SPEAKER-TARGETED AUDIO-VISUAL SPEECH RECOGNITION USING A HYBRID CTC/ATTENTION MODEL WITH INTERFERENCE LOSS

Ryota Tsunoda[1], *Ryo Aihara*[2], *Ryoichi Takashima*[1], *Tetsuya Takiguchi*[1], *Yoshie Imai*[2]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]Information Technology R&D Center, Mitsubishi Electric Corporation, Japan

ABSTRACT

Audio-visual (AV)-automatic speech recognition (ASR) can improve speech recognition accuracy by using lip images, especially in noisy environments. The recently proposed AV Align system integrates speech and image features based on a cross-modal attention mechanism, where attention weights for visual features are estimated by using acoustic features as queries. Although AV Align shows an improvement in recognition accuracy in background noise environments, we have observed that the recognition accuracy degrades significantly in interference speaker environments, where a target speech and an interfering speech overlap each other. In order to improve the speech recognition accuracy of the target speaker in such situations, we propose a method that combines the auxiliary loss function that maximizes the recognition accuracy of the interference speaker and the CTC loss function for training the AV-ASR model. The experimental results using the TCD-TIMIT dataset show that the use of these auxiliary loss functions improves the performance of target-speaker speech recognition in interference speaker environments.

Index Terms— Speech recognition, multi-modal, speaker targeted, cross-modal alignment, interference speaker

1. INTRODUCTION

Humans are able to recognize speech content even in noisy environments because they make use of a variety of information in an integrated manner. In particular, the influence of visual lip information on speech understanding is significant. For example, the McGurk effect [1] is known to cause people to misunderstand what is being said when they see a moving image in which the movement of the lips and the words being spoken do not match. Therefore, even when the speech signal degrades, humans are able to understand the speech content to some extent from visual lip information.

Based on the above background, audio-visual (AV) speech recognition has been studied in order to improve the accuracy of automatic speech recognition (ASR) by using both speech and lip video images [2, 3]. AV speech recognition has been shown to be effective, especially in noisy environments [4]. For systems used in noisy environments, such as car navigation systems and service robots, it is expected that the speech content can be recognized more robustly by using in-vehicle cameras and robot cameras, respectively.

In AV speech recognition, various methods have been proposed to integrate multimodal information from speech and lip images. In [5], a hybrid CTC/attention model is used for AV speech recognition, and feature-level and output-level fusions are compared. Since the frame rate of the video image is generally smaller than the sampling rate of the audio, feature-level fusion requires us to synchronize the alignment of the frames. As a conventional method for synchronization, upsampling by interpolating between video frames is often used [6, 7]. On the other hand, in [8, 9], AV Align, which integrates speech and visual features by using a cross-modal attention mechanism without upsampling, is proposed and shows better performance than conventional upsampling-based methods.

AV Align showed improved recognition accuracy in background noise environments; however, the recognition accuracy of the target speaker was greatly degraded in interference speaker environments, where multiple speakers were speaking simultaneously. One possible reason for this degradation is that AV Align does not have a mechanism to separate the target speaker from the interference speaker sufficiently. In addition, since attention weights are generally known to be difficult to estimate in noisy environments [10], it may be difficult to estimate the attention weights in interference speaker environments.

In this study, we investigate the performance improvement of AV speech recognition under interference speaker environments based on AV Align. First of all, while AV Align uses an attention decoder, the proposed method uses a hybrid CTC/attention decoder. In this decoder, the auxiliary loss of the connectionist temporal classification (CTC), which makes it easier to learn the attention mechanism by constraint of monotonic alignment, is used during training. This is expected to improve learning efficiency in interference environments. In addition, in order to improve the ability of AV Align to separate the target speaker from the interference speaker, we add a network to recognize interfering speech in the model and introduce an auxiliary loss function based on interfering speech recognition. Since the proposed method learns to recognize not only the target utterance but also the interfering

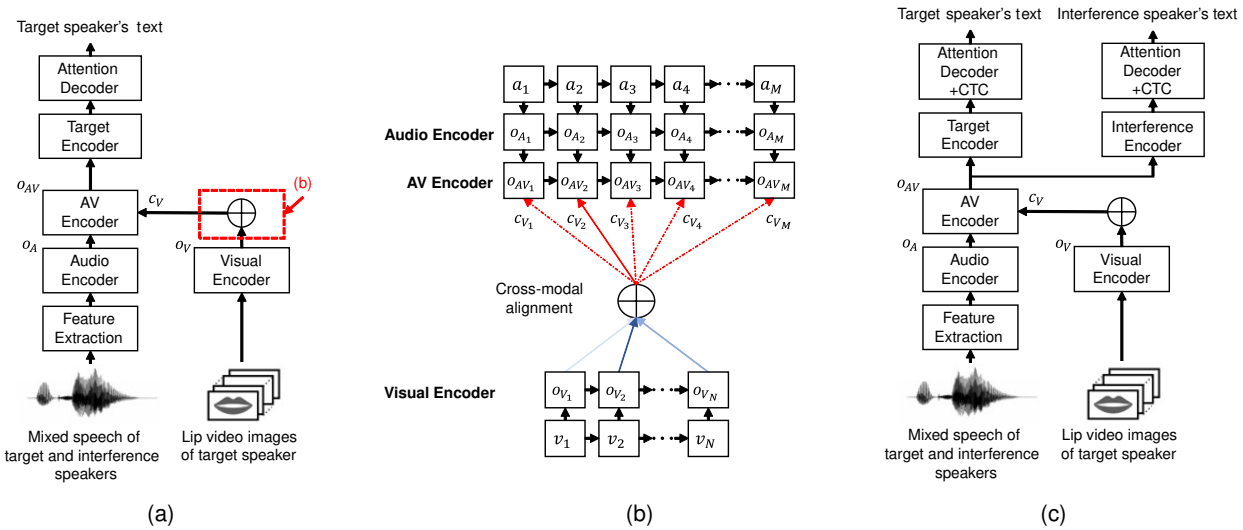

Fig. 1. (a) AV Align, (b) Cross-modal attention mechanism, (c) Proposed model architecture

utterance correctly, it is expected to improve the ability of speech recognition systems to separate the two types of utterances.

2. AV ALIGN

AV Align, as shown in Fig. 1a, consists of an encoder-decoder model. The encoder converts input sequences of acoustic and visual features into high-level features. The decoder estimates the character sequence related to the input speech from the encoder output.

In AV Align, a cross-modal attention that estimates cross-modal alignment is proposed for integrating speech and image features. Since the frame rate of video images is generally lower than the sampling rate of audio signals, in order to integrate them in feature-level, it is necessary to obtain the frame correspondence (alignment) between the audio and visual features. By using the cross-modal attention mechanism, AV Align can automatically learn the alignment between speech and images. A diagram of the cross-modal attention mechanism is shown in Fig. 1b. First, a sequence of audio features $\boldsymbol{a} = \{a_1, a_2, \ldots, a_M\}$ and a sequence of visual features $\boldsymbol{v} = \{v_1, v_2, \ldots, v_N\}(M > N)$ are input to the audio encoder and visual encoder, respectively. Each feature is transformed into a high-level feature $\boldsymbol{o}_A = \{o_{A_1}, o_{A_2}, \ldots, o_{A_M}\}$, $\boldsymbol{o}_V = \{o_{V_1}, o_{V_2}, \ldots, o_{V_N}\}$ through the corresponding encoder.

$$o_{A_i} = \mathrm{Encoder_A}(a_i, o_{A_{i-1}}) \quad (1)$$

$$o_{V_j} = \mathrm{Encoder_V}(v_j, o_{V_{j-1}}) \quad (2)$$

$\mathrm{Encoder_A}$ and $\mathrm{Encoder_V}$ are the audio encoder and visual encoder, respectively. By applying the cross-modal attention mechanism with high-level audio features h_i as queries to the high-level visual features o_V, the visual features having high relevance to the acoustic features at each frame are extracted, and the context vector $\boldsymbol{c}_V$ is calculated.

$$h_i = \mathrm{Encoder_{AV}}([o_{A_i}; o_{AV_{i-1}}], h_{i-1}) \quad (3)$$

$$c_{V_i} = \mathrm{attention}(h_i, o_V) \quad (4)$$

$\mathrm{Encoder_{AV}}$ is the AV encoder. h_i is the hidden state of the AV encoder at time i, and $o_{AV_{i-1}}$ is the output of the AV encoder at time $i-1$. The output o_{AV_i} of the AV encoder at time i is calculated using the following equation:

$$o_{AV_i} = W_{AV}[h_i; c_{V_i}] + b_{AV} \quad (5)$$

W_{AV} and b_{AV} are the weight matrix and bias vector of the linear layer, respectively. Then, the o_{AV_i} passes through the target encoder[1] and attention decoder, and the character sequence is output.

3. PROPOSED METHOD

Our proposed model is shown in Fig. 1c. Compared with the conventional model shown in Fig. 1a, the proposed model uses a hybrid CTC/attention decoder and an additional network that recognizes interference speaker's speech to calculate the auxiliary loss.

[1]The original AV Align in [8, 9] did not have a target encoder. However, we add this block for matching conditions with the proposed method described in Chapter 3. We have also confirmed that adding the target encoder improves the recognition accuracy of the original AV Align system.

3.1. Hybrid CTC/attention decoder

In this study, we use the hybrid CTC/attention model [11], which combines the CTC model [12] and attention encoder-decoder model [13]. This model is trained so as to minimize the following multi-task loss function:

$$L_{mtl} = \alpha L_{ctc} + (1 - \alpha) L_{att} \quad (6)$$

Here, L_{ctc} and L_{att} are the loss functions for the CTC model and attention encoder-decoder model, respectively, and α is the weight parameter for multi-task learning. Unlike the attention mechanism, the CTC loss function explicitly learns the monotonic alignment between input and output during training. Combining the CTC loss function with the attention mechanism has the advantage of allowing only monotonic alignment in the attention mechanism, leading to improved convergence performance. In this study, we expect that the addition of the CTC loss function to AV Align will enable robust speech recognition in interference speaker environments.

3.2. Interference speaker loss

In order to enhance the ability of AV Align to separate the target speech in interference speaker environments, this study introduces an auxiliary function based on interference speech recognition when training the model. For single-modal speech recognition, a method using an auxiliary loss based on the interference speech recognition has been proposed [14]. In this previous work, the auxiliary loss is used with a lattice-free maximum mutual information loss function [15] and showed improvement in the target speaker recognition accuracy.

The speech-visual features $\boldsymbol{o}_{AV}$ obtained by cross-modal attention mechanism described in Chapter 2 are input into the target encoder and the interference encoder for recognizing the target speaker and the interference speaker, respectively. The features processed by the two encoders are input to the respective decoders, which output the recognition results of the target speaker and the interference speaker. The loss function for speech recognition of the target and interference speakers are expressed as follows:

$$L_{traget} = \alpha L_{t\_ctc} + (1 - \alpha) L_{t\_att} \quad (7)$$

$$L_{interference} = \alpha L_{i\_ctc} + (1 - \alpha) L_{i\_att} \quad (8)$$

The overall loss function used for training in this study is expressed by the following equation:

$$L_{all} = \lambda_1 L_{traget} + \lambda_2 L_{interference} \quad (9)$$

λ_1 and λ_2 are the weight parameters for the respective loss functions. In this study, we set $\lambda_1 = \lambda_2 = 1.0$. In this way, we expect that the AV encoder will train feature representations for speaker separation.

4. EVALUATION EXPERIMENT

4.1. Experimental conditions

We used TCD-TIMIT [16] as a dataset for our AV speech recognition experiments. TCD-TIMIT consists of the audio and video images of 62 speakers uttering a total of 6,913 sentences. Since each speech in the dataset is uttered by a single speaker, in this experiment, we created an interference speaker environment by superimposing the speech of two speakers. First, for training data, we picked out a subset of 3,752 utterances from the dataset, and then we created 26,264 mixed speech utterances by superimposing two utterances randomly selected from the subset for each mixed utterance. In similar way, for evaluation data, we created 1,736 mixed speech utterances using different utterances from those used for training data. The signal-to-noise ratio (SNR) was set according to Loudness [17], which is based on human auditory characteristics. The SNR for the training data was randomly selected from -10 dB and 10 dB for each sentence, and the evaluation data was created for the five SNR conditions: -10, -5, 0, 5, and 10 dB.

We implemented our AV speech recognition model by modifying ESPnet toolkit [18], which implements a hybrid CTC/attention model. A total of 26-dimensional features, including 23-dimensional mel filter bank features and 3-dimensional pitch features, were used as input audio features. For the visual features, we used a 3-channel color image that was resized to 36×36 by cropping the lip area after detecting the face region from the video image using OpenFace [19]. Note that only the image features corresponding to the target speaker were input. The number of output dimensions was 31, including 26 types of English characters plus apostrophes, unknown characters, spaces, and start of sequence and end of sequence symbols.

The audio encoder consists of a 5-layer bi-directional GRU [20] with 320 hidden units in each layer. The visual encoder consists of an 11-layer Resnet CNN, which was used in [9], and a one-layer unidirectional LSTM [21] with 320 hidden units. For the AV encoder, we used a one-layer unidirectional LSTM with 320 hidden units. A three-layer bi-directional GRU with 320 hidden units in each layer was used for the target encoder and interference encoder. The decoder consists of a one-layer unidirectional LSTM with 320 hidden units, followed by a softmax layer with 31-dimensional output units. We used the coverage mechanism location aware attention as the attention mechanism, and AdaDelta [22] was used to optimize the model. Coverage mechanism location aware attention is a combination of coverage mechanism attention [23] and location aware attention [24]. We set the weight of the loss function of the CTC to 0.5 during both

Table 1. Character Error Rates (CERs) of each experimental condition

Model	CER[%]				
	10dB	5dB	0dB	-5dB	-10dB
Audio only	29.7	38.9	49.6	57.2	60.5
L_{t_att}[8]	29.5	32.0	34.0	36.2	38.5
$L_{t\_att} + L_{i\_att}$	29.2	30.4	32.4	34.4	36.9
$L_{t\_att} + L_{t\_ctc}$	18.4	21.1	23.9	27.5	30.6
$L_{t\_att} + L_{t\_ctc} + L_{i\_att}$	17.9	20.3	23.4	26.8	30.1
$L_{t\_att} + L_{t\_ctc} + L_{i\_att} + L_{i\_ctc}$	15.7	17.8	20.3	23.1	26.4

training and recognizing.

Before training a model using two-speaker mixed speech, a single-speaker model was first trained using the original TCD-TIMIT data, and then a two-speaker mixed speech model was trained using that as the initial model. At this time, the parameters of interference encoder were set to the initial value of the parameters of target encoder.

4.2. Experimental results

Table 1 shows the character error rates (CERs) for each condition. L_{t_att} corresponds to the conventional AV Align shown in Fig. 1a. In the case of Audio only[2], recognition is considered to be difficult because it is not possible to distinguish between the speech of the target speaker and the interference speaker. Compared to the results of audio-only, the recognition accuracy was improved by using lip video images as auxiliary information for the target speaker. When L_{t_ctc} was added to AV Align, the performance improved up to 37.6% relatively. It is thought that the addition of the CTC loss function was effective in training the model in the interference speaker environment, and, thus, improved the recognition accuracy. Also, when $L_{interference}$ was added to L_{target}, the accuracy improved. When L_{i_att} was added to $L_{t\_att} + L_{t\_ctc}$, the accuracy improved, especially when $L_{i\_att} + L_{i\_ctc}$ was added to $L_{t\_att} + L_{t\_ctc}$, the performance improved up to 16.0% relative to $L_{t\_att} + L_{t\_ctc}$. This indicates that the use of an auxiliary loss function based on the speech recognition of the interference speaker is effective in improving the speech recognition accuracy of the target speaker. In addition, the addition of the CTC loss function was also effective in training relatively complex models using the network for recognizing interference speech.

Table 2 shows the CERs of the target speaker speech recognition when the genders of the target and interference speakers in the test data were the same ("Same" in Table 2) and when they were different ("Diff" in Table 2). Whether $L_{interference}$ was used or not, there was no significant difference in performance between the Same and Diff conditions.

[2]The hybrid CTC/attention model was used to model the case where only an audio signal was used.

Table 2. CERs for each gender condition

SNR	L_{target}		$L_{target} + L_{interference}$	
	Same	Diff	Same	Diff
10 dB	18.3	18.6	15.8	15.5
5 dB	21.2	21.1	17.7	17.8
0 dB	24.1	23.9	20.5	19.9
-5 dB	27.0	27.7	23.1	23.0
-10 dB	30.8	30.5	27.0	26.0

Table 3. Experimental results of the target speaker ASR and the interference speaker ASR

SNR of target spk	SNR of interference spk	target spk	interference spk
10 dB	-10 dB	15.7	56.0
5 dB	-5 dB	17.8	47.1
0 dB	0 dB	20.3	38.7
-5 dB	5 dB	23.1	31.5
-10 dB	10 dB	26.4	27.0

In previous research [25], it was reported that using image information as auxiliary information enables robust source separation even for speech involving speakers of the same gender. Similarly, in this study, it is considered that those gender conditions did not affect the recognition performance.

Our proposed model can recognize the interference speaker's speech as well as the target speaker's speech. Table 3 shows the CERs of the recognition results of the interference speaker's speech. Compared with the speech recognition accuracy of the target speaker, the recognition accuracy is low because no auxiliary information is given for the interference speaker. Nevertheless, the recognition accuracy is higher than that of the audio-only case (see Audio only in Table 1), where no auxiliary information is given. This result suggests that the proposed model is able to learn feature expressions to separate the speech of the target speaker from the interference speaker by using the interference speaker loss, and as a result, it is able to recognize the interference speaker's speech to some extent.

5. CONCLUSION

In this study, we improved the accuracy of AV target-speaker speech recognition by using two auxiliary loss functions during training. We confirmed that the addition of the CTC loss can greatly improve the recognition accuracy. Furthermore, by adding a loss function based on the recognition of the interference speaker, the recognition accuracy can be further improved for all SNR, and the speech of the interference speaker can be recognized as well.

6. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

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

[3] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, Dec. 2018.

[4] K. Paleček and J. Chaloupka, "Audio-visual speech recognition in noisy audio environments," in *Proc. TSP*, July 2013, pp. 484–487.

[5] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," in *Proc. SLT*, Dec. 2018, pp. 513–520.

[6] J.S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, July 2017, pp. 3444–3453.

[7] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. ICASSP*, Apr. 2018, pp. 6548–6552.

[8] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proc. ACM ICMI*, Oct. 2018, pp. 111–115.

[9] G. Sterpu, C. Saam, and N. Harte, "How to teach dnns to pay attention to the visual modality in speech recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, pp. 1052–1064, Jan. 2020.

[10] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, Mar. 2017, pp. 4835–4839.

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

[12] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, Jun 2014, vol. 32, pp. 1764–1772.

[13] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. NIPS*, Dec. 2014.

[14] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, and S. Watanabe, "Auxiliary interference speaker loss for target-speaker speech recognition," in *Proc. INTERSPEECH*, Sept. 2019, pp. 236–240.

[15] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, Sept. 2016, pp. 2751–2755.

[16] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[17] ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," Aug. 2012.

[18] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Yalta, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, Sept. 2018, pp. 2207–2211.

[19] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. FG*, May 2018, pp. 59–66.

[20] K. Cho, B.V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, Oct. 2014, pp. 1724–1734.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[22] M.D. Zeiler, "ADADELTA: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, Dec. 2012.

[23] A. See, P.J. Liu, and C.D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. ACL*, July 2017, pp. 1073–1083.

[24] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, Dec. 2015, p. 577–585.

[25] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues," in *Proc. INTERSPEECH*, Sept. 2019.