The MLLP ASR Systems for IWSLT 2015

Miguel Ángel del-Agua, Adrià Martínez-Villaronga, Santiago Piqueras
Adrià Giménez, Alberto Sanchis, Jorge Civera, Alfons Juan

Machine Learning and Language Processing
Departament de Sistemes Informàtics i Computaciò
Universitat Politècnica de València, Spain

mdelagua@dsic.upv.es

Abstract

This paper describes the Machine Learning and Language Processing (MLLP) ASR systems for the 2015 IWSLT evaluation campaing. The English system is based on the combination of five different subsystems which consist of two types of Neural Networks architectures (Deep feed-forward and Convolutional), two types of activation functions (sigmoid and rectified linear) and two types of input features (fMLLR and FBANK). All subsystems perform an speaker adaptation step based on confidence measures the output of which is then combined with ROVER. This system achieves a Word Error Rate (WER) of 13.3% on the 2015 official IWSLT English test set.

1. Introduction

TED is a global set of conferences around the world carried out by the non-profit organisation *Sapling Foundation*. Its talks cover a wide range of different topics such as science, culture, economics or politics, always keeping in mind the slogan "ideas worth spreading". The speakers are given a maximum of 18 minutes to present their ideas in the most appealing way they can, typically in a storytelling format.

In order to ensure the maximum spread of these talks, turns out to be essential their transcription and translation. Big efforts have been devoted to this task, such as *The Open Translation Project* (OTP), which aims to reach out to the 4.5 billion people on the planet who do not speak English. Nevertheless, the OTP utilises crowd-based subtitling platforms, powered by volunteers to translate and caption the videos, which is still a very time-consuming task.

TED talks conform a very appropriate case study where new technologies can be applied. Particularly from the machine learning community, the International Workshop on Spoken Language Translation (IWSLT) organises a yearly challenge which aims at evaluating the core technologies in spoken language translation: automatic speech recognition (ASR), machine translation (MT) and spoken language translation (SLT). Automatically transcribing this kind of videos is still a challenging task due to the spontaneous nature of the speech; variety in acoustic conditions, the presence of disfluencies, hesitations and different accents states a great challenge even for cutting-edge technology in automatic automatic speech recognition.

This paper describes the English and German ASR systems developed in the MLLP group for the IWSLT 2015 evaluation campaign. Most effort went into the development of the English recognition system which is based on the ROVER combination of five subsystems. Each of those subsystems was based on hybrid Deep Neural Networks Hidden Markov Models (DNN-HMM) [1] with different input features (MFCCs and filter bank), activation functions (sigmoid and rectified linear) as well as various architectures such as Deep Convolutional Neural Networks (CNN). It is worth noting that all of these systems were entirely trained using our own software; the transLectures-UPV toolkit.

The rest of this paper is organised as follows. Section 2 describes the ASR toolkit used for the experiments. In Section 3 the automatic audio segmentation technique is introduced. Section 4 is devoted to the English transcription system. Similarly, in Section 5 the German ASR system is described. Finally, conclusions are given in Section 6.

2. Translectures-UPV Toolkit

The transLectures-UPV toolkit (TLK) is composed by a set of tools that allows the development of an end-to-end speech recognition system. Its application range extends from feature extraction to HMM and DNN training and decoding. Since last state published of the toolkit [2] new state-of-the-art techniques have been added:

- DNN training and decoding hybrid based systems.
- Support to Convolutional NNs.
- Support to Multilingual NNs.
- DNN speaker adaptation techniques such as output-feature discriminant linear regression (oDLR) [3].
- DNN sequence discriminative training based on Maximum Mutual Information (MMI).

3. Audio Segmentation

The audio segmentation step performed by the MLLP group for English and German can be viewed as a simplified case of ASR, in which the system vocabulary is constituted by the power set of segment classes: speech and background noise.

Provided an audio stream $\boldsymbol{x}$, the segmentation problem can be stated from a statistical point of view as the search of a sequence of class labels $\hat{\boldsymbol{c}}$ so that

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c} \in \mathcal{C}^*}{\operatorname{argmax}}\, p(\boldsymbol{x} \mid \boldsymbol{c})\, p(\boldsymbol{c}) \tag{1}$$

where, as in ASR, $p(\boldsymbol{x} \mid \boldsymbol{c})$ and $p(\boldsymbol{c})$ are modeled by acoustic and language models, respectively. In our case, it should be noted that each word is composed by a single phoneme.

Acoustic models were trained on MFCC feature vectors computed from acoustic samples using TLK. We used a 0.97 coefficient pre-emphasis filter and a 25 ms Hamming window that moves every 10 ms over the acoustic signal. From each 10ms frame, a feature vector of 12 MFCC coefficients is obtained using a 26 channel filter bank. Finally, the energy coefficient and the first and second time derivatives of the cepstrum coefficients are added to the feature vector.

Each segment class is represented by a single-state Hidden Markov Model (HMM) without loops, and its emission probability is modeled by a Gaussian Mixture Model (GMM). Acoustic HMM-GMM models were also trained using TLK, which implements the conventional Baum-Welch algorithm.

A 5-gram back-off language model with constant discount was trained on the sequence of class labels using the SRILM toolkit [4]. Finally, the segmentation process (search) was also carried out by the TLK toolkit.

4. English Transcription System

4.1. Acoustic Modeling

In this section the acoustic modeling process for the English system is described. First, the data selected for training is showed as well as the techniques used for its collection. Then, the training procedure is detailed along with all the subsystems associated.

4.1.1. Data Collection

This year, the IWSLT challenge allowed the use of any publicly available data for acoustic modeling, including TED talks without publication date restrictions (except those listed as disallowed). Given these requirements, roughly 400 hours of TED talks were downloaded from its official web-page [5].

The subtitles attached to a large part of the talks neither match the speaker's speech nor the timings. Therefore, a data filtering process is needed, in which those segments with a deficient or non-existent transcription must be removed. This process was performed in a similar manner to the data filtering performed for building the TEDLIUM corpus [6].

First of all, the input audio was segmented and preprocessed according to the caption timings. Secondly, a recognition step was performed using an out-of-domain acoustic model and a finite state language model. This finite state language model was built using the sequence of words from the reference with silence in-between, allowing loops (hesitations), initial state to any word transitions and from any word to final state transitions.

This way, those segments whose recognition does not match the reference suggest that either the timings are wrongly set or the system is unable to recognise the segment due to non-speech audio. Therefore, after decoding, all of these incorrectly recognised segments were removed, which left us a total of 245 hours of clean speech distributed among 1900 talks.

4.1.2. Training

Regarding feature extraction, two types of acoustic features were extracted. The first type of features are Mel-frequency cepstral coefficients (MFCC), which were extracted with a Hamming window of 25 ms, shifted at 10 ms intervals. The MFCC feature consisted of 16 MFCCs and their first and second derivatives (48-dimensional feature vectors). These feature vectors were then normalised by mean and variance at speaker level. After that, a single feature-space Maximum Likelihood Linear Regression (fMLLR) transform for each training speaker was then estimated and applied to perform speaker-adaptive training (SAT). The second type of features are log Mel filter bank (FBANK) with first and second derivatives which left 120 dimension feature vectors.

Five different acoustic models were trained in our system using TLK. All of them consisted of context-dependent Deep Neural Networks (DNNs) following an hybrid approach. To train these models, we first trained a basic context dependent triphone HMM model, after which a second-pass feature-space Maximum Likelihood Linear Regression (fMLLR) was applied. This model yielded a total of 10492 tied states, estimated following a phonetic decision tree approach [7]. It is worth noting that, in order to obtain the best transcription as to better perform fMLLR, an standard DNN was trained using the MFCCs features. The five models were build on top of these HMM acoustic model and followed a three-pass recognition approach as shown in Fig. 1.

From Fig.1, the *fMLLR CD-DNN* module can be switched among the five different acoustic models. Three of them are feed-forward DNNs and the other two are Deep Convolutional Neural Networks (CNNs). From the first set, all models took as input MFCCs feature frames with a window size of 11. Moreover, all three subsystems shared the same topology: $528 - 2048 * 7 - 10492$, i.e., an input layer with 528 neurons, 7 hidden layers with 2048 neurons and an output layer of 10492 neurons. The pre-training phase technique is also shared, which consisted of the Discriminative Pretraining [8] approach. The first system was a DNN with sigmoid activation functions, trained with the cross-entropy

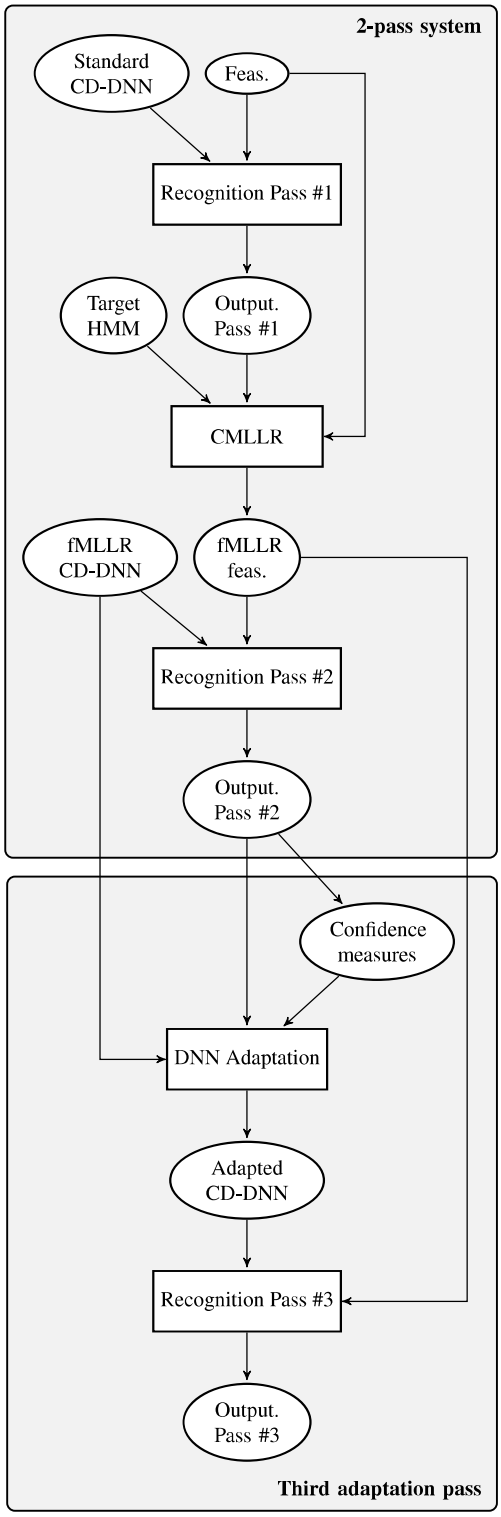

Figure 1: Overview of a multi-pass recognition system including DNN adaptation. Top: 2-pass recognition system using fMLLR features. Bottom: Third pass DNN adaptation.

(CE) criterion (10 epochs) and after that, with sequence discriminative training following the MMI criterion (hereafter DNN-mmi). The second model was a DNN with rectified linear activation functions, trained following the CE criterion during 45 epochs (hereafter DNN-relu). And the third model was a DNN with sigmoid activation functions trained with the CE criterion during 45 epochs (hereafter DNN-sigm).

Two models belong to the second set of acoustic models. Both take as input FBANK features with a window size of 11 and share the same topology. It consist of one convolution layer followed by a max pooling operation, 6 feed-forward hidden layers of 2048 units each, and an output layer of 10492. The convolutional layer is composed of 128 filters with a filter size of 9 and shift of 1. Meanwhile, the max-pooling layer was configured with a pooling width and shift of 2. The difference between both models is the type of activation functions used for the feed-forward layers: sigmoid (CNN-sigm) and rectified linear (CNN-relu).

4.1.3. DNN Speaker Adaptation

The output from the second recognition step was used to carry out speaker adaptation of DNNs (as indicated at the lower box of Fig. 1). The technique used consisted of a conservative training approach, using a very small learning rate and early stopping [9].

Moreover, we made use of confidence measures at word level to exploit inexpensive yet reliable unsupervised speech data. Specifically, confidence measures are estimated from the output of the second recognition pass in order to improve the DNN adaptation step. Although there are many different ways to estimate confidence measures, here we will resort to the conventional approach by which these measures are computed as word posterior probabilities [10].

In order to take advantage of confidence measures, we decided to use them to weight the samples during the adaptation. In this approach, all samples are taken into account, but the contribution of each sample is weighted by its corresponding confidence measure. The rationale behind this method is that only samples with high confidence measures are relevant for the adaptation process, whereas those with low confidence can be neglected. In some way, this method can be seen as a refinement of taking away those samples behind an specified threshold, avoiding the need of estimating that threshold.

Formally, adaptation with weighted samples is based on a modified cross entropy training criterion:

$$\sum_{n=1}^{N} c_n \log p(s_n \mid \mathbf{x}_n)\,, \tag{2}$$

where $\mathbf{x}_1^N$ is the set of frames, s_n is the senone (label) according to the output from the second pass, and $c_n \in [0, 1]$ is its confidence measure. This modified criterion leads to a different way to estimate errors in the Back-Propagation algorithm. In particular, the error for the nth frame δ^n is

Table 1: Stats of the different LM training corpora. The poliMedia [11], VideoLectures.NET and VL.NET subtitles [12] corpora were generated during transLectures project.

Corpus	Sentences	Words	Perplexity
Europarl	2.2M	53M	454.3
Europarl TV	128K	1.2M	454.5
Giga 10^9	22M	557M	296.9
Google Ngrams	-	303B	1871.1
NewsCrawl	53M	1.1B	151.7
poliMedia	4K	95K	1393.1
VideoLectures.NET	5K	127K	871.4
VL.NET subtitles	85K	1.7M	371.5
Wikipedia	82M	1.5B	200.1
TED train	520K	3.7M	218.2

estimated as follows

$$\delta^n = (\mathbf{y}_n - \mathbf{s}_n) \cdot c_n \,, \tag{3}$$

where $\mathbf{y}^n$ is the output of the last layer, and $\mathbf{s}^n$ are the target labels.

4.2. Language Modeling

We used several different text corpora to train the language models. They were preprocessed to normalise capitalisation, remove punctuation marks and transliterate numbers. We can distinguish two different types of corpora, out of domain corpora (OOD), most of them, and in domain corpora (ID), in this case only TED train set. Table 1 summarises the main figures of all the corpora used.

The vocabulary for the language models have been obtained by selecting the 200K most frequent words of a 1-gram LM interpolation of the OOD corpora. The words form the ID corpus are added to this selection, obtaining a final vocabulary of 209 660 words.

With this vocabulary, we trained standard Kneser-Ney smoothed n-gram models for each one of the corpora using the SRILM toolkit [4]. The order of each model is adjusted to 3 or 4 depending on the size of the corpus. The last column of Table 1 shows the perplexity obtained with all these models on the English development set.

All the resulting models are linearly interpolated to obtain a final powerful model adapted to the characteristics of the task, optimising the interpolation weights on the development set [13]. To reduce the size of the final model, it is pruned by removing those n-grams ($n > 1$) whose removal causes (training set) perplexity of the model to increase by less than 2×10^{-10}. This model obtained a perplexity of 126.1.

4.3. Experimental Results

In this section all the recognition experiments performed for the English transcription system are described. Recognition experiments were carried out on the IWSLT 2015 English ASR development and evaluation sets, the statistics of which are shown in Table 2.

Table 2: Statistics of the English ASR development and evaluation sets.

Set	# Talks	Time
tst2013	28	4h:39m
tst2014	15	2h:22m
tst2015	12	2h:25m

Following the IWSLT evaluation requirements, tst2013 was used as development set, tst2014 as progressive evaluation set and tst2015 as evaluation.

The decoding was performed for all the subsystems following the scheme from Fig. 1. The first step was common and its output was used to perform fMLLR speaker adaptation. After that, each subsystem performed the second recognition step, the output of which was used to perform DNN speaker adaptation using confidence measures. Results from these two steps are shown in Table 3.

Table 3: Effect of DNN Speaker Adaptation on each subsystem in terms of WER. Results are shown on tst2013 data-set.

Subsystem	Non-Adapt	Adapt	R. Improvement
DNN-mmi	16.9	16.7	1.2%
DNN-sigm	17.1	16.7	2.3%
DNN-relu	18.5	17.8	3.8%
CNN-sigm	19.4	18.8	3.1%
CNN-relu	18.7	18.0	3.7%

It is worth mention that none of the above results has been subjected to a process of spelling normalisation by means of a global mapping file. As we can observe, the DNN-mmi adaptation has not performed as the rest of system's adaptations. To our knowledge this is because there is not so much room for improvement as occurs in the other systems, and also to the change in the training criterion (from MMI to CE during adaptation).

Finally, a recogniser output voting error reduction (ROVER) algorithm was applied to combine the subsystem's output and further improve the recognition results. The combination weights were estimated based on the development set, giving 2:2:1:1:1 for DNN-mmi, DNN-sigm, DNN-relu, CNN-sigm and CNN-relu. The final scoring results are shown in Table 4. At the time of writing this paper results on the progress test set tst2014 were not provided.

5. German Transcription System

In this section the German ASR system is described. The first section details the data and training procedure, while the second section shows the results obtained by the system.

Table 4: The final results of the English system in terms of WER. (* means official result)

Set	ROVER
tst2013	16.2
tst2015	13.3*

5.1. Training

For the acoustic modelling, we decided not to use the Euronews ASR provided corpus due to processing power constraints and its acoustic conditions being far from target conditions. Instead, we downloaded and processed the German Speechdata Corpus (GSC) [14], an open source corpus recorded and released by the LT and the Teleccoperation group from the Technical University of Darmstadt. This corpus contains 180 different speakers and 36 hours of speech, recorded under controlled conditions with many microphones in parallel. The whole corpus was used as train data. The grapheme-to-phoneme conversion was performed with the help of MaryTTS software [15].

The training procedure for German was the same as the DNN-MFCC used in the English system (Sec. 4.1.2). 48-dimensional MFCC acoustic vectors were extracted and normalised by speaker. A single acoustic model was estimated for German, which consists of a feed-forward DNN with a window size of 11 and 4 hidden sigmoid layers with 2048 neurons each. The output layer features 12237 senones. The network initialisation was performed with the DPT approach, and then the network was trained using the Cross-Entropy error criterion for 10 epochs.

The training and recognition follow the same three-step approach of the English system. An speaker-independent model is used in the first step. The output transcription is then used to perform unsupervised fMLLR adaptation. This second transcription is employed to perform DNN Speaker adaptation (Sec. 4.1.3). In the case of German, no confidence measures have been used for this third step.

The language model for our German system is made up by a standard linear interpolation of 4-gram language models. These models were estimated from different open corpus downloaded from the Internet. The corpora were normalised by lower-casing, removing punctuation marks and transliterating numbers. The corpus statistics after this process can be found in Table 5.

Table 5: Statistics of the German LM corpus.

Corpus	Sentences	Words	Perplexity
Europarl	2M	46M	515.5
News-crawl	135M	2B	352.0
Wikipedia	31M	326M	423.4

When training, the vocabulary was restricted to 200k words, selected with the same procedure described in Section 4.2. The interpolation weights were set to optimise the perplexity of the dev set. In order to improve recognition time, the interpolated model was pruned with a prune factor of 2×10^{-9}. The perplexity of the language model is 290.4.

5.2. Experimental Results

We tested our system on the tst2013 corpus, which was set as the official development corpus of the 2015 challenge. This corpus contains 9 videos from the TEDx website, with varying acoustic conditions. The results are summarised in Table 6. At the time of writing this work results on tst2014 set were not provided.

Table 6: The final results of the German system in terms of WER. (* means official result)

Set	WER
tst2013	43.6
tst2015	43.3*

Unlike the English task, we were not able to obtain state-of-the-art results for the German task. We attribute this result to the lack of relevant in-domain acoustic resources and the simplicity of the approaches employed.

6. Conclusions

In this paper we have described the English and German ASR systems developed for the IWSLT 2015 evaluation campaign. For the first participation of the MLLP group, the presented systems make use of the hybrid approach of HMM-DNN. Particularly, the decoding step of the English system is based on the combination of five different transcription subsystems. Each one built as a three pass recognition system and combining different types of NNs architectures, input features and activation functions. Meanwhile, the German system constitutes our first large scale speech recognition system on this language and it is based on a three pass recognition system with DNN speaker adaptation.

7. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures) and ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no 621030 (EMMA), the Spanish MINECO Active2Trans (TIN2012-31723) research project, the Spanish Government with the FPU scholarship FPU13/06241 and the Generalitat Valenciana with the VALi+d scolarship ACIF/2015/082.

8. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[2] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan, "The translectures-upv toolkit," in *Proc. of VIII Jornadas en Tecnologa del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*, Las Palmas de Gran Canaria (Spain), 2014. [Online]. Available: http://www.mllp.upv.es/wp-content/uploads/2015/04/IberSpeech2014-TLK-camready1.pdf

[3] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. of the SLT*, 2012, pp. 366–369.

[4] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP'02*, September 2002, pp. 901–904.

[5] "TED: Ideas worth spreading," https://www.ted.com.

[6] A. Rousseau, P. Deléglise, and Y. Estève, "Ted-lium: an automatic speech recognition dedicated corpus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[7] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. of HLT*, 1994, pp. 307–312.

[8] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[9] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. of the ICASSP*, 2013, pp. 7893–7897.

[10] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.

[11] "poliMedia," https://polimedia.upv.es/.

[12] "Videolectures.NET: Exchange ideas and share knowledge," http://www.videolectures.net/.

[13] F. Jelinek and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *In Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May 1980, pp. 381–397.

[14] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvêa, S. Radomski, M. Mühlhäuser, and C. Biemann, "Open source german distant speech recognition: Corpus and acoustic model," in *Text, Speech, and Dialogue*. Springer, 2015, pp. 480–488.

[15] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.