IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

GTM-UVIGO System for Albayzin 2018 Speech-to-Text Evaluation

Laura Docio-Fernandez, Carmen Garcia-Mateo

atlanTTic Research Center, Multimedia Technologies Group, University of Vigo, Spain

ldocio@gts.uvigo.es, carmen.garcia@uvigo.es

Abstract

This paper describes the Speech-to-Text system developed by the Multimedia Technologies Group (GTM) of the atlanTTic research center at the University of Vigo, for the Albayzin Speech-to-Text Challenge (S2T) organized in the Iberspeech 2018 conference. The large vocabulary automatic speech recognition system is built using the Kaldi toolkit. It uses an hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) for acoustic modeling, and a rescoring of a trigram based word-lattices, obtained in a first decoding stage, with a fourgram language model or a language model based on a recurrent neural network. The system was evaluated only on the open set training condition.

Index Terms: automatic speech recognition, deep neural networks, language model, speech activity detection

1. Introduction

Automatic Speech Recognition (ASR) is essential in many applications such as: dictation and transcription or captioning apps, speech-controlled interfaces, search engines for large multimedia archives, speech-to-speech translation, etc.

These days, with increases in computing power, the use of deep neural networks has spread to many fields including the ASR. With their use the performance of automatic speech recognition has been greatly improved. The current ASR systems are predominantly based on acoustic Hybrid Deep Neural Network–Hidden Markov Models (DNN-HMMs) [1] and the n-gram language model [2] [3]. However, in recent years, Neural Network Language Models (NNLM) have begun to be applied [4] [5] [6]. In these, words are embedded in a continuous space, in an attempt to map the semantic and grammatical information present in the training data, and in this way to achieve better generalization than n-gram models. The depth of the created network (the number of hidden layers), together with the ability to model a large number of context-dependent states, results in a reduction in Word Error Rate (WER). The type of neural networks most often used in language modelling are recurrent neural networks (RNNLMs). The recurrent connections present in these networks allow the modelling of long-range dependencies that improve the results obtained by n-gram models. In more recent work, recurrent network topologies such as LSTM (Long Short-Term Memory) [7] have also been applied [8][9][10].

In this paper, we present a summary of the GTM efforts in developing speech-to-text technology for Spanish language and its evaluation in the IberSPEECH-RTVE Speech to Text Transcription challenge. The submitted ASR system was evaluated on the *open set* training condition.

The paper is organized as it follows: in Section 2 the ASR system is described. Section 3 presents the data used to train the acoustic and language models. Section 4 describes the submitted systems specific characteristics, and finally Section 5 offers some final conclusions.

2. The GTM-UVIGO ASR system

This section describes the main blocks that comprises the ASR system.

2.1. Speech activity detection

The speech activity detection activity detection approach developed in the proposed system has four main stages. First, a voice activity detection (VAD) based on gaussian mixture models (GMMs) is applied to the audio signal in order to discard the non-speech intervals. Next, a two-step audio segmentation approach, based on the Bayesian information criterion (BIC) [11], is carried out. Once the audio segmentation output is obtained, those segments that are classified as music by a logistic regression classifier are discarded; this classifier relies on the i-vector paradigm for audio representation, as done in [12][13].

The above stages use as acoustic features 19 Mel-frequency cepstral coefficients (MFCCs) plus energy, and a cepstral mean subtraction using a sliding window of 300 ms is applied.

2.2. Acoustic modeling

The acoustic models use a hybrid DNN-HMM modeling strategy with a neural network based on Dan Povey's implementation in Kaldi [14]. This implementation uses a multi-spliced TDNN (Time Delay Neural Network) feed-forward architecture to model long-term temporal dependencies and short-term voice characteristics. The inputs to the network are 40 Mel-frequency cepstral coefficients extracted in the Feature Extraction block with a sampling frequency of 16 kHz. In each frame, we aggregate a 100-dimensional iVector to a 40-dimensional MFCC input.

The topology of this network consists of an input layer followed by 5 hidden layers with 1024 neurons with RELU activation function. Asymmetric input contexts were used, with more context in the background, which reduces the latency of the neuronal network in on-line decoding, and also because it seems to be more efficient from a WER perspective. Asymmetric contexts of 13 frames were used in the past, and 9 frames in the future. Figure 1 shows the topology used and Table 1 the layerwise context specification corresponding to this TDNN.

Table 1: *Context specification of TDNN in Figure 1*

Layer	Input context
1	[-2,+2]
2	[-1,2]
3	[-3,3]
4	[-7,2]
5	{0}

10.21437/IberSPEECH.2018-58

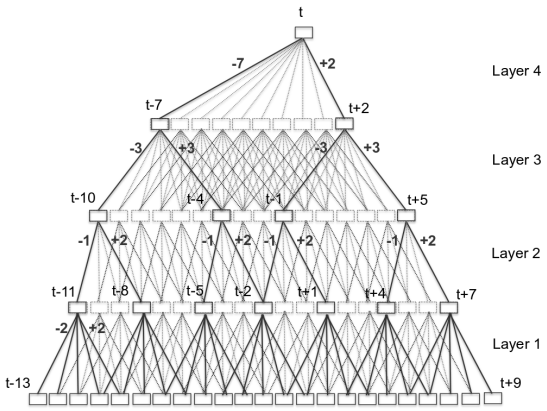

Figure 1: *TDNN used in acoustic models [14]*

2.3. Language modeling

In terms of the language models, when working with n-gram LMs the model was trained using the SRI Language Modeling Toolkit. N-gram models of order 3 and 4 were used, that is, trigrams and fourgrams. A modified Kneser-Ney discounting of Chen and Goodman has also been applied, together with a weight interpolation with lower orders [15].

For training the RNNLMs, the Kaldi RNNLM [16] software was also used. The neural network language model is based on a RNN with 5 hidden layers and 800 neurons, where TDNN layers with activation function RELU, and LSTM layers are combined. The training is performed using Stochastic Gradient Descent (SGD), and in several epochs (in our case, 20 epochs). All RNNLM models have been trained with the same material as in the case of n-gram statistical models. Figure 2 shows the topology of the network used.

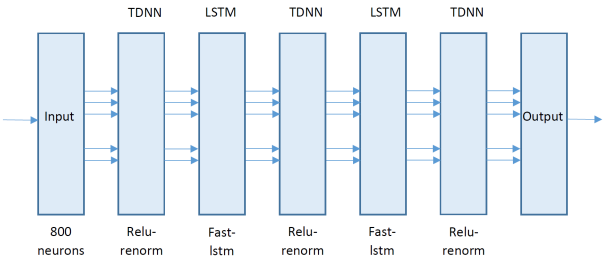

Figure 2: *RNNLM topology used in the ASR system.*

2.4. Recognition process

The recognition process is developed using the Kaldi toolkit [16]. The ASR system is based on two decoding stages to obtain the text transcription of the input speech signal. In the first decoding stage a lattice is obtained. This lattice is created using a 3-gram language model. In the second decoding stage, a language model rescoring is applied on this lattice. Figure 3 shows a block diagram of the recognition process.

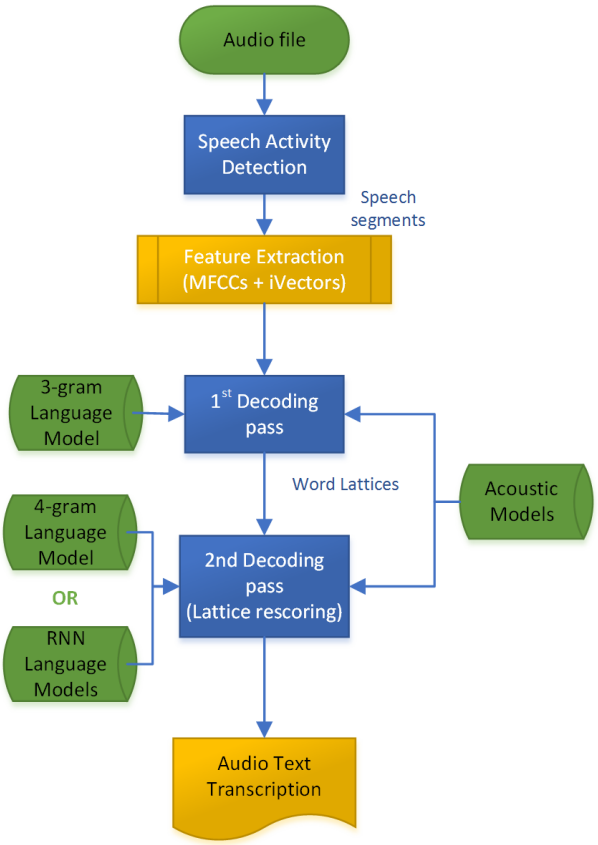

Figure 3: *GTM-UVIGO ASR System*

3. Data resources

As stated in Section 1, the submitted ASR system was evaluated on the *open set* training condition. It uses acoustic models trained with data not in the RTVE2018DB, and language models trained with both text data in the RTVE2018DB and extern text data.

Next, the data resources used for training the models are described.

3.1. Audio corpora

The data used for acoustic model training came from the following corpora:

- 2006 TC-STAR speech recognition evaluation campaign[17]: 79 hours of speaking in Spanish
- Galician broadcast news database Transcrigal [18]: 30 hours of speaking in Galician.

It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences and short speech utterances were discarded, so in the end the acoustic training material consisted of approximately 109 hours (79 hours in Spanish and 30 hours in Galician).

3.2. Text corpora

The following text corpora were used to train the LMs.

- A text corpus of approximately 90M words composed of material from several sources: transcriptions of Eu-

ropean and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses and the transcriptions of the Mavir sessions included in the development set of the Albayzin 2016 Spoken Term Detection Evaluation [1]. The vocabulary size of this corpus is approximately 250K words.

- The RTVE subtitles provided by the organizers of the evaluation. This text comprises approximately 60M words and its vocabulary size is approximately 173K words.

Table 2 shows the main characteristics of these text resources.

Table 2: *Characteristics of the text resources*

	Training text size	Vocabulary size
TC-STAR and Others	90M	250K
RTVE subtitles	60M	173K

From these text corpora, three LMs have been trained:

- **3-gram LM:** A trigram language model obtained by linear interpolation of two single trigram models. Each single model was trained using one of the above text resources.
- **4-gram LM:** A fourgram language model obtained by linear interpolation of two single fourgram models. Each single model was trained using one of the above text resources.
- **RNNLM:** A RNN language model trained with all the text described above.

The lexicon size of these models was approximately 312K words. The phonetic transcription of the lexicon words was automatically generated using the phonetic transcriber that forms part of the Cotovía GTM-UVIGO Text-to-Speech system [19].

4. Recognition results

This section presents the performance of the GTM-UVIGO ASR system on the development data provided by the organizers of the competition. Two systems that differ in the language model used in the rescoring stage were evaluated. The primary system uses the 4-gram LM and the contrastive system uses the RNNLM, both described in Sections 2 and 3. The results obtained in the development set with the primary and contrastive systems are shown in Table 3. The table shows the average Word Error Rate (WER) by TV show and also the global average WER.

5. Conclusions and future work

In this paper, we have presented two ASR systems to IberSPEECH-RTVE Speech to Text Transcription challenge. The difference between these two systems lies in the language models used in the decoding stage. The primary system uses a 4-gram language model for rescoring and the contrastive system uses a language model based on recurrent neural networks.

As a future line of development we plan to improve the acoustic and language models of the ASR system, as well as to enrich their output with punctuation marks and correct capitalization.

[1]MAVIR was a project funded by the Madrid region that coordinated several research groups and companies working on information retrieval (http://www.mavir.net)

Table 3: *Average WER on Development set.*

	Primary		Contrastive	
TV show	Dev1	Dev2	Dev1	Dev2
20H	12.13%	–	12.86%	–
AP	22.82%	–	24.30%	–
CA	53.95%	–	54.35%	–
LM	30.81%	–	32.48%	–
LN24H	27.22%	28.85%	28.48%	29.85%
millennium	–	25.52%	–	24.78%
Average	26.65%	25.30%	27.61%	26.47%
Average	26.37%		27.37%	

6. Acknowledgements

This work has received financial support from the Spanish Ministerio de Economía y Competitividad through project 'TraceThem' (TEC2015-65345-P), from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019) and the European Union (European Regional Development Fund ERDF).

7. References

[1] G. Hinton, L. Deng, D. Yu, and Y. Wang. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, vol. 9, no. 3, pp. 82-97.

[2] J. T. Goodman. 2001. A bit of progress in language modeling. Computer Speech and Language, vol. 15, no. 4, pages 403434.

[3] D. Jurafsky and J.H. Martin. 2008. Speech and Language Processing: An Introduction to Language Processing, Computational Linguistics, and Speech Recognition.

[4] T. Mikilov, S. Kombrink, A. Deoras, L. Bruget, and J. Cernicky. 2011. RNNLM-recurrent neural network language modeling toolkit, in Proc. of the 2011 ASRU Workshop, pages 196-201.

[5] E. Arisoy, T.N. Sainath, B. Kingsbury, and B. Ramabhadran. 2012. Deep Neural Network Language Models. In NAACL-HLT Workshop on the Future of Language Modeling for HLT, pages 20-28, Stroudsburg, PA, USA. Association for Computational LInguistics.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Juavi. 2003. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155.

[7] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. 2016. Exploring the limits of language modeling. In arXiv preprint arXiv:1602.02410.

[8] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur. 2018. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition, In ICASSP.

[9] M. Sundermeyer, Z. Tuske, R. Schluter, and H. Ney. 2014. Lattice decoding and rescoring with long-span neural network language models. En Fifteenth Annual Conference of the International Speech Communication Association.

[10] X. Chen, X. Liu, A. Ragni, Y. Wang and M. Gales. 2017. Future word contexts in neural network language models. ArXiv preprint arXiv:170805592.

[11] M. Cettolo and M. Vescovi. 2003. Efficient audio segmentation algorithms based on the BIC. In Proceedings of ICASSP. vol. VI, pp. 537540 (2003).

[12] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo. 2014. GTM-UVigo system for Albayzin 2014 audio segmentation evaluation. In Iberspeech 2014: VIII Jornadas en Tecnologa del Habla and IV Iberian SLTech Workshop (2014).

[13] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo. 2016. GTM-UVigo System for Albayzin 2016 Speaker Diarisation Evaluation. In Third International Conference Iberspeech (IberSpeech 2016).

[14] V. Peddinti, D. Povey and S. Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of INTERSPEECH 2015.

[15] A. Stolcke. 2002. SRILM An extensible language modeling toolkit. Proceedings of the International Conference on Statistical Language Processing, Denver, Colorado.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlcek, Y. Quian, P. Schwarz, J. Silovsk, G. Stemmer, and K. Vesel. 2011. The Kaldi Speech Recognition Toolkit. In ASRU.

[17] L. Docío, A. Cardenal and C. García. 2006. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. In Proc. Of TC-STAR Workshop on Speech-to-Speech Translation, ELRA, París, France.

[18] C. García, J. Tirado, L. Docío and A. Cardenal. 2004. Transcrigal: A bilingual system for automatic indexing of broadcast news. In IV International Conference on Language Resources and Evaluation.

[19] E. Rodríguez Banga, C. García Mateo, F.J. Méndez Pazó, M. González, C. Magariños Iglesias. 2012. Cotovía: an open source TTS for Galician and Spanish. In IberSPEECH 2012 – VII Jornadas en Tecnologa del Habla and III Iberian SLTech Workshop.