The MLLP system for the 4th CHiME Challenge

Miguel Ángel del-Agua, Adrià Martínez-Villaronga, Adrià Gimènez, Alberto Sanchis, Jorge Civera, Alfons Juan

MLLP, DSIC, Universitat Politècnica de València (UPV), Spain.
{mdelagua,amartinez1,agimenez,josanna,jcivera,ajuan}@dsic.upv.es

Abstract

The MLLP CHiME-4 system is presented in this paper. It has been built using the transLectures-UPV toolkit (TLK) developed by the MLLP research group which makes use of state-of-the-art automatic speech recognition techniques. Our best system built for the CHiME-4 challenge consists on the combination of two different sub-systems in order to deal with the variety of acoustic conditions. Each sub-system in turn, follows a hybrid approach with different acoustic models, such as Deep Neural Networks or BLSTM Networks.

1. Introduction

The CHiME Speech Separation and Recognition Challenge [1] encourage participants to develop innovative ASR approaches capable of dealing with challenging noisy environments that rely in speech processing, signal separation or machine learning. It is based on the Wall Street Journal corpus sentences, spoken by talkers located in real noisy environments, such as in a street junction, on the bus, or in a pedestrian area. All the audios have been recorded using a common 6-channel tablet microphone array.

In previous years, the challenge consisted of obtaining the best possible transcription from the 6 channels simultaneously, but given the successful results achieved, this year the challenge proposes two more tracks: 1-channel and 2-channels tracks. Each track only differs in the number of available channels for testing. Thus, the 6-channels track is the easiest since more favorable audio enhancement techniques can be applied. In the case of the 1-channel and 2-channels tracks, the audio enhancement techniques cannot exploit channel information at all which makes this tasks harder to deal with.

The MLLP CHiME-4 system has been developed focusing on the acoustic modeling aspect. Specifically, two different acoustic models have been trained following the hybrid approach. On the one hand, a Context-Dependent Deep Neural Network Hidden Markov Model (CD-DNN-HMM) and on the other hand, a Bidirectional Long Short Term Memory Neural Network (BLSTM). Both acoustic models will be trained on the same data and their output combined. From the proposed three tracks, this global back-end system have been tested in the 1-channel and 2-channel tracks.

The rest of this work is divided as follows. Section 2 describes the ASR toolkit used for the experiments. In Section 3 the proposed system is described and the conclusions are given in section 5.

2. The TransLectures-UPV Toolkit

The MLLP CHiME-4 system has been developed using the transLectures-UPV Toolkit (TLK) [2]. TLK comprises a set of tools for audio processing, feature extraction, HMM and DNN training and decoding. The main latest features added to the toolkit are the following:

- Multilingual and Convolutional NNs.
- Different DNN speaker adaptation techniques: output-feature discriminant linear regression (oDLR) [3] or Kullback-Leibler Divergence based [4].
- DNN sequence discriminative training based on Maximum Mutual Information (MMI).
- Online decoding.
- Gammatone feature extraction.

TLK has demonstrated to provide competitive results in challenging and well-known tasks. In [5] the TLK-based system dealt with TED video lectures, and in [6] the TLK system provided good results in the LibriSpeech [7] corpus.

3. Proposed System

The system proposed by the MLLP group is based on the TLK toolkit. It is composed of two transcription sub-systems that are combined following a recognizer output voting error reduction (ROVER). Each of those sub-systems are based on the HMM-NN hybrid approach. The only difference is that for the first sub-system a classical DNN is used whereas for the second sub-system a BLSTM NN is employed.

Each of those sub-systems perform a three step recognition process as can be observed in Fig. 1. The first and second steps are shown in the upper box. Regarding the first step, it is shared between both sub-systems, cepstral mean and variance normalization (CMVN) is applied and the decoding is performed using a standard DNN which provides the best possible transcription and a better feature-space Maximum Likelihood Linear Regression (fMLLR) transform. For the second step, each sub-system makes use of their own acoustic model (DNN or BLSTM) taking as input the transformed fMLLR features. The output of this system is used to perform a final third-pass recognition (the lower box of Fig. 1). During this step, an unsupervised speaker adaptation technique is applied to both, the DNN and the BLSTM. Specifically, the technique used in this work consisted of a conservative training approach using a very small learning rate and early stopping [4]. This means that a very small learning rate is estimated for a fixed number of epochs as to minimize the Word Error Rate (WER) and then this learning rate is used in evaluation. To the best of our knowledge, it is the

first time that this kind of technique is applied to BLSTM NNs for acoustic modeling.

TLK allows to perform decoding efficiently with large vocabulary language models applying pruning techniques: beam search, histogram pruning, word end pruning and look-ahead. Thus, the provided 5-gram language model has been used to obtain the recognition outputs along all the steps. Once the last step is performed, the output lattices are re-scored using also the provided RNN-based language model.

BLSTM NNs have been built using TensorFlow [8]. With this purpose, a new feature has been added to TLK for decoding using TensorFlow-based graphs.

4. Experimental evaluation

The data used for training the acoustic models belong to the multi-condition training set defined by the CHiME-4 challenge. In our case, all data from channels 1,3,4,5 and 6 have been used to train the DNN and the BLSTM sub-systems.

Regarding feature extraction, classical Mel-frequency cepstral coefficients (MFCC) were extracted with a Hamming window of 25 ms. shifted at 10 ms. intervals. This MFCC features consisted of 16 MFCCs and their first and second derivatives (48-dimensional feature vectors). The resulting feature vectors were then normalized by mean and variance at speaker level. And after that, a single fMLLR transform for each training speaker was then estimated and applied to perform speaker-adaptive training (SAT).

In order to train the DNN and BLSTM based acoustic models, we first trained a basic context dependent triphone HMM model up to 64 component Gaussian mixtures, after which a second-pass fMLLR was applied. This model yielded a total of 9079 tied states, estimated following a phonetic decision tree approach.Both models were built on top of these HMM acoustic model. On one hand, the DNN-based acoustic model took as input the fMLLR features with a window size of 11, 5 hidden layers, sigmoid activation functions and an output layer of 9079. It was applied a discriminative pre-training stage and after that, the network was trained as to obtain the best frame accuracy on a validation set. On the other hand, the BLSTM acoustic model was trained with fMLLR input features (without windowing) with 4 hidden layers of 500 units each (both forward and backward directions) and an output layer of 9079. In this case, dropout was applied at the output of each cell with a probability of 0.1, and the Newbob strategy was also applied in order to reduce the learning rate by 0.8 each time the frame accuracy improved less than 3% relative on the validation set. Both networks were trained by minimizing the cross-entropy loss function, following the classical stochastic gradient descent algorithm. This two acoustic models were used for the 1-channel and 2-channels tracks. It is worth mentioning, that in the case of the 2-channel track, the audio enhancement beamformit was applied.

In Table 1 the results after each recognition step from the 1 channel track are shown, and similarly in Table 2 the results from the 2-channels track. As can be observed, the first recognition step is common to both sub-systems and tracks. With respect to the rest of recognition passes, very similar behaviors are observed in both tracks; the DNN performs better in all recognition steps and the BLSTM obtains a huge gain after the third step. For the first statement, we argue that the DNN is far more complex in terms of number of parameters, as we have trained a 5 hidden layer neural network of 2048 units per layer, while the BLSTM consist of 4 hidden layers of 500 units each

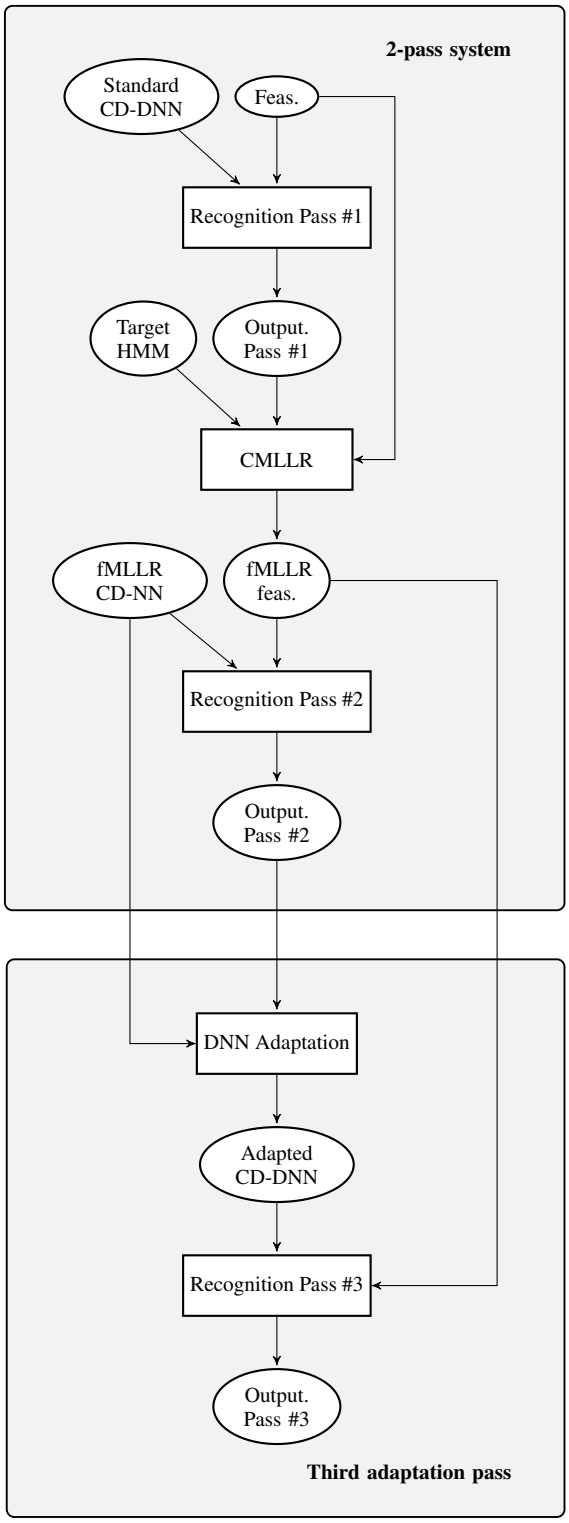

Figure 1: Multi-Pass recognition system with DNN adaptation. Top: 2-pass decoding using fMLLR features. Bottom: Third pass DNN adaptation.

Table 1: WER (%) per step for the 1-channel track.

System	Rec. Pass	Dev		Test	
		real	simu	real	simu
DNN	1	16.03	17.63	24.87	24.47
	2	12.66	14.52	19.80	19.92
	3	11.93	13.19	18.34	17.73
	+RNNLM	10.45	11.98	17.20	16.56
BLSTM	1	16.03	17.63	24.87	24.47
	2	15.10	17.18	23.09	23.56
	3	13.40	14.46	19.30	18.47
	+RNNLM	11.96	12.79	17.78	17.03

Table 2: WER (%) per step for the 2-channels track.

System	Rec. Pass	Dev		Test	
		real	simu	real	simu
DNN	1	13.83	14.35	21.14	20.80
	2	10.39	11.49	16.26	15.75
	3	9.60	10.46	14.77	13.71
	+RNNLM	8.45	9.29	13.71	12.57
BLSTM	1	13.83	14.35	21.14	20.80
	2	12.81	14.22	19.09	19.64
	3	11.63	12.67	15.50	14.93
	+RNNLM	10.12	11.36	14.31	13.46

one. Regarding the second statement, the huge WER improvement from the BLSTM at the third step comes from the fact that we are using the best transcription obtained during the previous step, i. e. the DNN, as to better perform speaker adaptation to the NN during the third step.

Once the output from both systems has been obtained, ROVER technique is applied as to combine both transcriptions. As can be seen in Table 3, the DNN system systematically outperforms the BLSTM-based. However, the combination of both systems yields the best result in both tracks. If we take a look to the real test set, the baseline provided by the organizers for the 1-channel track yielded 23.70% WER points whereas our system obtains 16.11%. This represents 32% relative reduction in WER for the 1-channel track. In the case of the 2-channels track, the baseline system achieved 16.58% average WER whereas our system achieves 12.82%. This represents a 22.7% relative reduction in WER for the 2-channel track. These improvements seems quite competitive, taking into account the simplicity of our system.

Table 4 summarizes the results obtained by the best system per environment. As shown, the most challenging has been the bus environment in all tracks for the real test set. In fact, the baseline system achieved 35.8%, while our system 21.61, which means almost 40% of relative improvement in the 1-channel track. In the case of the 2-channels track, the improvement is about 37% (from 25.37 to 16.00).

5. Conclusions

In this work we have described the MLLP ASR system developed for the CHiME-4 challenge built using TLK. The system is based on the combination of two sub-systems which make use of different acoustic models: DNNs and BLSTMs. The final system obtains 32% and 22.7% relative improvements over the 1-channel and 2-channels tracks compared to the baseline. This represents a good enough result taking into account the simplicity of our approach.

Table 3: Average WER (%) for the tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
1ch	DNN	10.45	11.98	17.20	16.56
	BLSTM	11.96	12.79	17.78	17.03
	Combined	9.95	11.13	16.11	15.72
2ch	DNN	8.45	9.29	13.71	12.57
	BLSTM	10.12	11.36	14.31	13.46
	Combined	7.96	8.93	12.82	12.06

Table 4: WER (%) per environment for the best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
1ch	BUS	11.74	9.04	21.61	10.95
	CAF	11.18	14.68	18.12	19.57
	PED	7.42	9.35	13.25	15.37
	STR	9.45	11.46	11.47	16.98
2ch	BUS	8.84	7.73	16.00	8.67
	CAF	8.70	11.55	13.78	14.34
	PED	6.27	7.45	11.17	11.77
	STR	8.02	9.00	10.31	13.47

6. Acknowledgments

The work leading to this invention has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no 287755 (transLectures). Also, it has received funding from the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme under grant agreement no 621030 (EMMA). In addition, this work has been supported by the Spanish research project MORE TIN2015-68326-R (MINECO/FEDER) and the Spanish Government with the FPU scholarship FPU13/06241.

7. References

[1] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language, to appear*, 2016.

[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 IberSpeech*, Las Palmas de Gran Canaria (Spain), 2014.

[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] 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.

[5] M. A. del Agua, A. Martínez-Villaronga, S. Piqueras, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "The mllp asr systems for iwslt 2015," in *Proc. of 12th IWSLT*, Da Nang (Vietnam), 2015.

[6] M. A. del Agua, S. Piqueras, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Asr confidence estimation with speaker-adapted recurrent neural networks," in *Proc. of InterSpeech*, San Francisco (USA), 2016, in press.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[8] M. Abadi and et al, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/