Acoustic-to-Word Models with Conversational Context Information

Suyoun Kim[1] **and Florian Metze**[2]
[1]Electrical & Computer Engineering
[2]Language Technologies Institute, School of Computer Science
Carnegie Mellon University
{suyoung1, fmetze}@andrew.cmu.edu

Abstract

Conversational context information, higher-level knowledge that spans across sentences, can help to recognize a long conversation. However, existing speech recognition models are typically built at a sentence level, and thus it may not capture important conversational context information. The recent progress in end-to-end speech recognition enables integrating context with other available information (e.g., acoustic, linguistic resources) and directly recognizing words from speech. In this work, we present a direct acoustic-to-word, end-to-end speech recognition model capable of utilizing the conversational context to better process long conversations. We evaluate our proposed approach on the Switchboard conversational speech corpus and show that our system outperforms a standard end-to-end speech recognition system.

1 Introduction

Many real-world speech recognition applications, including teleconferencing, and AI assistants, require recognizing and understand long conversations. In a long conversation, there exists the tendency of semantically related words or phrases reoccur across sentences, or there exists topical coherence. Thus, such conversational context information, higher-level knowledge that spans across sentences, provides important information that can improve speech recognition. However, the long conversations typically split into short sentence-level audios to make building speech recognition models computationally feasible in current state-of-the-art recognition systems (Xiong et al., 2017; Saon et al., 2017).

Over the years, there have been many studies have attempted to inject a longer context information into language models. Based on a recurrent neural network (RNNs) language models (Mikolov et al., 2010), (Mikolov and Zweig, 2012; Wang and Cho, 2015; Ji et al., 2015; Liu and Lane, 2017; Xiong et al., 2018), proposed using a context vector that would encode the longer context information as an additional network input. However, all of these models have been developed on text data, and therefore, it must still be integrated with a conventional acoustic model which is built separately without a longer context information, for speech recognition on long conversations.

Recently, new approaches to speech recognition models integrate all available information (e.g. acoustic, linguistic resources) in a so-called end-to-end manner proposed in (Graves et al., 2006; Graves and Jaitly, 2014; Hannun et al., 2014; Miao et al., 2015; Bahdanau et al., 2014; Chorowski et al., 2014, 2015; Chan et al., 2015; Kim et al., 2017). In these approaches, a single neural network is trained to recognize graphemes or even words from speech directly. Especially, the model using semantically meaningful units, such as words or sub-word (Sennrich et al., 2015), rather than graphemes have been showing promising results (Audhkhasi et al., 2017b; Li et al., 2018; Soltau et al., 2016; Zenkel et al., 2017; Palaskar and Metze, 2018; Sanabria and Metze, 2018; Rao et al., 2017; Zeyer et al., 2018).

In this work, motivated by such property of the end-to-end speech recognition approaches, we propose to integrate conversational context information within a direct acoustic-to-word, end-to-end speech recognition to better process long conversations. Thus far, the research in speech recognition systems has focused on recognizing sentences and to the best of our knowledge, there have been no studies of word-based models incorporating conversational context information. There has been recent work attempted to use the conversational context information from the preceding graphemes (Kim and Metze, 2018), however, it is

Proceedings of NAACL-HLT 2019, pages 2766–2771
Minneapolis, Minnesota, June 2 - June 7, 2019.

limited to encode semantically meaningful context representation. Another recent work attempted to use a context information (Pundak et al., 2018), however, their method requires a list of phrases at inference (i.e. personalized contact list). We evaluate our proposed approach on the Switchboard conversational speech corpus (Godfrey and Holliman, 1993; Godfrey et al., 1992), and show that our model outperforms the sentence-level end-to-end speech recognition model.

2 Models

2.1 Acoustic-to-Words Models

We perform end-to-end speech recognition using a joint CTC/Attention-based approach (Kim et al., 2017; Watanabe et al., 2017). The neural network is trained by both CTC (Graves et al., 2006) and Attention-based sequence-to-sequence (seq2seq) objectives (Bahdanau et al., 2014) to combine the strength of the two. With CTC, it preserves left-right order between input and output and with attention-based seq2seq, it learns the language model jointly without relying on the conditional independence assumption.

As an output, we use word-level symbols which generated from the bite-pair encoding (BPE) algorithm (Sennrich et al., 2015). This method creates the target units based on the frequency of occurrence in training sets. Similar to (Zeyer et al., 2018; Palaskar and Metze, 2018; Sanabria and Metze, 2018), we use BPE-10k which contains roughly 10k units (9,838), including 7,119 words and 2719 sub-words.

2.2 Conversational Context Representation

In order to use conversational context information within the end-to-end speech recognition framework, we extend the decoder sub-network to predict the output additionally conditioning on conversational context. To do so, we encode the preceding sentence into a single vector, a conversational context vector, then inject to decoder network as an additional input at every output step.

Let we have K sentences in a conversation. For k-th sentence, s^k, we have T^k-length input acoustic feature ($\boldsymbol{x^k}$) and U^k-length output words. Our proposed decoder generates the probability distribution over words (y_u^k), conditioned on 1) high-level representation ($\boldsymbol{h^k}$) of input ($\boldsymbol{x^k}$) generated from encoder, and 2) all the words seen previously ($y_{1:u-1}^k$), and 3) previous decoder state (d_{u-1}^k) 4)

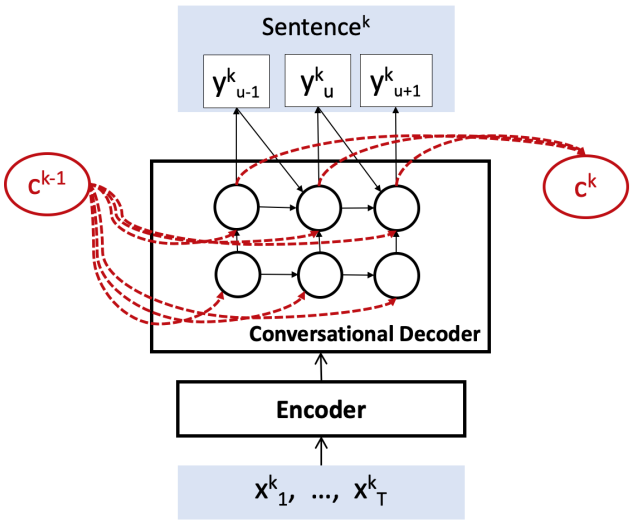

Figure 1: The architecture of our end-to-end speech recognition model with conversational context information. The c^{k-1} is the conversational context vector generated from the preceding $k-1$ sentence red curved dashed line represents the context information flow within the same conversation.

additionally conditioning on conversational context vector (c^{k-1}), which represents the information of the preceding sentence ($k-1$):

$$\boldsymbol{h^k} = Encoder(\boldsymbol{x^k}) \quad (1)$$

$$y_u^k \sim Decoder(\boldsymbol{h^k}, y_{1:u-1}^k, d_{u-1}^k, c^{k-1}) \quad (2)$$

We represent the context vector, c^{k-1}, from the preceding sentence in two different ways: (a) mean of word embedding, and (b) attentional word embedding. We first generate one-hot word vectors, and then we simply take the mean over word vectors to obtain a single vector in method (a), or we use attention mechanism over word vectors to obtain the weight over the words and then perform the weighted-sum. The parameter of the attention mechanism is optimized towards minimizing the conversation ID classification error similar to (Kim and Metze, 2018). The context vector is merged with a decoder state at every output step as follows:

$$\hat{d}_{u-1}^k = \tanh(W d_{u-1}^k + V c^{k-1} + b) \quad (3)$$

$$y_u^k \sim \text{softmax}(\text{LSTM}(\hat{d}_{u-1}^k, \boldsymbol{h_u^k}, y_{1:u-1}^k))) \quad (4)$$

where W, V, b are trainable parameters.

In order to learn and use the conversational-context during training and decoding, we serialize the sentences based on their onset times and their

conversations rather than the random shuffling of data. We shuffle data at the conversation level and create mini-batches that contain only one sentence of each conversation.

3 Experiments

3.1 Datasets

We investigated the performance of the proposed model on the Switchboard LDC corpus (97S62) which has a 300 hours training set. We split the Switchboard data into two groups, then used 285 hours of data (192 thousand sentences) for model training and 5 hours of data (4 thousand sentences) for hyper-parameter tuning. The evaluation was carried out on the HUB5 Eval 2000 LDC corpora (LDC2002S09, LDC2002T43), which have 3.8 hours of data (4.4 thousand sentences), and we show separate results for the Callhome English (CH) and Switchboard (SWB) evaluation sets. We denote train_nodup, train_dev, SWB, and CH as our training, development, and two evaluation datasets for CH and SWB, respectively. There are 2,402 conversations in training sets and 20 conversations in CH, and 20 conversations in SWB.

We sampled all audio data at 16kHz, and extracted 80-dimensional log-mel filterbank coefficients with 3-dimensional pitch features, from 25 ms frames with a 10ms frame shift. We used 83-dimensional feature vectors to input to the network in total. We used 9,840 distinct labels: 9,838 word-level BPE units, start-of-speech/end-of-speech, and blank tokens. Note that no pronunciation lexicon was used in any of the experiments.

3.2 Training and decoding

We used joint CTC/Attention end-to-end speech recognition architecture (Kim et al., 2017; Watanabe et al., 2017) with ESPnet toolkit (Watanabe et al., 2018). We used a CNN-BLSTM encoder as suggested in (Zhang et al., 2017; Hori et al., 2017). We followed the same six-layer CNN architecture as the prior study, except we used one input channel instead of three since we did not use delta or delta delta features. Input speech features were downsampled to (1/4 x 1/4) along with the time-frequency axis. Then, the 6-layer BLSTM with 320 cells was followed by CNN. We used a location-based attention mechanism (Chorowski et al., 2015), where 10 centered convolution filters of width 100 were used to extract the convolutional features.

The decoder network of both our proposed models and the baseline models was a 2-layer LSTM with 300 cells. Our proposed models additionally require linear projection layer in order to encode the conversational context vector and merge with decoder states.

We also built an external RNN-based language model (RNNLM) on the same BPE-10k sets on the same Switchboard transcriptions. The RNNLM network architecture was a two-layer LSTM with 650 cells. This network was used only for decoding.

The AdaDelta algorithm (Zeiler, 2012) with gradient clipping (Pascanu et al., 2013) was used for optimization. We used $\lambda = 0.5$ for joint CTC/Attention training. We bootstrap the training our proposed conversational end-to-end models from the baseline end-to-end models. When we decode with RNNLM, we used joint decoder which combines the output label scores from the AttentionDecoder, CTC, and RNNLM by using shallow fusion (Hori et al., 2017):

$$\begin{aligned} \boldsymbol{y}* = \text{argmax}\{ & \log p_{att}(\boldsymbol{y}|\boldsymbol{x}) \\ & + \alpha \log p_{att}(\boldsymbol{y}|\boldsymbol{x}) \\ & + \beta \log p_{rnnlm}(\boldsymbol{y})\} \end{aligned} \tag{5}$$

The scaling factor of CTC, and RNNLM scores were $\alpha = 0.3$, and $\beta = 0.3$, respectively. We used a beam search algorithm similar to (Sutskever et al., 2014) with the beam size 10 to reduce the computation cost. We adjusted the score by adding a length penalty, since the model has a small bias for shorter utterances. The final score $s(\boldsymbol{y}|\boldsymbol{x})$ is normalized with a length penalty 0.5.

The models were implemented by using the PyTorch deep learning library (Paszke et al., 2017), and ESPnet toolkit (Kim et al., 2017; Watanabe et al., 2017, 2018).

4 Results

We evaluated both the end-to-end speech recognition model which was built on sentence-level data and our proposed end-to-end speech recognition model which leveraged conversational context information.

Table 1 shows the WER of our baseline, proposed models, and several other published results those were only trained on 300 hours Switchboard training data. As shown in Table 1, we obtained a performance gain over our baseline by using the

Table 1: Comparison of word error rates (WER) on Switchboard 300h with standard end-to-end speech recognition models and our proposed end-to-end speech recognition models with conversational context.

Model	Output Units	LM	SWB (WER %)	CH (WER %)
Prior Models				
LF-MMI (Povey et al., 2016)	context-dependend phones	O	9.6	19.3
CTC (Zweig et al., 2017)	Char	O	19.8	32.1
CTC (Audhkhasi et al., 2017a)	Word (Phone init.)	O	14.6	23.6
CTC (Sanabria and Metze, 2018)	Char, BPE-{300, 1k, 10k}	O	12.5	23.7
Seq2Seq (Palaskar and Metze, 2018)	BPE-10k	O	21.3	35.7
Seq2Seq (Zeyer et al., 2018)	BPE-1k	O	11.8	25.7
Our Sentence-level Baseline				
Our baseline	BPE-10k	x	17.6	30.6
Our baseline	BPE-10k	O (only swb)	17.0	29.7
Our Proposed Conversational Model				
w/ Context *(a) mean*	BPE-10k	O (only swb)	16.3	29.0
w/ Context *(b) att*	BPE-10k	O (only swb)	16.4	29.2
w/ Context *(b) att* + pre-training	BPE-10k	O (only swb)	16.0	28.9

Table 2: Perplexities on a held-out set of our proposed conversational context LM and baselines.

Models	Fisher text	PPL
Baseline LM	x	74.15
Baseline LM	o	72.81
Proposed Conversational LM	x	67.03
Proposed Conversational LM	o	64.30

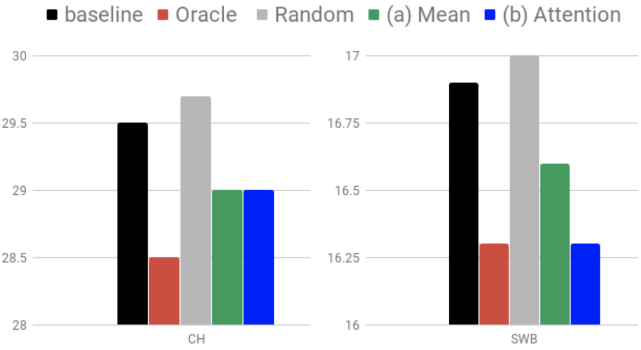

Figure 2: The architecture of our end-to-end speech recognition model with conversational context information. The c^{k-1} is the conversational context vector generated from the preceding $k-1$ sentence red curved line represents the context information flow within the same conversation.

conversational context information. Our proposed model *(a) mean* shows 4.1% and 2.4% relative improvement over our baseline on SWB and CH evaluation set, respectively. Our proposed model *(b) att* shows 3.5% and 1.7% relative improvement over our baseline on SWB and CH evaluation set, respectively. We also found that we can obtain further accuracy improvement by pre-training the decoder part only with transcription. With this pre-training technique, the *(b) att* shows 5.9% and 2.7% relative improvement. Unlike the previous work (Renduchintala et al., 2018), we did not use any additional encoder for the text data.

We also build the language model with or without the conversational context information. Table 2 shows the perplexity on a held-out set of our baseline LM and our conversational LM. We observed that incorporating the conversational context improves performance showing that 9.6% and 11.7% relative improvement on *SWBD only* and *SWBD + Fisher*. Note that the Fisher (LDC2004T19) parts (Cieri et al., 2004) of transcriptions is only used in these experiments.

We performed analyses in order to verify the conversational vector helps to improve recognition accuracy. We generate the context vector from an oracle preceding sentence and a random sentence, in addition to our predicted sentence. As described in Figure 2, the model using the oracle context performed best and the model using the random context was even worse than the baseline. Our model outperformed over the baseline and the model using the random context, we can conclude that the benefit from our proposed method is coming from the conversational context information.

5 Conclusion

We proposed an acoustic-to-word model capable of utilizing the conversational context to better process long conversations. A key aspect of our model is that the whole system can be trained with conversational context information in an end-to-

end framework. Our model was shown to outperform previous end-to-end speech recognition models trained on isolated utterances by incorporating preceding conversational context representations.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work also used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This research was supported by a fellowship from the Center for Machine Learning and Health (CMLH) at Carnegie Mellon University.

References

Kartik Audhkhasi, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Michael Picheny. 2017a. Building competitive direct acoustics-to-word models for english conversational speech recognition. *arXiv preprint arXiv:1712.03133*.

Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo. 2017b. Direct acoustics-to-word models for english conversational speech recognition. *arXiv preprint arXiv:1703.07754*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent NN: First results. *arXiv preprint arXiv:1412.1602*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pages 577–585.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium, Philadelphia*, LDC97S62.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *arXiv preprint arXiv:1706.02737*.

Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document context language models. *arXiv preprint arXiv:1511.03962*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4835–4839. IEEE.

Suyoun Kim and Florian Metze. 2018. Dialog-context aware end-to-end speech recognition. *arXiv preprint arXiv:1808.02171*.

Jinyu Li, Guoli Ye, Amit Das, Rui Zhao, and Yifan Gong. 2018. Advancing acoustic-to-word ctc model. *arXiv preprint arXiv:1803.05566*.

Bing Liu and Ian Lane. 2017. Dialog context language modeling with recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5715–5719. IEEE.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. *SLT*, 12:234–239.

Shruti Palaskar and Florian Metze. 2018. Acoustic-to-word recognition with sequence-to-sequence models. *arXiv preprint arXiv:1807.09597*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep context: end-to-end contextual speech recognition. *arXiv preprint arXiv:1808.02480*.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 193–199. IEEE.

Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. 2018. Multi-modal data augmentation for end-to-end asr. *arXiv preprint arXiv:1803.10299*.

Ramon Sanabria and Florian Metze. 2018. Hierarchical multi task learning with ctc. *arXiv preprint arXiv:1807.07104*.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. 2017. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Hagen Soltau, Hank Liao, and Hasim Sak. 2016. Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition. *arXiv preprint arXiv:1610.09975*.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tian Wang and Kyunghyun Cho. 2015. Larger-context language modelling. *arXiv preprint arXiv:1511.03729*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. The microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5255–5259. IEEE.

Wayne Xiong, Lingfeng Wu, Jun Zhang, and Andreas Stolcke. 2018. Session-level language modeling for conversational speech. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2764–2768.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Thomas Zenkel, Ramon Sanabria, Florian Metze, and Alex Waibel. 2017. Subword and crossword units for ctc acoustic models. *arXiv preprint arXiv:1712.06855*.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018. Improved training of end-to-end attention models for speech recognition. *arXiv preprint arXiv:1805.03294*.

Yu Zhang, William Chan, and Navdeep Jaitly. 2017. Very deep convolutional networks for end-to-end speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4845–4849. IEEE.

Geoffrey Zweig, Chengzhu Yu, Jasha Droppo, and Andreas Stolcke. 2017. Advances in all-neural speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4805–4809. IEEE.