both of these setups, the last (output) layer is a softmax layer whose output dimension equals the number of context-dependent states of a pre-trained HMM-GMM system. The neural net is trained

to predict the posterior probability of each context-dependent HMM state [13, 14]. During decoding the posterior probabilities are divided by the prior probability of each state to form a pseudo-likelihood that is used in place of the state emission probabilities in the HMM. Depending on which of the two setups is used the performance is different because of many differences in the recipes. For example, Karel's setup uses pre-training but Dan's setup does random initialization; Karel's setup uses early stopping using a validation set but Dan's setup uses a fixed number of epochs and averages the parameters over the last few epochs of training. Many other aspects of the training procedure are also different (nonlinearity types, learning rate schedules, etc.). Two speaker-adaptive DNN-HMM systems were developed by using the Dan's and Karel's setups.

3.4.1. Acoustic modeling

For acoustic modeling 13 mel-frequency cepstral coefficients (MFCCs), including the zero order coefficient, are extracted from the signal every 10ms by using a Hamming window of 25ms length. These features are then mean/variance normalized on a speaker-by-speaker basis, spliced by +/- 3 frames next to the central frame and projected down to 40 dimensions using linear discriminant analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). A single feature-space Maximum Likelihood Linear Regression (fMLLR) transform for each training speaker is then estimated and applied to train speaker-adaptively trained (SAT) triphone HMMs. These SAT triphone HMM have 6,349 tied-states and 130,000 Gaussians. The speaker-adaptive DNN-HMM hybrid systems are built on top of LDA-MLLT-fMLLR features and SAT triphone HMMs.

A first DNN is trained using the Karel's setup. An eleven frames context window of LDA-MLLT-fMLLR features (5 frames at each side) is used as input to form 440 dimensional feature vector. The DNN have 6 hidden layers each with 2048 neurons, the resulting architecture can be summarized as follows: 440x2048x2048x2048x2048x2048x2048x6349. The DNN is trained in several stages including Restricted Boltzmann Machines (RBM) pre-training, mini-batch Stochastic Gradient Descent training, and sequence-discriminative training such as Minimum Phone Error (MPE) and state-level Minimum Bayes Risk (sMBR).

A second DNN is trained based on the Dan's setup. A nine frames context window of LDA-MLLT-fMLLR features (4 frames at each side) is used as input to form 360 dimensional feature vectors. The DNN is a p-norm DNN with 5 hidden layers and p-norm (input, output) dimensions of (4000, 400) respectively, i.e. the nonlinearity reduces the dimension tenfold [12]. 12000 sub-classes are used, and the number of parameters is 11.0 million. The Dan's setup does not support RBM pretraining. Instead it performs something similar to the greedy layer-wise supervised training [15] or the layer-wise backpropagation of [14]. The network is initialized randomly with one hidden layer, trained for a short time (typically less than an epoch, meaning less than one full-pass through the data), then the layer of weights that go to the softmax layer is removed, a new hidden layer and two sets of randomly initialized weights are added, and trained again. This is repeated until we have four layers. The initial and final learning rates in our training setup are 0.08 and 0.0008 respectively, and during training is decreased exponentially, except for a five epochs at the end during which it is kept fixed. Dan's setup was originally written to support parallel training on multiple CPUs or GPUs. During training, a data-parallel method based on a periodic averaging the parameters of separate Stochastic Gradient Descent runs.

3.4.2. Decoding process

At recognition stage, LDA-MLLT-fMLLR features are first generated by using auxiliary HMMs. To this end, a decoding pass with speaker-independent GMM-HMM is conducted to produce a word lattice for each utterance. A single fMLLR transform for each speaker is then estimated from sufficient statistics collected from word lattices with respect to SAT triphone HMMs. These transforms are hence used in the second decoding pass with SAT HMM to produce new word lattices. A second set of fMLLR transforms is estimated from new word lattices and combined with the first set of transforms. Then a decoding pass is conducted on the obtained fMLLR adapted acoustic features with the DNN-HMM hybrid system, where the DNN is trained to provide posterior probability estimates for the SAT triphone HMM tied-states.

All decoding passes make use of a decoding graph built using a "pruned" version of the **two-mix** LM introduced above. The word lattice generated for each utterance by the DNN-HMM hybrid system is rescored with the "non pruned" **two-mix** LM in order to produce the final ASR hypothesis. Alternatively, as mentioned in Section 3.2, N-best (N=100) list is generated and rescored. In this case, rescoring consists in recomputing, for each hypothesis in the list, the corresponding LM probability as a linear interpolation of the probabilities given by the **all-mix**, **ted13**, **sel172M**, **RNNLM1** and **RNNLM2** LMs. Its worthwhile to mention that the interpolation weights are estimated in order to minimize the perplexity over all the 1-best hypotheses.

3.5. Development of complementary ASR systems

In view of system combination with ROVER, we explored the way to develop complementary systems. To this end, acoustic models of the HMM-GMM system for the IWSLT 2013 ASR English evaluation [3] were used to provide tied-state alignment to train two additional DNN-HMM hybrid systems which are described below.

3.5.1. Two-pass HMM-GMM system

The two decoding pass HMM-GMM system developed for the IWSLT 2013 evaluation uses the **wmt13** LM [3]. A first complementary systems developed for the IWSLT 2014 eval-

uation is obtained using the **two-mix** LM instead.

3.5.2. DNN-HMM systems

The first DNN-HMM system was trained on the tied-state alignment obtained with the SAT triphone HMMs used in the first decoding pass by the 2013 HMM-GMM system. The DNN was however trained on unnormalized acoustic features. The second DNN-GMM system was trained on the tied-state alignment obtained with the SAT triphone HMMs used in the second decoding pass by the 2013 HMM-GMM system and on SAT features. At recognition stage, a two pass DNN-HMM decoding system is obtained when word transcriptions generated by the DNN-HMM system using unnormalized acoustic features are used to supervise the extraction of the SAT acoustic features for a second decoding pass with the SAT DNN-HMM system.

First-pass DNN-HMM
The first DNN is trained on 13 MFCC, including the zero coefficient, without speaker normalization. A 31-frame context window is applied, the 403-dimensional features vector is then decorrelated with discrete cosine transform (DCT) and projected on a 208-dimensional feature vector. Average and covariance normalisations are applied to this later feature vector and the resulting, normalized, vector is used as input to the DNN. The DNN is composed of 5 hidden layers with 1500 elements per layer. The DNN is trained with cross-entropy on 10021 triphone tied-states targets obtained from time alignment with the first pass models of the 2013 HMM-GMM system. The resulting architecture can be summarized as follows: 208x1500x1500x1500x1500x1500x10021.

The TNet software package [16] is used for training. The training set for the DNN is composed only of TED data as explained above. The training set is split into two sets with non-overlapping speaker: training (90%) and cross-validation (10%). The DNN weights are initialized randomly and pre-trained with RBM [17, 18]. The first layer is pre-trained with a Gaussian-Bernoulli RBM trained during 10 iterations with a learning rate of 0.005. The following layers are pre-trained with a Bernoulli-Bernoulli RBM trained during 5 iterations with a learning rate of 0.05. Mini-batch size is 500. For the back propagation training the learning rate is kept to 0.08 as long as the frame accuracy on the cross-validation set progresses by, at least, 0.5% between successive epochs. The learning rate is then halved at each epoch until the frame accuracy on the cross-validation set fails to improve by at least 0.1%. The mini-batch size is 1024. In both pre-training and training, a first-order momentum of 0.9 is applied.

Second-pass SAT DNN-HMM
The second DNN is trained on the 39 SAT features as generated for the second pass triphone HMM of the 2013 HMM-GMM system. A 31-frame context window is applied. The resulting 1209-dimensional features vector is decorrelated with DCT and projected on a 468-dimensional feature vector. Average and covariance normalization is applied and the resulting, normalized, vector is used as input to the DNN. The DNN is composed of 5 hidden layers with 1500 elements per layer. It is trained with cross-entropy on 10021 triphone tied-states targets obtained from time alignment with the second pass models of the HMM-GMM baseline. The resulting architecture can be summarized as follows: 468x1500x1500x1500x1500x1500x10021. The training was conducted following the same set up as for the first-pass DNN above.

4. German and Italian transcription systems

For this evaluation, we decided to focus our efforts mostly on English and to dedicate a limited attention to German and Italian. For both languages we wanted to compare our in-house proprietary system with the Kaldi recognizer, but due to the aforementioned limitations, at the end we did the following submissions:

- **Italian primary** in-house SAT HMM-GMM system (see [3] for details);
- **Italian contrastive1** SAT Subspace Gaussian Mixture Model (SGMM) system developed with Kaldi [4];
- **German primary** SAT SGMM system developed with Kaldi.

4.1. Acoustic data

Concerning *Italian*, we could use the following corpora:

- **Euronews Italian Data** provided by the organizers, amounting to about 76h:38m of reliable speech. The corresponding transcription was obtained after a further step of light supervision training, using the domain dependent AMs trained on the originally provided data.
- **Italian Internal data**: about 216h:31m of reliably transcribed (partly manually, partly with light supervision techniques) speech collected in the previous years and belonging to 3 domains: *Apasci*, a phonetically balanced corpus; *Italian Parliament* recordings, *TV news* recorded from RAI. All this data were recorded before June 30th, 2011.

This data amounted to slightly more than 293 hours, but in order to speed up Kaldi experiments we decided to sample the data, by keeping only the first 100 sentences for each audio file. This resulted in about 154h:19m of speech (74h:30m Euronews data + 79h:49m Internal data).

Instead, the in-house proprietary system was trained on the Italian Internal data only (216h:31m).

Concerning *German*, we could use the following corpora:

- **Euronews German data** provided by the organizers, amounting to about 72h:18m of reliable speech. The

corresponding transcription was obtained after a further step of light supervision training, using the domain dependent AMs trained on the originally provided data.

- **German WEB data**: about 158h:47m of speech data transcribed using light supervision techniques, collected before July 2012.

The effective material used for training consisted in about 206h:54m of speech (66h:45m Euronews data + 140h:09m German WEB data).

4.2. Textual data

To build the **German LM** we used text data from various sources, including Europarl data, news from 2005 to June 30th, 2012, and of course the **ASR LM Training Data German** provided by IWSLT organizers. The total amount of words was about 1,130 million words. These data were processed in order to perform a normalization including in particular number and compound words splitting, which was performed in a fully automatic way described in [3]. After normalization, a 4-grams language model was built, resulting in about 481.3 millions of 4-grams. A pruned version of this LM, including about 9,7 millions of 4-grams, was used to build the FST used to build the lattices during decoding, while the full LM was used to rescore the lattices. The lexicon was fixed to the most frequent 200K words; the phonetic transcription was generated by our in-house system.

To build the **Italian LM** we used text data coming from news collected from 2005 to June 30th, 2011, in addition to the **ASR LM Training Data Italian** provided by IWSLT organizers. The total amount of words was about 985 million words. After text normalization and number splitting, a 4-grams language model was built, resulting in about 427,6 millions of 4-grams. For the contrastive system using Kaldi, a pruned version of this LM, including about 6,5 millions of 4-grams, was used to build the FST used to build the lattices during decoding, while the full LM was used to rescore the lattices. For the primary in-house system a static FSN was built using a pruned version of the LM, including was built 15,3 million 4grams. In both cases, the lexicon was fixed to the most frequent 200K words; the phonetic transcription was generated by our in-house system.

4.3. Decoding process

Both for German and Italian, we performed a two stage recognition. For the two Kaldi SGMM systems (German primary and Italian contrastive1) the two stage recognition was followed by a linguistic rescoring stage, obtained using the full LM over the generated lattices.

For the in-house system (Italian primary), no final LM rescoring was performed. Details about the AM adaptation performed for the second step decoding are described in [3].

5. Recognition Experiments

5.1. Results on English TED talks

Recognition experiments were carried out on the IWSLT 2014 English ASR development and evaluation data sets listed in Table 2. These data sets were released over several IWSLT evaluation campaigns. Recognition experiments on dev2012, tst2013 and tst2014 were always conducted in fully automatic mode. Instead, recognition experiments on all the other data sets (dev2010, tst2011 and tst2012) were conducted exploiting the provided manual segmentation.

Data Set	N. of Talks	Duration
dev2010	19	4h:00m
tst2011	8	1h:07m
dev2012	10	1h:57m
tst2012	11	1h:45m
tst2013	28	4h:38m
tst2014	15	2h:24m

Table 2: Details of the IWSLT 2014 English ASR development (dev) and evaluation (tst) data sets.

As a reference, Table 3 reports results achieved with the 2013 HMM-GMM system [3]. Column "Pruned LM" gives results obtained by the second decoding pass (see Section 3.5.1) using a pruned version of the LM, that is the **wmt13** LM introduced in Section 3.2. Column "Rover" in Table 3 reports results achieved with a combination, using ROVER, of 4 recognition outputs resulting from rescoring the word lattices generated by the second decoding pass by using different unpruned LMs [3]. The 23.7% WER reported on tst2013 data set is the result achieved by the FBK 2013 primary system in the IWSLT 2013 ASR evaluation campaign.

Data Set	System 2013	
	Pruned LM	Rover
dev2010	17.5	16.1
tst2011	15.6	13.6
dev2012	19.3	-
tst2012	17.6	16.1
tst2013	25.2	23.7

Table 3: % WER achieved by the HMM-GMM 2013 system on several English data sets. Results were obtained by: decoding with the pruned **wmt13** LM and performing ROVER combination of 4 different rescored outputs.

5.1.1. Experiments with Kaldi systems

Table 4 reports results with two SAT DNN-HMM systems developed with the Kaldi toolkit. "Dan" and "Karel" indicate the recipe, provided within the Kaldi toolkit, used to train the DNN.

Data Set	Kaldi DNN implementation	
	"Dan" (Pruned LM/Rescoring)	"Karel" (Pruned LM/Rescoring)
dev2010	14.8/13.4	13.4/12.5
tst2011	12.8/11.5	11.5/10.7
dev2012	18.0/17.0	16.3/15.3
tst2012	12.7/11.7	11.7/10.8
tst2013	19.4/18.0	17.5/16.4

Table 4: Results, in % WER, achieved by two different DNN-HMM systems on several English data sets. For each system and data set, it is reported the result achieved by: decoding with the pruned **two-mix LM** and performing rescoring of word lattices with the corresponding unpruned LM.

From results reported in Table 4 we can conclude that the "Karel" recipe allows to train a DNN which is consistently more effective than the DNN trained with the "Dan" recipe. In addition, performing rescoring of word lattices with the unpruned LM provides tangible benefit, for example dropping the WER, on the dev2010 data set, from 13.4% to 12.5% when using the the "Karel" DNN-HMM system.

The comparison of results reported in Tables 3 and 4, allows to appreciate the net improvements of the 2014 DNN-HMM systems over the 2013 HMM-GMM system. We believe that this major improvement can be attributed to the adoption of the deep learning paradigm for acoustic modeling, a better LM and a more comprehensive training procedure offered by the Kaldi development toolkit.

Data Set	Kaldi DNN implementation	
	"Dan" (N-best rescoring)	"Karel" (N-best rescoring)
dev2010	12.9	11.9
tst2011	10.7	10.0
dev2012	15.5	14.2
tst2012	11.0	10.4
tst2013	16.5	15.2

Table 5: % WER achieved by two different DNN-HMM systems on several English data sets by performing N-best (N=100) list rescoring using an interpolation of 4-gram and RNN LMs.

Table 5 reports results performing N-best list rescoring using an interpolation of 4-gram and RNN LMs, as described in section 3.4.2. By comparing these results with those in Table 4, we can notice the effectiveness of the N-best list rescoring method.

5.1.2. Experiments with complementary systems

Table 6 reports recognition results obtained with the 2014 HMM-GMM and DNN-HMM systems described in Section 3.5 without performing word lattice rescoring. Rows p1-GMM and p1-GMM+p2-GMM report results achieved performing one and two passes of decoding with the 2013 HMM-GMM system (see Section 3.5.1). Performing a single decoding pass 17.8% and 25.7% WER are achieved on the dev2010 and tst2013 data sets, respectively. While performing two decoding passes 16.3% and 23.4% WER are achieved on the dev2010 and tst2013 data sets, respectively. These latter results can be directly compared with the 17.5% and 25.2% WER, achieved by the 2013 HMM-GMM system as reported in the "Pruned LM" column of Table 3. The performance improvement can be attribute at the use of a better LM (that is **two-mix** Vs. **wmt13** LM). Results achieved performing one and two decoding passes with the DNN-HMM systems are reported in rows p1-DNN and p1-DNN+p2-DNN, respectively. We can see that performing a single decoding pass 16.5% and 21.9% WER are achieved on the dev2010 and tst2013 data sets, respectively. While performing two decoding passes 15.4% and 20.7% WER are achieved on the dev2010 and tst2013 data sets, respectively. These results confirm, once again, the effectiveness of the DNN-HMM hybrid approach. However, they are not as good as those obtained with DNN-HMM systems developed with the Kaldi toolkit and reported in Table 4 ("Pruned LM" condition).

Complementary System	dev2010	tst2013
p1-GMM	17.8	25.7
p1-GMM+p2-GMM	16.3	23.4
p1-GMM+p2-DNN	15.6	20.1
p1-DNN	16.5	21.9
p1-DNN+p2-GMM	15.5	22.0
p1-DNN+p2-DNN	15.4	20.7

Table 6: Results, in % WER, with different complementary system configurations on the dev2010 and tst2013 English data sets.

Table 6 reports also results obtained alternating recognition passes conducted with HMM-GMM and DNN-HMM systems. For example, row p1-DNN+p2-GMM reports results obtained performing the first pass with the DNN-HMM system and the second pass with the HMM-GMM system, this results in 15.5% and 22.0% WER on the dev2010 and tst2013 sets, respectively. In the following, we will refer to this system as "AltSystem1". One additional combination we have tried was as follows. The AltSystem1 was used to generate a word lattice which was acoustically rescored using the p2-DNN systems: we will refer to this system as "AltSystem2". The AltSystem2 resulted in 15.2% and 20.3% WER on the dev2010 and tst2013 data sets, respectively. The output of systems AltSystem1 and AltSystem2 were considered for system combination in the hope that they were different one each other enough.

Sub-systems	dev2010	tst2013
DNN-HMM "Dan"	12.9	16.5
DNN-HMM "Karel"1	11.9	15.3
DNN-HMM "Karel"2	11.9	15.2
AltSystem1	15.5	22.0
AltSystem2	15.2	20.3
	ROVER	
	11.7	14.7

Table 7: Results, in % WER, achieved by individual sub-systems, and performing ROVER-based system combination, on the dev2010 and tst2013 English data sets.

5.1.3. System combination

The 2014 primary system for English is based on the principle of system combination by means of ROVER. Table 7 reports recognition results achieved by the 2014 primary system, which combines the outputs of 5 sub-systems previously introduced. Results achieved by individual sub-systems are also reported. DNN-HMM "Karel"1 and DNN-HMM "Karel"2 denotes two sub-systems that differ only for the number of iterations in training of the corresponding DNN.

For the tst2013 data set we can see that an improvement of 0.5% WER is achieved with the ROVER combination w.r.t. the best sub-system entering in the combination: from 15.2% to 14.7% WER. The obtained 14.7% WER can be directly compared with the 23.7% WER obtained by the 2013 primary system on the same data (see Table 3). This represents a substantial improvement in terms of performance.

On the 2014 IWSLT English test set the official evaluation result achieved by the primary system is 11.4% WER, with an improvement of 0.7% WER w.r.t. the performance of the best sub-system entering in the ROVER combination, that is 12.1% WER.

5.2. Results on German TED talks

The subspace Gaussian mixture model system developed for German achieves 39.5% WER on the 2014 IWSLT German test sets.

5.3. Results on Italian TED talks

For Italian, the primary transcription system was based on hidden Markov models and achieves 23.8% WER on the 2014 IWSLT Italian test set. The contrastive1 transcription system, based on SGMM, achieves 24.6% WER.

6. Conclusions

In this paper we have presented the systems we developed for the participation in the IWSLT 2014 ASR evaluation campaign: we developed systems for the English, German and Italian ASR tracks.

For English, substantial progress, with respect to our primary system submission to IWSLT 2013 campaign [3], was demonstrated. This progress is due to the switching from the pure HMM-GMM approach to the adoption of DNN-HMM hybrid systems, the adoption of a better n-gram language model, and an N-best list rescoring strategy based on an interpolation of n-gram and RNN language models. In addition, we took advantage by using the Kaldi open source toolkit for system development.

7. Acknowledgements

This work was partially funded by the European project EU-BRIDGE, under the contract FP7-287658.

8. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 10th IWSLT Evaluation Campaign," in *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, 2013.

[2] J. Fiscus, "A post-processing system to yield reduced error rates: recognizer output voting error reduction (ROVER)," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1997, pp. 347–354.

[3] D. Falavigna, R. Gretter, F. Brugnara, D. Giuliani, and R. H. Serizel, "FBK@ IWSLT 2013-ASR track," in *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, 2013.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselỳ, "The Kaldi Speech Recognition Toolkit," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2011.

[5] M. Cettolo, "Segmentation, Classification and Clustering of an Italian Broadcast News Corpus," in *Proc. of Content-Based Multimedia Inf. Access Conf. (RIAO)*, Paris, France, 2000, pp. 372–381.

[6] L. Lamel, J.-L. Gauvain, and G. Adda, "Investigating Lightly Supervised Acoustic Model Training," in *Acoustics, Speech and Signal Processing, 2001, IEEE International Conference on*, Salt Lake City, UT, 2001.

[7] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proc. of ICSLP*, Brisbane, Australia, September 2008, pp. 1618–1621.

[8] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.

[9] D. Falavigna and G. Gretter, "Focusing language models for automatic speech recognition," in *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, HK, December 2012.

[10] T. Mikolov, S. Kombrink, L. Burget, J. H. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5528–5531.

[11] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence discriminative training of deep neural networks," in *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, 2013, pp. 2345–2349.

[12] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of Deep Neural Networks with Natural Gradient and Parameter Averaging," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Diego, CA, USA, 2015, To appear.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[14] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Proc. of 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, 2011, pp. 437–440.

[15] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. of the Conference on Advances in Neural Information Processing Systems (NIPS)*, vol. 19, 2007, pp. 153–160.

[16] K. Veselỳ, L. Burget, and F. Grézl, "Parallel training of neural networks for speech recognition," in *Text, Speech and Dialogue*. Springer, 2010, pp. 439–446.

[17] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.