INTERSPEECH 2020: Show & Tell Contribution
October 25–29, 2020, Shanghai, China

End-to-End Deep Learning Speech Recognition Model for Silent Speech Challenge

Naoki Kimura, Zixiong Su, Takaaki Saeki

The University of Tokyo, Japan
{kimura-naoki, zxsu}@g.ecc.u-tokyo.ac.jp, takaaki_saeki@ipc.i.u-tokyo.ac.jp

Abstract

This work is the first attempt to apply an end-to-end, deep neural network-based automatic speech recognition (ASR) pipeline to the Silent Speech Challenge dataset (SSC), which contains synchronized ultrasound images and lip images captured when a single speaker read the TIMIT corpus without uttering audible sounds. In silent speech research using SSC dataset, established methods in ASR have been utilized with some modifications to use it in visual speech recognition. In this work, we tested the SOTA method of ASR on the SSC dataset using the End-to-End Speech Processing Toolkit, ESPnet. The experimental results show that this end-to-end method achieved a character error rate (CER) of 10.1% and a WER of 20.5% by incorporating SpecAugment, demonstrating the possibility to further improve the performance with additional data collection.

Index Terms: silent speech interface, visual speech recognition, deep learning

1. Introduction

A silent speech interface is a technology that enables us to speak or use voice interfaces without uttering audible sounds. Since silent speech does not require vocalization, it can be used in situations or environments where we cannot speak aloud. It also provides a new means of communication for people with vocal impairments. The Silent Speech Challenge (SSC) [1] is a dataset containing synchronized ultrasound images and lip images captured when a single speaker read the TIMIT corpus without uttering any sound. Our task is to decode phonemes and characters from these images and finally estimate the sentences. As test data, it provides 100 sentences of utterances, a subset of WSJ0 [2] independent of TIMIT.

For the SSC dataset, established methods of automatic speech recognition (ASR), like GMM-HMM [3] and DNN-HMM [4], have been introduced with some modifications for visual speech recognition. The benchmark score is a word error rate (WER) of 17.4% with GMM-HMM [3] and 6.45% with DNN-HMM [4], using the 5000 words closed-vocabulary language model. While an end-to-end model using connectionist temporal classification (CTC)[5] and an attention-based encoder-decoder network[6] is becoming a state-of-the-art (SOTA) in ASR, the end-to-end models have not been applied to the SSC dataset, because the SSC dataset's training data contains only 2342 utterances by a single speaker. However, if we can get some decent results with the current amount of data, it will give us strong motivation to collect more data and incorporate data augmentation methods for the end-to-end solutions.

This work is the first attempt to apply an end-to-end ASR pipeline to the SSC dataset. We utilize the End-to-End Speech Processing Toolkit, ESPnet [7], to test the SOTA method of ASR on the SSC dataset with high reproducibility. The experimental results show that the end-to-end method achieves a character error rate (CER) of 10.1% and a WER of 20.5% with SpecAugment [8], demonstrating the possibilities of its further improvement with SSC data extension.

2. Silent Speech Challenge Dataset

The SSC dataset contains the 320 × 240 pixels tongue images and 640 × 480 pixels lip images synchronized at 60 fps. The training set consists of the images from a single native English speaker speaking 2342 utterances of the TIMIT corpus without vocalization. The test set includes 100 utterances from the WSJ0 5000-word corpus [2] read by the same speaker as the training set. It also provides features extracted with Discrete Cosine Transform (DCT) and that with deep autoencoder (DAE) used in the previous studies [3, 4].

3. ESPnet: end-to-end speech processing toolkit

To build our silent speech recognition system, we introduce the ASR parts of ESPnet [7], the open-source end-to-end speech processing toolkit. As shown in Figure 1, we utilize the DCT and AE features described in Section. 2 as inputs to the network instead of mel-frequency cepstral coefficients (MFCC). The features are from high sampling rate (60 fps) lip and tongue motion and thus considered capable of substituting for MFCC. Our end-to-end method is based on the VoxForge recipe of ESPnet [7], which uses a hybrid CTC / attention based end-to-end ASR model.

4. Experimental Results

4.1. Data Augmentation

End-to-end approaches usually require a large-scale dataset. WSJ0, a common linguistic dataset used for end-to-end ASR model, includes 12720 utterances. On the other hand, the training set of the SSC consists of only 2342 utterances and this problem usually leads to a high error rate. Therefore, we applied SpecAugment [8] to expand our data and thus obtained a four times larger dataset. This augmentation method implements prepossessing of time warp, frequency mask and time mask directly to the features.

4.2. Experimental condition

As the input feature of the network, we used 20 and 30 dimensions DCT features, and 20, 30 and 60 dimensions DAE features. As described in Section 3, we utilized the hybrid CTC / attention based end-to-end ASR model. The encoder had 12 layers with 4–header self-attention jointly attending to information from different positions. The 1st–layer CTC and attention decoder shared the encoder and performed joint decoding by combining both scores in a one-pass beam search

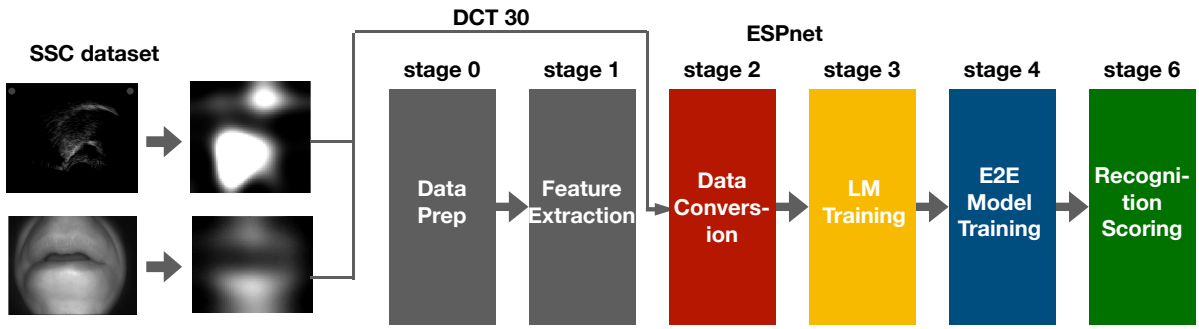

Figure 1: *The workflow of our method for silent speech recognition.*

Table 1: *Recognition results (CER) with and without SpecAugment*

SpecAug	Sub (%)	Del (%)	Ins (%)	CER (%)
Yes	4.2	5.9	1.9	**10.1**
No	8.1	10.5	4.0	18.7

Table 2: *Recognition results (CER and WER) of different features with SpecAugment*

Feature	Dimension	CER (%)	WER (%)
DCT	20	17.1	33.8
	30	**10.1**	**20.5**
AE	20	19.9	36.3
	30	17.0	33.2
	60	21.4	42.0

algorithm, where the joint training parameter α was set to 0.3. We trained the network for 200 epochs and set the learning rate of Adam optimizer to 20 with a Noam scheduler (25000 warm-up steps). We applied dropout layer (dropout rate was 0.1) and uniform label smoothing (penalty was 0.1) to avoid overfitting. In the test step, we used a model averaging approach that averaged the model parameters of the last 10 epochs for stable performance. For decoding to word-level, we used WSJ-based train _rnnlm_pytorch_lm_word 65000 as the language model, which is available on ESPnet [7] by default. More detailed experimental conditions are available at `https://github.com/espnet/espnet/blob/master/egs/timit_ssc/ssr1/conf/tuning/train_pytorch_transformer.yaml`. The evaluation in this section is based on commit a4b96c9.

4.3. Results and Discussion

Table 1 shows the comparison of recognition results with and without SpecAugment [8]. The CER with SpecAugment [8] was almost half of that without the augmentation. Even though the SSC dataset does not have enough amount of data to fully make use of an end-to-end model, our experimental results show that the model achieves promising performance only with data augmentation. We can expect that collecting more raw data improves the recognition accuracy further.

Table 2 shows the results on different features prepared by the SSC dataset. The best result, CER of 10.1%, was obtained by 30-dimension DCT features. Considering the lack of the amount of data, this result is quite promising. The Word Error Rate (WER) was 20.5%, and this is higher than the result of 6.45% from previous research [4]. However, this is actually a decent result considering that their 5000 words closed vocabulary language model was not available for us and we used a larger, less task specific 65,000 words language model.

5. Conclusions

In this paper, we reported the first case of applying the end-to-end deep learninng ASR model to the SSC dataset. Even though the SSC's amount of data is too small for the end-to-end approaches, our experiments using ESPnet yielded promising results of CER 10.1%, 20.5% only with the data augmentation. Our results give a strong reason to extend the SSC-formatted dataset and show the prospects of the end-to-end model approach.

6. Acknowledgements

This work was supported by JST, ACT-X Grant Number JPMJAX190B, Japan.

7. References

[1] B. Denby, T. Hueber, J. Cai, P. Roussel, L. Crevier-Buchman, S. Manitsaris, G. Chollet, M.Stone, and C.Pillot, "The silent speech challenge archive," https://ftp.espci.fr/pub/sigma/, 2013.

[2] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete - Linguistic Data Consortium," https://catalog.ldc.upenn.edu/LDC93S6A, May 2007.

[3] J. Cai, B. Demby, P.Roussel, G. Dreyfus, and L.Crevier-Buchman, "Recognition and real time performance of a lightweight ultrasound based silent speech interface employing a language model," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 1005–1008.

[4] Y. Ji, L. Liu, H. Wang, Z. Liu, Z. Niu, and B. Denby, "Updating the silent speech challenge benchmark with deep learning," *Speech Communication*, vol. 98, pp. 42–50, 2018.

[5] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, Beijing, China, June 2014, pp. 1764–1772.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, Shanghai, China, March 2016, pp. 4960–4964.

[7] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," *arXiv*, vol. abs/1804.00015, 2018.

[8] D. S. Park, W. Chen, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 2613–2617.