ATTENTION-BASED SPEECH RECOGNITION USING GAZE INFORMATION

Osamu Segawa

Chubu Electric Power Co., Inc.
Segawa.Osamu@chuden.co.jp

Tomoki Hayashi, Kazuya Takeda

Graduate School of Informatics
Nagoya University
hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp

ABSTRACT

We assume that there is a correlation between an utterance and a corresponding gaze object, and propose a new paradigm of multi-modal end-to-end speech recognition using multi-modal information, namely, utterances and corresponding gaze points. In our method, the system extracts acoustic features and corresponding images around gaze points, and inputs the information into the proposed attention-based multiple encoder-decoder networks. This makes it possible to integrate the two different modalities, and the performance of speech recognition is improved. To evaluate the proposed method, we prepared a simulation task of power-line control operations, and built a corpus that contains utterances and corresponding gaze points in the operations. We conducted an experimental evaluation using this corpus, and the results showed the reduction in the CER, suggesting the effectiveness of the proposed method in which acoustic features and gaze information are integrated.

Index Terms— end-to-end speech recognition, attention, multi-modal, gaze-point

1. INTRODUCTION

In this paper, we propose a new paradigm of multi-modal end-to-end speech recognition using multi-modal information, namely, utterances and corresponding gaze points. In some task-oriented operations, gaze points and linguistic information are closely related to each other. Therefore, a system should be able to estimate the mutual synchronization relationship and use the information for complementation and prediction. In our method, the system extracts acoustic features and corresponding images around gaze points, and uses the information as attentions [1] in encoder-decoder networks. To realize this concept, we propose an attention-based multiple encoder-decoder architecture that estimates character sequences directly using both acoustic features and image sequences. Applying the attention mechanism to both a sequence of acoustic features and a sequence of image features, the system can compensate for differences in the time resolution. Moreover, it is expected to automatically learn the difference between the occurrences of both types of information. As a result, owing to the use of the feature in which two different modalities are integrated, the improvement of recognition performance is expected. To evaluate the proposed method, we prepared the simulation task: power-line control operations, and built a corpus that contains utterances and corresponding gaze points in the operations. Utilizing this corpus, we evaluated the effectiveness of using gaze information to improve recognition performance.

2. RELATED WORKS

Owing to advances in sensing technologies, various kinds of signal, such as image, speech, biometric-signals and gaze points, can now be acquired easily at the same time. With this background, a large number of studies using multi-modal information have been carried out to improve recognition performance.

2.1. Using image features

Image data is a typical example of a multi-modal signal. Several researchers, dealing with multi-modal speech recognition, proposed the use of both lip images and speech signals [2, 3, 4] to improve speech recognition performance in noisy environments. Mroueh et al. [2] proposed a method in which two different neural networks use both acoustic features and lip images as input signals. Their method integrates the posterior probability of each network to improve recognition performance. Noda et al. [3] proposed a method in which the system estimates a clean acoustic feature from a noisy acoustic feature using denoising autoencoder (DAE). Their method calculates the phoneme posterior probability using a CNN, which is trained as a phoneme classification model using lip images. They built a speech recognition system based on GMM-HMM using a clean acoustic feature and the posterior probability to improve the recognition performance in a noisy environment. Petridis et al. [4] proposed an end-to-end model in which speech signals and lip images are processed in a single network to estimate the word class directly.

These studies demonstrated improved recognition performance using image information. However, the utilization of images other than lip images has not been considered.

2.2. Using gaze points

Another approach using multi-modal information is to use "gaze points" acquired by an eye-tracking device. For example, Nguyen et al. [5] proposed an assistant system for describing lecture notes that generates annotations automatically using the user's gaze point and extracts important sentences. Moreover, Vasudevan et al. [6] proposed an end-to-end deep learning method in which the system estimates the user's attention area (bounding box) corresponding to the utterances using several pieces of information: texts of utterances, gaze points, depth images, and body movement. The aim of their study is to improve the performance of object detection in video images using both gaze points and linguistic information. Furthermore, texts of utterances are the given information.

In contrast, there is another approach using both gaze points and speech as multi-modal information to improve recognition performance. For example, Rasmussen et al. [7] proposed a method for tracking reading process using a combination of gaze points and speech recognition. Cooke [8] proposed a probabilistic model that combine eye movement and speech recognition in a small simulation task. However, these approaches are based on HMM-based framework.

3. SPEECH RECOGNITION BASED ON ATTENTION

Recently, end-to-end speech recognition algorithms have been actively developed. In the early stage, the end-to-end approach was a method based on connectionist temporal classification (CTC) [9, 10]. However, once the effectiveness of the attention-based method [1, 11] was proven, a number of further studies were carried out.

In this approach, an encoder-decoder network architecture [12] is used to perform a direct mapping from a sequence of input features into a sequence of texts. The encoder network converts the sequence of input features to that of discriminative hidden states, and the decoder network uses the attention mechanism to obtain the alignment between each element of the output sequence and the encoder hidden states. Then it estimates the output symbol using weighted averaged hidden states, which is based on the alignment, as the inputs of the decoder network. In contrast to the CTC-based approach, the attention-based method does not require any conditional independence assumptions including the Markov assumption, language models, and complex decoding. However, a non-causal alignment problem is caused by a too flexible alignment of the attention mechanism [13]. To address this issue, the study [13] combines the objective function of the attention-based model with that of CTC to constrain flexible alignments of the attention mechanism.

In another study [14], multi-head attention (MHA) is used to obtain more suitable alignments. In MHA, multiple attentions are calculated, then integrated into a single attention.

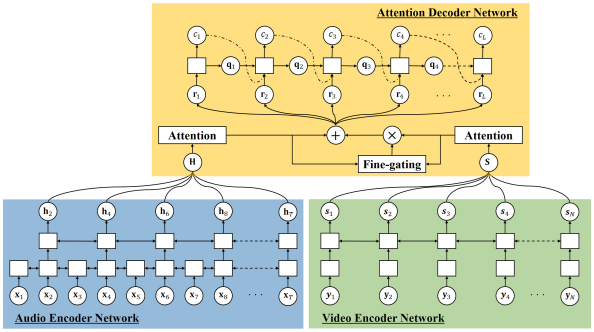

Fig. 1. *Overview of attention-based multiple encoder-decoder architecture.*

The use of MHA enables the model to jointly focus on information from different representation subspaces at different positions [15], leading to the improvement of the recognition performance. Although these studies demonstrated the improvement of the recognition performance using a new network architecture, they did not consider the use of multi-modal signals.

4. ATTENTION-BASED MULTIPLE ENCODER-DECODER

4.1. Overview of proposed method

An overview of the proposed method is shown in Fig.1. In our method, two encoder networks are employed: one is assigned to a sequence of acoustic features and the other is assigned to images around gaze points, and an attention mechanism is applied to each sequence of hidden states. Then, character-wise hidden features are calculated using the attention weight and the sequence of hidden states mentioned above. The acoustic features are time-series features such as a mel-filterbank, and the images around gaze points represent square regions, which are cropped from subjective images based on the coordinates of gaze points acquired by an eye-tracking device. Thus, a sequence of images around gaze points is an interrelated time-series of human gaze regions.

4.2. Formulation

The attention-based method directly estimates a posterior, $p(\mathbf{C}|\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ represents a sequence of speech features, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ represents a sequence of image features, and $\mathbf{C} = \{c_1, c_2, \ldots, c_L\}$ represents a sequence of output characters. The posterior $p(\mathbf{C}|\mathbf{X}, \mathbf{Y})$ is factorized with a probabilistic chain rule as follows:

$$p(\mathbf{C}|\mathbf{X}, \mathbf{Y}) = \prod_{l=1}^{L} p(c_l|c_{1:l-1}, \mathbf{X}, \mathbf{Y}) \quad (1)$$

where $c_{1:l-1}$ represents a subsequence $\{c_1, c_2, \dots\ c_{l-1}\}$. The posterior $p(c_l|c_{1:l-1}, \mathbf{X}, \mathbf{Y})$ is calculated as follows:

$$\mathbf{H} = \text{AudioEncoder}(\mathbf{X}) \tag{2}$$

$$\mathbf{S} = \text{VideoEncoder}(\mathbf{Y}) \tag{3}$$

$$a_{lt} = \text{LocationAttention}(\mathbf{q}_{l-1}, \mathbf{h}_t, \mathbf{a}_{l-1}) \tag{4}$$

$$b_{ln} = \text{LocationAttention}(\mathbf{q}_{l-1}, \mathbf{s}_n, \mathbf{b}_{l-1}) \tag{5}$$

$$\bar{\mathbf{h}}_l = \Sigma_{t=1}^{T} a_{lt}\mathbf{h}_t \tag{6}$$

$$\bar{\mathbf{s}}_l = \Sigma_{n=1}^{N} b_{ln}\mathbf{s}_n \tag{7}$$

$$\mathbf{g}_l = \sigma(\mathbf{W}_g[\bar{\mathbf{h}}_l^\top, \bar{\mathbf{s}}_l^\top]^\top + \mathbf{b}_g) \tag{8}$$

$$\mathbf{r}_l = \bar{\mathbf{h}}_l + \mathbf{g}_l \odot \bar{\mathbf{s}}_l \tag{9}$$

$$p(c_l|c_{1:l-1}, \mathbf{X}, \mathbf{Y}) = \text{Decoder}(\mathbf{r}_l, \mathbf{q}_{l-1}, c_{l-1}) \tag{10}$$

where equations (2) and (3) represent the audio encoder and the video encoder, respectively, and equation (10) represents the decoder network. $\mathbf{h}_t$, $\mathbf{s}_n$, and $\mathbf{q}_l$ represent the hidden state vectors of the audio encoder, video encoder, and decoder networks, respectively. $\mathbf{H}$ and $\mathbf{S}$ represents a sequence of hidden state vectors $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T\}$ and $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$, respectively. a_{lt} and b_{ln} represents the attention weights for the sequence of hidden state vectors in the audio encoder and video encoder network, respectively. $\bar{\mathbf{h}}_l$ and $\bar{\mathbf{s}}_l$ represents hidden state vector for audio encoder and video encoder, which are weighted sum of the attention weights, respectively. $\mathbf{g}_l$ represents the gate vector with a role of determining the priority of hidden vector of the video encoder, which is inspired by the fine-gating introduced in Cold Fusion [16]. Finally, the input vector $\mathbf{r}_l$ for the decoder is calculated by the summation of hidden states $\bar{\mathbf{h}}_l$ in the character-wise audio encoder and hidden states $\bar{\mathbf{s}}_l$ in the character-wise video encoder which is weighted by the gate vector $\mathbf{g}_l$.

The audio encoder network is a neural network that converts an acoustic feature sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ into a discriminative hidden vector sequence $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T\}$, and the network is modeled by a CNN (VGG [17]) and bidirectional long short-term memory (BLSTM), where the VGG is a simplified network with four convolution layers and two pooling layers.

$$\text{AudioEncoder}(\mathbf{X}) = \text{BLSTM}(\text{VGG}(\mathbf{X})) \tag{11}$$

In the case of speech recognition, the sequence length significantly varies, therefore, the input sequence is shortened to a quarter of its original length using max-pooling in the VGG.

The video encoder network converts an input image sequence $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ into a discriminative hidden vector sequence $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$, and it is modeled by AlexNet [18] and a recurrent neural network (BLSTM).

$$\text{VideoEncoder}(\mathbf{Y}) = \text{BLSTM}(\text{AlexNet}(\mathbf{Y})) \tag{12}$$

Because of insufficient training data, AlexNet was retrained using the ImageNet [19], one of a large image dataset.

The attention weight a_{lt} represents the alignment between an element c_l in the output sequence and the hidden state vector $\mathbf{h}_t$ in the audio encoder. $\text{LocationAttention}(\cdot)$ represents location-based attention [1] and is calculated as follows:

$$\mathbf{F}_l = \mathbf{K} * \mathbf{a}_{l-1} \tag{13}$$

$$e_{lt} = \mathbf{g}^{\mathrm{T}} \tanh(\mathbf{W}_q\mathbf{q}_l + \mathbf{W}_h\mathbf{h}_t + \mathbf{W}_f\mathbf{f}_{lt} + \mathbf{b}) \tag{14}$$

$$\mathbf{a}_l = \text{Softmax}(\mathbf{e}_l) \tag{15}$$

where $\mathbf{F}_l$ represents the vector sequence $\{\mathbf{f}_{l1}, \mathbf{f}_{l2}, \dots, \mathbf{f}_{lT}\}$, $\mathbf{K}$ represents a trainable convolutional filter, and the attention weight b_{ln} represents the alignment between elements c_l in the output sequence and the hidden state vector $\mathbf{s}_n$. These are calculated in the same manner as for a_{lt}.

The decoder network estimates the next character c_l from the previous character c_{l-1}, the hidden state vector of itself $\mathbf{q}_{l-1}$, and the character-wise hidden state vector $\mathbf{r}_l$, similarly to the RNN language model (RNNLM) [20]. It is typically modeled using LSTM as follows:

$$\mathbf{q}_l = \text{LSTM}(c_l, \mathbf{q}_{l-1}, \mathbf{r}_l) \tag{16}$$

$$\text{Decoder}(\cdot) = \text{Softmax}(\mathbf{W}\mathbf{q}_l + \mathbf{b}) \tag{17}$$

where $\mathbf{W}$ and $\mathbf{b}$ represent trainable matrix and vector parameters, respectively.

Finally, all of the above networks are optimized using back-propagation through time (BPTT) [21] to minimize the following objective function:

$$\begin{aligned}\mathcal{L} &= -\log p(\mathbf{C}|\mathbf{X}, \mathbf{Y}) \\ &= -\log\left(\sum_{l=1}^{L} p(c_l|c^*_{1:l-1}, \mathbf{X}, \mathbf{Y})\right)\end{aligned} \tag{18}$$

where $c^*_{1:l-1} = \{c^*_1, c^*_2, \dots, c^*_{l-1}\}$ represents the ground truth of the previous characters.

5. EXPERIMENT

5.1. Building Corpus

To evaluate the proposed method, we built a corpus that contains both utterances and corresponding images around gaze points in the operations. We used a glass type eye-tracking device, Tobii Glass2 [1] , and prepared the simulation task of power-line control operations. The images are cropped to 128 $\times$ 128 pixels from subjective images referring to the coordinates of gaze points. In Tobii Glass2, the sampling rate for recording the gaze points is 50 Hz, and the frame rate of subjective images (1920 $\times$ 1080 pixels, MP4) using the built-in camera is 25 fps. Therefore, we resampled the gaze points to 25 Hz.

In the following, we explain the simulation task of power-line operations. In this task, the subjects execute the sequence

[1] https://www.tobiipro.com

Fig. 2. *Example of simulation panel and gaze point.*

Table 1. Example of operation sequence (these utterances were originally in Japanese).

(1) Declaration of the operation procedure
"At *IDAKA, IDAKA-INOKOSHI* power-line No. 1,
turn on line-switch 782."
(2) Selection of the button corresponding to the operation
"Select line-switch 782." (push the button)
(3) Confirmation of the operation procedure
"Line-switch 782 was selected correctly."
(4) Execution of the operation
"Turn on the line-switch." (push the button)
(5) Confirmation of the operation result
"At *IDAKA, IDAKA-INOKOSHI* power-line No. 1,
line-switch 782 was turned on correctly."
(6) Confirmation of the time
"It's 15:35."

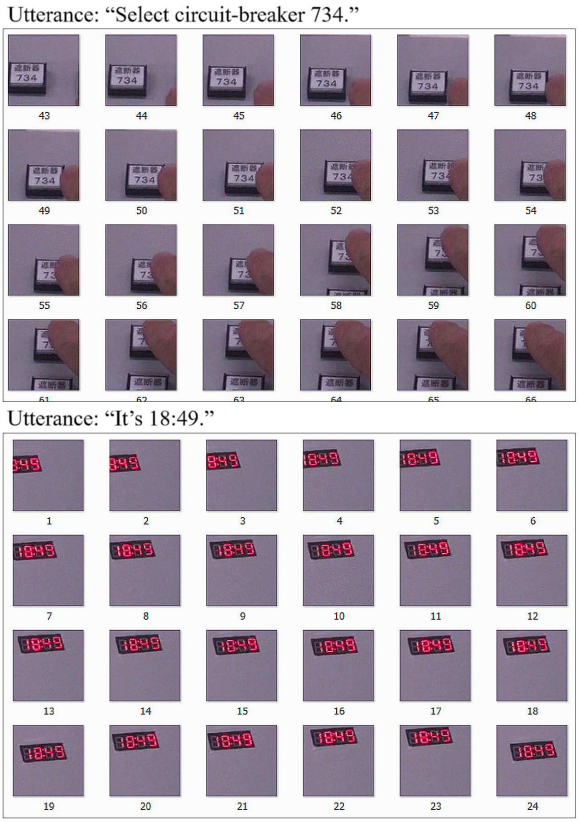

Fig. 3. *Examples of image sequences corresponding to utterances.*

of operations. In each operation, the subjects utter the operation content and confirm the operation results. The simulation panel and an example of a gaze point are shown in Fig.2; the center of the circle represents the coordinates of the gaze point. To build the corpus, the subjects executed the task in several sessions comprising the sequence of operations shown in Table 1 (including six utterances). The subjects were four males with no work experience related to the task. In each recording session, both the label of button and the power-line name were changed. Examples of image sequences corresponding to utterances are shown in Fig.3.

5.2. Experimental Setup

An overview of the collected corpus is shown in Table 2. Using the corpus, we conducted an experiment based on leave-one-subject-out (LOSO), that is, the data of two of the four speakers were used for the training, the data of one of the remaining speakers were used for the validation, and the data of the fourth speakers were used for evaluation. In this scheme, the character error rate (CER) was calculated for each evaluation speaker, and the average CER is the final evaluation result.

The experimental conditions are shown in Table 3. For the experiment, we implemented the proposed method on the basis of the hybrid CTC/Attention architecture [13] (ESPnet [22]). Because of the insufficient training data, we used the Corpus of Spontaneous Japanese [23] (CSJ) for pretraining of the initial model. First, in this procedure, the model is pretrained, except for the video encoder, using the whole set of CSJ data. Second, the whole parameter of model including the video encoder is retrained using the task corpus and the initial weights of both the audio encoder and the decoder. In this training, we used the initial weight trained by ImageNet [19] for the feature extraction module (CNN) of the video encoder. In addition, we used the Joint CTC Attention multi-task learning [13], which uses the CTC objective function together to learn the audio encoder. For the evaluation

Table 2. Overview of corpus.

Speaker ID	Num. of Sessions	Num. of Utterances
SPK01	23	138
SPK02	20	120
SPK03	20	120
SPK04	20	120

measure, the following CER was used:

$$\mathrm{CER} = \frac{S + D + I}{N} \tag{19}$$

where S, D, I and N are the number of substitution errors, deletion errors, insertion errors, and the total number of recognized characters, respectively. To evaluate the effectiveness of the proposed method, we compared the following two models (without the video encoder).

1. The model trained by the whole set of CSJ data.
2. The model retrained with the data of the evaluation task using the initial weight learned with all the CSJ data.

5.3. Experimental Results

The experimental results are shown in Table 4. First, the result for the initial model (CSJ pretrain) is very low, because of the effect of unknown words in the task. Second, the result for the model (+fine-tuning) is improved significantly. This improvement is mainly due to the reduction in the number of unknown words and acoustic adaptation.

The CER for the model (+video encoder) is reduced from 7.2 % to 6.9 %. Thus, the results suggest that our proposed method using gaze information is effective for improving recognition performance. In addition, the CER for each speaker in the LOSO experiment is shown in Table 5. The difference in performance may be due to individuality in both utterances and gaze points. We will analyze the issue with a larger corpus.

6. CONCLUSION

In this study, we assume that there is a correlation between an utterance and a corresponding gaze object, and proposed a new paradigm of multi-modal end-to-end speech recognition using multi-modal information, namely, acoustic features and gaze points. To evaluate the proposed method, we prepared a simulation task of power-line control operations, and built a corpus that contains utterances and corresponding gaze points in the operations. We conducted an experimental evaluation using this corpus. The results show the reduction in the CER suggesting the effectiveness of the proposed method in which

Table 3. Experimental condition.

# CSJ training data	445,068
# utterances in task corpus	498
# unique characters	3,260
sampling rate	16,000 Hz
window size	25 ms
shift size	10 ms
audio encoder type	VGG-BLSTM
# audio encoder BLSTM layers	4
# audio encoder BLSTM units	2,048
# audio encoder projection units	1,024
video encoder type	AlexNet-BLSTM
# video encoder BLSTM layers	1
# video encoder BLSTM units	2,048
# video encoder projection units	1,024
audio encoder attention type	Location-based
kernel size in audio encoder attention	100
# filters in audio encoder attention	10
video encoder attention type	Location-based
kernel size in video encoder attention	20
# filters in video encoder attention	10
decoder type	LSTM
# decoder layers	1
# decoder units	1,024
learning rate	1.0
dropout	0.2
gradient clipping norm	5
batch size	20 (pretrain) 8 (fine-tuning)
maximum epoch	15
beam size	20
MTL alpha	0.5
CTC weight in decoding	0.3

Table 4. Experimental results.

Model	S [%]	D [%]	I [%]	CER [%]
CSJ pretrain	23.6	2.5	7.0	33.0
+ fine-tuning	5.3	0.9	1.0	7.2
+ video encoder	5.2	1.1	0.7	**6.9**

Table 5. Experimental results (for each speaker).

Model (Speaker)	S [%]	D [%]	I [%]	CER [%]
CSJ pretrain				
(SPK01)	31.4	4.8	7.4	43.7
(SPK02)	21.0	1.5	10.4	32.9
(SPK03)	19.4	2.7	7.6	29.7
(SPK04)	22.4	0.9	2.6	25.8
+ fine-tuning				
(SPK01)	9.5	2.2	1.7	13.3
(SPK02)	0.6	0.0	0.2	0.8
(SPK03)	5.2	1.1	1.1	7.4
(SPK04)	6.0	0.2	0.9	7.1
+ video encoder				
(SPK01)	9.5	3.2	1.0	13.7
(SPK02)	0.8	0.1	0.2	1.1
(SPK03)	4.4	0.7	0.9	**6.1**
(SPK04)	6.0	0.2	0.6	**6.7**

acoustic features and gaze information are integrated. In future work, we will analyze the attention weights of image sequences around gaze points and evaluate the performance using a larger corpus.

7. REFERENCES

[1] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[2] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2130–2134.

[3] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[4] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," *arXiv preprint arXiv:1802.06424*, 2018.

[5] C. Nguyen and F. Liu, "Gaze-based notetaking for learning from lecture videos," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 2093–2097.

[6] A. B. Vasudevan, D. Dai, and L. Van Gool, "Object referring in videos with language and human gaze," *arXiv preprint arXiv:1801.01582*, 2018.

[7] M. H. Rasmussen and Z-H.Tan, "Fusing eye-gaze and speech recognition for tracking in an automatic reading tutor - a step in the right direction?" in *Proc. SLaTE*, 2013, pp. 112–115.

[8] N. J. Cooke, "Gaze-contingent automatic speech recognition," *Ph.D thesis, Univ. of Birmingham*, 2006.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[10] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[11] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[14] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[16] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," in *Proc. Interspeech 2018*, 2018, pp. 387–391.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," 2009.

[20] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[21] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[23] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.