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

End-to-end audio-visual speech recognition for overlapping speech

Richard Rose, Olivier Siohan, Anshuman Tripathi, and Otavio Braga

Google Inc., New York, USA
{rickrose, siohan, anshumant, obraga}@google.com

Abstract

This paper investigates an end-to-end audio-visual (A/V) modeling approach for transcribing utterances in scenarios where there are overlapping speech utterances from multiple talkers. It assumes that overlapping audio signals and video signals in the form of mouth-tracks aligned with speech are available for overlapping talkers. The approach builds on previous work in audio-only multi-talker ASR. In that work, a conventional recurrent neural network transducer (RNN-T) architecture was extended to include a masking model for separation of encoded audio features and multiple label encoders to encode transcripts from overlapping speakers. It is shown here that incorporating an attention weighted combination of visual features in A/V multi-talker RNN-T models significantly improves speaker disambiguation in ASR on overlapping speech relative to audio-only performance. The A/V multi-talker ASR systems described here are trained and evaluated on a two speaker A/V overlapping speech dataset created from YouTube videos. A 17% reduction in WER was observed for A/V multi-talker models relative to audio-only multi-talker models.

1. Introduction

It is well known that automatic speech recognition (ASR) from human-human interaction is a far more difficult problem than ASR from utterances arising from limited domain human machine interaction [1]. While there are many reasons for this, one major issue is the existence of overlapping speech in conversations. Studies of human-human interaction in meetings scenarios have shown that talker overlap typically occurs in from 10 to 30 percent of talkers' utterances [2]. Recent work on overlapping utterances taken from client-customer interactions in a call center application has shown that overlapping speech dramatically increases ASR word error rates (WERs), especially in regions where talkers overlap [3].

This paper presents techniques for addressing the problem of transcribing overlapping utterances from each of multiple talkers. The contributions of the paper include the following. First, the paper is the first to present techniques for exploiting both audio and visual features for decoding overlapping utterances from multiple speakers. Visual features in the form of mouth tracks aligned with speech have been used in audio-visual end-to-end models for single-talker ASR [4] [5] [6]. Two approaches are presented here for integrating audio and visual features in multi-talker ASR. The second contribution is an experimental study that is performed to evaluate the degree to which visual features are able to improve the WER of decoded text obtained from two overlapping speakers. It is shown here that the A/V multi-talker approach presented here has a larger impact on separating and recognizing speech from multiple overlapping talkers than the impact shown for using A/V features in single-talker ASR.

There has been a great deal of recent work on audio-only end-to-end approaches to multi-talker ASR [3] [7][8][9]. The A/V multi-talker techniques in this paper are motivated by the work in [3], which extends a single label encoder RNN-T by applying a masking model to the encoded audio input from an overlapping speech utterance. That system was developed and applied to utterances taken from a call center domain, and is summarized here in Section 2. Many of the recently developed audio-only multi-talker approaches are similar in that they involve an extension of the single label encoder end-to-end model with a training procedure that aligns overlapping speech with transcriptions from multiple speakers.

Work on end-to-end multi-talker ASR has been preceded by work on speech separation where the goal is to recover a target speech signal from overlapped speech [10][11][12]. This includes a recent approach that fuses audio-visual features for speech separation in videos [13]. Explicit speech separation systems generally optimize criteria related to signal-to-background distortion and overall signal fidelity. However, it has been difficult for these techniques to demonstrate large improvements in ASR word error rate (WER), especially when compared to multi-talker ASR systems that are trained to optimize fully end-to-end multi-talker criteria. There has also been recent work on A/V ASR in the presence of background speech. The goal of the work in [14] and [15] is to fuse audio and visual signals in an effort to improve speech recognition from a target speaker in the presence of background speech. The goal of the work presented here is different in that it uses the visual signal to improve speech recognition from multiple speakers and decode transcriptions from each speaker.

One important issue in extending the above audio-only multitalker ASR approach to audio-visual multi-talker ASR is the problem of associating one of multiple on-screen faces with each overlapping talker's audio signal. It is necessary to identify the on-screen face that is associated with the talker so the mouth track associated with that talker can be synchronized with the audio features and input to the ASR system. There are a number of techniques that have been proposed for selecting the mouth-track that is associated with the talker. This can be done prior to combining A/V features for ASR using any of several techniques including those relying on measurements of time synchronization between audio and visual signals [16]. It can also be done as part of end-to-end soft face selection that is integrated with the ASR model [17]. The end-to-end approach relies on an attention mechanism for selecting an attention weighted combination of mouth tracks for each time instant to be input to the multi-talker system with the audio features. Both of these techniques will be investigated in the A/V multi-talker systems described in Section 2.

A simulated overlapping speech A/V corpus was created for training A/V multi-talker models. The corpus is derived from the YouTube A/V corpus created for training A/V ASR models in [6]. This consists of a collection of short utterances where the audio matches the transcripts uploaded by the user with

http://dx.doi.org/10.21437/Interspeech.2021-1621

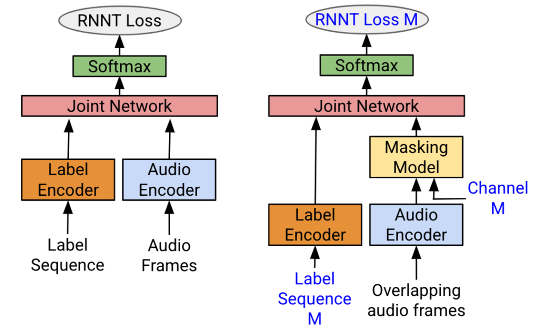
the YouTube video[18], and the video mouth track is aligned with the audio [6]. Each utterance in the simulated overlapped speech corpus was created by combining two of these short A/V utterances with randomly selected overlap interval ranging from one to five seconds. This simulated overlapping speech corpus is described in more detail in Section 3, and the results of an experimental study based on this corpus is presented in Section 4.

2. System Description

This section describes the end-to-end RNNT based approach to multi-talker modeling. First, the basic multi-talker model is described as a multi-channel RNNT model with an added mask layer to produce separate activations for the two overlapping speaker's utterances. Second, two approaches for integrating visual features into the multi-talker model are presented.

2.1. Audio-only multi-talker model

The extension of the single label encoder recurrent neural network transducer (RNNT) [19] to an audio-only multi-talker RNNT [3] is illustrated by the block diagrams in Figure 1. Figure 1a displays the RNNT as a encoder-decoder framework that can be trained end-to-end to map discrete audio input sequences to target label sequences. In this work the audio encoder is a five layer 1024 cell bidirectional LSTM, the label encoder is a two layer 2048 cell LSTM network, and the joint network is a 640 dimensional feed-forward neural network. The input audio features, $\mathbf{X}_a = \{x_{at}\}_{t=1}^{T}$, for a T length utterance are $D_a = 240$ dimensional vectors containing three stacked 80 dimensional mel-frequency filter-bank vectors. All parameters are trained end-to-end with the CTC loss function [19].



a) Single label encoder RNNT b) Multi-talker RNNT

Figure 1: *Extending single label RNNT to audio-only multi-talker RNNT.*

Figure 1b shows how the single label encoder RNNT can be extended to the multi-talker case by adding an LSTM masking model as shown in the figure. It is assumed in the figure that the audio input can contain up to M overlapping utterances. In training, it is assumed that a separate reference label sequence exists for each of the M overlapping utterances. Multi-talker training is performed by separately aligning the overlapped audio frames to each of the M label sequences. A unique channel sequence index is appended to the audio encoder embedding for each label sequence before inputting the embedding to the masking model. This serves to disambiguate speech associated with label sequence m from competing speech.

Separate RNNT losses are computed for each of the M label sequences, and the overall RNNT loss is the sum of channel specific RNNT losses. This is referred to as 2Chan-RNNT loss in Section 4. All parameters in the model are trained using audio signals containing simulated overlapping speech utterances.

A masking loss is also defined to inhibit the alignment of labels associated with one speaker to the opposite speaker's utterance. For the two speaker case, it is defined as:

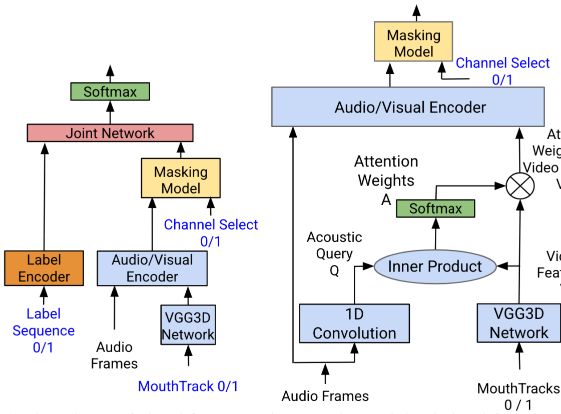
$$MaskLoss = L2\left(\mathcal{M}^{0}{}_{t=T_{End}}^{T}\right) + L2\left(\mathcal{M}^{1}{}_{t=0}^{T_{Start}}\right)$$

where $\mathcal{M}^0$ and $\mathcal{M}^1$ are the masking model activations for channels 0 and 1, and T_{Start} and T_{End} represent the start and end frames respectively of the speaker overlap interval in the T frame input utterance. The goal of the masking loss is to suppress the masking layer outputs for a given channel in those regions of the utterances where speech from that channel is not present. The total loss is computed by summing the RNNT losses for each channel with the mask loss, and will be referred to as 2Chan-RNNT+Mask in Section 4. During evaluation, M strings are decoded on overlapping speech by running decoding with each of M settings of the channel sequence index.

2.2. Audio-visual multi-talker model

The block diagrams in Figure 2 describe two approaches for extending the audio-only multi-talker model to include visual features. In both cases, it is assumed that there are M mouth tracks associated with each of the M speakers that are time synchronized with the speech in the overlapped input utterance. The input video frames, $\mathbf{X}_v : \{x_{vt}^{m}\}_{t=1, m=1}^{T, \quad M}$, for each of M overlapping speakers in a T length utterance are 128 × 128 × 3 thumbnail images. Visual features, $\mathbf{V}^m = \{v_t^m\}_{t=1, m=1}^{T, \quad M}$, are $D_v = 512$ dimensional vectors computed from the input video frames using a 3 dimensional 5 layer convolutional neural network. A detailed description of the video model can be found in [6] and [20].

The first approach for integrating video features in the multi-talker framework, shown in Figure 2a, appends visual features obtained directly from mouth-tracks aligned with the two speaker's utterances. In this case, multi-talker training is performed by appending the video features associated with the mth speaker with the audio features obtained from the overlapped speech and aligning the A/V features with the mth label sequence. The configuration in Figure 2a assumes that, during decoding, the mouth-track that is associated with a given speaker is known. This implies that there needs to be some form of active speaker detection that exists prior to decoding.



a) Direct input of visual features b) Attention weighted visual features

Figure 2: *Audio-visual multi-talker RNNT: a) Direct input (DI) and b) Attention weighted (AW) input of visual features.*

The second approach, shown in Figure 2b, computes attention weighted visual features, $\mathbf{V}'$, as a weighted combination of the visual features $\mathbf{V}^m$ from M overlapping speakers. The attention network is similar to the attention weighted approach for dealing with multiple on-screen faces in A/V ASR for non-overlapping speech [17], and represents a simplified version of a more general attention network [21]. It is trained end-to-end with the RNNT model, and produces a T length sequence of attention weights, $\mathbf{A}^m = \{a_t^m\}_{t=1,m=1}^{T,\ M}$. The attention weight, a_t^m, represents a normalized measure of similarity between acoustic and the visual features for the mth speaker at time t:

$$a_t^m = Softmax(InnerProduct(q_t, v_t^m)),$$

where the acoustic query vector, q_t, is a 512 dimensional embedding generated from the input audio features using a 1 dimensional convolutional network:

$$q_t = 1DConvNet(x_{at}).$$

The advantage of the attention weighted (AW) input of visual features, as depicted in Figure 2b, is that there is no need to determine which set of M video features corresponds to each of the M label sequences as is needed for the direct input (DI) of visual features in Figure 2a. The WERs obtained for both of these approaches on the simulated overlapping speech corpus described in Section 3 are given in Section 4.

3. Experimental Study

This section describes the experimental study performed to evaluate the performance of the audio-only and audio-visual multi-talker models presented in Section 2. The experiments are limited to the case where there are overlapping utterances from $M = 2$ speakers. The overlapping speech datasets, model parameterizations, and the evaluation measures are described.

3.1. Simulated audio-visual overlapping speech corpus

A training corpus of simulated overlapping speech utterances was created by combining short audio-visual utterances extracted from YouTube videos. There are three major steps to obtain this corpus. First, a process of unsupervised mining of short audio utterances from YouTube videos with user provided captions [18] is performed. Audio utterances are derived from islands of confidence where there is agreement between an ASR decoded result for a segment and the force-aligned user-provided transcript.

Second, face-tracking technology is leveraged to select utterances with matching speaking face [6][20]. Video snippets corresponding to the selected utterances are extracted, and face tracking is performed to locate all on-screen faces. A visual speech classifier is used to identify the speaking face, if any, that spans each audio segment, followed by an audiovisual synchrony detector to reject dubbed videos. The result of this process is a collection of short utterances (from a one to two seconds to tens of seconds long) totaling 31k hours of data, where with high confidence, the audio matches the user-uploaded transcripts, and the selected face video track matches the audio.

Finally, a training set of simulated overlapped utterances was created from the above confidence island utterances with aligned face and mouth tracks. The audio portion of the overlapped utterances was created by taking two of the above single speaker utterances, offsetting one in time with respect to the other, and adding the two audio signals. The video portion of the overlapped utterances consists of two mouth tracks where one of the mouth tracks has been offset to be aligned with the corresponding offset audio signal. However, shifting the video frames creates a situation where there are no video frames associated with a given speaker in those regions where that speaker is not speaking. To deal with this issue, video frames for non-speech regions were filled with forward-backward repetitions of video frames where speech was present. This provides video features from the same speaker, but that are not synchronized with the audio.

The offset used in shifting the audio signals was chosen to provide overlap intervals randomly selected with a uniform distribution between 1 and 5 seconds. Each overlapped speech utterance was stored with two reference transcriptions, two mouth tracks, and overlap interval start and end times which were used for computing the masking loss described in Section 2. The resulting training corpus contains 18k hours of training data.

3.2. Evaluation measures

In the two speaker multi-talker case, multi-talker decoding is run separately for two channels generating hypothesized strings h_1, h_2. These must be scored against the reference strings r_1, r_2 for the two overlapping utterances, and the scoring method must decide which hypothesized string is assigned to a given reference string. In this work the minimum permuted reference (prWER), was used:

$$\begin{aligned} prWER(h_1, h_2, r_1, r_2) &= \min(Err(h_1, r_1) + Err(h_2, r_2), \\ &\qquad Err(h_1, r_2) + Err(h_2, r_1)), \end{aligned}$$

where $Err()$ corresponds to the standard WER measure. A comparison between prWER and the more well known concatenated minimum permutation word error rate (cpWER) [22][3] was performed, and the two measures were found to differ by less than one percent relative.

3.3. AI Principles

The work presented in this paper abides by Google AI Principles [23]. We are hoping that this work, by improving the robustness of speech recognition systems, will increase the reach of ASR technology to a larger population of users, as well as the development of assistive technology. It should also be noted that the data and models developed in this work are restricted to a small group of researchers working on this project and are handled in compliance with the European Union General Data Protection Regulation [24].

4. Experimental Results

Results are presented comparing WERs obtained with audio-only and A/V multi-talker end-to-end RNNT models on an overlapping speech test set. The test set was obtained from human transcribed utterances with aligned mouth tracks taken from YouTube videos, and not the semi-supervised procedure described in Section 3. However, the process of forming overlapped utterances as a combination of single speaker utterances is the same as described above. The test set contains 3135 utterances ranging in length from 3.2 to 13.6 seconds. WERs obtained for the multi-talker models are also compared with those obtained for single channel A/V and audio-only models.

4.1. Audio-only multi-talker results

Table 1 provides a comparison between audio-only multi-talker and single-channel RNNT performance on the YouTube test set. The single taker (Single) and overlapped (Overlap) test sets contain the same number of utterances. The overlapped set contains the same utterances as the single speaker set except waveforms were offset and added to these utterances as described in Section 3.1. The first row of the table gives the WER for a baseline audio-only single channel RNNT model (SingleChan). The second row gives the prWER computed for that same model evaluated on the overlapped test set. The third and fourth rows of the table give the prWER for the audio-only multi-talker RNNT model (MultiTalker) trained with summed RNNT (2Chan-RNNT) loss with and without mask loss (Mask).

There are several observations that can be made from the results in Table 1. First, comparing rows one and two, simply decoding on the Overlap test set with the SingleChan model gives a large increase in WER, which is a result of a large number deletions in the overlap intervals. Second, comparing the WERs from the second and third rows, the WER for the MultiTalker model decreases by a factor of two compared to the SingleChan model. This is primarily due to a dramatic reduction in the number of deletions. Third, comparing rows three and four, adding the Mask loss reduces the WER by about 4 percent relative to training with the 2Chan-RNNT loss alone.

Table 1: *WERs for audio-only single channel RNNT (SingleChan) and audio-only multi-talker RNNT (MultiTalker) models on single talker (Single) and 2 talker overlapped speech (Overlap) test sets. MultiTalker models are trained with summed RNNT (2Chan-RNNT) loss with and without mask loss.*

Audio-only RNNT Model Performance (WER%)			
RNNT Model	**Loss**	**Test Set**	**WER**
SingleChan	RNNT	Single	14.8
SingleChan	RNNT	Overlap	44.8
MultiTalker	2Chan-RNNT	Overlap	21.6
MultiTalker	2Chan-RNNT+Mask	Overlap	20.7

4.2. A/V multi-talker results

The impact of the audio-visual multi-talker models is summarized by the results in Table 2. The first row of the table gives the WER for the baseline audio-visual RNNT model on the single speaker test set. The second row of Table 2 gives the prWER computed for the audio-visual multi-talker model with direct input (DI) of visual features (MultiTalker-DI), as depicted in Figure 2a, evaluated on the overlapped A/V test set. This represents an ideal case since, as mentioned in Section 2.2, this scenario assumes that decoding is preceded by an active speaker detection module. It is assumed in this result that this module is error free. The third row of the table gives the prWER for the audio-visual multi-talker RNNT model with attention weighted (AW) visual features (MultiTalker-AW), as shown in Figure 2b, also evaluated on the overlapped A/V test set. The fourth and fifth rows show the WERs for the multi-talker models trained with mask loss (Mask). Comparing the WER for the A/V baseline in the first row of Table 2 to the audio-only WER in the first row of Table 1, there is a 6.7% reduction in WER with respect to the audio-only baseline. This is consistent with results obtained for the impact of using A/V features in [6].

Several observations can be made about A/V multi-talker ASR performance from this table. First, comparing the DI and AW A/V multi-talker performance in rows 3 and 4 of Table 2, the WER for the DI A/V scenario is about 8% lower than for the AW case. Second, the relative decrease in WER associated with Mask loss shown in Table 2 is slightly less than that obtained for the audio-only multi-talker model. Third, comparing the WERs in the fourth row of Table 1 and the fifth row of 2, there is a 17% decrease in WER for the AW A/V multi-talker system relative to the audio-only multi-talker system. This suggests that the integration of visual features in the A/V multi-talker models has a significant impact in disambiguating overlapping speakers relative to audio-only models.

Table 2: *WERs for A/V single channel RNNT and A/V multi-talker RNNT models using direct input of visual features (MultiTalker-DI) and attention weighted visual features (MultiTalker-AW).*

Audio-visual RNNT Model Performance (WER%)			
A/V Model	**Loss**	**Test Set**	**WER**
SingleChan	RNNT	Single	13.8
MultiTalker-DI	2Chan-RNNT	Overlap	16.4
MultiTalker-AW	2Chan-RNNT	Overlap	17.8
MultiTalker-DI	2Chan-RNNT+Mask	Overlap	16.2
MultiTalker-AW	2Chan-RNNT+Mask	Overlap	17.2

Figure 3 provides a comparison between audio-visual and audio-only models according to how the WERs for the two models vary across the percentage overlap for the overlapping utterances. The percentage overlap is defined here as $(T_{End} - T_{Start})/T$. Note that the WER for the audio-visual model is nearly uniform across the range of percent overlap, while the WER for the audio-only model shows a significant increase as the degree of overlap increases.

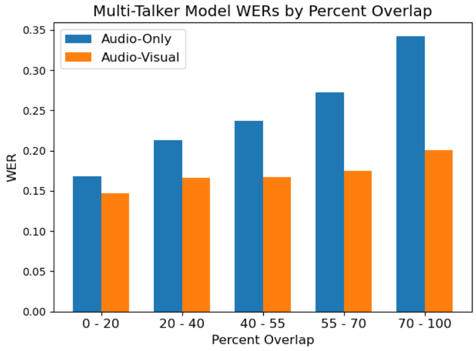


Figure 3: *WER by utterance overlap interval for audio-only and audio-visual multi-talker models.*

5. Summary and Conclusions

The work in this paper represents the first attempt at integrating the visual modality in end-to-end multi-talker ASR on overlapping speech utterances. The experimental study has demonstrated that integrating visual features in multi-talker ASR has a bigger impact on performance relative to the audio-only case than the impact of integrating the visual modality in single label encoder ASR. This was demonstrated by the fact that an attention based A/V multi-talker system resulted in a reduction in WER of 17% on a two speaker simulated overlapping speech corpus, while the A/V single-talker system was responsible for a decrease in WER of 6.7% relative to audio-only ASR. Further work is being directed towards replacing the existing end-to-end encoders with transformer transducers [25] and investigating alternative architectures for fusing A/V features.

6. Acknowledgments

The authors would like to thank Basi Garcia for help with tools in model evaluation, and Takaki Makino and Hank Liao for their contribution to A/V speech corpus development.

7. References

[1] M. Meteer and R. Iyer, "Modeling conversational speech for speech recognition," in *EMNLP*, 1996.

[2] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *InterSpeech*, 2006.

[3] Anshuman Tripathi, Han Lu, and Hasim Sak, "End-to-end multi-talker overlapping speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6129–6133.

[4] 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*, 2018.

[5] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic, "Audio-visual speech recognition with a hybrid CTC/Attention architecture," in *IEEE SLT*, 2018.

[6] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 905–912.

[7] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, 2020.

[8] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 02 2020.

[9] Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong, "Streaming end-to-end multi-talker speech recognition," *arXiv e-prints*, p. arXiv:2011.13148, Nov. 2020.

[10] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP 2016 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

[11] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: dataset and analysis," in *ICASSP 2020*.

[12] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[13] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug 2018.

[14] Jianwei Yu, Bo Wu, Rongzhi Gu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu. Meng Yu, Dan Su, Dong Yu, Xunying Liu, and Helen Meng, "Audio-visual multi-channel recognition of overlapped speech," in *InterSpeech*, 2020.

[15] Guan-Lin Chao, William Chan, and Ian Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," in *InterSpeech*, 09 2016, pp. 2120–2124.

[16] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," *ICASSP 2019*, May 2019.

[17] Otavio Braga, Takaki Makino, O. Siohan, and H. Liao, "End-to-end multi-person audio/visual automatic speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6994–6998, 2020.

[18] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 368–373.

[19] Alex Graves, "Sequence Transduction with Recurrent Neural Networks," *International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning*, Nov. 2012.

[20] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas, "Large-scale visual speech recognition," in *InterSpeech*, 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, NIPS'17, p. 6000–6010.

[22] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. Subramanian, J. Trmal, B. Ben Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings," in *Proc. The 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

[23] Google, "Artificial intelligence at Google: Our principles," `https://ai.google/principles/`.

[24] European Union Law, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation)," `https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX`.

[25] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020*, pp. 7829–7833.