INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Audio-Visual Multi-Talker Speech Recognition in A Cocktail Party

Yifei Wu[1], *Chenda Li*[1], *Song Yang*[2], *Zhongqin Wu*[2], *Yanmin Qian*[1†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]TAL Education Group, China

{yifei.wu,lichenda1996,yanminqian}@sjtu.edu.cn,{yangsong1,wuzhongqin}@tal.com

Abstract

Speech from microphones is vulnerable in a complex acoustic environment due to noise and reverberation, while the cameras are not. Thus, utilizing the visual modality in the "cocktail party" scenario with multi-talkers has become a promising and popular approach. In this paper, we have explored the incorporating of visual modality into the end-to-end multi-talker speech recognition task. We propose two methods based on the modality fusion position, which are encoder-based fusion and decoder-based fusion. And for each method, advanced audio-visual fusion techniques including attention mechanism and dual decoder have been explored to find the best usage of the visual modality. With the proposed methods, our best audio-visual multi-talker automatic speech recognition (ASR) model gets almost ∼50.0% word error rate (WER) reduction compared to the audio-only multi-talker ASR system.

Index Terms: audio-visual, multi-talker ASR, cocktail party, attention model

1. Introduction

Multi-talker automatic speech recognition (ASR) is one of the techniques for solving the "cocktail party problem" [1]. Thanks to the permutation invariant training (PIT) [2] and the advances of end-to-end ASR [3, 4, 5, 6] systems, researchers are able to train multi-talker ASR systems in an end-to-end manner [7, 8, 9, 10, 11, 12, 13, 14].

Compared to the single-talker ASR systems, the main challenge of multi-talker ASR in the "cocktail party" comes from the more complex acoustic environment. On the one hand, in that complex "cocktail party" environment, more than one talker may talk simultaneously, making it more difficult to track the talkers. On the other hand, noise and reverberation may also be involved, which will make the situation more complicated. Thus, the speech signal collected from the microphone will be heavily distorted in these complex conditions, and consequently, the performance of the multi-talker ASR system will be degraded. Some speech separation and enhancement methods can be incorporated into the multi-talker ASR system to tackle the complex acoustic environment [8, 15, 16, 17]. However, these techniques focus on the speech signal itself while more information may be utilized in reality.

In real scenarios, information other than speech can be utilized to solve the "cocktail party" problem. We humans are masters at this. In humans' selective hearing [18], the visual modality sometimes plays an important role [19] in addition to the audio modality. An extreme example is that lip-reading without auditory perception can also achieve fair performance [20]. Besides, the visual modality is usually not disturbed by the acoustic environments, i.e. the quality of visual modality is immune to poor acoustic environments [21]. Thus, utilizing the visual modality in the multi-talker scene is an intuitive and promising approach.

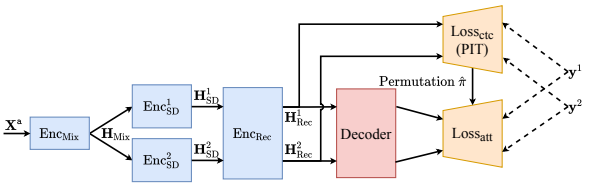


Figure 1: *The backbone and baseline end-to-end single-channel multi-talker ASR model adopted in this paper. K is assumed to be 2 in the figure.*

The advantages of incorporating the visual modality into multi-talker ASR are twofold. First, due to the high correlation with speech, the visual modality (e.g. lip movement, facial expression) contains the knowledge for speech recognition. Researches have been conducted to investigate the possibility to introduce visual information into speech recognition [22, 23, 24, 25]. Second, visual is also a practical guide for solving the permutation problem (also known as the label ambiguity problem) [26, 2] in the multi-talker disentangling training. It is well-known that the PIT [2] criterion is first presented to solve the label ambiguity problem in the speech separation task. However, with each talker's visual information introduced, we could reasonably assume that the separation model would output the results with a permutation prejudged by the visual inputs. Thus we can obtain the speech of the talker we care about, training the model with fixed label order. This idea has been verified by experiments conducted on speech enhancement tasks [27, 28] and speech separation tasks [29, 30].

A recent work [31] also focused on this idea and proposed a streamlined and integrated audio-visual speech recognition (AVSR) system to recognize the target speech out of an overlapped one. The primary approach is to mask the hidden audio features using a gate calculated by both the audio mixture and the target talker's visual input. However, this work focuses on one-talker case only, while the vision to multiple talkers might further improve the system performance.

In this paper, we extend the audio-visual speech recognition task to the multi-talker application, to recognize multiple talkers' speech in the cocktail party problem utilizing visual information of each talker. We also present several practical approaches to introduce visual information into a Transformer-based [32] end-to-end multi-talker ASR system [33]. With the

†corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2128

proposed method, an approaching relative ∼50.0% reduction on WER is achieved compared to the baseline, which only utilizes audio information.

The rest of this paper is organized as follows. Section 2 describes the assumptions and the goal of the task and the baseline model. Section 3 proposes two approaches to introduce visual information into the encoder part of the model and two for the decoder part. The experimental details and the results are presented and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Task Definition and Baseline

The multi-talker ASR aims to recognize the speech for each talker from speech mixtures of K ($K \geq 2$) talkers. Let $\mathbf{X}^{\mathrm{a}} = (\mathbf{x}_1, \cdots, \mathbf{x}_{T_{\mathrm{a}}})$ denotes T_{a} frames of input features extracted from the speech mixture where K talkers talk simultaneously. And the transcriptions of the k-th talker can be denoted as $\mathbf{y}^k = (y_1^k, \cdots, y_{U^k}^k)$, where U^k is the length of the tokens for the k-th talker, and $y_u^k \in \{1, \cdots, W\}$ is distinct labels in the dictionary of W tokens. The task of multi-talker ASR is to estimate $\hat{\mathbf{y}}^k, k = 1, \cdots, K$ from the mixture input $\mathbf{X}^{\mathrm{a}}$:

$$\{\hat{\mathbf{y}}^k : k = 1, \cdots, K\} = f_{\mathrm{a}}(\mathbf{X}^{\mathrm{a}}) \tag{1}$$

where $f_{\mathrm{a}}(\cdot)$ is the mapping function of the audio-only multi-talker ASR system.

In this paper, the baseline audio-only multi-talker ASR system is mainly adopted from [33] , which is based on the joint CTC/attention encoder-decoder model [5]. The backbone of baseline model is shown in Fig.1, which could be divided into the *encoder* part and the *decoder* part. In the *encoder* part, there are two kinds of layers named speaker-different encoder ($\mathrm{Enc}_{\mathrm{SD}}^k$) layers and recognition encoder ($\mathrm{Enc}_{\mathrm{Rec}}$) layers. The former one is deemed to extract the target talker from the mixture, while the latter one performs the regular function of the ASR encoder layer. Each of $\mathrm{Enc}_{\mathrm{SD}}^k$, and $\mathrm{Enc}_{\mathrm{Rec}}$ is a stack of Transformer encoder layers, while Decoder is a stack of Transformer decoder layers with an input embedding layer and a feed-forward output layer for sequence inference. Formally, the encoding process could be written as follows:

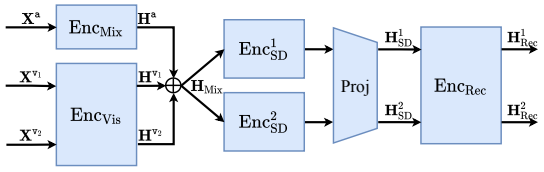
$$\begin{aligned}
\mathbf{H}_{\mathrm{Mix}} &= \mathrm{Enc}_{\mathrm{Mix}}(\mathbf{X}^{\mathrm{a}}), \\
\mathbf{H}_{\mathrm{SD}}^k &= \mathrm{Enc}_{\mathrm{SD}}^k(\mathbf{H}_{\mathrm{Mix}}), k = 1, \cdots, K, \\
\mathbf{H}_{\mathrm{Rec}}^k &= \mathrm{Enc}_{\mathrm{Rec}}(\mathbf{H}_{\mathrm{SD}}^k), k = 1, \cdots, K.
\end{aligned} \tag{2}$$

where $\mathbf{X}^{\mathrm{a}}$ is the feature sequence of the input speech mixture and $\mathbf{H}_{\mathrm{Rec}}^k$ is the encoded feature of talker k. For each $\mathbf{H}_{\mathrm{Rec}}^k$, the decoding process of Decoder is

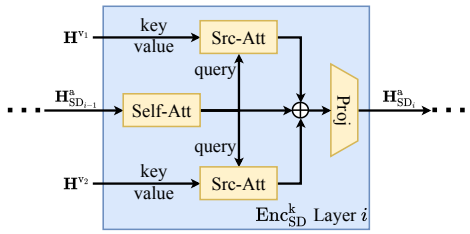
$$\begin{aligned}
\mathbf{c}_n^k &= \mathrm{Attention}(\mathbf{e}_{n-1}^k, \mathbf{H}_{\mathrm{Rec}}^k), \\
\mathbf{e}_n^k &= \mathrm{Update}(\mathbf{e}_{n-1}^k, \mathbf{c}_{n-1}^k, y_{n-1}^k), \\
y_n^k &\sim \mathrm{Decoder}(\mathbf{e}_n^k, y_{n-1}^k).
\end{aligned} \tag{3}$$

where $\mathbf{c}_n^k$ and $\mathbf{e}_n^k$ respectively denotes the context vector with dimension D and the hidden state of decoder of talker k at decoding step n.

In the training phase, the correspondence between the output label sequences and the reference label sequences are determined using the permutation presented by PIT on the connectionist temporal classification (CTC) [34] loss.



(a) *Concat*



(b) *Query Vision*

Figure 2: *Two approaches to introduce visual information into the encoder part. K is assumed to be 2 in the figure. $\mathbf{X}^{v_1} \in \mathbb{R}^{T_{v_1} \times D_v}$, $\mathbf{X}^{v_2} \in \mathbb{R}^{T_{v_2} \times D_v}$ are the visual features of talker 1 and 2 respectively. Fig.2b shows the modified structure of each layer in the Transformer encoder Enc_{SD}^k. $\oplus$ represents concatenation. Modules of the same name share parameters.*

3. Audio-Visual Multi-Talker ASR

In this section, we introduce the audio-visual multi-talker ASR. Based on the position of incorporating visual information, the proposed audio-visual multi-talker ASR model can be divided into two types, the *audio-visual encoder* approach and *audio-visual decoder* approach. For both approaches, we have explored two variants of model structure respectively.

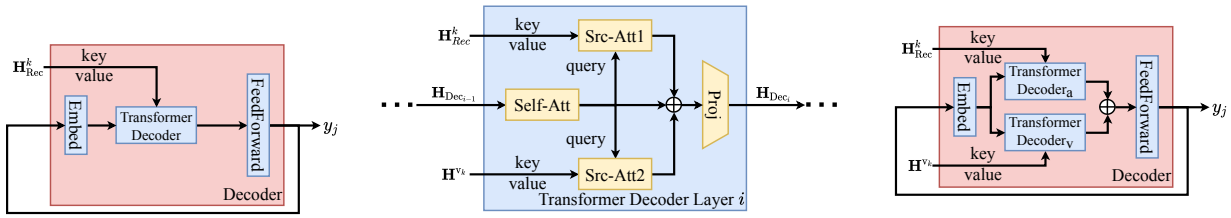
3.1. Audio-Visual Encoder

We describe two approaches to introduce visual information into the *encoder* part of the end-to-end multi-talker ASR model introduced in Section 2. The approaches are shown in Fig.2. A module named $\mathrm{Enc}_{\mathrm{Vis}}$ is included to transform the visual features $\mathbf{X}^{\mathrm{v}_k}$ into deep embedding $\mathbf{H}^{\mathrm{v}_k}$ of shape $T_{\mathrm{v}_k} \times D$.

Variant 1: Concat. The first approach named *Concat* in Fig.2a is based on the idea that the visual embedding and speech embedding could be concatenated. The visual embeddings are firstly resampled on the time dimension to match the length of the audio embedding. Then the concatenation is performed on the feature dimension D. After concatenation, each $\mathrm{Enc}_{\mathrm{SD}}^k$ processes the fusion embedding $\mathbf{H}_{\mathrm{Mix}}$ and the outputs $\mathbf{H}_{\mathrm{SD}}^{k'}$ for each talker is projected to get dimension-compressed embedding $\mathbf{H}_{\mathrm{SD}}^k$. $\mathbf{H}_{\mathrm{SD}}^k$ is regarded as the visual-aware deep embedding, and the following pipeline is the same as the baseline system. Eq.4 formally describes the calculation processes.

$$\begin{aligned}
\mathbf{H}_{\mathrm{Mix}} &= \mathrm{Concat}(\mathrm{Enc}_{\mathrm{Mix}}(\mathbf{X}^{\mathrm{a}}), \{\mathbf{H}^{\mathrm{v}_k} : k = 1, \cdots, K\}), \\
\mathbf{H}_{\mathrm{SD}}^k &= \mathrm{Proj}(\mathrm{Enc}_{\mathrm{SD}}^k(\mathbf{H}_{\mathrm{Mix}})), k = 1, \cdots, K, \\
\mathbf{H}_{\mathrm{Rec}}^k &= \mathrm{Enc}_{\mathrm{Rec}}(\mathbf{H}_{\mathrm{SD}}^k), k = 1, \cdots, K.
\end{aligned} \tag{4}$$

Variant 2: Query Vision. Inspired by [24], we present another approach named *Query Vision* as the second approach. As shown in Fig.2b, it only modifies the inner structure of each layer of $\mathrm{Enc}_{\mathrm{SD}}^k$. The main idea is to enable $\mathrm{Enc}_{\mathrm{SD}}^k$ to learn to



(a) *Baseline Decoder* (b) *Dual Attention* (c) *Dual Decoder*

Figure 3: *The baseline decoder structure and two approaches to introduce visual information into the decoder part.* $\mathbf{H}^{v_k}$ *and* $\mathbf{H}^k_{Rec}$ *are the transformed visual features and the output of* Enc_{Rec} *of talker* k*, respectively. Fig.3b shows the modified structure of each layer in Transformer decoder.* $\oplus$ *represents concatenation.*

extract the visual information needed by the current frame of speech feature. The speech of the target talker could then be focused on and extracted. First, we create 2 attention modules [32] named *Src-Att* inside the layer. Then we take the output of the self-attention module from the audio feature to query the key-value pairs generated by visual features, utilizing the attention mechanism. Finally, the outputs of all the attention modules are concatenated and projected to the size of the input speech feature. The whole process could be written as:

$$\begin{aligned}
\mathbf{H}^{\text{a}}_{\text{SD}_0} &= \mathbf{H}_{\text{Mix}}, \\
\mathbf{H}^{\text{q}}_{\text{SD}_i} &= \text{SelfAttention}(\mathbf{H}^{\text{a}}_{\text{SD}_{i-1}}), \\
\mathbf{H}^{\text{att}_k}_{\text{SD}_i} &= \text{Attention}(\mathbf{H}^{\text{q}}_{\text{SD}_i}, \mathbf{H}^{\text{v}_k}), k = 1, \cdots, K, \\
\mathbf{H}^{\text{a}}_{\text{SD}_i} &= \text{Proj}(\text{Concat}(\mathbf{H}^{\text{q}}_{\text{SD}_i}, \{\mathbf{H}^{\text{att}_k}_{\text{SD}_i} : k = 1, \cdots, K\})), \\
\mathbf{H}_{\text{SD}} &= \mathbf{H}^{\text{a}}_{\text{SD}_n}.
\end{aligned} \tag{5}$$

where $i = 1, \ldots, n$, n is the number of layers in Enc^k_{SD}. $\mathbf{H}^{\text{a}}_{\text{SD}_i}$ represents the output of the i-th layer of Enc^k_{SD}. Although *Query Vision* shares a common structure with *Dual Attention* proposed in [24], their aims differ. *Query Vision* uses the audio modality to fetch relevant visual information related to the target talker, while *Dual Attention* utilizes predicted symbol to query both modalities to obtain information related to the next prediction.

3.2. Audio-Visual Decoder

In this subsection, we describe two approaches named *Dual Attention* and *Dual Decoder* to introduce visual information into the decoder part of the model introduced in Section 2. They are shown in Fig.3b and Fig.3c. To clarify the modification we make, the structure of Decoder module in the baseline network is also presented in Fig.3a. Note that these two approaches could be used together with the ones described in Section 3.1.

Variant 1: Dual Attention. The first approach is named *Dual Attention*, and this mechanism is originally proposed in [24] to combine audio and visual information during decoding. It duplicates the attention module inside the decoder layer to query both audio and visual features. After the context features are calculated, they are concatenated and projected to a size similar to the input $\mathbf{H}_{\text{Dec}_{i-1}}$ of the layer. Eq.6 gives the formal definition of each Transformer layer in *Dual Attention*.

$$\begin{aligned}
\mathbf{H}^{\text{q}}_{\text{Dec}_i} &= \text{SelfAttention}(\mathbf{H}_{\text{Dec}_{i-1}}), \\
\mathbf{H}^{\text{att}_\text{a}}_{\text{Dec}_i} &= \text{Attention}_1(\mathbf{H}^{\text{q}}_{\text{Dec}_i}, \mathbf{H}^k_{\text{Rec}}), \\
\mathbf{H}^{\text{att}_\text{v}}_{\text{Dec}_i} &= \text{Attention}_2(\mathbf{H}^{\text{q}}_{\text{Dec}_i}, \mathbf{H}^{\text{v}_k}), \\
\mathbf{H}_{\text{Dec}_i} &= \text{Proj}(\text{Concat}(\mathbf{H}^{\text{q}}_{\text{Dec}_i}, \mathbf{H}^{\text{att}_\text{a}}_{\text{Dec}_i}, \mathbf{H}^{\text{att}_\text{v}}_{\text{Dec}_i})).
\end{aligned} \tag{6}$$

where $i = 1, \ldots, m$, m is the number of layers in the Transformer decoder. $\mathbf{H}_{\text{Dec}_i}$ represents the output of the i-th layer of the Transformer decoder. $\mathbf{H}_{\text{Dec}_0}$ is the embedded output symbol, and $\mathbf{H}_{\text{Dec}_n}$ is used to calculate the next output symbol.

Variant 2: Dual Decoder. Inspired by [35], we propose a variant of Transformer-based *Dual Decoder* as the second decoder-based approach. It duplicates the Transformer decoder for the other modality and concatenates the decoders' output before predicting output symbols by FeedForward. Formally,

$$\begin{aligned}
\mathbf{H}^{\text{a}}_{\text{Dec}} &= \text{TransformerDecoder}_\text{a}(\text{Embed}(y_{j-1}), \mathbf{H}^k_{\text{Rec}}), \\
\mathbf{H}^{\text{v}}_{\text{Dec}} &= \text{TransformerDecoder}_\text{v}(\text{Embed}(y_{j-1}), \mathbf{H}^{\text{v}_k}), \\
\mathbf{H}_{\text{Dec}} &= \text{Concat}(\mathbf{H}^{\text{a}}_{\text{Dec}}, \mathbf{H}^{\text{v}}_{\text{Dec}}).
\end{aligned} \tag{7}$$

4. Experiment

4.1. Data Preparation

Experiments were done on LRS2 dataset [35]. It consists of synchronized video and audio pairs collected from BBC television. The videos are all 25fps and the sample rate of the audios is 16kHz. The dataset is split into four subsets: pretrain, train, val, and test. Here we combine the pretrain set and train set for our model training. Two-talker audio mixtures are generated by normalizing and summing two audios selected randomly with different talker labels and less than 20% difference in lengths. The signal-to-noise ratio (SNR) is randomly chosen in [-10, 10]. 80-dimensional log filterbank features are extracted for each mixture with Hann window length 25ms and hop length 10ms.

A lip-reading model was trained on LRW dataset [36] following the recipe described in [20, 22]. To obtain the 512-dimensional visual features as input, we crop the videos to the mouth region and process them with the 3D ResNet frontend.

4.2. Experimental Setup

Our proposed model is built and evaluated on ESPnet2 [37] framework. A VGG-like module with 256 channels in each of the two 2D convolution layers is used as Enc_{Mix} for subsampling the length of audio feature sequence by 0.5. Enc_{Vis}, Enc^1_{SD}, Enc^2_{SD}, Enc_{Rec} are all Transformer encoders with 2, 4, 4, 8 layers respectively. Transformer decoders in the model are all with 6 layers. When using *Concat* as encoder part, Enc^1_{SD} and Enc^2_{SD} has attention feature dimension 768, while length of the visual features are upsampled by 2 using nearest interpolating strategy to match the audio feature's length. In other cases, every layer in Transformer encoder or decoder has attention feature dimension of 256, feedforward layer dimension of 2048, and the attention heads number is 4. The training loss is computed by interpo-

lating the CTC loss and attention decoder loss with factor 0.3. For the models that use audio-only encoders, PIT is used for the CTC loss calculation, and the obtained minimum permutation is used for the attention decoder. Models are trained until convergence by using Adam optimizer with batch size set to 240. The learning rate is set to 10^{-3} and 25000 warmup steps [32]. An RNN language model for decoding is also trained with the text of our training set with the default configuration.

4.3. Results

We evaluated different combinations of the audio-visual encoders and decoders. The performance on test set is listed in Table.1. For baseline and configurations with only an audio-visual decoder introduced, PIT is necessary since visual information could not efficiently guide the separation. In these cases, both $\mathbf{H}^{v1}$ and $\mathbf{H}^{v2}$ are inputted to the audio-visual decoder and the context vectors of them are extracted separately using the same parameters for calculating attention scores. In *Dual Attention* the extracted visual context vectors are concatenated with $\mathbf{H}^{q}_{\text{Dec}_i}$ and $\mathbf{H}^{\text{att}_a}_{\text{Dec}_i}$ in each decoder layer i, while in *Dual Decoder* the features are concatenated just before the FeedForward layer. Another approach named audio-visual modality driven gated fusion was proposed by [31] to recognize the target speech out of an overlapped one by gating the hidden audio and visual features, and its performance on LRS2 is also shown in Table.1 for reference. It is noted that this result may not be totally comparable with our results due to the differences in data preparation.

Table 1: *Performance of different configuration combinations. Columns titled "Fixed" contain error rates computed by the outputs with fixed permutation similar to the one of input visions, while columns titled "Min" try to find a permutation with minimum error rates. * The best system proposed in [31]. †With audio-visual inputs rather than audio-only input.*

Configuration	CER(%)		WER(%)	
	Fixed	Min	Fixed	Min
AV⊗A+concat[31]*	-	-	10.31	-
baseline (PIT)	49.59	11.28	64.49	17.49
+DualAtt (PIT)†	49.36	11.27	64.18	17.31
+DualDec (PIT)†	49.56	11.07	64.46	17.20
Concat	9.76	9.75	15.49	15.48
+DualAtt	10.43	10.42	16.31	16.30
+DualDec	10.67	10.66	16.87	16.84
QueryVision	5.32	5.32	9.47	9.47
+DualAtt	5.50	5.50	9.69	9.69
+DualDec	**5.04**	**5.04**	**9.10**	**9.10**

Table.1 shows that the best approach proposed in this paper is to combine *Query Vision* and *Dual Decoder*, which outperforms the baseline system by a relative ∼50.0% WER reduction. This improvement strongly demonstrates the advantage of introducing visual information into a multi-talker ASR system.

By comparing the WERs of the audio-visual encoders, it turns out that *Query Vision* behaves much better than *Concat*, which implies the effectiveness of *Query Vision* in extracting related visual features while maintaining a relatively small information loss. For audio-visual decoders, it is interesting to notice that both *Dual Attention* and *Dual Decoder* increase the test WER when combined with *Concat*. We believe it indicates that Enc_{Vis} trained inside model with *Concat* could not extract visual features compatible with the audio-visual decoders.

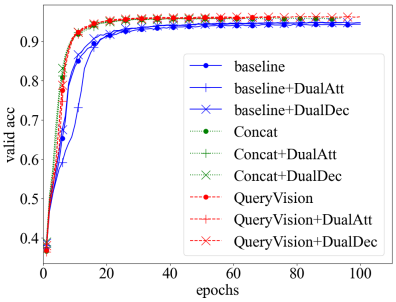


Figure 4: *The accuracy curves on validation set when training the system with different configurations listed in Table.1. For clarity, we mark every 5 data points on each curve.*

With *QueryVision* or no audio-visual encoder introduced, *Dual Decoder* slightly outperforms *Dual Attention*. It could be noticed that *Dual Attention* always increases the WER when an audio-visual encoder is included, compared with the configuration without it. However, it brings a slight decrease to the WER of the baseline system. This is possibly because of the poor similarity between the visual features and the embedded symbol. According to [35, 24, 31], recent lip-reading systems still provide unsatisfying results on LRS2 dataset. Thus, it is reasonable that the relationship between the visual features and the embedded symbol is hard to be discovered by the attention mechanism. We believe this factor causes the decrease in system performance when combining *Dual Attention* with audio-visual encoders. Better audio-visual decoders might be designed following this idea, which could be included in our future work.

It is expected that for configurations without an audio-visual encoder, the CERs and WERs computed with fixed output permutation are considerably high, because they lack additional information to guide the separation process. For the other configurations, the error rates computed with fixed permutation and the minimum permutation (permutation with the minimum error rate) are almost the same, which implies that both the proposed two audio-visual encoders could utilize the visual inputs to extract the corresponding speech from the mixture.

Fig.4 shows how accuracy on validation set changes during training for each configuration listed in Table.1. We observe that the curves of configurations with audio-visual encoders increase faster and converge earlier than others. This indicates that introducing visual information into the encoder accelerates the model's learning on separating and recognizing the mixture.

5. Conclusions

In this paper, we have explored the visual modality utilization in the multi-talker speech recognition task. We have also presented several approaches to introduce visual information into a Transformer-based multi-talker ASR system. By evaluating the system on LRS2 dataset, we have demonstrated that the additional visual information greatly contributes to multi-talker ASR systems by solving the label ambiguity problem, speeding convergence and improving the system performance.

6. Acknowledgements

This work was supported by National Key R&D Program of China, under Grant No. 2020AAA0104500 and the China NSFC projects (No. 62071288 and No. U1736202).

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953, publisher: acoustical society of America.

[2] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, Mar. 2017, pp. 241–245.

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

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

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

[6] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A Comparative Study on Transformer vs RNN in Speech Applications," in *Proc. IEEE ASRU*, Dec. 2019, pp. 449–456.

[7] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A Purely End-to-End System for Multi-speaker Speech Recognition," in *Proc. ACL*, Jul. 2018, pp. 2620–2630.

[8] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, Nov. 2018.

[9] S. Settle, J. L. Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-End Multi-Speaker Speech Recognition," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 4819–4823.

[10] X. Chang, Y. Qian, and D. Yu, "Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks," in *Proc. ISCA Interspeech*, Sep. 2018, pp. 1586–1590.

[11] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end Monaural Multi-speaker ASR System without Pretraining," in *Proc. IEEE ICASSP*, May 2019, pp. 6256–6260.

[12] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition," in *Proc. IEEE ASRU*, Dec. 2019, pp. 237–244.

[13] A. Tripathi, H. Lu, and H. Sak, "End-To-End Multi-Talker Overlapping Speech Recognition," in *Proc. IEEE ICASSP*, May 2020, pp. 6129–6133.

[14] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving End-to-End Single-Channel Multi-Talker Speech Recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1385–1394, 2020.

[15] T. v. Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-Talker ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR," in *Proc. ISCA Interspeech*, Oct. 2020, pp. 3097–3101.

[16] T. v. Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-End Training of Time Domain Audio Separation and Recognition," in *Proc. IEEE ICASSP*, May 2020, pp. 7004–7008.

[17] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: End-To-End Speech Enhancement and Separation Toolkit Designed for ASR Integration," in *Proc. IEEE SLT*, Jan. 2021, pp. 785–792.

[18] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[19] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976.

[20] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. ISCA Interspeech*, Aug. 2017, pp. 3652–3656.

[21] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *arXiv:2008.09586 [cs, eess]*, Mar. 2021.

[22] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in *Proc. IEEE ICASSP*, Apr. 2018, pp. 6548–6552.

[23] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture," in *Proc. IEEE SLT*, Dec. 2018, pp. 513–520.

[24] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-visual Speech Recognition," *IEEE Trans. PAMI.*, pp. 1–1, 2018.

[25] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, "Modality Attention for End-to-end Audio-visual Speech Recognition," in *Proc. IEEE ICASSP*, May 2019, pp. 6565–6569.

[26] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 31–35.

[27] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang, "Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[28] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Proc. ISCA Interspeech*, Sep. 2018, pp. 3244–3248.

[29] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug. 2018.

[30] C. Li and Y. Qian, "Deep Audio-Visual Speech Separation with Attention Mechanism," in *Proc. IEEE ICASSP*, May 2020, pp. 7314–7318.

[31] J. Yu, S.-X. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset," in *Proc. IEEE ICASSP*, May 2020, pp. 6984–6988.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-To-End Multi-Speaker Speech Recognition With Transformer," in *Proc. IEEE ICASSP*, May 2020, pp. 6134–6138.

[34] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE ICASSP*, May 2013, pp. 6645–6649.

[35] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *Proc. IEEE CVPR*, Jul. 2017, pp. 3444–3453.

[36] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in *Computer Vision – ACCV*, 2016, vol. 10112, pp. 87–103.

[37] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. ISCA Interspeech*, Sep. 2018, pp. 2207–2211.