# Listen, Attend and Spell

**William Chan et al., cs.CL, 2015**

**translated by. triplet02**

## 0. 요약(Abstract)

우리는 음성 발성을 문자로 변환하도록 학습하는 신경망 구조인 Listen, Attned and Spell(LAS) 모델을 제안한다. 전통적인 HMM-DMM 모델과 달리, 이 모델은 음성 인식기(speech recognizer)의 모든 요소들(components)을 공동으로(jointly) 학습한다. 이 시스템은 두 가지 요소, 리스너(Listener)와 스펠러(Speller)로 구성된다. 리스너는 필터 뱅크 스펙트럼들(filter bank spectra)을 입력으로 하는, 피라미드형으로 쌓은 순환신경망(recurrent network)이다. 스펠러는 문자(characters)를 출력으로 하는 어텐션 기반(attention-based) 순환신경망이다. 이 신경망은 문자들 사이에 어떠한 독립적인 가정(independence assumptions)도 하지 않은 상태로 문자 시퀀스를 생성한다. 이것이 LAS 모델이 기존의 엔드투엔드 CTC 모델에서 핵심적으로 개선된 부분이다. 구글 음성 탐색 과제(Google voice search task)에서 LAS는 사전이나 언어 모델 없이 14.1%의 단어 오류율(Word Error Rate, WER)을 기록하였고, 상위 32개 빔(Beam)으로 재조정(rescoring)되는 언어 모델을 사용하여 10.3%의 WER을 기록하였다. 본 모델과 비교하면, 최고 성능(state-of-the-art) CLDNN-HMM 모델은 8.0%의 WER을 기록하였다.

## 1. 도입(Introduction)

## 2. 관련 연구(Related Work)

## 3. 모델(Model)

### 3.1. Listen

### 3.2. Attend and Spell

### 3.3. 학습(Learning)

### 3.4. 디코딩과 재조정(Decoding and Rescoring)

## 4. 실험(Experiments)

### 4.1. 어텐션 시각화(Attention Visualization)

### 4.2. 빔 너비의 효과(Effects of Beam Width)

### 4.3. 발화 길이의 효과(Effects of Utterance Length)

### 4.4. 단어 빈도(Word Frequency)

### 4.5. 흥미로운 디코딩 예시(Interesting Decoding Examples)

## 5. 결론(Conclusion)

### 감사의 말

### References

[1] Nathaniel Morgan and Herve Bourlard. Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models. In IEEE International Conference on Acoustics, Speechand Signal Processing, 1990.

[2] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey E. Hinton. Deep belief networks forphone recognition. In Neural Information Processing Systems: Workshop on Deep Learningfor Speech Recognition and Related Applications, 2009.

[3] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Large vocabulary continuous speechrecognition with context-dependent dbn-hmms. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

[4] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing,20(1):14–22, 2012.

[5] Navdeep Jaitly, Patrick Nguyen, Andrew W. Senior, and Vincent Vanhoucke. Applicationof Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In INTERSPEECH, 2012.

[6] Tara Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. DeepConvolutional Neural Networks for LVCSR. In IEEE International Conference on Acoustics,Speech and Signal Processing, 2013.

[7] Kanishka Rao, Fuchun Peng, Hasim Sak, and Francoise Beaufays. Grapheme-to-phonemeconversion using long short-term memory recurrent neural networks. In IEEE InternationalConference on Acoustics, Speech and Signal Processing, 2015.

[8] Kaisheng Yao and Geoffrey Zweig. Sequence-to-Sequence Neural Net Models for Graphemeto-Phoneme Conversion. 2015.

[9] Tomas Mikolov, Karafiat Martin, Burget Luka, Eernocky Jan, and Khudanpur Sanjeev. Recurrent neural network based language model. In INTERSPEECH, 2010.

[10] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmiduber. ConnectionistTemporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In International Conference on Machine Learning, 2006.

[11] Alex Graves. Sequence Transduction with Recurrent Neural Networks. In International Conference on Machine Learning: Representation Learning Workshop, 2012.

[12] Alex Graves and Navdeep Jaitly. Towards End-to-End Speech Recognition with RecurrentNeural Networks. In International Conference on Machine Learning, 2014.

[13] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, RyanPrenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Ng. Deep Speech:Scaling up end-to-end speech recognition. In http://arxiv.org/abs/1412.5567, 2014.

[14] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results. In Neural Information Processing Systems: Workshop Deep Learning and Representation Learning Workshop, 2014.

[15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio.Attention-Based Models for Speech Recognition. In http://arxiv.org/abs/1506.07503, 2015.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation byJointly Learning to Align and Translate. In International Conference on Learning Representations, 2015.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to Sequence Learning with NeuralNetworks. In Neural Information Processing Systems, 2014.

[18] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares,Holger Schwen, and Yoshua Bengio. Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation. In Conference on Empirical Methods in NaturalLanguage Processing, 2014.

[19] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In http://arxiv.org/abs/1506.03099, 2015.

[20] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak. Convolutional, Long ShortTerm Memory, Fully Connected Deep Neural Networks. In IEEE International Conference onAcoustics, Speech and Signal Processing, 2015.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with DeepConvolutional Neural Networks. In Neural Information Processing Systems, 2012.

[22] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of FiniteState Markov Chains. The Annals of Mathematical Statistics, 37:1554–1563, 1966.

[23] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In International Conference onMachine Learning, 2001.

[24] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In Association for Computational Linguistics, 2015.

[25] Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using VeryLarge Target Vocabulary for Neural Machine Translation. In Association for ComputationalLinguistics, 2015.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A NeuralImage Caption Generator. In IEEE Conference on Computer Vision and Pattern Recognition,2015.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov,Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generationwith Visual Attention. In International Conference on Machine Learning, 2015.

[28] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton.Grammar as a foreign language. In http://arxiv.org/abs/1412.7449, 2014.

[29] Oriol Vinyals and Quoc V. Le. A Neural Conversational Model. In International Conferenceon Machine Learning: Deep Learning Workshop, 2015.

[30] Sepp Hochreiter and Jurgen Schmidhuber. Long Short-Term Memory. Neural Computation,9(8):1735–1780, November 1997.

[31] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid Speech Recognition withBidirectional LSTM. In Automatic Speech Recognition and Understanding Workshop, 2013.

[32] Salah Hihi and Yoshua Bengio. Hierarchical Recurrent Neural Networks for Long-Term Dependencies. In Neural Information Processing Systems, 1996.

[33] Jan Koutnik, Klaus Greff, Faustino Gomez, and Jurgen Schmidhuber. A Clockwork RNN. InInternational Conference on Machine Learning, 2014.

[34] Navdeep Jaitly, Vincent Vanhoucke, and Geoffrey Hinton. Autoregressive product of multiframe predictions can improve the accuracy of hybrid models. In INTERSPEECH, 2014.

[35] Hasim Sak, Andrew Senior, Kanishka Rao, and Francoise Beaufays. Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. In INTERSPEECH, 2015.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Neural Information ProcessingSystems, 2013.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, NagendraGoel, Mirko Hannenmann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, GeorgStemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In Automatic SpeechRecognition and Understanding Workshop, 2011.

[38] Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z.Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. LargeScale Distributed Deep Networks. In Neural Information Processing Systems, 2012.